

LEHRBUCH

Nicola Döring · Jürgen Bortz

Forschungsmethoden und Evaluation

in den Sozial- und Humanwissenschaften

5. Auflage

EXTRAS ONLINE

Springer

Springer-Lehrbuch

Nicola Döring
Jürgen Bortz

Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften

5. vollständig überarbeitete, aktualisierte und erweiterte Auflage

Mit 194 Abbildungen und 167 Tabellen
Unter Mitarbeit von Sandra Pöschl

Springer

Nicola Döring
Institut für Medien und Kommunikationswissenschaft
Technische Universität Ilmenau
Ilmenau, Deutschland

Sandra Pöschl
Institut für Medien und Kommunikationswissenschaft
Technische Universität Ilmenau
Ilmenau, Deutschland

Jürgen Bortz†

Zusätzliches Material zu diesem Buch finden Sie auf http://www.lehrbuch-psychologie.de

ISSN 0937-7433
ISBN 978-3-642-41088-8 ISBN 978-3-642-41089-5 (eBook)
DOI 10.1007/978-3-642-41089-5

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer
© Springer-Verlag Berlin Heidelberg 1984, 1995, 2002, 2006, 2016
Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Produkthaftung: Für Angaben über Dosierungsanweisungen und Applikationsformen kann vom Verlag keine Gewähr übernommen werden. Derartige Angaben müssen vom jeweiligen Anwender im Einzelfall anhand anderer Literaturstellen auf ihre Richtigkeit überprüft werden.

Planung: Joachim Coch, Heidelberg
Projektmanagement: Judith Danziger, Kerstin Kindler, Heidelberg
Lektorat: Christine Bier, Nußloch
Projektkoordination: Michael Barton, Heidelberg
Umschlaggestaltung: deblik Berlin
Fotonachweis Umschlag: © deblik Berlin
Herstellung: le-tex publishing services GmbH, Leipzig

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Springer-Verlag GmbH Berlin Heidelberg ist Teil der Fachverlagsgruppe Springer Science+Business Media
(www.springer.com)

Vorwort zur fünften Auflage

Konzept und Zielgruppen des Buches

Ein **Lehrbuch zum Einstieg** und ein **Handbuch zum Nachschlagen** – beides möchte die vorliegende Monografie „Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften" bieten. Studierende und Forschende sollen darin unterstützt werden, ihre eigenen empirischen Studien mit Blick auf das Forschungsproblem sowie die vorhandenen Ressourcen realistisch zu planen, gemäß dem methodischen „State of the Art" umzusetzen und der Wissenschaftsgemeinschaft sowie der breiten Öffentlichkeit zu präsentieren. Zudem will das Buch das nötige Hintergrundwissen liefern, das zum Verständnis und zur kritischen Beurteilung vorliegender Untersuchungen bzw. zur Lektüre der sozial- und humanwissenschaftlichen Fachliteratur notwendig ist. Nicht zuletzt sollen Lehrende im Methodenfach unterstützt werden, indem sie einzelne Kapitel als Basislektüre für ihre Lehrveranstaltungen nutzen können: Jedes Kapitel ist jetzt in sich abgeschlossen und jeweils mit Lernquiz, Übungsaufgaben und einem eigenen Literaturverzeichnis versehen.

Die vorliegende fünfte Auflage gliedert sich in drei Teile: Im **Grundlagen-Teil** wird erklärt, was gute empirische Forschung in den Sozial- und Humanwissenschaften ausmacht. Dabei geht es um ein Grundverständnis empirischer Sozialforschung, um Wissenschaftstheorie, Qualitätskriterien empirischer Studien sowie um Forschungs- und Wissenschaftsethik. Im **Anwendungs-Teil** – dem Herzen des Buches – wird der gesamte empirische Forschungsprozess von der Wahl des Forschungsthemas über Datenerhebungs- und Auswertungsmethoden bis zur Ergebnispräsentation in Form einer Abschlussarbeit, eines Konferenzvortrags oder eines Zeitschriftenartikels anschaulich und praxisnah behandelt. Der **Vertiefungs-Teil** widmet sich spezifischen Fragen; hier werden z. B. die Arbeitsschritte einer Metaanalyse geschildert, Strukturgleichungsmodelle vorgestellt und nicht zuletzt die Besonderheiten der Forschung im Bereich Evaluation verdeutlicht. Generell zielt das Buch darauf ab, Forschungsmethodik relativ breit gefächert und anwendungsnah zu behandeln. Es will Neugier wecken, Einstieg und Orientierung erleichtern und wichtige Hinweise für die praktische Forschungstätigkeit liefern. Für spezifische Detailfragen wird auf Spezialliteratur verwiesen.

Der Schwerpunkt des Buches liegt in der **quantitativen Forschungstradition**. Dementsprechend werden Fragen der Operationalisierung und Messung, experimentelle Designs, repräsentative Stichproben, strukturierte Datenerhebungsverfahren und statistische Datenauswertungsmethoden detailliert behandelt. Doch auch **qualitative Methoden** sind vertreten, insbesondere im Zusammenhang mit Datenerhebung und Datenauswertung: So werden etwa ethnografische Feldforschung, Leitfaden-Interview und Gruppendiskussion sowie interpretative Datenanalyseverfahren (etwa im Kontext der Grounded-Theory-Methodologie) anschaulich dargestellt. Dass quantitative und qualitative Forschungsansätze nicht in einem Konkurrenz-, sondern in einem fruchtbaren Ergänzungsverhältnis zueinander stehen können und sollen, ist schließlich in vielen Sozial- und Humanwissenschaften sowie in der Evaluationsforschung eine heute weitgehend konsensfähige Position, der sich das vorliegende Buch anschließt. Die aktuelle Entwicklung einer **Mixed-Methods-Forschung**, die qualitative und quantitative Vorgehensweisen wissenschaftstheoretisch reflektiert miteinander verzahnt, wird aufgegriffen.

Fachlich beheimatet ist das Buch in der **Psychologie**. Dies spiegelt sich unter anderem darin wider, dass wir psychologische Testverfahren und physiologische Messungen als Datenerhebungsmethoden behandeln, die in manchen angrenzenden Fächern gut etabliert sind (z. B. psychologische Tests in der Erziehungswissenschaft; physiologische Messungen in der Medizin), in anderen sozialwissenschaftlichen Disziplinen dagegen kaum Beachtung finden (z. B. in Soziologie oder Politikwissenschaft). Bei der Wahl der Studienbeispiele haben wir uns um ein breites Themenspektrum bemüht, in dem sich diverse **sozial- und humanwissenschaftliche Disziplinen** wiederfinden.

Auswahl und Darstellung der Studienbeispiele

Auswahl und Darstellung der im Buch präsentierten Beispiele für empirische Studien orientieren sich strikt an **didaktischen Kriterien**. Einzelne Studien wurden ausgewählt und selektiv in jenen Ausschnitten beschrieben, die das im jeweiligen Lehrbuch-Kapitel dargestellte Thema gut illustrieren. Es wird mit den Studienbeispielen ausdrücklich nicht das Ziel verfolgt, die jeweilige Studie in ihren theoretischen Hintergründen, methodischen Besonderheiten oder empirischen Ergebnissen umfassend zu würdigen. Ebenso ist mit der Auswahl oder Nicht-Auswahl einzelner Studien bzw. Autorinnen und Autoren keinerlei Bewertung ihrer wissenschaftlichen Bedeutung im jeweiligen Forschungsgebiet intendiert. Zuweilen werden **klassische Studien** herangezogen (z. B. das Stanford-Prison- oder das Milgram-Experiment), es werden jedoch auch zahlreiche **unbekannte Untersuchungen** angeführt, sei es, weil in ihnen ein bestimmter methodischer Aspekt besonders prototypisch zum Ausdruck kommt, sei es, weil die betreffende Publikation das methodische Vorgehen besonders transparent darstellt, sei es, weil das Untersuchungsthema fächerübergreifend für eine breite Leserschaft nachvollziehbar und spannend sein könnte.

Themen, die für Fachleute einer bestimmten wissenschaftlichen Teildisziplin besonders faszinierend sind, erscheinen für Außenstehende nicht selten speziell und abstrakt. Umgekehrt mögen Themen, die von großem öffentlichem Interesse sind, aus fachdisziplinärer Sicht unbedeutend oder abseitig anmuten. In dem Bestreben, Forschungsmethoden anschaulich und zuweilen auch unterhaltsam zu präsentieren, wurde bewusst das eine oder andere **einprägsame und „exotische" Beispiel** einbezogen und immer wieder auch auf öffentliche Debatten rund um sozialwissenschaftlich untersuchte Phänomene Bezug genommen.

Änderungen gegenüber der vierten Auflage

Gegenüber der letzten Auflage gibt es vor allem drei grundlegende Änderungen:

1. Die Inhalte des Buches wurden **stringenter strukturiert**. Ergebnis ist die Gesamtgliederung in drei Teile (Grundlagen, Anwendung, Vertiefung) sowie die verbesserte Binnengliederung der einzelnen Kapitel, die jetzt in sich abgeschlossen mit jeweils eigenem Literaturverzeichnis nutzbar sind.
2. **Aktuelle Themen** wurden umfassend eingearbeitet (z. B. Strukturgleichungsmodelle, Mixed Methods Designs, Wissenschaftsethik, qualitative und quantitative Datenanalyse-Software, Online-Methoden etc.), überholte Inhalte dagegen gestrichen (darunter auch viele alte Literaturquellen) und Spezialthemen auf das buchbegleitende Webangebot verlagert (z. B. Coombs-Skala, Signalentdeckungs-Paradigma, Parameterschätzungen bei unterschiedlichen Typen probabilistischer Stichproben). Die Aktualisierung der Inhalte hat dazu geführt, dass für die vorliegende 5. Auflage rund 70 % des Textes vollkommen neu geschrieben wurden, was dann auch eine Änderung der Autorenreihenfolge nach sich zog.
3. Auf eine zeitgemäße **didaktische Aufbereitung** wurde noch mehr Wert gelegt als bisher: Zahlreiche internationale Studienbeispiele sollen die dargestellten Methoden veranschaulichen, Tabellen, Grafiken und Cartoons den Text auflockern und Definitions-Boxen, Wissens-Checks sowie Übungsaufgaben beim Lernen helfen. Zentrale Fachbegriffe werden in deutscher und englischer Sprache eingeführt. Ein begleitendes Webangebot (▶ Webangebot zu diesem Buch) auf www.lehrbuch-psychologie.de mit Lösungen zu jedem Lernquiz, mit Tabellen für statistische Prüfgrößen, mit zusätzlichen Textteilen und Glossar runden das Buch ab.

Gedenken

Es war schmerzlich und schwierig, nach dem Tod von Prof. Dr. Jürgen Bortz im September 2007 die Arbeit an der Neuauflage in alleiniger Verantwortung fortzuführen. Einerseits wollte ich dem bisherigen Werk treu bleiben und Bewährtes erhalten, andererseits aber auch formale und inhaltliche Neuerungen einbringen.

Forschungsmethoden zugänglich zu machen und Begeisterung für empirische Forschung zu wecken – das war immer ein zentrales Anliegen von Jürgen Bortz, der sich als Autor einschlägiger Methoden- und Statistik-Lehrbücher ebenso engagierte wie als Hochschullehrer, dessen lebendige Lehrveranstaltungen bei den Studierenden an der Technischen Universität Berlin großen Anklang fanden. Diese anwendungsbezogene Herangehensweise spiegelt sich in der noch stärker didaktisch gestalteten 5. Auflage wider und fände, so denke ich, seine Zustimmung.

In inhaltlichen und methodischen Fragen haben wir nicht immer übereingestimmt. Insbesondere war Jürgen Bortz den qualitativen Methoden gegenüber distanzierter eingestellt. So manche Passage in dieser Neuauflage, die sich qualitativen Methoden der Datenerhebung und Datenanalyse ausführlicher widmet, wäre vermutlich für ihn verzichtbar gewesen. Anhängern und Fans der klassischen „Bortz-Schule“ möchte ich deswegen nahelegen, sich auf die Textpassagen zu quantitativen Methoden zu konzentrieren. Diese wurden ebenfalls ausgebaut und aktualisiert. Das betrifft nicht zuletzt die Thematisierung der Grenzen des Nullhypothesen-Signifikanztests, die Jürgen Bortz sehr am Herzen lag, und die sich zunehmend etablierende stärkere Beachtung von Effektgröße und Teststärke neben der statistischen Signifikanz.

Ich hoffe, die fünfte Auflage des Buches ist zeitgemäß und bewahrt gleichzeitig das geistige Erbe ihres geschätzten und zu früh verstorbenen Mitautors. Alle Fehler, Lücken und Ungereimtheiten dieser Neuauflage habe ich allein zu verantworten.

Dank

Die Arbeit an der vorliegenden 5. Auflage wurde von vielen Kolleginnen und Kollegen unterstützt, denen ich zu großem Dank verpflichtet bin.

An erster Stelle danken möchte ich meiner langjährigen Mitarbeiterin Dr. Sandra Pöschl, die inhaltlich und redaktionell von Anfang an mit unermüdlichem Engagement und großer Expertise mitgewirkt und auch in schwierigen Arbeitsphasen nie die Geduld verloren hat.

Dr. Christina S. Werner (Universität Zürich), Apl. Prof. Dr. Karin Schermelleh-Engel, Dipl.-Psych. Carla Gerhard und Dipl.-Psych. Jana C. Gäde (Goethe-Universität Frankfurt am Main) haben dankenswerterweise ein „Gastkapitel“ zu Strukturgleichungsmodellen beigesteuert.

Sehr dankbar bin ich all denjenigen, die Teile des Manuskripts im Sinne eines Peer Review kritisch gelesen und teilweise sehr detailliert kommentiert haben: Prof. Dr. Andreas Beelmann (Friedrich-Schiller-Universität Jena), Prof. Dr. Gary Bente (Universität zu Köln), Prof. Dr. Franz Breuer (Westfälische Wilhelms-Universität Münster), Dr. Tim Blumer (ehemals Universität der Künste Berlin), Prof. Dr. Michael Bosnjak (Freie Universität Bozen), Prof. Dr. Martin Emmer (Freie Universität Berlin), Prof. Dr. Edgar Erdfelder (Universität Mannheim), Prof. Dr. Gisela Erdmann (Technische Universität Berlin), Prof. Dr. Andreas Fahr (Universität Freiburg), Prof. Dr. Joachim Funke (Universität Heidelberg), Prof. Dr. Anja Geigenmüller (Technische Universität Ilmenau), Prof. Dr. Mario Gollwitzer (Philipps-Universität Marburg), Prof. Dr. Andreas Hadjar (Universität Luxemburg), Dr. Jan Hense (Ludwig-Maximilians-Universität München), Dr. Marcus Ising (Max-Planck-Institut für Psychiatrie München), Prof. em. Dr. Helmut Kromrey (Freie Universität Berlin), Dr. Jan Kruse (Albert-Ludwigs-Universität Freiburg), Prof. Dr. Udo Kuckartz (Philipps-Universität Marburg), Dr. Christoph Kuhlmann (Technische Universität Ilmenau), Dr. Konrad Leitner (Technische Universität Berlin), Dr. Emanuel Maxl (Context-Research, Österreich), Prof. Dr. Günter Mey (Hochschule Magdeburg-Stendal), Dr. Stefan Rädiker (Philipps-Universität Marburg), Apl. Prof. i.R. Dr. Jürgen Rost (IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik Kiel), Prof. Michael Schäfer (Hochschule Ruhr West), Dr. Christian Schendera (Method Consult, Schweiz), Apl. Prof. Dr. Karin Schermelleh-Engel (Goethe-Universität Frankfurt am Main), Dr. Frank Schneider (Universität Koblenz-Landau), Prof. Dr. Christof Schuster (Justus-Liebig-Universität Gießen), Prof. Dr. Renate Soellner (Universität Hildesheim), Prof. Dr. Werner Sommer (Humboldt-Universität zu Berlin), Dipl.-Päd. Claus Stefer (Philipps-Universität Marburg), Dr. Ines Steinke (Siemens AG München), Apl. Prof. Dr. Udo Thiedeke (Johannes Gutenberg-Universität Mainz), Prof. Dr. Helmut Thome (Martin-Luther-Universität Halle-Wittenberg), PD Dr. Thomas Widmer (Universität Zürich), Jun.-Prof. Dr. Jeffrey Wimmer (Technische Universität

Ilmenau), Prof. Dr. Markus Wirtz (Pädagogische Hochschule Freiburg), Prof. Dr. Jens Wolling (Technische Universität Ilmenau). Die Rückmeldungen waren von unschätzbarem Wert und wurden so gut wie möglich im Rahmen des Gesamtkonzepts des Buches umgesetzt.

Nicht zuletzt möchte ich mich im Voraus bei allen interessierten Leserinnen und Lesern bedanken, die sich mit Feedback melden und uns auf Verbesserungsmöglichkeiten für die nächste Auflage hinweisen.

Ilmenau, im Februar 2014 Nicola Döring

Vorwort zur ersten Auflage

Empirische Forschung kann man nicht allein durch die Lektüre von Büchern erlernen. Praktische Erfahrungen im Umgang mit den Instrumenten der empirischen Sozialforschung sind durch kein auch noch so vollständig und detailliert abgefasstes Lehrbuch ersetzbar. Dass hier dennoch der Versuch unternommen wurde, die wichtigsten in den Sozialwissenschaften gebräuchlichen Untersuchungsvarianten sowie zahlreiche Methoden der Datenerhebung in einem Buch zusammenzufassen und zu diskutieren, geschah in der Absicht, dem Studenten Gelegenheit zu geben, sich parallel zu praktisch-empirischen Übungen einen Überblick über empirische Forschungsmöglichkeiten zu verschaffen. Ich hoffe, dass das „Lehrbuch der empirischen Forschung" dem Studenten hilft, für seine Diplomarbeit, Magisterarbeit o. Ä. ein geeignetes Thema zu finden, einen für sein Thema angemessenen Untersuchungsplan zu entwickeln sowie häufig begangene Fehler bei der Untersuchungsdurchführung, Auswertung und Interpretation zu vermeiden.

Das Buch wendet sich in erster Linie an Psychologiestudenten, kann aber darüber hinaus auch anderen sozialwissenschaftlichen bzw. empirisch orientierten Fachvertretern (Soziologen, Pädagogen, Medizinern, Wirtschaftswissenschaftlern etc.) viele Anregungen und Hilfen geben. Es ist als Studienbegleiter konzipiert und enthält deshalb Passagen, die sich explizit an den Studienanfänger richten (z. B. Kapitel 1) sowie Abschnitte, die den fortgeschrittenen Studenten bei seinem Untersuchungsvorhaben konkret anleiten.

Der Aufbau des Buches ist der Überzeugung verpflichtet, dass das methodische Vorgehen dem wissenschaftlichen Status der inhaltlichen Frage nachgeordnet ist. Moderne Sozialwissenschaften, deren Fragen teilweise wissenschaftliches Neuland betreten oder auf bereits vorhandenes Wissen zurückgreifen, benötigen beschreibende Untersuchungen und hypothesenprüfende Untersuchungen gleichermaßen. Dementsprechend behandelt Kapitel 3 beschreibende Untersuchungsvarianten, die in erster Linie der Anregung neuartiger inhaltlicher Hypothesen oder Ideen dienen, und Kapitel 4 Untersuchungen, mit denen Populationen oder Grundgesamtheiten anhand von Stichproben beschrieben werden. Knüpft eine Forschungsfrage hingegen an eine bereits entwickelte Forschungstradition an, aus deren Theorienbestand begründete Hypothesen ableitbar sind, ist die Konzeption und Durchführung einer hypothesenprüfenden Untersuchung geboten. Auch hier sind es inhaltliche Überlegungen, die darüber entscheiden, ob das Forschungsgebiet bereits genügend entwickelt ist, um die Überprüfung einer Hypothese mit vorgegebener Effektgröße (Kapitel 6) zu rechtfertigen oder ob die bereits bekannten Theorien und Forschungsinstrumente noch so ungenau sind, dass die in der Hypothese behaupteten Unterschiede, Zusammenhänge oder Veränderungen bestenfalls ihrer Richtung nach, aber nicht hinsichtlich ihrer Größe vorhersagbar sind (Kapitel 5, Untersuchungen zur Überprüfung von Hypothesen ohne Effektgrößen).

Die Inhalte der beiden ersten Kapitel sind für alle vier Hauptarten empirischer Untersuchungen gleichermaßen bedeutsam. Kapitel 1 befasst sich mit allgemeinen Prinzipien der Untersuchungsplanung und -durchführung und Kapitel 2 mit Methoden der empirischen Datenerhebung (Zahlen, Urteilen, Testen, Befragen, Beobachten und physiologische Messungen).

Empirische Forschung erfordert nicht nur Erfahrung in der Anlage von Untersuchungen und im Umgang mit sozialwissenschaftlichen Forschungsinstrumenten, sondern auch profunde Statistikkenntnisse, die in diesem Buch nicht vermittelt werden. Ich habe in diesem Text auf die Behandlung statistischer Probleme bewusst weitgehend – bis auf einige Ausführungen, die spezielle, in der Standardstatistikliteratur nicht behandelte Verfahren sowie die Grundprinzipien des statistischen Schließens und Testens betreffen – verzichtet; sie sind an anderer Stelle (Bortz, 1979) zusammengefasst. In dieser Hinsicht ist der vorliegende Text als Ergänzung des Statistiklehrbuches (bzw. umgekehrt, das Statistiklehrbuch als Ergänzung dieses Empirielehrbuches) zu verstehen.

Mein Dank gilt vor allem meinem Mitarbeiter, Herrn Dipl.-Psych. D. Bongers, der mit mir die Konzeption zu diesem Buch diskutierte, Vorlagen zu den Kapiteln 1.4.6 (Messtheoretische Probleme), 2.5 (Beobachten) und zu Kapitel 3 (Untersuchungen zur Vorbereitung der Hypothesengewinnung) aufarbeitete und der – wie auch Herr cand. psych. D. Widowski, dem ich ebenfalls herzlich danke – den

gesamten Text kritisch überprüfte. Ich danke ferner Frau Dipl.-Psych. D. Cremer für ihre Anregungen zur Gestaltung des ersten Kapitels, meinem Kollegen Herrn A. Upmeyer und Herrn Dipl.-Psych. K. Leitner für ihre ständige Bereitschaft, mit mir über Probleme der empirischen Forschung zu diskutieren, sowie Frau cand. psych. Y. Kafai für die Überprüfung der Korrekturabzüge. Schließlich sei Frau K. Eistert, meiner Sekretärin Frau W. Otto und auch meiner Frau für die oftmals schwierige Manuskriptanfertigung gedankt sowie den Mitarbeitern des Springer-Verlages für ihr Entgegenkommen bei der Umsetzung der Wünsche des Autors.

Berlin, Frühjahr 1984 Jürgen Bortz

Döring, Bortz: Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften
Der Wegweiser zu diesem Lehrbuch

Was erwartet mich? **Lernziele** zeigen, worauf es im Folgenden ankommt

Griffregister: Zur schnellen Orientierung

Definitionen: Zentrale Fachbegriffe werden in deutscher und englischer Sprache eingeführt sowie kurz und knapp erläutert

Wichtig: Hervorgehobene Merksätze

Fit für die Prüfung? **Lernquiz** und **Übungsaufgaben** – Die Lösungen der Lernquiz finden Sie im Web

10.4 Psychologischer Test

Lernziele
- Wissen, was wissenschaftliche psychologische Tests von nicht-wissenschaftlichen „Psychotests" unterscheidet.
- Verschiedene Arten projektiver und psychometrischer psychologischer Tests im Hinblick auf Zielsetzung und Anwendungsfelder voneinander abgrenzen können.

Anstelle des Mittelwertes wird auch bei kardinalskalierten Variablen der Schwierigkeitsindex gern als Prozentwert angegeben. Dieser **prozentuale Schwierigkeitsindex** P_i eines **intervallskalierten Items** i ist (sofern die Antwortskala mit dem Messwert 0 beginnt) einfach der Quotient aus der bei diesem Item tatsächlich erreichten Punktsumme aller n Testpersonen und der maximal erreichbaren Punktsumme, multipliziert mit 100 (Kelava & Moosbrugger, 2012, S. 76f.) und berechnet sich nach (Gl. 10.12)

$$P_i = \frac{\sum_{v=1}^{n} x_{vi}}{n \cdot \max(x_i)} \cdot 100. \qquad (10.12)$$

$\sum x_{vi}$ = Punktsumme aller Testpersonen beim Item i
$\max(x_i)$ = maximal erreichbare Punktsumme

Beginnt die Antwortskala des Items nicht mit Wert 0, sondern mit einem anderen Wert (z. B. 1), so muss eine verallgemeinerte Umrechnungsformel genutzt werden, um kardinalskalierte Itemantworten in einen prozentualen Schwierigkeitsindex zu überführen (Kelava & Moosbrugger, 2012, S. 77):

$$P_i = \frac{\sum_{v=1}^{n} [x_{vi} - \min(x_i)]}{n \cdot [\max(x_i) - \min(x_i)]} \cdot 100. \qquad (10.13)$$

x_{vi} = Punktsumme aller Testpersonen beim Item i
$\max(x_i)$ = maximal erreichbare Punktsumme
$\min(x_i)$ = minimal erreichbare Punktsumme

Die Klassische sowie die Probabilistische Testtheorie lassen sich nur anwenden, wenn *mehrere Items* verwendet werden, die alle *dasselbe Konstrukt* messen und zu einem Test- bzw. Skalenwert zusammengefasst werden. Nicht anwendbar ist die Testtheorie, wenn als Messinstrumente Einzelitems bzw. Einzelindikatoren verwendet werden.

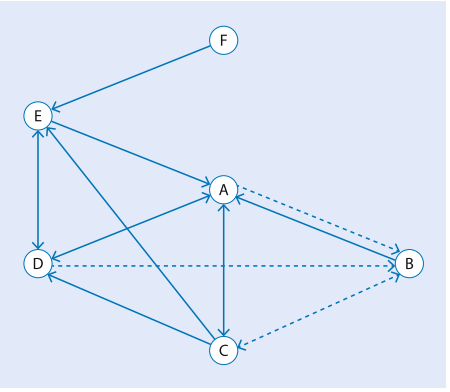


Abb. 10.13 Soziogramm einer 6-köpfigen Gruppe; die *gestrichelten Pfeile* stellen negative Wahlen dar

10.4.4 Testtheorie

Eine Pumpe füllt einen Behälter, der 40 Liter fasst, in 5 Minuten. Wie lange benötigt die Pumpe, um einen Behälter mit 64 Liter zu füllen? Auf diese Frage gibt eine Testperson die richtige Antwort: 8 Minuten. Kann man aufgrund dieser einen Antwort behaupten, die Testperson verfüge über eine gute mathematische Denkfähigkeit? Sicherlich nicht. Es leuchtet intuitiv ein, dass diese Informationsbasis nicht ausreicht, um entscheiden zu können, ob diese Frage „mathematische Denkfähigkeit" oder etwas anderes misst. Es bleibt offen, wie viel Zeit die Testperson zur Lösung dieser Aufgabe beanspruchte, ob sie nur zufällig eine richtige Schätzung abgab, ob sie ähnliche oder auch schwerere Aufgaben lösen könnte und vieles mehr.

Wissenschaftliche Beobachtung – Unter einer wissenschaftlichen Beobachtung („scientific observation") versteht man die zielgerichtete, systematische und regelgeleitete Erfassung, Dokumentation und Interpretation von Merkmalen, Ereignissen oder Verhaltensweisen mithilfe menschlicher Sinnesorgane und/oder technischer Sensoren zum Zeitpunkt ihres Auftretens.

Lernquiz
1. Nennen Sie die Haupt- und Nebengütekriterien psychometrischer Tests!
2. Welche Techniken zur Abschätzung der Reliabilität eines psychometrischen Tests werden verwendet?
3. Was versteht man unter einer Item-Charakteristik und wie wird sie dargestellt?

Itemanalyse der Neurotizismus-Subskala aus dem NEO-FFI

Die **Neurotizismus-Subskala** aus dem Persönlichkeitstest NEO-FFI (Borkenau & Ostendorf, 2008; ▶ Abschn. 10.4.3) erfasst mit emotionaler Labilität eine der fünf zentralen Persönlichkeitseigenschaften. Alle Items sind jeweils auf einer **5-stufigen Ratingskala** zu beantworten (starke Ablehnung: 0, Ablehnung: 1, neutral 2; Zustimmung: 3; starke Zustimmung: 4). Die mittels Ratingskala erhobenen Messwerte werden als **intervallskaliert** aufgefasst (▶ Abschn. 8.4.4; mithilfe der Probabilistischen Testtheorie könnte diese Skalierungs-Annahme auch empirisch geprüft werden).

Im Folgenden soll an einem empirischen Datensatz demonstriert werden (Datensatz aus Pöschl, 2010), wie die Schwierigkeiten und Trennschärfen der einzelnen Items der Testskala ausfallen und inwiefern die Itemanalyse zur Verbesserung der Testgüte genutzt werden kann.

Studienbeispiel

Erhebung persönlicher Netzwerke mithilfe standardisierter Fragebögen

Fragestellung: Angenommen man interessiert sich dafür, wie sich die **persönlichen Netzwerke** von Personen verschiedener Altersgruppen unterscheiden und zwar hinsichtlich ihrer **Größe** (Anzahl der Personen im Netzwerk), ihrer **Zusammensetzung** (z. B. Anteil von Verwandten im Netzwerk, Anteil von Personen mit ähnlichen oder abweichenden politischen Meinungen) und ihrer **Dichte** (Intensität der Beziehungen zwischen den Netzwerkmitgliedern).

Tabelle 10.9 Anwendung des Schwedenschlüssels auf einen Beispielhaushalt mit sechs Personen – das Verhältnis der Haushaltsmitglieder zum männlichen Haushaltsvorstand ergibt laut Kish-Verfahren die Erwachsenen-Nr. von 1 bis 6, die zur Auswahl genutzte Erwachsenen-Nr. ergibt sich aus den vorgefertigten Auswahltabellen (die bei einem Haushalt mit 6 Personen z. B. Person 4 vorgeben)

Beziehung	Geschlecht	Alter	Erwachsenen-Nr. laut Kish-System	Auswahl laut Auswahltabelle
Mann („Haushaltsvorstand“)	M		2	
Frau	W	40	5	
Vater des Mannes	M		1	
Sohn	M		3	
Tochter	W		6	
Tante der Frau	W	44	4	X

Übungsaufgaben

1. Recherchieren Sie verschiedene Intelligenztests für Erwachsene – z. B. aus dem Brickenkamp Handbuch psychologischer und pädagogischer Tests (Brähler et al., 2002a, 2002b) oder bei der Hogrefe Testzentrale (http://www.testzentrale.de). Vergleichen Sie deren Validität, Reliabilität, Testdauer, Verfügbarkeit von Normwerten sowie Kosten und treffen Sie eine Entscheidung, welchen Test Sie anhand dieser Kriterien in einer eigenen Studie einsetzen würden. Begründen Sie Ihre Auswahl!

Weiterführende Literatur

- Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg: Springer.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.

Literatur

Adams, J., Parkinson, L., Sanson-Fisher, R. W., & Walsh, R. A. (2008). Enhancing Self-Report of Adolescent Smoking: The Effects of Bogus Pipeline and Anonymity. *Addictive Behaviors, 33*(10), 1291–1296.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1995). *Standards for educational and psychological testing* (4th ed.). Washington: AERA, APA, NCME.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing* (5th ed.). Washington: AERA, APA, NCME.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing* (6th ed.) Washington: AERA, APA, NCME.

Navigation: Mit Seitenzahl und Kapitelnummer

Boxen: Exkurse und Anwendungsbeispiele veranschaulichen das Wissen

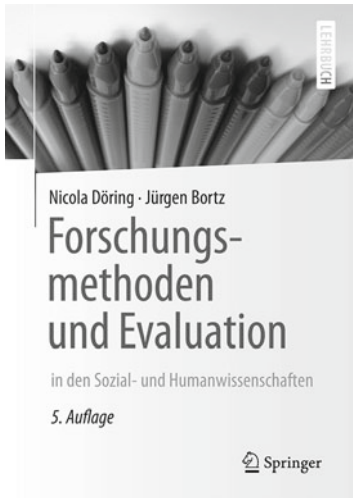
Studienbeispiele: Internationale Beispiele für empirische Studien

Anschaulich: Mit 194 **Abbildungen** und 167 **Tabellen**

Noch nicht genug? Tipps für die **Weiterführende Lektüre**

Website zum Buch auf **www.lehrbuch-psychologie.de**

Lernmaterialien zu *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* im Internet – www.lehrbuch-psychologie.de



- **Alles für die Lehre – fertig zum Download: Foliensätze, Abbildungen und Tabellen** für Dozentinnen und Dozenten zum Download
- Schnelles Nachschlagen: **Glossar** mit zahlreichen Fachbegriffen und englischer Übersetzung
- **Zusammenfassungen aller 18 Buchkapitel:** Das steckt drin im Lehrbuch
- **Memocards** und **Verständnisfragen** mit Antworten
- PDF mit Lösungen zu den **Lernquiz** des Buches zum Download
- **Tabellen statistischer Prüfgrößen**
- **Web-Links**

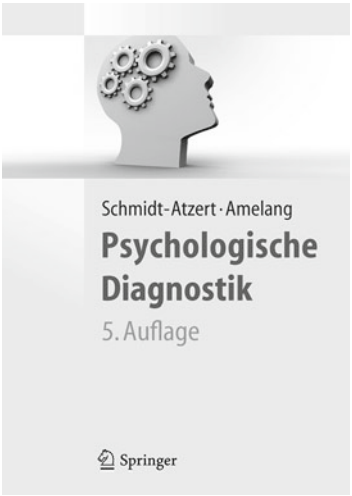
Weitere Websites unter ▸ www.lehrbuch-psychologie.de



- Rechnen mit SPSS und R: Syntax-Kommentare zur Berechnung der Software
- Glossar der wichtigsten Fachbegriffe
- Zusammenfassungen der 28 Buchkapitel
- Memocards: Überprüfen Sie Ihr Wissen
- Dozentenmaterialien: Vorlesungsfolien, Abbildungen und Tabellen



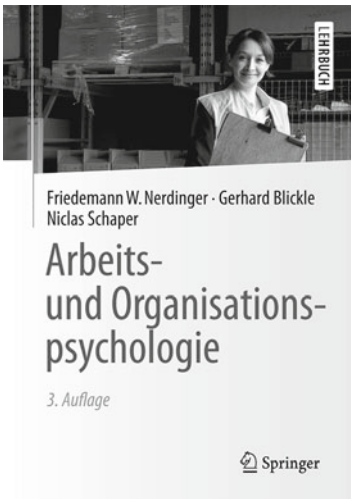
- Hinweise und Anleitungen zur Nutzung von SPSS oder LISREL als kostenlose Zusatzkapitel mit Datensätzen
- Glossar mit zahlreichen Fachbegriffen
- Memocards und Verständnisfragen
- Kommentierte Linksammlung
- Dozentenmaterialien: Prüfungsfragen, Abbildungen und Tabellen



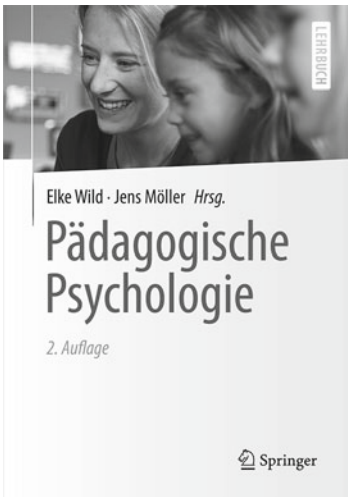
- Zusammenfassungen der 11 Buchkapitel
- Glossar der wichtigsten Fachbegriffe
- Memocards und Verständnisfragen: Überprüfen Sie Ihr Wissen
- Kommentierte Linksammlung
- Dozentenmaterialien: Vorlesungsfolien, Prüfungsfragen, Abbildungen und Tabellen



- Glossar mit zahlreichen Fachbegriffen
- Memocards: Überprüfen Sie Ihr Wissen
- Kapitelzusammenfassungen
- Prüfungsfragen & Antworten: Üben Sie für die Prüfung
- Dozentenmaterialien: Abbildungen und Tabellen



- Deutsch-englisches Glossar mit zahlreichen Fachbegriffen
- Memocards: Fachbegriffe pauken
- Kommentierte Linksammlung
- Kleine Phraseologie des Business-Neusprech
- Dozentenmaterialien: Abbildungen und Tabellen



- Verständnisfragen und -antworten
- Glossar mit zahlreichen Fachbegriffen
- Memocards
- Kapitelzusammenfassungen
- Dozentenmaterialien: Folien, Abbildungen und Tabellen

Einfach lesen, hören, lernen im Web – ganz ohne Registrierung!

Fragen? redaktion@lehrbuch-psychologie.de

Inhaltsverzeichnis

I Grundlagen

1 **Empirische Sozialforschung im Überblick** ... 3
Nicola Döring, Jürgen Bortz

1.1 **Sozialwissenschaftlicher Erkenntnisgewinn** ... 4

1.1.1 Sozialwissenschaftliche Forschungs- und Methodenkompetenz ... 4

1.1.2 Sozial- und Humanwissenschaften zwischen Geistes- und Naturwissenschaften ... 12

1.1.3 Forschungsansätze in der empirischen Sozialforschung ... 14

1.2 **Empirische Studien verstehen und beurteilen** ... 19

1.2.1 Primär- und Sekundärquellen für empirische Studien ... 19

1.2.2 Wissenschaftliche Zeitschriftenartikel lesen ... 20

1.3 **Empirische Studien planen und durchführen** ... 22

1.3.1 Der Forschungsprozess in der quantitativen Sozialforschung ... 23

1.3.2 Der Forschungsprozess in der qualitativen Sozialforschung ... 25

1.3.3 Forschungsprozesse in der Mixed-Methods-Sozialforschung ... 26

Literatur ... 29

2 **Wissenschaftstheoretische Grundlagen der empirischen Sozialforschung** ... 31
Nicola Döring, Jürgen Bortz

2.1 **Wissenschaftstheorie im Überblick** ... 34

2.1.1 Wissenschaftstheoretische Grundbegriffe ... 34

2.1.2 Explizite und implizite Wissenschaftstheorien ... 36

2.2 **Das quantitative Paradigma und der Kritische Rationalismus** ... 36

2.2.1 Rationalismus statt Empirismus ... 37

2.2.2 Wahrheitssuche statt Wahrheitsbehauptung ... 38

2.2.3 Kritischer statt naiver Realismus ... 40

2.2.4 Grad der Falsifizierbarkeit und Informationsgehalt von Theorien ... 43

2.2.5 Raffinierter statt einfacher methodologischer Falsifikationismus ... 43

2.2.6 Entdeckungs-, Begründungs- und Verwendungszusammenhänge von Theorien ... 45

2.2.7 Ethische Verantwortung und Objektivität in der Forschung ... 46

2.2.8 Deduktiv-nomologische Erklärungen der Realität ... 48

2.2.9 Datengewinnung und statistische Hypothesenprüfung ... 52

2.2.10 Prüfung und Weiterentwicklung von Theorien ... 54

2.2.11 Grenzen des Kritischen Rationalismus ... 57

2.2.12 Neuere wissenschaftstheoretische Ansätze im quantitativen Paradigma ... 62

2.3 **Das qualitative Paradigma und der Sozialkonstruktivismus** ... 63
2.3.1 Erstes Prinzip: Ganzheitliche und rekonstruktive Untersuchung lebensweltlicher Phänomene ... 64
2.3.2 Zweites Prinzip: Reflektierte theoretische Offenheit zwecks Bildung neuer Theorien ... 66
2.3.3 Drittes Prinzip: Zirkularität und Flexibilität des Forschungsprozesses zwecks Annäherung an den Gegenstand ... 67
2.3.4 Viertes Prinzip: Forschung als Kommunikation und Kooperation zwischen Forschenden und Beforschten 68
2.3.5 Fünftes Prinzip: Selbstreflexion der Subjektivität und Perspektivität der Forschenden ... 70
2.3.6 Der Sozialkonstruktivismus als wissenschaftstheoretische Basis des qualitativen Paradigmas ... 71
2.4 **Das Mixed-Methods-Paradigma und der Pragmatismus** ... 72
2.4.1 Fünf Positionen zu Mixed-Methods als drittem Paradigma ... 73
2.4.2 Der Pragmatismus als wissenschaftstheoretische Basis eines Mixed-Methods-Paradigmas ... 74
Literatur ... 78

3 **Qualitätskriterien in der empirischen Sozialforschung** ... 81
Nicola Döring, Jürgen Bortz
3.1 **Wissenschaftlichkeit und wissenschaftliche Qualität** ... 84
3.1.1 Vier Standards der Wissenschaftlichkeit ... 85
3.1.2 Vier Kriterien der wissenschaftlichen Qualität ... 89
3.2 **Gütekriterien für quantitative Forschung** ... 93
3.2.1 Quantitative Gütekriterien im Überblick ... 93
3.2.2 Vier Typen der Validität in der Campbell-Tradition ... 93
3.3 **Gütekriterien für qualitative Forschung** ... 106
3.3.1 Kontroversen über Gütekriterien qualitativer Forschung ... 106
3.3.2 Vier Kriterien der Glaubwürdigkeit von Lincoln und Guba (1985) ... 108
3.3.3 Sieben Kernkriterien zur Bewertung qualitativer Forschung nach Steinke (1999) ... 111
3.4 **Gütekriterien für Mixed-Methods-Forschung** ... 114
Literatur ... 117

4 **Forschungs- und Wissenschaftsethik** ... 121
Nicola Döring, Jürgen Bortz
4.1 **Forschungsethik** ... 123
4.1.1 Ethischer Umgang mit Untersuchungspersonen ... 123
4.1.2 Ethische Richtlinien der einzelnen Wissenschaftsdisziplinen ... 128
4.1.3 Einhaltung ethischer Richtlinien ... 130
4.2 **Wissenschaftsethik** ... 132
4.2.1 Regeln guter wissenschaftlicher Praxis ... 132
4.2.2 Wissenschaftsfälschungen ... 133
4.2.3 Plagiarismus in der Wissenschaft ... 135
4.2.4 Wissenschaftliche Autorschaft ... 136
Literatur ... 138

II Anwendung

5 **Forschungsthema** ... 143
Nicola Döring, Jürgen Bortz

5.1 **Vom Forschungsthema zum Forschungsproblem** ... 144
5.2 **Forschungshypothesen und Forschungsfragen** ... 145
5.3 **Sechs Kriterien für die Wahl des eigenen Forschungsthemas** ... 149
5.3.1 Persönliches Interesse am Thema ... 149
5.3.2 Theoretische und methodische Vorkenntnisse zum Thema ... 150
5.3.3 Wissenschaftliche Relevanz des Themas ... 150
5.3.4 Praktische Relevanz des Themas ... 150
5.3.5 Empirische Untersuchbarkeit des Themas ... 151
5.3.6 Betreuungsangebot zum Thema ... 153
5.4 **Planung des eigenen Forschungsprojektes** ... 153
Literatur ... 155

6 **Forschungsstand und theoretischer Hintergrund** ... 157
Nicola Döring, Jürgen Bortz

6.1 **Literaturrecherche** ... 158
6.1.1 Suchbegriffe ... 158
6.1.2 Datenbanken ... 159
6.1.3 Datenbankabfrage und Schneeballsystem ... 160
6.1.4 Sichtung der Fundstellen ... 161
6.1.5 Literaturbeschaffung ... 162
6.2 **Forschungsstand** ... 163
6.3 **Theoretischer Hintergrund** ... 165
6.3.1 Durchführung einer reinen Theoriearbeit ... 166
6.3.2 Empirische Überprüfung und Modifikation einer einzelnen Theorie ... 168
6.3.3 Formalisierung und Überprüfung einer Theorie mittels Computersimulation ... 169
6.3.4 Verbindung mehrerer Theorien zu einem integrativen Theoriemodell ... 170
6.3.5 Gegenstandsverankerte Theoriebildung ... 170
6.3.6 Unterschiedliche Strategien zur Entwicklung neuer Theorien ... 172
6.4 **Ableitung von Forschungsfragen und Forschungshypothesen** ... 173
6.4.1 Ableitung von Forschungsfragen aus dem Forschungsstand ... 174
6.4.2 Ableitung von Forschungsfragen aus einer oder mehreren Theorien ... 174
6.4.3 Ableitung von Forschungsfragen und/oder Forschungshypothesen aus Alltagstheorien ... 175
6.4.4 Ableitung von Forschungshypothesen aus einer Theorie ... 175
6.4.5 Ableitung von Forschungshypothesen aus mehreren Theorien ... 176
Literatur ... 177

7 **Untersuchungsdesign** ... 181
Nicola Döring, Jürgen Bortz
7.1 **Quantitative, qualitative oder Mixed-Methods-Studie** ... 184
7.2 **Grundlagen- oder anwendungswissenschaftliche Studie** ... 185
7.3 **Empirische, Methoden- oder Theoriestudie** ... 186
7.4 **Primär-, Sekundär- oder Metaanalyse** ... 191
7.5 **Explorative, deskriptive oder explanative Studie** ... 192
7.6 **Experimentelle, quasi-experimentelle oder nicht-experimentelle Studie** ... 193
7.6.1 Die experimentelle Studie ... 194
7.6.2 Die quasi-experimentelle Studie ... 199
7.6.3 Die nicht-experimentelle Studie ... 201
7.7 **Labor- oder Feldstudie** ... 205
7.8 **Studie mit oder ohne Messwiederholungen** ... 208
7.8.1 (Quasi-)Experimentelle Studien mit und ohne Messwiederholungen ... 209
7.8.2 Nicht-experimentelle Studien mit und ohne Messwiederholungen ... 210
7.9 **Gruppen- oder Einzelfallstudie** ... 214
Literatur ... 218

8 **Operationalisierung** ... 221
Nicola Döring, Jürgen Bortz
8.1 **Theoretische Konzepte in der qualitativen und quantitativen Forschung** ... 222
8.1.1 Induktive Bildung theoretischer Konzepte ... 222
8.1.2 Deduktive Definition theoretischer Konzepte ... 223
8.2 **Deduktive Konzeptspezifikation und Variablen** ... 223
8.2.1 Nominaldefinition von theoretischen Konzepten ... 224
8.2.2 Bedeutungsanalyse von etablierten theoretischen Konzepten ... 226
8.2.3 Dimensionale Analyse von neuen theoretischen Konzepten ... 226
8.3 **Grundlagen zur Durchführung einer Operationalisierung** ... 228
8.3.1 Operationale Definition von theoretischen Konzepten ... 228
8.3.2 Operationalisierung von abhängigen Variablen ... 230
8.3.3 Operationalisierung von unabhängigen Variablen ... 231
8.3.4 Fehlinterpretation von Operationalisierungen ... 232
8.4 **Messung und die vier Skalenniveaus** ... 232
8.4.1 Messung ... 235
8.4.2 Nominalskala ... 237
8.4.3 Ordinalskala ... 239
8.4.4 Intervallskala ... 244
8.4.5 Verhältnisskala ... 256
8.4.6 Skalentransformationen ... 257

8.5 **Messung mit Einzelindikatoren** ... 264
8.5.1 Messung manifester Variablen mit Einzelindikatoren ... 265
8.5.2 Messung latenter Variablen mit Einzelindikatoren ... 265
8.6 **Messung mit psychometrischen Skalen und Skalenbildung** ... 267
8.6.1 Likert-Skala ... 269
8.6.2 Guttman-Skala ... 272
8.6.3 Thurstone-Skala ... 274
8.6.4 Semantisches Differenzial ... 276
8.7 **Messung mit Indizes und Indexbildung** ... 277
8.7.1 Auswahl und Art der Indikatoren ... 278
8.7.2 Standardisierung der Indikatoren ... 279
8.7.3 Zusammenfassung der Indikatoren ... 280
8.7.4 Gewichtung der Indikatoren ... 281
8.7.5 Index als standardisierter Wert ... 282
Literatur ... 283

9 **Stichprobenziehung** ... 291
Nicola Döring, Jürgen Bortz
9.1 **Population und Stichprobe** ... 292
9.1.1 Vollerhebung der Population ... 292
9.1.2 Teilerhebung der Population: Stichprobenverfahren ... 294
9.1.3 Repräsentativität von Stichproben ... 297
9.2 **Stichproben für qualitative Studien** ... 302
9.2.1 Theoretische Stichprobe ... 302
9.2.2 Stichprobe gemäß qualitativem Stichprobenplan ... 303
9.2.3 Gezielte Auswahl bestimmter Arten von Fällen ... 304
9.3 **Stichproben für quantitative Studien** ... 305
9.3.1 Nicht-probabilistische Stichproben ... 305
9.3.2 Probabilistische Stichproben ... 310
Literatur ... 318

10 **Datenerhebung** ... 321
Nicola Döring, Jürgen Bortz
10.1 **Beobachtung** ... 323
10.1.1 Wissenschaftliche Beobachtung ... 323
10.1.2 Qualitative Beobachtung ... 332
10.1.3 Quantitative Beobachtung ... 342
10.2 **Interview** ... 356
10.2.1 Wissenschaftliches Interview ... 356
10.2.2 Qualitatives Interview ... 365
10.2.3 Quantitatives Interview ... 381

10.3 **Fragebogen** ... 398
10.3.1 Die wissenschaftliche Fragebogenmethode ... 398
10.3.2 Qualitative Fragebogenmethode ... 401
10.3.3 Quantitative Fragebogenmethode ... 405
10.4 **Psychologischer Test** ... 430
10.4.1 Wissenschaftliche psychologische Tests ... 431
10.4.2 Projektive Tests ... 450
10.4.3 Psychometrische Tests ... 453
10.4.4 Testtheorie ... 461
10.5 **Physiologische Messungen** ... 499
10.5.1 Wissenschaftliche physiologische Messung ... 500
10.5.2 Hirnaktivität ... 508
10.5.3 Herz-Kreislauf-Aktivität ... 514
10.5.4 Elektrodermale Aktivität ... 517
10.5.5 Muskuläre Aktivität ... 519
10.5.6 Augenaktivität ... 522
10.5.7 Messung sexueller Erregung ... 524
10.6 **Dokumentenanalyse** ... 533
10.6.1 Wissenschaftliche Dokumentenanalyse ... 536
10.6.2 Qualitative Dokumentenanalyse ... 540
10.6.3 Quantitative Dokumentenanalyse ... 552

11 **Datenaufbereitung** ... 579
Nicola Döring, Jürgen Bortz
11.1 **Definition der Datenaufbereitung** ... 580
11.2 **Funktionen und Bedeutung der Datenaufbereitung** ... 581
11.3 **Aufbereitung qualitativer Daten** ... 582
11.3.1 Erstellung und Kommentierung qualitativer Datensätze ... 582
11.3.2 Umfang, Art und Überprüfung der Transkription ... 583
11.3.3 Anonymisierung qualitativer Datensätze ... 584
11.4 **Aufbereitung quantitativer Daten** ... 584
11.4.1 Erstellung und Kommentierung quantitativer Datensätze ... 585
11.4.2 Anonymisierung quantitativer Datensätze ... 589
11.4.3 Datenbereinigung bei quantitativen Datensätzen ... 589
11.4.4 Datentransformation bei quantitativen Datensätzen ... 590
Literatur ... 594

12 **Datenanalyse** ... 597
Nicola Döring, Jürgen Bortz
12.1 **Qualitative Datenanalyse** ... 599
12.1.1 Verschiedene Ansätze der qualitativen Datenanalyse ... 600
12.1.2 Zentrale Arbeitsschritte der qualitativen Datenanalyse ... 603
12.1.3 Computerprogramme für die qualitative Datenanalyse ... 607
12.2 **Quantitative Datenanalyse** ... 612
12.2.1 Verschiedene Ansätze der quantitativen Datenanalyse ... 612
12.2.2 Zentrale Arbeitsschritte der quantitativen Datenanalyse ... 616
12.2.3 Computerprogramme für die quantitative Datenanalyse ... 618
12.3 **Quantitative Datenanalyse bei explorativen Studien** ... 621
12.3.1 Deskriptivstatistische und explorative Datenanalyse ... 621
12.3.2 Inferenzstatistische Überprüfung von Forschungsfragen ... 627
12.3.3 Data-Mining ... 628
12.4 **Quantitative Datenanalyse bei deskriptiven Studien** ... 631
12.4.1 Punktschätzung von Populationsparametern ... 631
12.4.2 Intervallschätzung von Populationsparametern ... 640
12.4.3 Präzisierung von Parameterschätzungen ... 652
12.5 **Quantitative Datenanalyse bei explanativen Studien** ... 657
12.5.1 Hypothesenprüfung mittels klassischem Signifikanztest ... 659
12.5.2 Signifikanztests für Zusammenhangshypothesen ... 677
12.5.3 Signifikanztests für Unterschiedshypothesen ... 705
12.5.4 Signifikanztests für Veränderungshypothesen ... 727
12.5.5 Signifikanztests für Einzelfallhypothesen ... 765

13 **Ergebnispräsentation** ... 785
Nicola Döring, Jürgen Bortz
13.1 **Zeitschriftenartikel** ... 787
13.1.1 Auswahl der Zeitschrift ... 787
13.1.2 Autorenrichtlinien ... 788
13.1.3 Gliederung des Artikels ... 788
13.1.4 Auswahl der Inhalte ... 788
13.1.5 Schreibprozess ... 789
13.1.6 Literaturverwaltung und Zitation ... 789
13.1.7 Formatierung ... 790
13.1.8 Tabellen und Abbildungen ... 790
13.1.9 Korrektur des Manuskriptes ... 790
13.1.10 Artikeleinreichung und Peer Review ... 791
13.1.11 Verbreitung des Artikels ... 792

13.2 **Konferenzvortrag** ... 792
13.2.1 Auswahl der Konferenz ... 793
13.2.2 Abstract ... 794
13.2.3 Beitragseinreichung und Peer Review ... 794
13.2.4 Vortragsplanung ... 794
13.2.5 Präsentationsfolien und Zusatzmaterial ... 795
13.2.6 Üben des Vortrags ... 796
13.2.7 Vortrag und Diskussion ... 797
13.2.8 Networking ... 798
13.2.9 Auszeichnungen für Konferenzvorträge ... 799
13.3 **Wissenschaftliches Poster** ... 799
13.3.1 Inhalte des Posters ... 799
13.3.2 Gestaltung des Posters ... 799
13.3.3 Zusatzmaterial zum Poster ... 800
13.3.4 Erstellung des Posters ... 800
13.3.5 Präsentation des Posters ... 800
13.3.6 Auszeichnungen für wissenschaftliche Poster ... 800
13.3.7 Weiterverwertung des Posters ... 801
13.4 **Öffentlichkeitsarbeit** ... 801
13.4.1 Wissenschaftskommunikation im Internet ... 801
13.4.2 Umgang mit Journalisten ... 802
13.4.3 Beiträge in Transferzeitschriften ... 802
13.4.4 Populärwissenschaftliche Publikationen ... 802
13.4.5 Eingeladene Vorträge ... 803
13.4.6 Präsentation auf Messen ... 803
Literatur ... 804

III Vertiefung

14 **Bestimmung von Teststärke, Effektgröße und optimalem Stichprobenumfang** ... 807
Nicola Döring, Jürgen Bortz
14.1 **Teststärkeanalysen** ... 809
14.1.1 Die Post-hoc-Teststärkeanalyse ... 809
14.1.2 Die A-priori-Teststärkeanalyse ... 814
14.1.3 Weitere Typen der Teststärkeanalyse ... 815
14.2 **Ergebnisse quantitativer Studien in Form von standardisierten Effektgrößen berichten** ... 816
14.2.1 Verschiedene Typen von Effektgrößenmaßen ... 816
14.2.2 Klassifikation von Effektgrößen: kleine, mittlere und große Effekte ... 819
14.2.3 Standardisierte Effektgrößen der wichtigsten Signifikanztests und deren Konfidenzintervalle ... 821

14.3 **Quantitative Studien mit optimalen Stichprobenumfängen planen** ... 840
14.3.1 Vorgehen bei der Festlegung des optimalen Stichprobenumfanges ... 841
14.3.2 Optimale Stichprobenumfänge der wichtigsten Signifikanztests ... 842
14.4 **Beispiele für die Planung optimaler Stichprobenumfänge und das Berichten von Effektgrößen** 849
14.4.1 Vergleich von zwei Mittelwerten ... 850
14.4.2 Bivariate Korrelation ... 853
14.4.3 Korrelationsdifferenz ... 854
14.4.4 Differenz eines Anteilswertes von .50 ... 854
14.4.5 Differenz zweier Anteilswerte ... 855
14.4.6 χ^2-Test ... 856
14.4.7 Varianzanalysen ... 857
14.4.8 Multiple Korrelation ... 862
Literatur ... 864

15 **Minimum-Effektgrößen-Tests** ... 867
Nicola Döring, Jürgen Bortz
15.1 **Das Good-Enough-Prinzip** ... 869
15.2 **Überprüfung von Minimum-Effekt-Nullhypothesen** ... 870
15.2.1 Signifikanzschranken und Teststärkeanalysen ... 871
15.2.2 Hinweise zur Untersuchungsplanung ... 876
15.2.3 Transformation statistischer Test- und Kennwerte in die F-Statistik ... 878
15.3 **Zur Frage der „Bestätigung" von Nullhypothesen** ... 885
15.3.1 Beispiele für Wunsch-Nullhypothesen ... 888
15.3.2 Alternative Prüfung von Wunsch-Nullhypothesen ... 890
15.3.3 Modellanpassungstests und Wunsch-Nullhypothesen ... 890
Literatur ... 892

16 **Metaanalyse** ... 893
Nicola Döring, Jürgen Bortz
16.1 **Quantitative Metaanalyse** ... 895
16.1.1 Definition und Zielsetzung von Metaanalysen ... 895
16.1.2 Abgrenzung zwischen Metaanalyse und narrativem Review ... 898
16.1.3 Probleme bei der Durchführung von Metaanalysen und Gegenmaßnahmen ... 900
16.2 **Arbeitsschritte einer Metaanalyse** ... 904
16.2.1 Formulierung der Forschungsfrage(n) der Metaanalyse ... 905
16.2.2 Suche nach inhaltlich einschlägigen Primärstudien ... 906
16.2.3 Auswahl und Kodierung der methodisch geeigneten Primärstudien ... 910
16.2.4 Entscheidung für ein metaanalytisches Modell ... 913
16.2.5 Berechnung einheitlicher Effektgrößenmaße für alle Primärstudien ... 914
16.2.6 Berechnung und Signifikanzprüfung des Gesamteffektes ... 918
16.2.7 Prüfung der Homogenität der einzelnen Effektgrößen der Primärstudien ... 919
16.2.8 Moderatorvariablen-/Subgruppen-Analyse bei Heterogenität der Effektgrößen ... 921
16.2.9 Präsentation und Interpretation der Ergebnisse der Metaanalyse ... 922

16.3 **Aggregierung von Primärstudien ohne gemeinsame Effektgrößenschätzung** ... 924
16.3.1 Aggregierung statistisch signifikanter Ergebnisse ... 925
16.3.2 Aggregierung positiver (deskriptiv-hypothesenkonformer) Ergebnisse ... 926
16.3.3 Aggregierung exakter Irrtumswahrscheinlichkeiten ... 926
16.4 **Beispiel: Eine Fixed-Effect-Metaanalyse zum Lehrererwartungs-Effekt** ... 927
16.4.1 Die Primärstudien: Fünf Untersuchungen zum Lehrererwartungs-Effekt ... 928
16.4.2 Berechnung der Δ-Maße für alle fünf Studien ... 930
16.4.3 Berechnung und Signifikanzprüfung des Lehrererwartungs-Gesamteffektes ... 931
16.4.4 Prüfung der Homogenität der fünf Δ-Maße ... 934
16.4.5 Moderatorvariablen-Analyse wegen Heterogenität der einzelnen Effektgrößen ... 935
16.4.6 Interpretation der Ergebnisse der fiktiven Metaanalyse zum Lehrererwartungs-Effekt ... 937
16.5 **Software für Metaanalysen** ... 937
16.5.1 Metaanalyse-Add-Ons für Excel ... 938
16.5.2 Metaanalyse-Makros für Statistik-Programmpakete ... 938
16.5.3 Spezielle Metaanalyse-Programme ... 938
Literatur ... 940

17 **Strukturgleichungsmodelle** ... 945
Christina S. Werner, Karin Schermelleh-Engel, Carla Gerhard, Jana C. Gäde
17.1 **Motivation: Wozu Strukturgleichungsmodelle?** ... 946
17.2 **Aufbau von Strukturgleichungsmodellen** ... 949
17.2.1 Messung von Konstrukten: Faktorenanalyse ... 949
17.2.2 Effekte zwischen Konstrukten: Regressions-/Pfadanalyse ... 952
17.3 **Analyse von Strukturgleichungsmodellen** ... 955
17.3.1 Kovarianzstrukturanalyse ... 955
17.3.2 Schätzmethoden ... 956
17.3.3 Modellgüte und Test der Annahmen ... 957
17.4 **Anwendungsbeispiel: Ein Strukturgleichungsmodell zu Burnout** ... 958
17.5 **Praktischer Einsatz, Herausforderungen und Probleme** ... 963
17.5.1 Erklärung vs. Vorhersage von Variablen ... 963
17.5.2 Modellidentifikation ... 964
17.5.3 Schätzprobleme ... 965
17.5.4 Ergebnisinterpretation ... 965
17.5.5 Voraussetzungen ... 966
17.5.6 Kausalaussagen ... 968
17.6 **Software für Strukturgleichungsmodelle** ... 969
Literatur ... 972

18 **Evaluationsforschung** ... 975
Nicola Döring, Jürgen Bortz
18.1 **Evaluationsforschung im Überblick** ... 977
18.1.1 Abgrenzung von Grundlagen-, Interventions- und Evaluationsforschung ... 977
18.1.2 Definition der Evaluationsforschung ... 977
18.1.3 Der Evaluationsgegenstand ... 979
18.1.4 Die Anspruchsgruppen ... 981
18.1.5 Die Evaluationskriterien ... 983
18.1.6 Die Evaluationsfunktionen ... 987
18.1.7 Die Evaluationsnutzung ... 988
18.1.8 Wichtige Typen von Evaluationsstudien ... 989
18.1.9 Die Evaluationsstandards ... 991
18.1.10 Anforderungen an Evaluierende ... 992
18.1.11 Fachinformationen zur Evaluationsforschung ... 994
18.2 **Verschiedene Ansätze in der Evaluationsforschung** ... 995
18.2.1 Der ergebnisorientierte Evaluationsansatz ... 997
18.2.2 Der systemische Evaluationsansatz ... 1007
18.2.3 Der theorieorientierte Evaluationsansatz ... 1008
18.2.4 Der akteursorientierte Evaluationsansatz ... 1012
18.3 **Phasen der Evaluationsforschung** ... 1017
18.3.1 Initiierungsphase der Evaluation ... 1017
18.3.2 Konzeptionsphase der Evaluation ... 1019
18.3.3 Planungsphase der Evaluation ... 1021
18.3.4 Realisierungsphase der Evaluation ... 1026
18.3.5 Abschlussphase der Evaluation ... 1028
Literatur ... 1033

Serviceteil ... 1037
Anhang ... 1038
Sachverzeichnis ... 1042

Autorenverzeichnis

Bortz, Jürgen, Prof. Dr. †
ehemals Abteilung für Psychologische Methodenlehre am Institut für Psychologie, Technische Universität Berlin, Franklinstraße 28–29, 10587 Berlin, Deutschland

Döring, Nicola, Prof. Dr.
Fachgebiet Medienpsychologie und Medienkonzeption am Institut für Medien und Kommunikationswissenschaft, Technische Universität Ilmenau, Ehrenbergstraße 29, 98693 Ilmenau, Deutschland
nicola.doering@tu-ilmenau.de; www.nicola-doering.de

Gäde, Jana Caroline, Dipl.-Psych.
Methodenberatung für Bachelorstudierende, Goethe-Universität Frankfurt am Main, Campus Westend, Grüneburgplatz 1, 60323 Frankfurt am Main, Deutschland
gaede@psych.uni-frankfurt.de

Gerhard, Carla, Dipl.-Psych.
Abteilung Psychologische Methodenlehre, Evaluation und Forschungsmethodik, Goethe-Universität Frankfurt am Main, Campus Westend, Grüneburgplatz 1, 60323 Frankfurt am Main, Deutschland
gerhard@psych.uni-frankfurt.de

Pöschl, Sandra, Dr.
Fachgebiet Medienpsychologie und Medienkonzeption am Institut für Medien und Kommunikationswissenschaft, Technische Universität Ilmenau, Ehrenbergstraße 29, 98693 Ilmenau, Deutschland
sandra.poeschl@tu-ilmenau.de

Schermelleh-Engel, Karin, Apl. Prof. Dr.
Abteilung Psychologische Methodenlehre, Evaluation und Forschungsmethodik, Goethe-Universität Frankfurt am Main, Campus Westend, Grüneburgplatz 1, 60323 Frankfurt am Main, Deutschland
schermelleh-engel@psych.uni-frankfurt.de

Werner, Christina, Dr.
Psychologische Methodenlehre, Evaluation und Statistik, Psychologisches Institut der Universität Zürich, Binzmühlestrasse 14, Box 27, 8050 Zürich, Schweiz
c.werner@psychologie.uzh.ch

Grundlagen

Kapitel 1 **Empirische Sozialforschung im Überblick – 3**
Nicola Döring, Jürgen Bortz

Kapitel 2 **Wissenschaftstheoretische Grundlagen der empirischen Sozialforschung – 31**
Nicola Döring, Jürgen Bortz

Kapitel 3 **Qualitätskriterien in der empirischen Sozialforschung – 81**
Nicola Döring, Jürgen Bortz

Kapitel 4 **Forschungs- und Wissenschaftsethik – 121**
Nicola Döring, Jürgen Bortz

Empirische Sozialforschung im Überblick

Nicola Döring, Jürgen Bortz

1.1 Sozialwissenschaftlicher Erkenntnisgewinn – 4

1.2 Empirische Studien verstehen und beurteilen – 19

1.3 Empirische Studien planen und durchführen – 22

Literatur – 29

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_1, © Springer-Verlag Berlin Heidelberg 2016

1

Lernziele

- Darstellen können, was wissenschaftliche Erkenntnisse von anderen Formen des Wissens unterscheidet.
- Sozial- und Humanwissenschaften von Geistes- und Naturwissenschaften abgrenzen können.
- Wissen, was quantitative Forschung, qualitative Forschung und Mixed-Methods-Forschung als drei zentrale empirische Forschungsstrategien in den Sozialwissenschaften jeweils auszeichnet.
- Wissenschaftliche Zeitschriftenartikel als Primär- bzw. Originalquellen empirischer Studien verstehen und beurteilen können.
- Die typischen Phasen des empirischen Forschungsprozesses im quantitativen, qualitativen und Mixed-Methods-Ansatz nennen und charakterisieren können.

Wer ein sozial- oder humanwissenschaftliches Studium absolviert, erwirbt je nach Fachdisziplin die unterschiedlichsten Kenntnisse und Fähigkeiten. Unabhängig von allen fachlichen Inhalten verfolgt die akademische Bildung immer auch das Ziel, Studierende in die Lage zu versetzen, **wissenschaftliche Erkenntnisse verstehen, beurteilen und selbst produzieren** zu können. Für diese aktive Teilnahme am wissenschaftlichen Erkenntnisgewinn ist Methodenkompetenz notwendig. Denn ohne Forschungsmethoden ist keine Wissenschaft möglich.

Das vorliegende Kapitel widmet sich zunächst dem wissenschaftlichen Erkenntnisgewinn und den Merkmalen empirischer Forschung in den Human- und Sozialwissenschaften (► Abschn. 1.1). Anschließend geht es um ein wichtiges Anwendungsfeld von Methodenkompetenz, und zwar um die Fähigkeit, vorliegende empirische Studien anhand von Primär- bzw. Originalquellen verstehen und sachgerecht beurteilen zu können (► Abschn. 1.2). Es besteht nämlich ein sehr großer Unterschied, ob man einen wissenschaftlichen Zeitschriftenartikel als Primärquelle lesen und kritisch einschätzen kann, oder ob man sich nur aus Sekundärquellen wie Lehrbüchern, Vorlesungen oder Presseberichten über Forschungsergebnisse informiert. Die Abweichungen zwischen Originalquellen und Sekundärquellen sind oft so erheblich, dass alle Studierenden gut beraten sind, die Recherche nach und Lektüre von Originalquellen frühzeitig zu lernen und oft zu praktizieren.

Noch höhere Ansprüche an die Methodenkompetenz werden gestellt, wenn es darum geht, eigene empirische Studien zu planen und durchzuführen (► Abschn. 1.3). Gleichzeitig gibt es keine bessere Möglichkeit, Methodenkompetenz zu erwerben und auszubauen, als selbst alle Phasen des empirischen Forschungsprozesses zu durchlaufen und dabei aus Fehlern bzw. Fehlentscheidungen zu lernen.

1.1 Sozialwissenschaftlicher Erkenntnisgewinn

Das vorliegende Lehr- und Handbuch beschäftigt sich mit Forschungsmethoden, die in der sozialwissenschaftlichen Grundlagen- und Anwendungsforschung (zu letzterer gehört auch die Evaluationsforschung) zum Einsatz kommen. Was wissenschaftliche Erkenntnisse von anderen Formen des Wissens unterscheidet und wozu Methodenkompetenz dient, welche Wissenschaftsdisziplinen zu unterscheiden sind und worin die Besonderheiten der Sozialwissenschaften bestehen sowie welche Forschungsansätze innerhalb der empirischen Sozialforschung voneinander abzugrenzen sind, wird im Folgenden behandelt.

1.1.1 Sozialwissenschaftliche Forschungs- und Methodenkompetenz

Empirische **Sozial- und Human**forschung befasst sich mit **sozialen und humanen Sachverhalten**, d. h. mit dem Verhalten, Erleben und Zusammenleben von Menschen sowie mit körperlichen und seelischen Merkmalen des Menschen. „Human“ und „sozial“ sind hierbei also keine positiv-wertenden, sondern neutral-beschreibende Begriffe, die auf das Menschliche und Zwischenmenschliche Bezug nehmen. Fragestellungen, die untersucht werden, lauten z. B.:

- Welche Zusammenhänge bestehen zwischen Persönlichkeitseigenschaften einerseits und Berufserfolg andererseits?
- Sind Einzelkinder egoistischer als Geschwisterkinder? Und wenn ja, wie kommt dieser Effekt zustande?
- Ist das Mammographie-Screening zur Brustkrebsfrüherkennung sinnvoll oder überwiegen Fehldiagnosen und Nebenwirkungen des Verfahrens?
- Welches Wissen und welche Einstellungen gegenüber Stammzellforschung bestehen in der Bevölkerung unterschiedlicher europäischer Länder?
- Wie verbreitet ist Hyperaktivität unter Kindern und Jugendlichen in Deutschland und wie wird sie erfolgreich behandelt?
- Welche Kampagnen und Maßnahmen gegen das Zigarettenrauchen sind besonders effektiv und effizient?
- Steigert die Nutzung gewalthaltiger Videospiele die Aggressivität?
- Welchen Einfluss haben globale Finanzkrisen auf das Prestige der Wirtschaftselite?

- Wie kann das Stromsparen bei privaten Endverbrauchern gefördert werden?
- Welche Unterschiede und Gemeinsamkeiten bestehen zwischen herkömmlichem Mobbing und so genanntem Cybermobbing?
- Wie lässt sich die Glaubwürdigkeit von Zeugenaussagen vor Gericht beurteilen?
- Welche hirnphysiologischen Veränderungen gehen mit Meditation einher?
- Welche individuellen, situativen und gesellschaftlichen Faktoren tragen zu Ausländerfeindlichkeit bei und wie kann sie abgebaut werden?

Ziel der **Human- und Sozialwissenschaften** ist es, solche Sachverhalte wissenschaftlich zu untersuchen und auf diese Weise zu möglichst gesicherten Erkenntnissen zu gelangen. Da soziale Sachverhalte Elemente unserer **wahrnehmbaren Erfahrungswirklichkeit** sind, handelt es sich bei den Human- und Sozialwissenschaften um **empirische Wissenschaften** bzw. Erfahrungswissenschaften: Sie basieren auf der systematischen Sammlung, Aufbereitung und Analyse von **empirischen Daten**, d. h. von Informationen über die Erfahrungswirklichkeit (z. B. Testwerte, physiologische Messwerte, Interviewaufzeichnungen, Beobachtungsprotokolle).

Empirische Daten – Empirische Daten („empirical data") sind gezielt im Hinblick auf das Forschungsproblem ausgewählte und dokumentierte Informationen über die Erfahrungswirklichkeit. Sie werden mit wissenschaftlichen Datenerhebungsmethoden (Beobachtung, Interview, Fragebogen, psychologischer Test, physiologische Messung, Dokumentenanalyse) unter Nutzung entsprechender standardisierter oder nicht-standardisierter Erhebungsinstrumente (Beobachtungsplan, Interviewleitfaden, Fragebogen, Messgerät etc.) gesammelt. Aussagekräftig sind die Daten nur, wenn sie im Rahmen eines dem Forschungsproblem angemessenen Forschungsprozesses und Untersuchungsdesigns an einer passenden Stichprobe (oder an der gesamten Population) erhoben wurden, sachgerecht ausgewertet und theoriebezogen interpretiert werden.

Wissenschaftlicher Erkenntnisgewinn ist zum einen ein Wert an sich (Zuwachs an gesichertem Wissen), kann und soll zum anderen aber auch bei der Lösung praktischer Probleme helfen. Wissenschaftliche Erkenntnisse werden sowohl von Individuen als auch von Gruppen und Organisationen als Legitimations- und Entscheidungsgrundlage herangezogen (z. B. Einführung oder Weiterführung einer politischen, wirtschaftlichen oder medizinischen Maßnahme nach positivem Ergebnis einer wissenschaftlichen Evaluationsstudie). Wissenschaftliche Erkenntnisse bilden zudem nicht selten die Basis zur Entwicklung und Optimierung sowie zur Vermarktung von Maßnahmen und Produkten (z. B. Therapieansätze, Medienangebote, Lernprogramme, Konsumprodukte) und haben somit wirtschaftliche Bedeutung. Um den hohen Stellenwert wissenschaftlichen Erkenntnisgewinns in allen gesellschaftlichen Bereichen zu betonen, spricht man von der **Wissensgesellschaft** (Engelhardt & Kajetzke, 2010), damit ist vor allem die Bedeutung von Wissenschaft und Technologie als Produktionsfaktor angesprochen. Ebenso ist von der **Wissenschaftsgesellschaft** (Böschen & Schulz-Schäffer, 2003) die Rede; dabei geht es vor allem um die Verwissenschaftlichung alltäglicher Lebensbereiche und die Orientierung an bzw. Abhängigkeit von Expertenurteilen. Das betrifft auch den Bereich sozialer Sachverhalte und deren **Beschreibung, Erklärung, Vorhersage und Veränderung** auf der Basis sozialwissenschaftlicher Forschungsergebnisse.

Wissenschaftlicher Erkenntnisgewinn – Wissenschaftlicher Erkenntnisgewinn („scientific knowledge gain") basiert in Erfahrungswissenschaften wie den Sozial- und Humanwissenschaften auf der systematischen Sammlung, Aufbereitung und Analyse von empirischen Daten im Rahmen eines geordneten und dokumentierten Forschungsprozesses. Dabei kommen sozialwissenschaftliche Methoden der Untersuchungsplanung, Stichprobenziehung, Datenerhebung, Datenaufbereitung und Datenanalyse zum Einsatz. Des Weiteren ist der empirische Forschungsprozess theoriebasiert, d. h. in seinem Verlauf werden wissenschaftliche Theorien über den Forschungsgegenstand (sowie über die Forschungsmethodik) angewendet und geprüft oder gebildet und weiterentwickelt. Erst mit Bezug auf Theorien sind empirische Daten sinnvoll interpretierbar.

Die Sozial- und Humanwissenschaften haben kein Monopol auf Wissensproduktion zu sozialen Sachverhalten. Auch auf vielen anderen Wegen wird Wissen generiert, das unser Verständnis sozialer Wirklichkeit prägt und individuelle und kollektive Entscheidungen über soziale Sachverhalte beeinflusst. Verbreitete Wege der **nicht-wissenschaftlichen Produktion und Begründung von Wissen** stellt ◘ Tab. 1.1 dar (vgl. für ähnliche Darstellungen Gravetter & Forzano, 2009, S. 6ff.; Mitchell & Jolley, 2010, S. 21ff.).

Wissen über soziale Sachverhalte, das auf Tradition, Intuition oder sog. gesundem Menschenverstand basiert, kann im Alltag im Einzelfall zutreffend sein und/oder sich als nützlich erweisen. Nicht-wissenschaftlicher Erkenntnisgewinn geht aber auch mit erheblichen Problemen einher. So ist die Basis der Behauptungen oft unklar: Woher stammt tradiertes Wissen, woraus speist sich die Intuition? Der Einfluss und der Anteil von Fehlern, Mythen, Verzerrungen, Vorurteilen oder Eigeninteressen kann nicht bestimmt werden. Oft stehen sich widersprüchliche Aussagen unvereinbar gegenüber: Zur selben Sachfrage kann die eine Person ein sehr positives, die andere ein negatives und eine dritte ein gemischtes „Bauchgefühl" haben. Ebenso ist es mit persönlichen Erfahrungen: Diese gehen bei ein und demselben Sachverhalt oft auseinander – haben dann alle gleichermaßen recht oder unrecht? Auch ist nicht-wissenschaftliche Wissensproduktion sehr lückenhaft: Zu bestimmten Themen

1

Tabelle 1.1 Varianten nicht-wissenschaftlicher Produktion und Begründung von Wissen im Kontrast zu wissenschaftlichen Erkenntnissen

Nicht-wissenschaftliche Produktion und Begründung von Wissen durch Berufung auf …	Vorgehensweise	Grenzen	Beispiele	Wissenschaftliche Erkenntnisse
Autoritätspersonen	Man stützt sich auf die Aussagen von Autoritätspersonen bzw. Experten	Unterschiedlicher Grad an tatsächlicher Expertise, subjektive Meinungen/Interessen von Experten, widersprüchliche Aussagen unterschiedlicher Experten	Experte A: Vitamin C stärkt das Immunsystem und schützt vor Erkältungen Experte B: Vitamintabletten sind bei normaler Ernährung völlig überflüssig	Wissenschaftliche Studien zeigen, dass Vitamin C (täglich mind. 0.2 g) bei der Normalbevölkerung zwar nicht die Häufigkeit, aber die Dauer und Schwere von Erkältungen reduziert (Hemilä, Chalker, & Douglas, 2007)
Religion	Man stützt sich auf religiöse Dogmen und Schriften	Anerkennung göttlicher Offenbarung setzt entsprechenden Glauben voraus und steht im Widerspruch zur verbreiteten säkularen Weltanschauung	Kreationismus, wie er in den USA zum Teil an Schulen gelehrt wird: Laut biblischer Schöpfungslehre wurden das Universum und die Erde, der Mensch und alle anderen Lebewesen tatsächlich durch Gott erschaffen	Wissenschaftliche Theorien und Befunde aus der Physik (Urknalltheorie) und der Biologie (Evolutionstheorie) widersprechen dem Kreationismus und werden mittlerweile auch vom Mainstream der christlichen Kirchen anerkannt
Tradition	Man stützt sich auf überliefertes Wissen früherer Generationen	Tradiertes Wissen basiert oft auf Missverständnissen, Fehlern, Mythen etc.	Seit Generationen wird überliefert, dass man beim Husten und Niesen die Hand vor den Mund halten soll, um andere nicht anzustecken	Aus wissenschaftlicher Sicht wird heute abgeraten, in die Hand zu niesen oder zu husten, da Erreger an der Hand lange überleben und durch Händedruck, Berührung von Gegenständen etc. ständig weitergegeben werden (stattdessen sollte ein Taschentuch oder die Ellenbeuge genutzt werden)
Gesunder Menschenverstand	Man stützt sich auf den gesunden Menschenverstand („common sense“) als geteilte Überzeugung einer Gruppe	Was als „gesunder Menschenverstand“ angesehen wird, variiert zwischen sozialen Gruppen sehr stark und ist oft auch von Vorurteilen, Gruppeninteressen etc. geprägt	In einer Talkshow zum Thema Adoption durch gleichgeschlechtliche Paare werden verschiedene Positionen vertreten – alle berufen sich auf den gesunden Menschenverstand Position A: Es wird doch niemand bestreiten wollen, dass Kinder immer Vater und Mutter brauchen Position B: Es ist doch allgemein bekannt, dass Kinder stabile Bezugspersonen brauchen – unabhängig von Geschlecht, sexueller Orientierung oder Verwandtschaftsgrad	Ein empirischer Vergleich von $n = 106$ Adoptivkindern, die bei Frauen-, Männer- oder heterosexuellen Paaren aufwachsen, zeigte keinen Einfluss der sexuellen Orientierung der Eltern auf die Entwicklung der Kinder. Diese hing von anderen Faktoren ab (z. B. Erziehungsstil sowie Beziehungsqualität der Eltern; Farr, Forssell, & Patterson, 2010)

Tabelle 1.1 (Fortsetzung)

Nicht-wissenschaftliche Produktion und Begründung von Wissen durch Berufung auf …	Vorgehensweise	Grenzen	Beispiele	Wissenschaftliche Erkenntnisse
Intuition	Man stützt sich auf seine eigene Intuition („Bauchgefühl“, „Instinkt“)	Das „Bauchgefühl“ ist durch viele Einflussfaktoren, insbesondere auch durch Vorurteile, Wunschdenken etc. beeinflusst	Das Bauchgefühl sagt einem, ein internet-öffentliches Sexualstraftäter-Verzeichnis würde die Sicherheit im Wohnbezirk erhöhen	Evaluationsstudien zu Effekten auf die Sicherheit fehlen, allerdings liegen Befragungsstudien zur Einstellung von Laien und Experten gegenüber den Verzeichnissen vor, denen eine sehr skeptische Haltung von Experten zu entnehmen ist (Malesky & Keim, 2001; Salerno et al., 2010)
Anekdotische Evidenz	Man stützt sich auf eigene Lebenserfahrungen und/oder Beispiele aus dem Umfeld oder den Medien	Persönliche Lebenserfahrungen sind sehr stark verzerrt durch Merkmale der eigenen Person sowie des eigenen kulturellen und sozialen Umfeldes	Nach fünf enttäuschenden Verabredungen mit Internet-Bekanntschaften schlussfolgert man, dass Online-Partnersuche in Wirklichkeit gar nicht funktioniert	Wissenschaftliche Studien zum Online-Dating zeigen, dass und inwiefern es sowohl Vorteile als auch Nachteile birgt (Finkel et al., 2012)
Logik	Man stützt sich auf logische Argumente	In sozialen Zusammenhängen handeln Menschen und Organisationen oft nicht logisch oder rational	Gegner von Frauenförderprogrammen behaupten: Bei der Besetzung von Führungspositionen gibt es keine geschlechtsbezogene Benachteiligung von Frauen, denn es wäre doch unlogisch, wenn Unternehmen sich gute Bewerberinnen entgehen ließen	Der wissenschaftliche Forschungsstand zur Situation weiblicher Führungskräfte weist verschiedene Benachteiligungen von Frauen im Berufsleben nach (Eagly & Carlib, 2003)

hat fast jeder Mensch seine eigene Alltagstheorie konstruiert (z. B. zu Geschlechtsunterschieden im Kommunikationsverhalten), über andere Themen wird im Alltag kaum nachgedacht (z. B. Kulturunterschiede im räumlichen Denken).

Wissenschaftliche Forschung – Wer wissenschaftliche Forschung („scientific research“) betreibt, sucht mithilfe anerkannter wissenschaftlicher Methoden und Methodologien auf der Basis des bisherigen Forschungsstandes (d. h. vorliegender Theorien und empirischer Befunde) zielgerichtet nach gesicherten neuen Erkenntnissen, dokumentiert den Forschungsprozess sowie dessen Ergebnisse in nachvollziehbarer Weise und stellt die Studie in Vorträgen und Publikationen der Fachöffentlichkeit vor.

Ziel des Erkenntnisgewinns auf der Basis sozialwissenschaftlicher Forschung ist es, die Probleme nicht-wissenschaftlichen Erkenntnisgewinns zu überwinden und in systematischer, nachvollziehbarer und überprüfbarer Weise auf der Basis empirischer Daten zu (vorläufig) gültigen Ergebnissen zu Fragen oder Annahmen über soziale Sachverhalte zu kommen (▶ Zentrale Merkmale von human- und sozialwissenschaftlicher Forschung).

Ein erstes Grundverständnis von sozialwissenschaftlicher Forschung entsteht dadurch, dass man den wissenschaftlichen Erkenntnisgewinn von nicht-wissenschaftlicher Wissensproduktion abgrenzt, wie dies in Tab. 1.1 exemplarisch getan wurde. Es ist Aufgabe der **Wissenschaftstheorie** als einem Zweig der Philosophie, die Voraussetzungen und Möglichkeiten wissenschaftlichen Erkenntnisgewinns zu beschreiben und auch normativ festzulegen (zu wissenschaftstheoretischen Grundlagen der empirischen Sozialforschung ▶ Kap. 2). Dabei existieren innerhalb der Sozialwissenschaften unterschiedliche Wis-

1

Zentrale Merkmale von human- und sozialwissenschaftlicher Forschung

Abgrenzung zu nicht-wissenschaftlicher Wissensproduktion:

1. **Formulierung empirisch untersuchbarer und gut begründeter Forschungsfragen bzw. Forschungshypothesen**. Wissenschaftlicher Erkenntnisgewinn ist nur möglich, wenn man mit einem eingegrenzten Forschungsproblem startet, das sich auf bestimmte, untersuchbare Ausschnitte der Erfahrungswirklichkeit konzentriert (zur Auswahl und Formulierung von Forschungsproblemen ► Kap. 5). Negativbeispiel: Forschungsfragen beziehen sich auf Sachverhalte, die für eine empirische Untersuchung zu allgemein oder zu vage sind.
2. **Berücksichtigung des Forschungsstandes und ausdrücklicher Theoriebezug**. Empirische Forschung knüpft am aktuellen Forschungsstand an. Zudem sollen theoretische Vorannahmen oder theoriebezogene Interpretationen als solche explizit gemacht und diskutiert werden. Negativbeispiel: Eine Studie zu „Cybermobbing" greift nicht auf vorliegende Theorien und Befunde zu Mobbing und Online-Mobbing zurück, sondern arbeitet mit einer willkürlichen eigenen Begriffsdefinition.
3. **Systematische Erhebung, Aufbereitung und Analyse von empirischen Daten zum Forschungsproblem mithilfe wissenschaftlicher Methoden unter Einhaltung wissenschaftlicher Gütekriterien**. Wissenschaftlicher Erkenntnisgewinn ist in einer Erfahrungswissenschaft nur möglich auf der Basis empirischer Daten. Dazu notwendig sind ein angemessenes Untersuchungsdesign, geeignete Untersuchungsinstrumente, eine aussagekräftige Stichprobe (sofern keine Vollerhebung durchgeführt wird), die Anwendung begründeter Datenerhebungsmethoden, die systematische Datenaufbereitung sowie die Verwendung etablierter Datenanalysemethoden (► Teil II des Buches). Negativbeispiel: Es erfolgt eine unzureichende oder verzerrte Datenerhebung und Datenauswertung, etwa wenn die im Rahmen einer Interviewstudie tätigen Interviewerinnen und Interviewer mangels Schulung durch unangemessenes Verhalten (z. B. wertende Äußerungen, Suggestivfragen) die Datenqualität beeinträchtigen. Mit **wissenschaftlichen Methoden** sind im engeren Sinne Methoden der Datenerhebung (z. B. Technik des Leitfaden-Interviews) sowie der Datenauswertung (z. B. statistische Verfahren der Parameterschätzung) gemeint.
4. **Vorgehen gemäß etablierter wissenschaftlicher Methodologie unter Einhaltung von wissenschaftlichen Gütekriterien**. Wissenschaftlicher Erkenntnisgewinn ist nur möglich, wenn der Forschungsprozess in seinen verschiedenen Phasen sich an etablierten wissenschaftlichen Methodologien orientiert und dabei wissenschaftliche Gütekriterien in nachweisbarer Weise erfüllt (► Kap. 3). Im Unterschied zur wissenschaftlichen Methode als separater Technik der Datenerhebung oder Datenauswertung sind mit der **wissenschaftlichen Methodologie** Vorgaben zum Forschungsprozess und Untersuchungsdesign gemeint. Negativbeispiel: Es wird ein sozialwissenschaftliches Laborexperiment anhand einer Quoten-Stichprobe durchgeführt, die man möglichst heterogen zusammensetzt (z. B. hinsichtlich Alter, Bildung, Nationalität, Persönlichkeitseigenschaften) in dem Bestreben, damit die Aussagekraft der Studie zu erhöhen – tatsächlich aber mit dem Effekt, dass die gesamte Studie aufgrund zu hoher Fehlervarianz nicht aussagekräftig ist. Das fehlerhafte Design kommt hier durch eine Vermischung der Methodologien von Experimentalforschung (in der Regel kleine homogene Stichproben) und Umfrageforschung (in der Regel große heterogene Stichproben) zustande.
5. **Vorgehen gemäß Prinzipien der Forschungsethik sowie der Wissenschaftsethik**. Wissenschaftlicher Erkenntnisgewinn muss immer mit redlichen Mitteln erfolgen. So verbietet es die Wissenschaftsethik, Daten in Richtung auf ein Wunschergebnis zu manipulieren (► Abschn. 4.2). Zudem dürfen Untersuchungsteilnehmerinnen und -teilnehmer im Zuge der Forschung keinen unnötigen Risiken oder unvertretbaren Beeinträchtigungen ausgesetzt werden. So verlangt die Forschungsethik bei riskanten Studien eine gründliche Aufklärung und ausdrückliche Einwilligung der Untersuchungsteilnehmenden (► Abschn. 4.1). Negativbeispiel: Bei der Datenerhebung wird unzulässig die Privatsphäre der unwissentlichen Untersuchungsteilnehmenden verletzt, indem heimlich an nicht-öffentlichen Orten Daten gesammelt werden.
6. **Ausführliche schriftliche Dokumentation des Forschungsprozesses und Archivierung des Datenmaterials zwecks Nachvollziehbarkeit und Nachprüfbarkeit**. Wissenschaft verlangt eine detaillierte Nachvollziehbarkeit und Nachprüfbarkeit des Vorgehens im Forschungsprozess. Negativbeispiel: Einzelne Schritte im Forschungsprozess werden nicht genau beschrieben, die Daten werden nicht archiviert, so dass bei Zweifeln oder Rückfragen keine Klärung möglich ist und die Studie auch nicht in vergleichbarer Form wiederholt (repliziert) werden kann.
7. **Ausgewogene Ergebnisinterpretation mit Hinweisen auf widersprüchliche Befunde und auf Grenzen der Aussagekraft der Studie**. Wissenschaftlicher Erkenntnisgewinn ist nur möglich, wenn Forschende die Ergebnisse einer Studie in ihrer Gänze (und nicht nur in wünschenswerten Ausschnitten) betrachten, angemessen mit Bezug auf Theorien interpretieren und wenn sie die stets beschränkte Aussagekraft einer einzelnen Studie ausdrücklich reflektieren (anstatt unkritisch zu verallgemeinern). Negativbeispiel: Einzelbefunde werden als allgemeingültige Wahrheiten dargestellt.
8. **Wissenschaftliche Veröffentlichung der Studie – möglichst nach fachkundiger Begutachtung (Peer-Review).** Wissenschaftlicher Erkenntnisgewinn ist nur möglich, wenn eine Studie veröffentlicht und somit von Fachkolleginnen und Fachkollegen kritisch hinsichtlich ihrer Stärken und Schwächen sowie der Aussagekraft der Befunde diskutiert werden kann. Im Idealfall erfolgt die Publikation in Fachzeitschriften, die eine fachkundige Begutachtung (Peer-Review-Prozess) jedes Manuskriptes durchführen und bei Qualitätsmängeln zur Nachbesserung auffordern oder auf eine Veröffentlichung verzichten. Negativbeispiel: Ein spektakuläres Studienergebnis wird per Presseerklärung der breiten Öffentlichkeit vorgelegt, ohne dass zuvor eine detaillierte wissenschaftliche Publikation und Prüfung durch Fachkollegen stattfand.

senschaftsverständnisse bzw. **wissenschaftstheoretische Paradigmen** (zum Überblick z. B. Teddlie & Tashakkori, 2009). Ein wissenschaftstheoretisches Paradigma besteht zum einen aus Vorannahmen über den Gegenstand der Sozialwissenschaften, also über die Beschaffenheit sozialer Sachverhalte bzw. sozialer Erfahrungswirklichkeiten (**Ontologie**). Eine ontologische Frage ist dabei z. B., ob es eine einzige objektivierbare Erfahrungswirklichkeit gibt oder ob viele subjektive Erfahrungswirklichkeiten nebeneinander bestehen. Zum anderen gehören zum wissenschaftstheoretischen Paradigma Vorannahmen über die Erkennbarkeit der Erfahrungswirklichkeit **(Epistemologie):** Können wir soziale Sachverhalte besser beschreiben, verstehen und erklären, wenn wir sehr strukturiert vorgehen und theoriebasiert systematisch einzelne Aspekte erfassen und in ihren Relationen prüfen oder wenn wir bewusst unstrukturiert vorgehen und zunächst versuchen, den Untersuchungsgegenstand möglichst ganzheitlich zu erfassen? Auch die Bedeutung von Werten im Forschungsprozess ist Gegenstand wissenschaftstheoretischer Auseinandersetzungen **(Axiologie):** Können und sollen Forschende im Zuge ihrer wissenschaftlichen Arbeit eine unabhängige und neutrale Haltung einnehmen oder für bestimmte Bevölkerungsgruppen ausdrücklich Partei ergreifen?

Verschiedene wissenschaftstheoretische Paradigmen in der empirischen Sozialforschung – insbesondere das **quantitative und das qualitative Paradigma** – geben auf derartige Fragen sehr unterschiedliche Antworten, so dass auch von einem **Paradigmenstreit** („paradigm controversy"; z. B. Lincoln, Guba, & Lynham, 2012) oder gar **Paradigmenkrieg** („paradigm war"; z. B. Hammersley, 1992) die Rede ist. Doch genau wie inhaltliche Theorien und Erkenntnisse in der Wissenschaft einem Wandel unterliegen, ist dies auch im Bereich der Wissenschaftstheorie der Fall. So zeichnet sich in den letzten Jahren die Tendenz ab, quantitatives und qualitatives Paradigma nicht mehr so stark in einem Konkurrenzverhältnis, sondern stärker in einem Ergänzungsverhältnis zueinander zu betrachten. Dies schlägt sich in Diskussionen um einen integrativen dritten Ansatz – das sog. **Mixed-Methods-Paradigma** – nieder (▶ Abschn. 1.1.3 und ▶ Kap. 2). Ein Verständnis wissenschaftstheoretischer Fragen gehört zur sozialwissenschaftlichen Forschungs- und Methodenkompetenz dazu, um das eigene wissenschaftliche Vorgehen hinsichtlich seiner philosophischen Grundlagen begründen und reflektieren zu können.

Wissenschaftliche Methoden – Wissenschaftliche Methoden („scientific methods") sind in der empirischen Sozialforschung vor allem **einzelne Verfahren bzw. Techniken** der Datenerhebung und Datenanalyse. Wissenschaftliche Datenerhebungsmethoden in diesem Sinne sind beispielweise psychologische Testverfahren wie der Intelligenztest, physiologische Messungen wie die EKG-Messung oder Interviewtechniken wie das Leitfaden-Interview. Wissenschaftliche Datenanalysemethoden sind z. B. die qualitative Inhaltsanalyse oder die statistische Varianzanalyse.

Wissenschaftliche Methodologien oder Forschungsstrategien – Wissenschaftliche Methodologien („scientific methodologies") oder **Forschungsstrategien** („research strategies", „strategies of inquiry") sind Anweisungen zur **Gestaltung des Forschungsprozesses und des Untersuchungsdesigns.** Methodologien geben an, in welcher Weise einzelne Methoden in einen Untersuchungsplan zu integrieren sind und was beim Durchlaufen des Forschungsprozesses zu beachten ist, um Wissenschaftlichkeit und hohe Aussagekraft zu gewährleisten. So stehen z. B. hinter ethnografischer Feldforschung oder experimenteller Laborforschung unterschiedliche Methodologien, welche die Basis bilden für Entscheidungen über Untersuchungsbedingungen, Untersuchungszeitpunkte, Untersuchungsdauer, Auswahl von Untersuchungspersonen, Anzahl und Rolle der Forschenden, Art der Dokumentation der Daten etc. Im Rahmen einer bestimmten Methodologie können unterschiedliche Methoden einzeln oder kombiniert zum Einsatz kommen. In einer zweiten Begriffsbedeutung meint Methodologie die **wissenschaftstheoretische Begründung der Methoden** (▶ Abschn. 2.1).

Beim wissenschaftlichen Erkenntnisgewinn wird der **Forschungsprozess** in den Mittelpunkt gerückt: Er soll in seinen verschiedenen Phasen genau dokumentiert sein, etablierten Methodologien folgen und wissenschaftliche Methoden einsetzen, um **intersubjektive Nachvollziehbarkeit und Nachprüfbarkeit** sicherzustellen. Es reicht also nicht aus, dass Forschende in ihrer Rolle als Experten etwas behaupten oder sich pauschal auf ihre Forschungsergebnisse berufen, sie müssen stets gewillt und in der Lage sein, im Detail über die Merkmale und Grenzen des Forschungsprozesses einer Studie zu berichten. Nach wissenschaftlichen Maßstäben gewinnen eine Studie und Forschende in der Regel an Glaubwürdigkeit, wenn sie Widersprüche, Probleme und offengebliebene Fragen selbst ansprechen und ihre Befunde mit Vorsicht interpretieren. Vorschnelle Verallgemeinerungen, Überinterpretationen oder das unkritische Verbreiten von vermeintlichen Gewissheiten – all dies widerspricht dem wissenschaftlichen Selbstverständnis (zu entsprechenden Regeln guter wissenschaftlicher Praxis ▶ Abschn. 4.2.1).

Die **wissenschaftliche Vorsicht und Zurückhaltung** steht oft in krassem Widerspruch zu Anforderungen der Praxis sowie zu öffentlichen Erwartungen. Hier möchte man in der Regel schnelle, eindeutige und leicht verwertbare wissenschaftliche Ergebnisse vorgelegt bekommen. Dieses Spannungsverhältnis wird sich nicht auflösen lassen. Vielmehr geht es um einen bewussten Umgang mit diesen Problemen und einer Vermeidung der typischen Extremreaktionen, nämlich Wissenschaftsgläubigkeit einerseits oder Wissenschaftsfeindlichkeit andererseits:

- **Wissenschaftsgläubigkeit**: Das Merkmal von Wissenschaft, nicht nur fertige Ergebnisse vorzulegen, sondern den Forschungsprozess transparent zu machen und in seinen Grenzen fortlaufend zu problematisieren, erzeugt neben neuen Erkenntnissen immer auch Komplexität, Unsicherheit und Unübersichtlichkeit. Eine Reaktion darauf kann Wissenschaftsgläubigkeit sein. Hier werden die Aussagen einzelner bekannter Wissenschaftlerinnen und Wissenschaftler oder die Befunde einzelner spektakulärer Studien hervorgehoben und als „die" wissenschaftliche Erkenntnis stilisiert oder gar regelrecht vermarktet. Das Publikum hat dann das gute Gefühl zu wissen, was „die Wissenschaft" zu einem Sachverhalt sagt. Durch diese Vorgehensweise wird jedoch die nichtwissenschaftliche Methode der Berufung auf Autoritätspersonen wieder eingeführt (Tab. 1.1). Ebenso wird verkannt, dass die Aussagekraft einer einzelnen sozialwissenschaftlichen Studie sehr gering ist und der Forschungsstand erst unter Berücksichtigung vieler Studien und deren systematischer Synthese (z. B. durch Metaanalysen, ► Kap. 16) sich herauskristallisiert.
- **Wissenschaftsfeindlichkeit**: Die Unzufriedenheit mit der wissenschaftlichen Herangehensweise führt auf der anderen Seite häufig auch zu Wissenschaftsfeindlichkeit: Wenn „die Wissenschaft" offenbar nicht in der Lage ist, auf praktische soziale Fragen einfache und eindeutige Antworten zu geben, dann läuft dort etwas schief, so eine häufige Kritik: Womöglich ist man zu abgehoben und praxisfern im „Elfenbeinturm", untersucht die falschen Themen, manipuliert die Ergebnisse ohnehin nach Belieben, ist von Lobbyisten beeinflusst, nur auf die eigene Karriere bedacht etc. Eine solche offen negative oder gar zynische Haltung gegenüber dem Wissenschaftssystem und den Forschenden steht neben subtileren Formen der Wissenschaftsfeindlichkeit. So haben gerade die Sozialwissenschaften mit einem Glaubwürdigkeitsproblem zu kämpfen, weil ihre Untersuchungsthemen oft recht alltagsnah sind und sozialwissenschaftliche Befunde in öffentlichen Debatten auf vorwissenschaftliche Theorien bzw. Alltagsüberzeugungen treffen. Dementsprechend gibt es drei typische Ausgänge der Laienbewertung von sozialwissenschaftlichen Forschungsergebnissen, die allesamt zu Ungunsten der empirischen Sozialforschung ausgelegt werden können (vgl. auch Kromrey, 2009, S. 16ff.):
  1. **Vorwurf trivialer Forschung**: Eine Studie liefert Ergebnisse, die dem Alltagsverständnis (Tradition, Intuition, gesunder Menschenverstand etc.) der Bevölkerungsmehrheit oder großer Gruppen entsprechen. Dann wird kritisiert, das hätte man ja wohl vorher wissen können und diese triviale Studie sei überflüssig und Geldverschwendung. (Zuweilen freut sich die Öffentlichkeit natürlich auch, die eigenen Vorstellungen wissenschaftlich bestätigt zu sehen.)
  2. **Vorwurf fehlerhafter Forschung**: Eine Studie liefert Ergebnisse, die dem Alltagsverständnis und/oder den Interessen bestimmter Gruppen widersprechen. An dieser Stelle bezweifelt die Öffentlichkeit dann nicht selten die Aussagekraft der Studie. Ihr werden methodische Fehler vorgeworfen, meist eine „nicht repräsentative Stichprobe". (Zuweilen greifen es Massenmedien aber auch wiederum gerne als Thema auf, wenn es zu einer verbreiteten Überzeugung wissenschaftliche Gegenstimmen gibt.)
  3. **Vorwurf nutzloser Forschung**: Eine Studie liefert sehr differenzierte und komplexe Ergebnisse und/oder thematisiert alltagsferne Themen, die für Laien nicht sofort einleuchtend und nachvollziehbar sind. Sofern derartige Studien überhaupt jenseits der Fachcommunity öffentlich wahrgenommen werden, reagieren Fachfremde oft mit Desinteresse oder Ablehnung und stellen den Sinn und Nutzen derartiger „abgehobener" Projekte infrage. (Zuweilen wird anspruchsvolle Grundlagenforschung, insbesondere wenn sie mit aufwändigen Technologien betrieben wird, aber auch mit Respekt vor der wissenschaftlichen Leistung betrachtet, obwohl – oder besser: weil – die Befunde für Laien kaum verständlich sind).

Die hier nur angedeuteten Haltungen und Kontroversen rund um den Stellenwert von Sozialwissenschaften in der Gesellschaft bzw. die Aussagekraft und Nützlichkeit sozialwissenschaftlicher Forschung, sollten Studierende und Forschende für sich selbst reflektieren und eine eigene Haltung entwickeln. Verwiesen sei auch auf die empirische Forschung zum Laienverständnis von Wissenschaft sowie zur Repräsentation sozialwissenschaftlicher Forschung in den Massenmedien (z. B. Hammersley, 2006). Ein typischer – tendenziell wissenschaftsfeindlicher – Fehler von Studierenden besteht darin, aufgrund überhöhter Erwartungen an die Erklärungskraft einzelner Forschungsprojekte vorliegenden Studien angesichts von deren ganz normalen methodischen Limitationen pauschal die Aussagekraft abzusprechen (zum Lesen und Bewerten von Zeitschriftenartikeln ► Abschn. 1.2.2 Inhaltliche Merkmale von Zeitschriftenartikeln). Gleichzeitig sind Forschende, auch weil ihre Projekte oft mit öffentlichen Mitteln finanziert werden, zunehmend in der Pflicht, bei der Ergebnispräsentation nicht nur die Fachkollegen, son-

dern auch die breite Öffentlichkeit zu adressieren und sich um einen verbesserten Transfer wissenschaftlicher Ergebnisse in die Praxis sowie in die Massenmedien einzusetzen. Dies verlangt einen Spagat zwischen wissenschaftlicher Vorsicht und allgemeinverständlicher Vereinfachung (zur Öffentlichkeitsarbeit ▶ Abschn. 13.4).

Man spricht von **Wissenschaftsbildung** – oft aufgeteilt nach Naturwissenschaften („science literacy“) und Sozialwissenschaften (social science literacy; Weber, 2010) – als Sammelbegriff für die notwendigen Kenntnisse und Fähigkeiten, die Bürgerinnen und Bürger in einer Wissens- und Wissenschaftsgesellschaft benötigen, um sich eigenständig mit dem Wissenschaftssystem, dem Forschungsprozess und wissenschaftlichen Befunden auseinanderzusetzen und mit Bezug auf diese zu naturwissenschaftlich-technischen sowie sozialen Sachverhalten eigene Standpunkte zu entwickeln. Wissenschaftsbildung sollte möglichst vor Wissenschaftsgläubigkeit einerseits und Wissenschaftsfeindlichkeit andererseits schützen und zu ausgewogenen Einschätzungen hinsichtlich Leistungen und Grenzen wissenschaftlichen Erkenntnisgewinns befähigen.

Im engeren Sinne sprechen wir von **sozialwissenschaftlicher Forschungs- bzw. Methodenkompetenz** („social scientific research literacy“), um Kenntnisse und Fähigkeiten anzusprechen, die Mitglieder des Wissenschaftssystems benötigen, um sozialwissenschaftliche Originalquellen verstehen und beurteilen sowie eigene sozialwissenschaftliche Studien sachgerecht durchführen zu können. Dabei ist zu beachten, dass es bei wissenschaftlicher „Methoden“-Kompetenz nicht nur um Kenntnisse einzelner Methoden der Datenerhebung oder Datenauswertung geht, sondern vor allem auch um Kenntnisse über Methodologien und nicht zuletzt auch wissenschaftstheoretische Paradigmen. Für den Erwerb und Ausbau der eigenen sozialwissenschaftlichen Methodenkompetenz, den auch das vorliegende Buch unterstützen will, sprechen vor allem drei gute Gründe:

1. **Sozialwissenschaftliche Methodenkompetenz im Studium**: Alle Studierenden müssen in der Lage sein, Forschungsprozesse und Forschungsergebnisse, wie sie in wissenschaftlichen Publikationen ihres jeweiligen Faches dargestellt sind, zu verstehen und zu bewerten. Während in der Schule Lehrbuchwissen vermittelt wird, hat die Hochschule den Anspruch, Studierenden den Zugang zu den primären wissenschaftlichen Quellen (wissenschaftliche Fachvorträge und Fachartikel) zu erschließen. Darüber hinaus haben Hochschulen und insbesondere Universitäten mit ihrem Prinzip der **Einheit von Forschung und Lehre** den Anspruch, dass Studierende nicht nur vorhandene Forschungsergebnisse kritisch rezipieren, sondern selbst eigene Forschungsprojekte realisieren können. Hierzu ist nicht nur theoretische Methodenkenntnis, sondern auch Handlungskompetenz notwendig, die typischerweise im Zuge praktischen Übens entsteht. In Referaten, Hausarbeiten, Bachelor- und Master- sowie Doktorarbeiten muss die eigene Methodenkompetenz praktisch unter Beweis gestellt werden.
2. **Sozialwissenschaftliche Methodenkompetenz im Beruf**: Die Fähigkeit, wissenschaftliche Forschungsergebnisse zu rezipieren und zu produzieren wird von allen hauptberuflichen Wissenschaftlerinnen und Wissenschaftlern gefordert, die an Hochschulen, Forschungsinstituten oder in Forschungsabteilungen von Unternehmen tätig sind. Wissenschaftliche Methodenkompetenz ist darüber hinaus in vielen anderen **akademischen Berufen** gefragt, die eine gewisse Forschungsnähe aufweisen: Praktisch tätige Ärztinnen, Psychologen, Seelsorgerinnen, Berufsberater, Sozialarbeiterinnen, Lehrer etc. führen teilweise Studien durch, etwa um Projekte zu akquirieren oder ihre Arbeit zu evaluieren. Im Journalismus wird mit Forschungsergebnissen – oft mit Statistiken – ebenso operiert wie in der Politik oder in Wirtschaftsunternehmen. In vielen akademischen Berufen wird heute gefordert, dass man sich in der Praxis am aktuellen Forschungsstand orientiert und auch nach dem Studium fortlaufend weiterbildet. Evidenzbasierte Medizin und Psychotherapie verlangen es, dass sich die Praktikerinnen und Praktiker anhand aktueller Metaanalysen darüber informieren, welche Interventionsmethoden sich als wissenschaftlich wirksam oder unwirksam erwiesen haben (zu Metaanalysen ▶ Kap. 16). Hierbei sind Wissenschaftsgläubigkeit oder Wissenschaftsfeindlichkeit hinderlich, stattdessen ist eine kompetente Einordnung sozial- und humanwissenschaftlicher Studien und ihrer Befunde gefragt.
3. **Sozialwissenschaftliche Methodenkompetenz im Alltag**: In dem Maße, in dem wissenschaftliche Ergebnisse auch im Alltag als Entscheidungsgrundlage herangezogen werden, wird von mündigen Bürgerinnen, Patienten, Konsumentinnen, Wählern etc. gefordert, Forschungsbefunde sachgerecht bewerten zu können. Egal ob eine Partei mit den neuesten Arbeitslosenstatistiken Wahlkampf betreibt, der Beipackzettel eines Medikamentes über dessen Nebenwirkungen berichtet oder die Lebensmittelindustrie die gesundheitsfördernden Effekte von probiotischem Joghurt propagiert – in derartigen Diskussionen wird immer auch auf den wissenschaftlichen Forschungsstand Bezug genommen, dessen methodische Grundlagen also auch außerhalb von Wissenschaft und wissenschaftsnahen Berufen von Bedeutung sind. Über allgemeine Wissenschaftsbildung hinausgehend (wie sie bereits in der Schule vermittelt werden sollte) bietet im Stu-

dium erworbene vertiefte Forschungs- und Methodenkompetenz eine noch bessere Bewertungsgrundlage für die im Alltag herangezogenen wissenschaftlichen Befunde.

1.1.2 Sozial- und Humanwissenschaften zwischen Geistes- und Naturwissenschaften

Das vorliegende Buch befasst sich mit Methoden der empirischen Sozialforschung bzw. mit Forschungsmethoden für **Sozial- und Humanwissenschaften**, weshalb wir kurz auf die Abgrenzung und Einordnung von Wissenschaftsdisziplinen eingehen wollen: Das Wissenschaftssystem ist disziplinär organisiert. Einzelne **Wissenschaftsdisziplinen** bzw. Fächer (z. B. Psychologie, Erziehungswissenschaft, Wirtschaftswissenschaft, Medizin) zeichnen sich aus wissenschaftstheoretischer Sicht durch ihren jeweiligen Untersuchungsgegenstand, ihre Fragestellungen sowie ihre spezifischen Forschungsmethoden aus. Zudem konstituiert sich eine Wissenschaftsdisziplin aber auch durch eigene Institutionen (Fachgesellschaften, Fachzeitschriften, Forschungsinstitute, Studiengänge etc.) sowie durch ihre Fachkultur und das Selbstverständnis und Auftreten der Fachvertreterinnen und Fachvertreter. Um die Vielzahl der Disziplinen zu ordnen, werden sie oft zu Fächergruppen zusammengefasst (◘ Tab. 1.2).

Der **wissenschaftliche Fächerkanon** hat sich historisch entwickelt und ist in fortlaufender Veränderung begriffen. So bilden sich durch soziale Veränderungen und technische Innovationen immer wieder neue wissenschaftliche Disziplinen. Die im Mittelalter gegründeten Universitäten gliederten sich typischerweise zunächst in vier Fakultäten: **Philosophie, Theologie, Jurisprudenz und Medizin** (vgl. Grant, 1996). Eigene Fakultäten für **Natur- und Technikwissenschaften** wurden erst ab dem 18./19. Jahrhundert eingerichtet, als mit der Industrialisierung die Natur- und Technikwissenschaften an gesellschaftlicher Bedeutung gewannen. Die **Sozialwissenschaften** etablierten sich überwiegend erst ab dem 20. Jahrhundert durch eigene Institute und Fakultäten sowie Fachgesellschaften. So wurde z. B. die Deutsche Gesellschaft für Psychologie (DGPs) im Jahr 1904 und die Deutsche Gesellschaft für Soziologie (DGS) im Jahr 1909 gegründet.

Die im vorliegenden Buch behandelten Forschungsmethoden beziehen sich immer auf **empirische Forschung** („empirical research"), das heißt auf eine Forschung, die auf die Erfahrungswirklichkeit abzielt. **Empirische Wissenschaften** werden unterteilt in **Natur- und Technikwissenschaften** („natural and engineering sciences") einerseits und **Sozialwissenschaften** („social sciences") andererseits (◘ Tab. 1.2). In beiden Disziplinfamilien wird empirisch geforscht, indem methodisch angeleitet Daten über die Erfahrungswirklichkeit gesammelt und analysiert werden: So können z. B. das Fressverhalten von Tieren, die Ausfallhäufigkeit technischer Anlagen, das Auftreten von Symptomen bei Patienten oder die Häufigkeit bestimmter Straftaten in unterschiedlichen Gesellschaften gleichermaßen beobachtet und statistisch dargestellt werden.

Da in den **Sozialwissenschaften** das Verhalten, Erleben und Zusammenleben von Menschen im Fokus steht, sind – im Unterschied zur Beobachtungsmethode, die in allen empirischen Wissenschaften einsetzbar ist – noch diverse spezifische Methoden einschlägig. So ist z. B. die Interview- ebenso wie die Fragebogenmethode nur in den Sozialwissenschaften sinnvoll einsetzbar, sie ist nicht auf unbelebte Materie, Pflanzen, Tiere oder Maschinen anwendbar. Man spricht auch von **Methoden der empirischen Sozialforschung** („social research methods"), um diese spezifischen empirischen Forschungsmethoden abzugrenzen von Beobachtungs- bzw. Messmethoden, die in den Natur- und Technikwissenschaften zum Einsatz kommen.

Sozialwissenschaften heben sich durch ihren Untersuchungsgegenstand, der sich im weitesten Sinne auf das Verhalten und Erleben sowie die Merkmale von Menschen bezieht, einerseits von den Natur- und Technikwissenschaften ab, andererseits aber auch von den **Geisteswissenschaften**, die kulturelle Artefakte wie Kunstwerke, historische Dokumente, religiöse Schriften oder Gesetzestexte untersuchen und dafür eigene, vor allem textverstehende und interpretative Methoden einsetzen. Damit ergibt sich auch ein eigener sozialwissenschaftlicher Methodenkanon, der jedoch an natur- und geisteswissenschaftliche Methoden anknüpft. So steht die qualitative empirische Sozialforschung in der Tradition der Geisteswissenschaften, die quantitative empirische Sozialforschung dagegen in der Tradition der Naturwissenschaften (▶ Abschn. 1.1.3).

Eine **disziplinäre Einteilung der Wissenschaften**, wie sie in ◘ Tab. 1.2 dargestellt ist, liefert Orientierung im Wissenschaftssystem, ist aber auch im Detail strittig und widersprüchlich. So wird z. B. die Informatik teilweise den Formalwissenschaften zugeordnet (weil sie ebenso wie die Mathematik mit formalen Sprachen wie z. B. Programmiersprachen operiert), meist aber den Technikwissenschaften (weil Computersteuerung zentraler Bestandteil vieler moderner Technologien ist). Eine eindeutige Zuordnung ist auch bei geistes- und sozialwissenschaftlichen Fächern nicht immer möglich. So findet sich innerhalb der Erziehungswissenschaft als empirischer Wissenschaft eine starke geisteswissenschaftliche Strömung, umgekehrt wird innerhalb der Sprachwissen-

Tabelle 1.2 Einteilung von Wissenschaftsdisziplinen in Fächergruppen

Nicht-empirische Wissenschaften		Empirische Wissenschaften/Erfahrungswissenschaften		
Formalwissenschaften („formal sciences“)	**Geisteswissenschaften („humanities“)**	**Sozialwissenschaften auch: Humanwissenschaften, Gesellschaftswissenschaften („social sciences“)**	**Naturwissenschaften („natural sciences“)**	**Technikwissenschaften auch: Ingenieurwissenschaften („engineering sciences“)**
Philosophie	Theologie	Psychologie	Physik	Maschinenbau
Mathematik	Rechtswissenschaft	Medizin	Chemie	Elektrotechnik
etc.	Geschichte	Erziehungswissenschaft	Biologie	Bauingenieurwesen
	Literaturwissenschaft	Soziologie	Geowissenschaften	Verfahrenstechnik
	Sprachwissenschaft	Wirtschaftswissenschaft	Astronomie	Informatik
	Medienwissenschaft	Kommunikationswissenschaft	etc.	etc.
	etc.	etc.		

schaft als geisteswissenschaftlicher Disziplin teilweise empirisch gearbeitet. Auch der Begriff „Sozialwissenschaft“ selbst wird unterschiedlich gebraucht, etwa – wie in diesem Buch – als Oberbegriff für eine Gruppe von Wissenschaftsdisziplinen, aber auch als Einzelwissenschaft ähnlich der Soziologie.

Ein Versuch, derartige Einteilungsprobleme zu überwinden stellt das Konzept der **Humanwissenschaften** („human sciences“) dar, das alle geistes-, sozial-, natur- und technikwissenschaftlichen Fächer bzw. Subdisziplinen vereint, die sich mit dem Menschen als zentralem Untersuchungsgegenstand befassen – dies wären dann neben Psychologie und Humanmedizin z. B. auch Humanbiologie, Anthropologie, Ethnologie, Archäologie, Erziehungswissenschaft, Wirtschaftswissenschaft, Sportwissenschaft, Pflegewissenschaft, Soziale Arbeit, Soziologie, Politikwissenschaft, Biomedizintechnik etc. Allerdings wird das Konzept der Humanwissenschaften nicht nur als gegenstandsorientierter Oberbegriff verwendet, sondern teilweise – etwa in der Psychologie – auch mit einer eher geisteswissenschaftlichen Orientierung verknüpft (Barrell, Aanstoos, Rechards, & Arons, 1987; Smith, 1994). Zudem ist zu beachten, dass in einigen sozialwissenschaftlichen Disziplinen auch mit theoretischen Modellen gearbeitet wird, die ausdrücklich *nicht* den Menschen als Untersuchungseinheit betrachten (z. B. systemtheoretische Ansätze), so dass das Etikett der „Humanwissenschaft“ dann missverständlich wirken könnte.

Im vorliegenden Methodenlehrbuch verwenden wir die Bezeichnungen „Humanwissenschaften“ und „Sozialwissenschaften“ weitgehend synonym und als **Oberbegriffe für ein breites Spektrum an wissenschaftlichen Einzeldisziplinen und Subdisziplinen**, die sich empirisch im weitesten Sinne mit dem Menschen bzw. mit sozialen Sachverhalten befassen, wobei besonderes Augenmerk auf die Psychologie gelegt wird. Aber auch andere human- und sozialwissenschaftliche Einzeldisziplinen wie Soziologie, Erziehungswissenschaft, Medizin oder Kommunikationswissenschaft werden immer wieder durch fachspezifische Studienbeispiele angesprochen. Nicht zuletzt soll die inhaltliche und methodische Breite des Lehrbuchs auf interdisziplinäres Arbeiten vorbereiten.

Disziplinäre Einteilungen und Abgrenzungen sind für die Organisation des Wissenschaftsbetriebs notwendig (z. B. fachspezifische Institute und Studiengänge). Gleichzeitig ist für die umfassende Bearbeitung vieler Forschungsprobleme angesichts ihrer Komplexität ein Austausch über Disziplingrenzen hinweg sinnvoll und notwendig. Dabei sind drei Typen der Kooperation zu unterscheiden:

1. **Multidisziplinarität**: Eine Fragestellung wird von unterschiedlichen Fachdisziplinen parallel beleuchtet und die Ergebnisse werden additiv zusammengetragen. Beispiel: Das *Journal of Youth and Adolescence* bezeichnet sich als „multidisziplinäres“ Publikationsorgan, weil es wissenschaftliche Beiträge zum Jugendalter bzw. zu Jugendlichen veröffentlicht, die aus Psychologie, Psychiatrie, Biologie, Kriminologie, Erziehungswissenschaft etc. stammen können.
2. **Interdisziplinarität**: Eine Fragestellung wird von unterschiedlichen Fachdisziplinen bearbeitet, wobei es zu einem wechselseitigen Austausch und einer Kom-

bination von Theorien, Methoden und Ergebnissen unter Berücksichtigung der jeweiligen disziplinspezifischen Fragestellungen und Methoden kommt. Interdisziplinarität fordert von den Forschenden entweder eine Ausbildung in mehreren Disziplinen und/oder die Bildung fächerübergreifender Forschungsteams. Dabei ist eine intensive Kommunikation zwischen den Vertreterinnen und Vertretern verschiedener Fächer notwendig, um ein gemeinsames Vokabular zu finden und die fachspezifischen Theoriemodelle und Forschungsmethoden wechselseitig zu vermitteln und zu verzahnen (Blanckenburg, Böhm, Dienel, & Legewie, 2005; Jungert, Romfeld, Sukopp, & Voigt, 2010). Interdisziplinarität beginnt bei der Zusammenarbeit von benachbarten Subdisziplinen innerhalb eines Faches und reicht bis zur Kooperation zwischen sehr unterschiedlich geprägten Fächern, so dass bei interdisziplinärem Arbeiten auch zwischen den Kulturen der verschiedenen Disziplinen zu vermitteln ist (► Beispiele für interdisziplinäre Forschung zwischen sehr unterschiedlichen Disziplinen).

3. **Transdisziplinarität**: Transdisziplinarität meint ebenso wie Interdisziplinarität eine enge fächerübergreifende Zusammenarbeit. Bei der Transdisziplinarität liegt dabei zusätzlich der Akzent auf dem Wissenstransfer und der Kooperation mit Praxisvertretern bzw. politischen Entscheidungsträgern (Hirsch Hadorn et al., 2008; Sukopp, 2010).

1.1.3 Forschungsansätze in der empirischen Sozialforschung

Innerhalb der empirischen Sozialforschung existieren unterschiedliche Forschungsansätze. Es wird vor allem zwischen quantitativer und qualitativer Forschung unterschieden sowie zwischen Grundlagen- und Anwendungsforschung, die alle jeweils methodische und methodologische Besonderheiten aufweisen.

Quantitative und qualitative Forschung

Der quantitative Forschungsansatz (► Abschn. 2.2) ist in vielen human- und sozialwissenschaftlichen Disziplinen (z. B. Psychologie, Soziologie) der **historisch ältere und führende** (zur Geschichte der modernen Sozialwissenschaften s. Backhouse & Fontaine, 2010; Porter & Ross, 2003). So wird der Beginn der **Institutionalisierung der Psychologie** als wissenschaftliche Disziplin meist auf 1879 datiert, das Jahr der Einrichtung des „Laboratoriums für experimentelle Psychologie" an der Universität Leipzig durch **Wilhelm Wundt** (1832–1920). In dem Labor wurden psychophysikalische Experimente durchgeführt, etwa zu Reaktionszeiten bei der visuellen und akustischen Wahrnehmung unterschiedlich starker Stimuli. Nach dem Vorbild des Leipziger Instituts wurden national und international weitere Labore für Experimentalpsychologie eingerichtet.

Als Gegenströmung zu diesem quantitativ-naturwissenschaftlichen Paradigma in der Psychologie wurde eine qualitativ-geisteswissenschaftliche Psychologie propagiert, namhaft durch **Wilhelm Dilthey** (1833–1911), der zwischen einer quantitativen bzw. „zergliedernden" und einer qualitativen bzw. „beschreibenden" Psychologie unterschied (Dilthey, 1923). Im qualitativen bzw. geisteswissenschaftlichen Ansatz (► Abschn. 2.3) sollen soziale Sachverhalten umfassend beschrieben werden, d. h. nicht einzelne Variablen, sondern der „ganze Mensch" steht im Fokus. Zudem gehe es nicht um das **„Erklären"** von Sachverhalten aufgrund von Ursache-Wirkungs-Mechanismen wie in den Naturwissenschaften, sondern um das **„Verstehen"** des subjektiven Sinns menschlichen Handelns.

Auch die Institutionalisierung der **Soziologie** im deutschsprachigen Raum war geprägt von einem **Methoden-** bzw. **Paradigmenstreit,** der sich um die Orientierung an naturwissenschaftlichen oder geisteswissenschaftlichen Methoden dreht. Dieser Streit kulminierte in der deutschsprachigen Soziologie zu Beginn des 20. Jahrhunderts im sog. **Werturteilsstreit** und wurde mit anderer Akzentuierung in den 1960er-Jahren im **Positivismusstreit** aufgegriffen und fortgesetzt (für eine Zusammenfassung dieser wissenschaftstheoretischen Kontroverse s. Keuth, 1989; ► Abschn. 2.2.11).

Die Orientierung am naturwissenschaftlich-quantitativen Wissenschafts- und Methodenverständnis bedeutete historisch für die sich am Beginn des 20. Jahrhunderts herausbildenden Sozialwissenschaften wie Psychologie oder Soziologie die **Übernahme eines erfolgreichen und anerkannten Modells wissenschaftlichen Erkenntnisgewinns**. Gerade weil soziale Sachverhalte als relativ alltagsnahe Untersuchungsgegenstände oft auf der Basis subjektiver Einstellungen und anekdotischer Evidenzen beurteilt werden (► Abschn. 1.1.1), erscheint es sinnvoll, einen wissenschaftlichen Zugang zu sozialen Sachverhalten zu finden, der jenseits persönlicher Erfahrungen und Meinungen zu objektivierbaren und verallgemeinerbaren Erkenntnissen führt. Der Kausalnachweis im Experiment oder die Umfrageforschung auf der Basis bevölkerungsrepräsentativer Stichproben sind typische Beispiele quantitativer Methodologien. Die komplexe soziale Erfahrungswirklichkeit wird dabei durch eine theoriebasierte Erfassung einzelner Merkmale bzw. Variablen und der Prüfung ihrer Relationen zueinander beschrieben und erklärt (für aktuelle Handbücher zum quantitativen Ansatz s. z. B. Hancock & Mueller, 2010; Holling & Schmitz, 2010). Als wissenschaftstheoretische Basis fungiert der

Beispiele für interdisziplinäre Forschung zwischen sehr unterschiedlichen Disziplinen

Der britische Chemiker und Schriftsteller **Charles Percy Snow** (1905–1980) stellte Ende der 1950er-Jahre auf der Basis seiner persönlichen Erfahrungen im Umgang mit Naturwissenschaftlern sowie mit Literaten die **These der zwei Kulturen** auf (Snow, 1959). Demgemäß seien naturwissenschaftlich-technisch tätige Intellektuelle ganz anders geprägt hinsichtlich ihres Wissens sowie ihrer Einstellungen und Werte als geisteswissenschaftlich-literarisch arbeitende Intellektuelle – zwischen Vertretern beider Kulturen gäbe es kaum Verständigungsmöglichkeiten: Das geisteswissenschaftlich-literarische Lager sei z. B. eher historisch orientiert und kulturkritisch, das naturwissenschaftlich-technische Lager dagegen fortschrittsoptimistisch. Die einen hätten Shakespeare nie gelesen, den anderen seien die Gesetze der Thermodynamik völlig unbekannt. Snow kritisierte das britische Bildungssystem, das im Unterschied zum nordamerikanischen oder deutschen System Bildung zu einseitig geisteswissenschaftlich-literarisch belege und zu wenig naturwissenschaftlich-technische Bildung vermittle, die in technisch hochentwickelten Gesellschaften aber wichtig sei. Die These der „zwei Kulturen" wurde und wird intensiv diskutiert, wobei die inzwischen verstärkten inter- und transdisziplinäre Forschungsbemühungen zur Überwindung dieser Kluft beitragen sollen.

Ein Beispiel für die ausdrückliche Integration beider Kulturen ist die relativ junge Disziplin der **Digital Humanities** (digitale Geisteswissenschaften; auch: „e-Humanities", „humanities computing"), in der Computer- und Online-Technologien genutzt und weiterentwickelt werden, um geistes- und kulturwissenschaftliche Forschung zu betreiben. Weltweit existieren inzwischen Dutzende von Digital-Humanities-Forschungszentren. Die internationale Dachorganisation ist die 2005 gegründete Alliance of Digital Humanities Organizations (ADHO; http://www.digitalhumanities.org). Von Vertreterinnen und Vertretern dieser Disziplin wird erwartet, dass sie sowohl über geisteswissenschaftliche (z. B. Kunstgeschichte, Sprachwissenschaft, Literaturwissenschaft) als auch technikwissenschaftliche (v. a. Informatik, Computerwissenschaft) Kompetenzen verfügen und diese in der Forschung integrieren. Dass die Auseinandersetzung mit klassischen kulturellen Artefakten eben nicht im Widerspruch steht zur Beschäftigung mit modernen Computertechnologien, sondern dass beide Bereiche wissenschaftlich voneinander profitieren können, belegen die globalen Initiativen digitaler Archive und Museen. Sie tragen sowohl dazu bei, kulturelle Artefakte, die oft schwer zugänglich sind, in digitaler Form zu erhalten und für Öffentlichkeit und Forschung zugänglich zu machen, als auch die computertechnischen Grundlagen für entsprechende Archivierung und Dokumentation sowie Analyse weiterzuentwickeln.

Ein anderes Beispiel für die fruchtbare Verknüpfung scheinbar sehr unterschiedlicher Forschungskulturen ist die **neurowissenschaftliche Erforschung fernöstlicher Meditation**, wobei u. a. Kognitionsforscher und buddhistische Gelehrte zusammenarbeiten und z. B. die von Meditationslehren und -techniken angezielten Bewusstseinszustände hinsichtlich ihrer hirnphysiologischen Korrelate untersucht werden. Ein entsprechender Wissenschaftsdialog zwischen Kognitions-/Neurowissenschaft und dem Dalai Lama findet seit den 1980er-Jahren statt und wird heute u. a. über das gemeinnützige „Mind & Life-Institute" (http://www.mindandlife.org) organisiert. Hintergrund derartiger interdisziplinärer Kooperationen ist eben nicht die Betonung von kulturellen Unterschieden, sondern vor allem die Suche nach den verbindenden und integrierbaren Elementen naturwissenschaftlich geprägter Erforschung des Bewusstseins einerseits und traditioneller philosophisch-spiritueller Lehren andererseits.

Die neue Wissenschaftsdisziplin der Digital Humanities sowie Kooperationsprojekte zwischen Neurowissenschaft und Buddhismus sind Beispiele für Interdisziplinarität, bei denen sehr unterschiedliche Fächer und Traditionen aufeinander treffen. Interdisziplinäres Arbeiten findet oft aber auch zwischen unmittelbar **benachbarten Subdisziplinen** statt, etwa in der Psychologie zwischen Entwicklungs- und Sozialpsychologie oder in der Medizin zwischen Psychosomatik und Psychiatrie.

Kritische Rationalismus von Karl Popper (1934/1989), demgemäß wissenschaftlicher Erkenntnisgewinn in den Sozial- und Naturwissenschaften sich nicht prinzipiell voneinander unterscheidet.

Quantitative Sozialforschung – Quantitative Sozialforschung („quantitative social research") steht in der wissenschaftstheoretischen Tradition der **Naturwissenschaften**. Sie bedient sich im Zuge eines sequenziell strukturierten Forschungsprozesses quantitativer d. h. **strukturierter Methoden der Datenerhebung** (z. B. standardisierte Fragebogenerhebungen, psychologische Testverfahren, physiologische Messungen), aus denen quantitative bzw. **numerische Daten** (Messwerte) resultieren, die **statistischen Methoden der Datenanalyse** unterzogen werden.

Kritischer Rationalismus als Wissenschaftstheorie – Wissenschaftstheoretische Basis des quantitativen Paradigmas in der empirischen Sozialforschung ist vor allem der **Kritische Rationalismus** („critical rationalism"). Er geht davon aus, dass Erkenntnisgewinn dadurch zustande kommt, dass man zunächst Theorien formuliert, daraus Hypothesen ableitet und diese in nachvollziehbarer Weise anhand von Daten prüft. Dabei wird die Erfahrungswirklichkeit anhand ihrer einzelnen Merkmale (Variablen) und deren Relationen untereinander anhand größerer Stichproben von Untersuchungseinheiten analytisch untersucht.

Wichtige quantitative Methodologien – Wichtige quantitative Methodologien sind u. a. die **Umfrageforschung** („survey research"), die mit sehr großen und teilweise bevölkerungsrepräsentativen Stichproben arbeitet, welche anhand standardisierter Fragebögen untersucht werden, sowie die **Experimentalforschung** („experimental research"), bei der Ursache-Wirkungs-Relationen durch kontrollierte Variation der Bedingungen und Messung der Effekte im Feld oder im Labor geprüft werden. Typisch für quantitative Sozialforschung ist darüber hinaus der Einsatz von **psychometrischen Tests** („psychometric tests") sowie **physiologischen Messungen** („physiological measurements"), um psychische und physische Merkmale des Menschen präzise zu quantifizieren.

Was Anhängern des qualitativen Paradigmas an diesem naturwissenschaftlich-quantitativen Wissenschaftsverständnis missfiel, waren vor allem zwei Punkte. Auf der Ebene des **Wirklichkeitsverständnisses** (Ontologie; ► Abschn. 2.1) wurde die Vorstellung bestritten, es gäbe

eine objektiv beobachtbare soziale Erfahrungswirklichkeit, die sich durch Analysen auf Variablenebene erklären lässt. Vielmehr ging man davon aus, dass vielfältige soziale Erfahrungswirklichkeiten nebeneinander existieren und diese sich nur untersuchen lassen, indem man ganzheitlich und kontextbezogen die Handlungsweisen und Sichtweisen der einzelnen Akteurinnen und Akteure rekonstruiert. Anstelle der Messung vordefinierter Variablen sollten dazu offene Beobachtungs- und Befragungsmethoden eingesetzt werden, um auch Aspekte berücksichtigen zu können, die vorab von den Forschenden nicht erwartet wurden. Ein zweiter Hauptkritikpunkt des qualitativen Paradigmas am quantitativen Paradigma betrifft die **Rolle der Forschenden.** Im quantitativen Paradigma herrscht ein Verständnis der Forschenden vor, demgemäß diese bei der Hypothesenprüfung eine möglichst neutrale Haltung einnehmen, um auf dieser Weise theorie- und empiriebasiert zu ausgewogenen und gesicherten Erkenntnissen zu gelangen. Demgegenüber geht das qualitative Paradigma davon aus, dass Forschende immer durch ihren jeweiligen Standpunkt, ihre Erfahrungen und Werte geprägt sind und dass diese in alle Phasen des Forschungsprozesses maßgeblich einfließen. Statt diese Hintergründe zu verschleiern, fordert der qualitative Ansatz, dass die Forschenden ihre Standpunkte offenlegen und sich als Forschende möglichst ausdrücklich für gesellschaftliche Verbesserungen einsetzen (für aktuelle Handbücher zum qualitativen Ansatz s. z. B. Denzin & Lincoln, 2012; Mey & Mruck, 2010).

Qualitative Sozialforschung – Qualitative Sozialforschung („qualitative social research") steht in der wissenschaftstheoretischen Tradition der **Geisteswissenschaften.** Sie bedient sich im Zuge eines zirkulär bzw. iterativ strukturierten Forschungsprozesses qualitativer d. h. **unstrukturierter Methoden der Datenerhebung** (z. B. teilnehmende Feldbeobachtung, narratives Interview), aus denen qualitative bzw. **nichtnumerische Daten** (d. h. Text-, Bild-, Videomaterial) resultieren, die **interpretativen Methoden der Datenanalyse** unterzogen werden.

Sozialkonstruktivismus als Wissenschaftstheorie – Die wissenschaftstheoretische Basis des qualitativen Paradigmas in der empirischen Sozialforschung ist deutlich heterogener als im quantitativen Paradigma. Je nach qualitativem Forschungsfeld wird u. a. mit **sozialkonstruktivistischen**, interpretativen, poststrukturalistischen, symbolisch-interaktionistischen, phänomenologischen, dialektischen **Erkenntnistheorien** gearbeitet. Es existieren aber übergreifende wissenschaftstheoretische Prinzipien qualitativer Sozialforschung. So wird in der Regel davon ausgegangen, dass Erkenntnisgewinn dadurch zustande kommt, dass man zunächst den Untersuchungsgegenstand in seinem Kontext sowie die Sichtweisen der Beteiligten detailliert rekonstruiert, um daraus dann schrittweise Hypothesen und Theorien abzuleiten und in mehreren Überarbeitungsschleifen mit den Daten abzugleichen. Dabei wird die Erfahrungswirklichkeit anhand der Detailbetrachtung weniger Fälle möglichst ganzheitlich untersucht.

Wichtige qualitative Methodologien – Wichtige qualitative Methodologien sind u. a. die **ethnografische Feldforschung** („ethnographic field research"), bei der die Forschenden über längere Zeiträume hinweg in das untersuchte soziale Feld eintauchen, um durch Miterleben des Alltags ein umfassendes Verständnis zu gewinnen, sowie die **Biografieforschung** („biographical research"), bei der die interessierenden sozialen Sachverhalte im Kontext der Lebensgeschichten von Menschen betrachtet werden und im Zuge von narrativen Interviews Raum gegeben wird für ausführliche Erzählungen von Lebensereignissen. Ein populärer qualitativer Ansatz ist auch die **Grounded-Theory-Methodologie** („grounded theory methodology"), die durch theoretische Stichprobenbildung, die Kodiermethode des permanenten Vergleichs sowie systematisches Verfassen von Memos auf der Basis von qualitativem Datenmaterial schrittweise gegenstandsverankerte Theorien generiert.

Der Paradigmenstreit („paradigm controversy") in der empirischen Sozialforschung wurde bis in die 1980er-Jahre teilweise so heftig geführt, dass in der Literatur auch vom **Paradigmenkrieg** („paradigm war") die Rede ist (z. B. Lincoln, 2010). Denn beide Paradigmen sprachen sich teilweise die Existenzberechtigung ab. Der qualitative Ansatz warf dem dominierenden quantitativen Ansatz ein völlig verfehltes Wissenschaftsverständnis vor, auf dessen Basis nutzlose oder affirmative (d. h. die vorherrschenden gesellschaftlichen Machtverhältnisse bestätigende) Befunde produziert werden. Aus Sicht des quantitativen Paradigmas wurde dagegen die Wissenschaftlichkeit des qualitativen Ansatzes infrage gestellt. Mit dem Anspruch der Gesellschaftsverbesserung auf der Basis einiger weniger Fälle weitreichende Schlussfolgerungen zu ziehen, wie es der qualitative Ansatz propagierte, erschien der Mehrheit der quantitativ Forschenden fragwürdig.

Quantitative „Erbsenzählerei" versus qualitative „Kaffeesatzleserei" – diese Form des Paradigmenstreits, die den jeweils anderen Ansatz karikaturhaft darstellt, ist seit den 1990er-Jahren zunehmend abgelöst worden durch Sichtweisen, die beiden Paradigmen Stärken und Schwächen zuweisen und somit von einem fruchtbaren Ergänzungsverhältnis innerhalb der empirischen Sozialforschung ausgehen.

Auch haben die zahlreichen Forschungsprojekte, die – ungeachtet des Paradigmenstreits – qualitative und quantitative Methoden produktiv kombinieren, die Fronten teilweise aufgeweicht. Das gilt vor allem für die Anwendungsforschung, deren Fokus auf der praxisnahen Problemlösung liegt und die sich von wissenschaftstheoretischen Grundsatzstreits eher distanziert. Nicht zuletzt ist zu beobachten, dass sich innerhalb des quantitativen wie des qualitativen Paradigmas im Zuge des Paradigmenstreits auch Veränderungen ergeben haben. So wurden im quantitativen Paradigma Fragen nach Werten und Machtverhältnissen stärker aufgegriffen und im qualitativen Paradigma, bei aller Offenheit des bewusst unstrukturierten Vorgehens, eine stärkere Kanonisierung des Methodenspektrums sowie eine Konkretisierung methodischer Qualitäts- und Gütekriterien vorangetrieben (zu Gütekriterien in der qualitativen Forschung ► Abschn. 3.3).

Dass der qualitative Ansatz in der empirischen Sozialforschung heute keine Außenseiterrolle mehr hat, sondern – je nach Disziplin – mehr oder minder umfassend zum Mainstream dazugehört, zeigt sich anhand einer Reihe institutioneller Indikatoren, von denen einige exemplarisch aufgeführt seien (Mruck & Mey, 2010; Reichertz, 2007): Einschlägige sozialwissenschaftliche **Methodenlehrbücher** behandeln in den aktuellen Auflagen beide Ansätze (z. B. Bryman, 2008; Neumann, 2003) und auch in den **Curricula sozialwissenschaftlicher Studiengänge** sind inzwischen meist quantitative sowie qualitative Methodenveranstaltungen verankert. Führende inhaltliche **Fachzeitschriften** publizieren ausdrücklich Beiträge sowohl aus dem quantitativen als auch dem qualitativen Paradigma. Methoden-Fachzeitschriften wie *Quality & Quantity* (Springer) befassen sich mit fortgeschrittenen statistischen Verfahren ebenso wie mit qualitativen Methodologien. Eine Inhaltsanalyse und Herausgeberbefragung von 15 international führenden Psychologiezeitschriften zeigte eine – wenn auch noch ungenügend spezifizierte und kommunizierte – Offenheit für qualitative Beiträge (Kidd, 2002). Die Methoden-Division der **wissenschaftlichen Fachgesellschaft** American Psychological Association (APA) heißt „Evaluation, Measurement and Statistics“ und beinhaltet eine Sektion „Qualitative Methods“. Innerhalb der Deutschen Gesellschaft für Soziologie (DGS) wurde 2003 ergänzend zur Sektion „Methoden der empirischen Sozialforschung“ eine neue Sektion „Methoden der qualitativen Sozialforschung“ eingerichtet.

Auch wenn der Methoden- bzw. Paradigmenstreit teilweise fortlebt, ist inzwischen die Auffassung konsensfähiger geworden, dass es keine per se „guten“ oder „schlechten“ Forschungsmethoden bzw. -methodologien gibt, sondern dass der gewählte Forschungsansatz zum jeweiligen sozialwissenschaftlichen Forschungsproblem sowie den Kontextbedingungen des Forschungsprojektes passen muss. Die einstige Gegnerschaft der Paradigmen ist somit häufig einer friedlichen Koexistenz gewichen. Darüber hinaus wird in den letzten Jahren sogar eine Kooperation und Integration beider Ansätze zu einem **Mixed-Methods-Ansatz** („mixed methods approach“, „mixed methods research“) vorangetrieben (Creswell & Plano Clark, 2011; Kelle, 2008; Kuckartz, 2012; Tashakkori & Teddlie, 2010; Teddlie & Tashakkori, 2009).

Der **Nutzen** einer Mixed-Methods-Sozialforschung (► Abschn. 2.4) besteht darin, dass das jeweilige Forschungsproblem auf der Basis quantitativer und qualitativer Methoden umfassender bearbeitet werden kann und dass durch die Integration der verschiedenen Forschungsstrategien und Datenquellen eine bessere Absicherung der Ergebnisse möglich ist. Die Durchführung von Mixed-Methods-Studien erfordert Methodenkompetenzen im Bereich quantitativer, qualitativer und Mixed-Methods-Forschung. Der **Aufwand** ist bei Mixed-Methods-Studien in der Regel größer als bei quantitativen oder qualitativen Mono-Method-Studien. Wissenschaftstheoretisch wird aktuell noch kontrovers diskutiert, ob es sinnvoll ist, den Mixed-Methods-Ansatz als eigenes wissenschaftstheoretisches Paradigma aufzufassen, etwa auf der Basis des philosophischen **Pragmatismus** (► Abschn. 2.4.2).

Mixed-Methods-Sozialforschung – Der Mixed-Methods-Ansatz in der empirischen Sozialforschung („mixed methods approach“, „mixed-methods research“) **kombiniert und integriert quantitative und qualitative Forschungsstrategien innerhalb einer Studie** bzw. eines Forschungsprojekts, um dadurch den Erkenntnisgewinn zu erhöhen.

Pragmatismus als Wissenschaftstheorie – Es existieren unterschiedliche Auffassungen dazu, ob für erfolgreiche Mixed-Methods-Forschung eine eigenständige wissenschaftstheoretische Grundlage notwendig ist und wie diese aussehen sollte. Manche Beiträge erklären den Mixed-Methods-Ansatz zu einem dritten Paradigma innerhalb der empirischen Sozialforschung und sehen eine Fundierung im **philosophischen Pragmatismus**, der erkenntnistheoretische Positionen von Kritischem Rationalismus und Interpretionismus vereine.

Wichtige Mixed-Methods-Methodologien – Mixed-Methods-Forschung basiert immer auf einer dezidierten **Mixed-Methods-Methodologie** („mixed methods methodology“), die quantitative und qualitative Forschungsprozesse miteinander koppelt. Dazu stehen diverse **Kombinationsstrategien** zur Verfügung, die jeweils Verzahnungen in unterschiedlichen Phasen des Forschungsprozesses – etwa auf der Ebene der Forschungsfrage, des Untersuchungsdesigns, der Datenerhebung, der Datenanalyse und/oder der Interpretation – beinhalten.

Grundlagen- und Anwendungsforschung

Wenn eine Studie vor allem darauf ausgerichtet ist, zum wissenschaftlichen Kenntnisstand in einem bestimmten Forschungsbereich beizutragen, handelt es sich um so genannte **Grundlagenforschung** („basic research“, „pure research“). Man spricht auch von **akademischer Forschung** („academic research“), da sie vor allem an Universitäten betrieben wird. Ergebnisse der Grundlagenforschung sind zunächst innerhalb der Wissenschaftsgemeinschaft („scientific community“) von Interesse. Sie werden über Fachzeitschriften und Konferenzen national und international unter Forschenden ausgetauscht und regen weitere Forschungsaktivitäten an wie die folgenden:

- Man prüft eine etablierte Theorie und entwickelt sie weiter oder entwirft eine neue Theorie.
- Man versucht, spektakuläre oder überraschende empirische Befunde zu bestätigen bzw. zu widerlegen.
- Man konstatiert Forschungslücken und bemüht sich, diese zu schließen, indem bislang vernachlässigte Sachverhalte untersucht werden.

- Man fasst die Befunde unterschiedlicher Studien zum selben Forschungsthema zu einem Gesamtresultat zusammen.
- Man entwickelt eingeführte Forschungsmethoden weiter oder schlägt innovative qualitative, quantitative oder Mixed-Methods-Methoden bzw. -Methodologien vor.

Teilweise werden Ergebnisse der Grundlagenforschung auch in der breiten Öffentlichkeit bekannt, etwa vermittelt durch Wissenschaftsjournalismus. Aus Sicht von Laien wirken viele grundlagenwissenschaftliche Projekte in der empirischen Sozialforschung abstrakt und belanglos, wenn kein direkter Bezug zur eigenen Lebenswelt erkennbar ist. Wenn es Forschenden z. B. anhand der Befunde von psychologischen Laborexperimenten gelingt, die Überlegenheit einer Theorie gegenüber einer anderen zu begründen, dann kann das wissenschaftsintern ein wichtiges Ergebnis sein, auch wenn daraus zunächst keinerlei direkte Konsequenzen für die Praxis folgen.

Wie ist also der gesellschaftliche Nutzen der überwiegend staatlich finanzierten Grundlagenforschung zu begründen? Abgesehen davon, dass Erkenntnisgewinn per se als Wert zu betrachten ist und wissenschaftliche Weiterentwicklung vor allem durch die Grundlagenforschung vorangetrieben wird, zeigt sich bei ihr auch ein starker indirekter Praxisbezug: So greift die praxisorientierte Anwendungsforschung regelmäßig auf Theorien und Methoden zurück, die in der Grundlagenforschung entwickelt wurden und wäre somit ohne diese gar nicht möglich (zur Begründung und Messung des Nutzens der Grundlagenforschung s. auch Schibany, & Gassler, 2010).

Grundlagenforschung bzw. akademische Forschung – Die Grundlagenforschung („basic research", „pure research") bzw. akademische Forschung („academic research") zielt auf die Erweiterung des wissenschaftlichen Kenntnisstandes ab.

Anwendungsforschung – Die Anwendungsforschung („applied research") widmet sich praxisbezogenen Fragestellungen und erfolgt oft als Auftragsforschung.

Nicht allgemeiner wissenschaftlicher Erkenntnisgewinn, sondern die Beantwortung praxisorientierter Fragestellungen ist das Ziel der **Anwendungsforschung** („applied science"). Hier handelt es sich oft um **Auftragsforschung**, d. h. von Praxisvertretern wird die Durchführung einer entsprechenden angewandten sozialwissenschaftlichen Studie zur Lösung eines Problems oder Klärung einer Frage in Auftrag gegeben (▶ Abschn. 7.2). Auftragnehmer sind dabei teilweise Wissenschaftlerinnen und Wissenschaftler an Hochschulen sowie an außeruniversitären Forschungseinrichtungen, die auf diese Weise Drittmittel einwerben. Zum großen Teil wird angewandte Sozialforschung aber von privatwirtschaftlichen Markt-, Meinungs-, Wahl-, Zukunfts- oder Sozialforschungsinstituten durchgeführt. Nicht zuletzt realisieren große Unternehmen Anwendungsforschung in ihren betriebseigenen Forschungs- und Entwicklungs- sowie Marketingabteilungen.

Die Befunde einer angewandten Studie sind direkt praxisbezogen und sollen bei Entscheidungen im Praxisfeld helfen (z. B. Weiterführung, Veränderung oder Einstellung einer Maßnahme). Da sie auf die Besonderheiten des konkreten Anwendungsfalls zugeschnitten sind, ist es oft nicht möglich, verallgemeinerbare Schlussfolgerungen zu ziehen. Ihr Nutzen für den allgemeinen wissenschaftlichen Erkenntnisfortschritt ist deswegen oft begrenzt. Befunde von Auftragsstudien werden teilweise auch gar nicht publiziert bzw. unterliegen sogar ausdrücklich der Geheimhaltung.

Da Grundlagen- und Anwendungsforschung unterschiedliche Ziele verfolgen und die Anwendungsforschung theoretisch und methodisch auf die Grundlagenforschung aufbaut, stehen beide Varianten der empirischen Sozialforschung nicht in einem Konkurrenz-, sondern in einem Ergänzungsverhältnis zueinander (zur Gegenüberstellung von Grundlagen- und Anwendungsforschung ◘ Tab. 1.3).

Die Methodenausbildung im Studium ist meist zunächst an der Grundlagenforschung orientiert. Durch die Integration der Evaluationsforschung in die Curricula z. B. psychologischer oder erziehungswissenschaftlicher Studiengänge wurde der großen und wachsenden Bedeutung sozialwissenschaftlicher Anwendungsforschung Rechnung getragen. Hinsichtlich ihrer methodischen Ausrichtung unterscheiden sich sozialwissenschaftliche Grundlagen- und Anwendungsforschung graduell. So ist die Anwendungsforschung definitionsgemäß stark unter Druck, mit begrenzten Ressourcen praxisrelevante wissenschaftliche Befunde zu produzieren. Methodologische Grundsatzstreite oder einseitige Methodenpräferenzen sind hier weniger verbreitet, stattdessen sind Methodenkombinationen vor dem Hintergrund eines pragmatischen Wissenschaftsverständnisses üblicher. Das gilt z. B. im Anwendungsfeld der **Marktforschung** (vgl. Berekoven, Eckert & Ellenrieder, 2009; Buber & Holzmüller, 2009), wo neben standardisierten mündlichen, telefonischen oder Online-Umfragen auch qualitative Einzel- und Gruppeninterviews zum typischen Methodenarsenal gehören. Die **Evaluationsforschung** arbeitet häufig mit einem Mix von quantitativen und qualitativen Methoden und basiert auf Methodologien, die ausdrücklich für beide Ansätze offen sind (z. B. die sog. nutzenorientierte Evaluation; ▶ Kap. 18). Abgesehen von der Tendenz zur quantitativen und qualitativen Paradigmenintegration ist in der Anwendungsforschung nicht selten auch ein multi-, inter- oder transdisziplinäres Arbeiten notwendig,

Tabelle 1.3 Vergleich zwischen Grundlagen- und Anwendungsforschung. (Mod. nach Freeman & Rossi, 1984, S. 572. Copyright © 1984 by SAGE Publications. Reprinted by Permission of SAGE Publications; Neumann, 2003, S. 23. Neumann W. Lawrence, Social Research Methods: Qualitative and Quantitative Approaches, 5th, © 2003. Printed and Electronically reproduced by permission of Pearson Education, Inc., New York, New York)

Grundlagenforschung	Anwendungsforschung
1. Forschungsthemen können von den Forschenden meist relativ frei gewählt werden.	1. Forschungsthemen werden in der Regel von Auftraggebern vorgegeben.
2. Zielsetzung ist es, einen Beitrag zum wissenschaftlichen Erkenntnisgewinn zu leisten.	2. Zielsetzung ist es, für Auftraggeber und Praxis nützliche wissenschaftliche Befunde zu produzieren.
3. Grundlagenwissenschaftliche Studien orientieren sich theoretisch und methodisch am aktuellen Forschungsstand und streben größtmögliche wissenschaftliche Strenge an.	3. Anwendungswissenschaftliche Studien sind durch die Randbedingungen des Auftrags (Zeit- und Finanzrahmen) sowie die Praxisbedingungen limitiert; es müssen oft Abstriche bei der wissenschaftlichen Strenge gemacht werden.
4. Die Beurteilung der Forschungsergebnisse erfolgt durch Fachkollegen primär anhand wissenschaftsimmanenter Kriterien. Aber auch die praktische oder gesellschaftliche Relevanz der Befunde kann ein Kriterium sein.	4. Die Beurteilung der Forschungsergebnisse erfolgt durch Auftraggeber primär anhand praxisbezogener Erfordernisse. Auch an Anwendungsforschung werden jedoch Kriterien der Wissenschaftlichkeit angelegt.
5. Erfolgreiche grundlagenwissenschaftliche Studien werden in Fachzeitschriften publiziert und in der Fachcommunity zitiert. Ihre Theorien, Methoden und/oder Befunde sind zudem indirekt praxisrelevant, da Anwendungsforschung auf Grundlagenforschung aufbaut.	5. Erfolgreiche anwendungswissenschaftliche Studien liefern die Grundlage für Entscheidungen in der Praxis. Sie werden zuweilen aber auch publiziert und tragen ihrerseits zum wissenschaftlichen Erkenntnisgewinn bei.

wodurch sich ein Mix von Methoden und Methodologien unterschiedlicher (Sub-)Disziplinen ergibt.

Demgegenüber wird in der Grundlagenforschung häufig „puristischer“ vorgegangen. Man kann sich eher auf ein bestimmtes disziplinär verankertes Theoriemodell und eine bestimmte Methodologie konzentrieren und in diesem Rahmen mit großer methodischer Strenge arbeiten.

1.2 Empirische Studien verstehen und beurteilen

Wir haben schon betont, dass in der heutigen Wissenschaftsgesellschaft in vielen öffentlichen Debatten auf wissenschaftliche Befunde Bezug genommen wird. Dabei handelt es sich oft um sog. Sekundärquellen (z. B. Medienberichte über Studienergebnisse). In Studium und Forschung ist es dagegen notwendig, sich auch mit den Primärquellen empirischer Studien zu befassen und sich vor allem die Inhalte wissenschaftlicher Zeitschriftenartikel zu erarbeiten.

1.2.1 Primär- und Sekundärquellen für empirische Studien

Als **Primär- bzw. Originalquelle** („primary source“) einer empirischen Studie zählt deren erstmalige Darstellung durch die Forschenden selbst in einer Publikation. Besonderen Stellenwert haben dabei Publikationen in Fachzeitschriften, die zuvor einer unabhängigen und wechselseitig anonymen („doppelblinden“) fachlichen Begutachtung durch Kollegen („peer review“) unterzogen wurden. Ergebnisveröffentlichungen in Form von Buchkapiteln oder Büchern unterliegen dagegen meist nicht dem Peer Review. Konferenzvorträge sind zwar ebenso wie Zeitschriftenartikel in der Regel begutachtet, aber werden viel seltener als Quellen herangezogen, da oft nicht der gesamte Vortragstext, sondern nur die Kurzzusammenfassung (Abstract) verfügbar ist. **Begutachtete bzw. referierte Zeitschriftenartikel** gelten somit in den Human- und Sozialwissenschaften als die wichtigsten Primärquellen empirischer Studien.

Als **Sekundärquellen** („secondary sources“) sind Texte und Darstellungen zu werten, die inhaltlich auf Primärquellen Bezug nehmen und deren Inhalte summarisch, selektiv oder bewertend wiedergeben. In der Wissenschaft verbreitete Sekundärquellen sind z. B. **Lehrbücher**, in denen wichtige Studien dargestellt werden. Auch im vorliegenden Buch finden sich zahlreiche Darstellungen von Studienbeispielen. Diese Lehrbuchdarstellungen von Studien haben den Charakter von Sekundärquellen: Die Originalarbeiten werden jeweils unter ganz bestimmten didaktischen Aspekten aufbereitet und erheben nicht den Anspruch, alle oder auch nur die – aus Sicht der Forschenden – wichtigsten Aspekte abzudecken. Wer sich für die angesprochenen Studien genauer interessiert, sei also auf die jeweiligen Primärquellen verwiesen. **Vorlesungen**

und Skripte, die empirische Studien präsentieren, sind als Sekundärquellen einzuordnen und sollten z. B. nicht in Abschlussarbeiten zitiert werden, ohne dass man zuvor die Originalquelle konsultiert hat. Ebenso verhält es sich mit der Darstellung von Studien in populärwissenschaftlichen **Fachbüchern**, **Lexika** (z. B. Wikipedia-Artikel) oder in der **Presse** (z. B. BBC-Berichte oder Spiegel-Artikel über einzelne Studien). Derartige Sekundärquellen sind in eigenen wissenschaftlichen Arbeiten in der Regel nicht zitierfähig, wenn man etwas über die Originalstudie aussagen möchte. Nicht einmal auf **wissenschaftliche Artikel als Sekundärquellen** sollte man sich blindlings verlassen. Wird etwa in der Einleitung eines Zeitschriftenartikels auf Vorläuferstudien eingegangen, so werden diese jeweils unter bestimmten Aspekten relativ knapp dargestellt und nicht in ihrer Gänze gewürdigt. Aussagen über eine der dargestellten Vorläuferstudien sollten sich also auch hier nicht auf die Sekundärquelle stützen, sondern auf die Primärquelle.

Einen Sonderfall der wissenschaftlichen Sekundärquelle stellen **Übersichtsartikel** dar, die den Forschungsstand in einem Themenfeld zusammenfassen (z. B. narratives Review, systematisches Review). Ein noch spezielleren Sonderfall sind **Metaanalysen** („meta analysis"), die die Ergebnisse mehrerer früherer quantitativer Studien statistisch zu einem Gesamteffekt aggregieren (zu Reviews und Metaanalysen ► Kap. 16). Diese Artikel sind insofern Sekundärquellen, als sie keine eigenen neuen Daten präsentieren, sondern auf früheren Studien basieren. Sie sind in dieser Funktion als eigenständige wissenschaftliche Arbeiten zitierbar und haben durch ihre Leistung der Forschungssynthese besonders hohe Aussagekraft. Allerdings sind sie ebenfalls nicht als Quellen geeignet, um einzelne Studien umfassend zu diskutieren.

1.2.2 Wissenschaftliche Zeitschriftenartikel lesen

Im Folgenden werden einige Hinweise dazu gegeben, wie man wissenschaftliche Zeitschriftenartikel hinsichtlich ihrer formalen und inhaltlichen Merkmale erschließen und bewerten kann, sei es für ein Referat oder im Rahmen der Theoriearbeit für eine eigene empirische Studie (zur Literaturrecherche ► Kap. 6).

Formale Merkmale von Zeitschriftenartikeln

Bevor man sich auf die Inhalte des Beitrags einlässt, kann schon eine erste Beurteilung anhand seiner formalen Merkmale vorgenommen werden:

1. **Publikationsort**: Die Fachzeitschrift, in der der Beitrag erschienen ist, gibt Auskunft über seine Zuordnung zu einer Fachdisziplin oder einem Forschungsfeld. Zu unterscheiden ist zwischen nationalen (deutschsprachigen) und internationalen (englischsprachigen) Fachzeitschriften, wobei letztere oft ein höheres Ansehen genießen und einem strengeren **Peer Review** unterliegen. Schließlich wird die Bedeutung einer Zeitschrift noch in ihrem Impact Factor ausgedrückt, einer bibliometrischen Maßzahl, die angibt, wie oft Beiträge der entsprechenden Zeitschrift in anderen Zeitschriften zitiert wurden. Ein **Impact Factor** von 1.0 besagt, dass im Durchschnitt jeder in der Zeitschrift publizierte Artikel in einem definierten Zeitraum zumindest einmal zitiert wird. Das Ansehen einer Zeitschrift steigt mit dem Impact Factor, wobei dieser jedoch auch durch die Größe der jeweiligen Fachcommunity beeinflusst wird. Informationen über die Zeitschrift, ihren Leserkreis und Impact Factor ist ihrer jeweiligen Online-Präsenz (meist auf der Verlags-Website) zu entnehmen.
2. **Publikationsjahr**: Da sich soziale Phänomene fortlaufend verändern und die Wissenschaft sich weiterentwickelt (aktuelle Studien basieren auf früheren Studien bzw. dem bisherigen Forschungsstand), läuft man bei älteren Beiträgen Gefahr, dass diese inhaltlich, theoretisch und/oder methodisch mittlerweile überholt sind. Dennoch existieren in der empirischen Sozialforschung auch „klassische" Studien, die mehrere Jahrzehnte alt sind, aber auf deren Theorien, Methoden oder Befunde immer wieder zurückgegriffen wird.
3. **Autorinnen und Autoren**: Beiträge mit sehr langer Autorenliste weisen auf groß angelegte, interdisziplinäre und/oder interkulturelle Studien hin. Zudem liefern folgende Hintergrundinformationen über die Autorinnen und Autoren Hinweise auf die vermutete Qualität des Beitrags: a) disziplinäre Herkunft (aus dem Fach oder fachfremd); b) akademische Seniorität (noch im Studium vs. forschungserfahren) sowie c) inhaltliche Expertise (geringe oder hohe Zahl an einschlägigen Vorarbeiten zum Thema). Einem Zeitschriftenartikel ist in der Regel ein Hinweis auf die institutionelle Zugehörigkeit („affiliation") der Autorinnen und Autoren zu entnehmen, so dass man sich z. B. auf der Website der Institution über ihre sonstigen Forschungsaktivitäten informieren kann.
4. **Zitationen des Artikels**: In vielen Literaturdatenbanken (APA PsycInfo; Web of Knowledge; ► Abschn. 6.1.2) wird angegeben, wie oft und wo ein einzelner Artikel bereits zitiert wurde. Wenn ein Beitrag viele Zitationen aufweist, deutet dies auf die besondere Relevanz in der Fachcommunity hin. Die Zitationszahl ist jedoch am Thema und am Erscheinungsjahr zu relativieren: Beiträge zu sehr allgemeinen Themen (z. B. Methodenbeiträge) werden häufiger zitiert als Artikel

zu Spezialthemen; ältere Publikationen können mehr Zitationen sammeln als gerade erst erschienene.

Entsprechende formale Merkmale helfen bei der Einordnung eines Artikels, sind aber auch **nicht zu verabsolutieren**. Erst eine Betrachtung der Inhalte erlaubt es, die Substanz eines Beitrags wirklich einzuschätzen. Im Einzelfall kann sich z. B. ein älterer, abseitig publizierter Zeitschriftenartikel inhaltlich als sehr einschlägig und innovativ für das jeweilige Forschungsproblem erweisen, ein aktueller, hochrangig publizierter Beitrag dagegen als weniger relevant.

Inhaltliche Merkmale von Zeitschriftenartikeln

Um einen sozialwissenschaftlichen Text inhaltlich zu erschließen, ist ein strukturiertes Vorgehen einem einfachen sequenziellen Durchlesen des Textes in der Regel überlegen. Als Lesetechnik kann die **SQ3R-Methode** (Survey, Question, Read, Recite, Review; Robinson, 1978) eingesetzt werden, die das Erarbeiten eines Textes in fünf Phasen gliedert:

1. Überblick gewinnen („survey“)

Man überfliegt zunächst den Text und verschafft sich einen Gesamtüberblick über dessen Länge, Struktur und Hauptinhalte. Wichtige Anhaltspunkte dafür bieten der Titel des Beitrags sowie die Zwischenüberschriften. Zudem liefern wissenschaftliche Zeitschriftenartikel in der Regel am Anfang des Textes eine Kurzzusammenfassung (Abstract) sowie am Ende ein Fazit bzw. eine Diskussion – nur diese beiden Textelemente sollten zu Beginn gründlich gelesen werden, da sie die Hauptinhalte und Schlussfolgerungen präsentieren. Der restliche Text ist zunächst nur querzulesen. Ein typischer Anfängerfehler besteht darin, den Text von Anfang an ganz genau zu lesen, was bei anspruchsvollen wissenschaftlichen Beiträgen schon nach wenigen Absätzen oder Seiten dazu führt, dass man sich überfordert fühlt und in Details verliert. Das gilt insbesondere bei englischsprachigen Beiträgen, wenn während des Lesens noch ständig Vokabeln nachgeschlagen werden müssen. Deswegen sollte das Erarbeiten eben nicht mit einer kompletten Detail-Lektüre starten, sondern mit einem groben Überblick über den zu erarbeitenden Text. Im Ergebnis sollte man u. a. das Thema des Beitrags kennen und wissen, ob es sich um einen empirischen Beitrag handelt oder um einen Theorie- oder Methodenbeitrag, ob eine quantitative, qualitative oder Mixed-Methods-Studie vorliegt und ob es um Grundlagen- oder Anwendungsforschung geht.

2. Fragen stellen („question“)

Um den Beitrag nun detaillierter zu erarbeiten, stellen sich Fragen nach den einzelnen Aspekten der berichteten Studie, also nach 1. Forschungsthema, 2. theoretischem Rahmen, 3. Fragestellungen/Hypothesen, 4. Untersuchungsdesign, 5. Stichprobe, 6. Datenerhebungsmethoden, 7. Datenanalysemethoden, 8. Ergebnissen und Schlussfolgerungen. Dieses Fragenraster sollte man im Hinterkopf behalten bzw. anfangs schriftlich niederlegen.

3. Lesen („read“)

Das Fragenraster bildet dann den Hintergrund der Lektüre des Textes. Typischerweise sind wissenschaftliche Artikel bereits nach den im Fragenraster angesprochenen Hauptaspekten gegliedert. Während in quantitativen Studien der theoretische Rahmen und die Hypothesen am Anfang des Beitrags erläutert werden, da sie der quantitativen Datenerhebung vorausgehen, stehen sie bei qualitativen Studien typischerweise am Ende, da sie Ergebnisse der qualitativen Datenanalyse darstellen. Zur Beantwortung der Fragen wird der Text abschnittweise gelesen. Wichtige Stellen können farblich markiert, Fragen oder Ideen am Rand angemerkt werden, sei es auf dem eigenen Ausdruck oder im elektronischen Dokument.

4. Wiedergeben („recite“/„recall“)

Im Zuge des Lesens sollte jede Frage im Fragenraster kurz und bündig – und möglichst auch schriftlich – beantwortet werden. Auf diese Weise entsteht ein strukturiertes **Exzerpt**, eine Art Steckbrief der Studie. Der wissenschaftliche Text ist dann gründlich gelesen und verstanden worden, wenn zu allen gemäß Punkt 2 formulierten Kernfragen pointiert angegeben werden kann, was dazu im Artikel steht. Dies sollte in eigenen Worten erfolgen, allerdings unter Verwendung der einschlägigen Fachbegriffe.

5. Wiederholen und Bewerten („review“)

Abschließend ist die Lektüre des Zeitschriftenartikels anhand des ausgefüllten Fragenrasters bzw. strukturierten Exzerpts noch einmal zu rekapitulieren, damit sich die Merkmale der Studie einprägen. Sind einem die Merkmale der dargestellten Studie vertraut, kann abschließend noch eine Einordnung und Bewertung vorgenommen werden. Diese wird umso leichter fallen, je mehr wissenschaftliche Artikel man bereits gelesen hat und je besser man sich im jeweiligen Forschungsfeld auskennt. Generell werden wissenschaftliche Beiträge vor allem nach zwei Bündeln von Qualitätskriterien bewertet: Inhaltliche Relevanz („relevance“) und methodische Strenge („rigour“; ▶ Kap. 3).

Die **inhaltliche Relevanz** ist einzuschätzen vor dem Hintergrund des Forschungsstandes zum jeweiligen Untersuchungsthema und allgemein der Begründung des gewählten Forschungsproblems. **Relevanzkriterien** sind sehr vielfältig und hängen wesentlich von der eigenommenen Perspektive ab. So kann es grundlagenwissen-

schaftlich ebenso als wichtig gelten, einen bekannten Effekt erneut zu prüfen wie einen bislang unerforschten Gegenstand zu erkunden. Einer anwendungswissenschaftlichen Studie, die einen innovativen Lösungsansatz für ein praktisches Problem präsentiert, wird man Relevanz zusprechen, aber auch einer Studie, die aufzeigt, dass eine etablierte Maßnahme unter bestimmten Bedingungen versagt. Es kann als aufschlussreich gelten, ein Phänomen zu untersuchen, das massenhaft vorkommt, ebenso kann es den Kenntnisstand erweitern und ethisch wünschenswert sein, ein Minderheitenphänomen zu erforschen. Wenn man eine Studie als mehr oder weniger bedeutsam einstuft, sollte man immer reflektieren, welche Vorannahmen man dabei macht und welche Relevanzbewertungen aus anderen Perspektiven resultieren würden. Ob man persönlich eine Studie auf der Basis eigener Lebenserfahrungen und Themenpräferenzen subjektiv interessant findet oder nicht, ist dabei nicht mit der fachlichen Relevanz gleichzusetzen.

Das zweite Beurteilungskriterium ist die **methodische Strenge** einer Studie, also die Frage, inwiefern das gewählte methodische Vorgehen zu möglichst gut gesicherten wissenschaftlichen Ergebnissen führt. Dazu sind **wissenschaftliche Güte- und Qualitätskriterien** an alle Aspekte des methodischen Vorgehens anzulegen (► Kap. 3): An die Wahl des Untersuchungsdesigns, die Zusammensetzung und Größe der Stichprobe, die Entwicklung und Testung der Instrumente oder die Auswahl und Schulung von Interviewern. Sehr negativ auf die Beurteilung der wissenschaftlichen Strenge wirkt es sich aus, wenn methodische Fehler erkennbar sind (z. B. es werden statistische Datenauswertungsverfahren verwendet, deren Voraussetzungen vom Datenmaterial nicht erfüllt werden), wenn das methodische Vorgehen unvollständig dokumentiert ist (z. B. ungenügende Angaben über die Zusammensetzung der untersuchten Stichprobe), wenn die Erfüllung zentraler wissenschaftlicher Gütekriterien nicht sichergestellt wurde (z. B. es wird ein psychologisches Testverfahren eingesetzt ohne Überprüfung von dessen Messgenauigkeit) oder wenn der Forschungsprozess insgesamt keiner etablierten Methodologie folgt (z. B. es werden statistische Signifikanztests gerechnet, ohne dass vorab theoretisch begründete Hypothesen aufgestellt wurden).

Eine **kritische Lektüre** wissenschaftlicher Fachliteratur ist wünschenswert, in deren Zuge Studierende die inhaltliche Relevanz und methodische Strenge vorliegender Publikationen hinterfragen. Erfahrungsgemäß ist bei vielen Studierenden tendenziell keine unkritische, sondern eher eine **überkritische Haltung** gegenüber konkreten wissenschaftlichen Forschungsprojekten zu beobachten. Das mag daran liegen, dass die Methodenausbildung häufig stark grundlagenwissenschaftlich orientiert ist und idealtypisches methodisches Vorgehen vermittelt (z. B. in der quantitativen Forschung randomisierte Kontrollgruppen-Experimente oder bevölkerungsrepräsentative Umfragen). In der Forschungspraxis weisen Studien dagegen nicht selten beträchtliche Limitationen auf, die die methodische Strenge reduzieren (z. B. nicht-experimentelles Design, nicht-repräsentative Stichprobe). Häufig müssen aufgrund forschungsökonomischer Rahmenbedingungen Abweichungen vom idealen Vorgehen in Kauf genommen werden, um überhaupt eine Studie durchführen zu können. Eine faire Bewertung bemisst Studien also nicht pauschal an Methoden-Idealen, sondern vor dem Hintergrund ihrer selbst gesetzten Erkenntnisziele und unter Beachtung der forschungsökonomischen Rahmenbedingungen (etwa Zeit- und Kostenplan). Bevor man eine vorliegende Studie wegen deren methodischen Einschränkungen und Schwächen grundsätzlich verwirft, kann es aufschlussreich sein, sich die Frage zu stellen „Was ist/war die methodische Alternative?“. Limitationen sind dann zu akzeptieren, wenn eine strengere Methodik aus organisatorischen oder ethischen Gründen nicht realisierbar war und wenn die methodischen Grenzen bei der Ergebnisinterpretation ausdrücklich einbezogen werden.

Als Hilfsmittel zum Verständnis und zur Beurteilung internationaler empirischer Studien sind neben einem **Englisch-Wörterbuch** vor allem fachspezifische **Lexika, Handbücher** und **Glossare** zu nutzen, in denen einerseits die **inhaltlichen Konzepte und Theorien** des jeweils betrachteten Forschungsfeldes präsentiert und andererseits **quantitative und qualitative Methoden** erläutert werden (z. B. Hancock & Mueller, 2010; Wirtz 2013). Von einer einfachen Online-Recherche nach methodischen und inhaltlichen Fachbegriffen ist ausdrücklich abzuraten, da man dabei oft auf fehler- und lückenhafte Darstellungen stößt. So ist z. B. die deutschsprachige Wikipedia aktuell (noch?) keine zuverlässige Quelle für Fachbegriffe aus dem Bereich der sozialwissenschaftlichen Methodenlehre.

1.3 Empirische Studien planen und durchführen

Jede empirische Studie erfordert vor ihrer Durchführung eine gründliche Planung und Vorbereitung. Das betrifft:

- **inhaltliche Aspekte** (z. B. die Formulierung sinnvoller und überprüfbarer Forschungsfragen bzw. Forschungshypothesen auf der Basis des aktuellen Forschungsstandes, wofür sorgfältige Literaturrecherche und Theoriearbeit notwendig sind),
- **methodische Aspekte** (z. B. Auswahl eines geeigneten Forschungsdesigns und Stichprobenplans, Entwicklung und Überprüfung eines zum Forschungsproblem passenden Interviewleitfadens, eines stan-

dardisierten Fragebogens oder eines inhaltsanalytischen Kategoriensystems für die Studie, wofür eine Recherche etablierter Instrumente, die eigene Instrumentenentwicklung, mehrere Pretests und ggf. auch Interviewer- oder Kodiererschulungen notwendig sind),
- **forschungsökonomische und organisatorische Aspekte** (z. B. Eruierung praktikabler Wege, Orte, Zeiten, Kontaktpersonen, Ansprachemodalitäten, Anreize etc., um für die spätere Stichprobenziehung potenzielle Untersuchungspersonen in geeigneter Zahl zu erreichen und für die Teilnahme motivieren zu können) sowie
- **ethische Aspekte** (z. B. im Falle von Studien, bei denen die Untersuchungsteilnehmenden besonderen Belastungen ausgesetzt werden oder personalisierte Daten gesammelt werden, ist im Vorfeld die Konsultation von Fachliteratur zur Forschungsethik, von Fachkolleginnen und -kollegen, Datenschutzbeauftragen und Ethikkommissionen notwendig, um sicherzustellen, dass die geplante Vorgehensweise legitim ist und nicht die Rechte von Untersuchungsteilnehmenden in ethisch oder gar rechtlich illegitimer Weise beschneidet).

Die Planung einer Studie muss sich zum einen an wissenschaftlichen Standards orientieren und gemäß der gewählten Methodologie den typischen **Phasen des empirischen Forschungsprozesses** folgen. Zum anderen ist jede Studie als Forschungsprojekt, das in definierter Zeit mit begrenzten Ressourcen bestimmte Ergebnisse erzielen soll, immer auch unter **Projektmanagement-Perspektive** zu organisieren (z. B. realistische Zeit-, Kosten-, Personalplanung). Forschungspraktische bzw. forschungsökonomische Fragen nach Aufwand und Kosten einzelner Forschungsschritte sind bei der Untersuchungsplanung nicht zu vernachlässigen. Hier gilt es jeweils gemäß Anlass (z. B. Bachelorarbeit vs. Doktorarbeit) und Ressourcen (z. B. studentische Abschlussarbeit vs. gefördertes dreijähriges Forschungsverbundprojekt) die richtige Balance zu finden zwischen einer gemäß wissenschaftlicher Aussagekraft optimalen Vorgehensweise einerseits und einem mit vorhandenen Ressourcen und realistischem Aufwand machbaren Projekt andererseits.

Um selbst erfolgreich empirische Studien durchführen zu können, ist es wichtig, eine solche **Prozess- und Projektperspektive** zu entwickeln. Ein typischer Anfängerfehler besteht darin, eine Studie primär nur unter dem Aspekt der Datenerhebung zu betrachten und etwa zu glauben, eine Fragebogenstudie bestünde darin, einen Fragebogen zusammenzustellen und zu verteilen. Diese verkürzte Vorstellung vernachlässigt wichtige und zudem sehr zeitaufwändige Phasen des Forschungsprozesses (z. B. Literaturrecherche und Theoriearbeit, Formulierung von Forschungsfragen und -hypothesen, Konzeptspezifikationen, qualitative und quantitative Fragebogen-Pretests etc.), so dass die Gefahr besteht, dass die Studie am Ende wissenschaftlichen Kriterien nicht genügt (z. B. Verwendung eines Ad-hoc-Fragebogens, der weder theoretisch fundiert noch empirisch durch Pretests geprüft und optimiert worden ist) oder dass die Studie den zeitlichen Rahmen sprengt und somit nicht oder in unvollständiger Weise abgeschlossen wird (z. B. kommen bei nicht wenigen studentischen Arbeiten aufgrund suboptimalen Zeitmanagements – z. B. überambitionierte Sammlung von zu vielen Daten – am Ende die Datenauswertung, Interpretation und ansprechende Ergebnispräsentation zu kurz, was letztlich Sinn und Zweck der ganzen Forschungsbemühungen infrage stellt).

Im Folgenden werden wir den Forschungsprozess in der quantitativen sowie der qualitativen Sozialforschung in seinen wichtigsten Phasen vorstellen. Zudem werden wir kurz auf Mixed-Methods-Studien eingehen, die qualitative und quantitative Vorgehensweisen kombinieren.

1.3.1 Der Forschungsprozess in der quantitativen Sozialforschung

In der quantitativen Sozialforschung werden typischerweise mit standardisierten Messinstrumenten an relativ großen und möglichst repräsentativen Stichproben (Umfrageforschung) oder unter kontrollierten Bedingungen (Experimentalforschung) numerische Messwerte erhoben und statistisch ausgewertet, um theoretisch begründete Hypothesen zu prüfen.

Der **Forschungsprozess** lässt sich bei quantitativen empirischen Studien in neun Phasen einteilen (Tab. 1.4). Das in diesem Buch verwendete **Neun-Phasen-Modell** entspricht inhaltlich anderen in der Literatur gebräuchlichen Darstellungen (z. B. Bryman, 2008, S. 141; Neumann, 2003, S. 13). Unterschiede zwischen den verschiedenen Darstellungen des quantitativen Forschungsprozesses ergeben sich vor allem durch die Etikettierung der Phasen (statt von „Stichprobenziehung“ wird z. B. auch von „Sampling“, „Auswahlverfahren“ gesprochen) sowie durch die Abgrenzung/Zusammenlegung der Phasen (z. B. wird die Datenaufbereitung oft mit der Datenanalyse zusammengefasst).

Der quantitative Forschungsprozess wird in der Regel **sequenziell** abgearbeitet. So wird z. B. während der Datenerhebung das dabei verwendete standardisierte Datenerhebungsinstrument (z. B. der Fragebogen) nicht mehr verändert. Ebenso wird während der Datenanalyse der vorhandene Datensatz bzw. die verwendete Stichprobe nicht mehr durch nachträgliche weitere Daten-

Tabelle 1.4 Phasen des quantitativen Forschungsprozesses im Überblick

Phasen des empirischen Forschungsprozesses in der quantitativen Sozialforschung	Erläuterung
1. Forschungsthema (► Kap. 5)	Jede Studie beginnt mit der Entscheidung für ein **Forschungsthema** und der Konkretisierung eines relevanten **Forschungsproblems**. Zu diesem sind dann detailliertere **Forschungsfragen** bzw. **Forschungshypothesen** zu formulieren, wobei der aktuelle Forschungsstand sowie ein Theoriemodell zugrunde zu legen sind.
2. Forschungsstand und theoretischer Hintergrund (► Kap. 6)	Jede quantitativ-empirische Studie muss an den bisherigen **Forschungsstand** anknüpfen und einen **theoretischen Rahmen** vorgeben, um den interessierenden Sachverhalt zu untersuchen. Notwendige Voraussetzung ist eine gründliche Recherche und Aufarbeitung der Fachliteratur zum Forschungsthema. Zudem muss **Theoriearbeit** in dem Sinne geleistet werden, dass aus der Fülle theoretischer Ansätze die für das Forschungsproblem leistungsfähigsten ausgewählt, modifiziert und/oder miteinander zu einem Theoriemodell verknüpft werden, aus dem Forschungsfragen bzw. Forschungshypothesen abzuleiten sind.
3. Untersuchungsdesign (► Kap. 7)	Bei quantitativen sozialwissenschaftlichen Untersuchungen unterscheiden wir je nach Erkenntnisinteresse drei Typen von Studien: – **Hypothesengenerierende bzw. gegenstandserkundende** (explorative) Studien, die einen wenig erforschten Sachverhalt detailliert beschreiben und zur Theoriebildung beitragen sollen. – **Populationsbeschreibende** (deskriptive) Studien, die die Verteilung oder Ausprägung bestimmter Merkmale in definierten Bevölkerungsgruppen feststellen sollen. – **Hypothesenprüfende** (explanative) Studien, die theoretisch wohlbegründete Hypothesen anhand von Daten testen sollen. Je nach konkreten inhaltlichen Forschungsfragen bzw. Forschungshypothesen muss die Studie unterschiedlich angelegt sein; damit ist das Untersuchungsdesign gemeint. Es ist in verschiedener Hinsicht zu konkretisieren, etwa hinsichtlich Untersuchungsort (Feld- vs. Laborstudie), Untersuchungszeitpunkten (Querschnitt- vs. Längsschnittstudie), Herkunft der Daten (Primär- vs. Sekundäranalyse), Behandlung der Untersuchungspersonen (experimentelle vs. quasi-experimentelle vs. nicht-experimentelle Studie) etc. Die präzise und sachgerechte Ausarbeitung des Untersuchungsdesigns entscheidet darüber, ob die Studie a) praktisch mit den vorhandenen Mitteln durchführbar ist und b) zu aussagekräftigen Resultaten hinsichtlich des Forschungsproblems führt.
4. Operationalisierung (► Kap. 8)	Alle im Rahmen der zu prüfenden Hypothesen bzw. zu beantwortenden Forschungsfragen relevanten Merkmale der Erfahrungswirklichkeit müssen präzise **definiert** und in ihren wichtigen **Aspekten bzw. Dimensionen entfaltet werden** (**Konzeptspezifikation**). Auf dieser Basis muss dann festgelegt werden, wie die Merkmale bzw. ihre Ausprägungen zu **messen** sind, damit sich aussagekräftige quantitative Daten ergeben. Mit der Operationalisierung wird gleichzeitig das **Skalenniveau** der Variablen festgelegt und damit der Informationsgehalt der Daten sowie ihre statistischen Auswertungsmöglichkeiten.
5. Stichprobenziehung (► Kap. 9)	Zunächst ist zu entscheiden, ob die gesamte Population (**Vollerhebung**) oder nur ein Ausschnitt (**Stichprobenerhebung**) untersucht werden soll. Bei der Stichprobenziehung sind – die **Stichprobenart** (zufällige oder nicht-zufällige Auswahl) sowie – der **Stichprobenumfang** entscheidend für die Aussagekraft der Studie. Der Stichprobenplan muss durch eine geeignete Sammlung von Untersuchungseinheiten bzw. **Rekrutierung** von Untersuchungspersonen umgesetzt werden.
6. Datenerhebung (► Kap. 10)	Für die Erhebung quantitativer Daten stehen in den Human- und Sozialwissenschaften eine Reihe von **Datenerhebungsmethoden** zur Verfügung: – Strukturierte Beobachtung – Strukturierte mündliche Befragung (Interview) – Strukturierte schriftliche Befragung (Fragebogen) – Psychologischer Test – Physiologische Messung – Quantitative Dokumentenanalyse bzw. Inhaltsanalyse Um quantitative Daten auf diese Weise zu erheben, müssen zuvor im Zuge der Operationalisierung **standardisierte Erhebungsinstrumente** (Beobachtungsplan, Interviewleitfaden, Fragebogen, Kategoriensystem etc.) entwickelt worden sein. Zudem wird zur Datenerhebung in der Regel auch **geschultes Personal** (z. B. Interviewer, Kodierer) und mehr oder minder viel Zeit benötigt.

Tabelle 1.4 (Fortsetzung)

Phasen des empirischen Forschungsprozesses in der quantitativen Sozialforschung	Erläuterung
7. Datenaufbereitung (► Kap. 11)	Das erhobene **Rohdatenmaterial** muss vor der Datenanalyse sorgfältig aufbereitet werden (z. B. Sortierung, Bereinigung um Fehler, Anonymisierung etc.). Nach der Datenbereinigung stehen vollständige, fehlerfreie und kommentierte **Datensätze** zur Verfügung, meist in elektronischer Form.
8. Datenanalyse (► Kap. 12)	Die quantitative Datenanalyse erfolgt computergestützt über Tabellenkalkulationsprogramme oder spezielle **Statistik-Software** (z. B. SPSS). Die Analysestrategie hängt vom Untersuchungstyp ab: – Hypothesengenerierende bzw. gegenstandserkundende (explorative) Studien beschränken sich oft auf **deskriptiv-statistische und exploratorische (z. B. grafische) Auswertungen**. – Bei populationsbeschreibenden (deskriptiven) Studien steht die inferenzstatistische Schätzung von Populationsparametern mit Verfahren der **Punkt- und Intervallschätzung** im Zentrum. – Bei hypothesenprüfenden (explanativen) Studien geht es definitionsgemäß um die Testung von Hypothesen, wobei vor allem **klassische statistische Signifikanztests**, **Minimum-Effektgrößen-Tests** oder **Strukturgleichungsmodelle** zum Einsatz kommen. Die Datenanalyse mündet in eine **Interpretation der Befunde**: Die einzelnen Forschungsfragen werden beantwortet bzw. die Forschungshypothesen geprüft, eine Gesamtaussage über das zugrunde gelegte Theoriemodell und den Untersuchungsgegenstand wird getroffen, die Grenzen der Aussagekraft der Studie werden hervorgehoben und Schlussfolgerungen für die zukünftige Forschung sowie für die Praxis gezogen.
9. Ergebnispräsentation (► Kap. 13)	Die Ergebnisse empirischer Studien werden der Scientific Community mittels – **Zeitschriftenartikeln**, – **Konferenzvorträgen** und – **Postern** präsentiert, die jeweils einen **Peer-Review-Prozess** (Begutachtung durch Fachkollegen) durchlaufen. Darüber hinaus existieren weitere Veröffentlichungsmöglichkeiten (z. B. Bücher, Vorträge), die keiner Qualitätskontrolle durch Kollegenbegutachtung unterliegen. Die Kommunikation wissenschaftlicher Ergebnisse an die **breite Öffentlichkeit** etwa durch Internetpräsenzen, Presseinterviews oder populärwissenschaftliche Beiträge gewinnt an Bedeutung.

erhebungen vergrößert. Entscheidungen, die am Anfang des quantitativen Forschungsprozesses getroffen wurden, lassen sich also später nicht mehr revidieren. Deswegen ist die gründliche theoretische und empirische Vorarbeit (Literaturrecherche, Pretests von Instrumenten, Konsultation von Kollegen) in der quantitativen Forschung besonders wichtig.

Die sequenzielle Darstellung darf freilich nicht darüber hinwegtäuschen, dass alle Phasen inhaltlich eng miteinander **verzahnt** sind. So muss in der Phase der Operationalisierung, d. h. wenn festgelegt wird, wie die interessierenden Variablen gemessen werden sollen, die avisierte Datenerhebungsmethode berücksichtigt werden: Je nachdem, ob die Datenerhebung per mündlichem Interview, schriftlichem Fragebogen oder Online-Dokumentenanalyse erfolgen soll, ist z. B. ein Merkmal wie „Geschlecht" auf unterschiedliche Weise zu erfassen (z. B. Notiz des Interviewers, Frage im Fragebogen, Angaben im Online-Profil). Zudem müssen in der Phase der Operationalisierung auch die geplante statistische Datenanalyse und inhaltliche Ergebnisinterpretation berücksichtigt werden: Variablen, die für die Studie von großer Bedeutung sind und in viele statistische Analysen eingehen sollen, sind auf einem möglichst hohen Skalenniveau sowie mit möglichst validen Messinstrumenten zu erfassen. Bei weniger wichtigen Variablen sind aus forschungsökonomischen Gründen geringere Ansprüche an die Operationalisierung zu stellen (zu Details der Operationalisierung ► Kap. 8).

Den in Tab. 1.4 vorgestellten **neun Phasen des quantitativen Forschungsprozesses** ist in Teil II des vorliegenden Buches jeweils ein eigenes Kapitel gewidmet.

1.3.2 Der Forschungsprozess in der qualitativen Sozialforschung

In der qualitativen Sozialforschung (zum Überblick s. z. B. Denzin & Lincoln, 2012; Mey & Mruck, 2010) wird in bewusst nicht-strukturierter Weise zu relativ wenig Fällen umfassendes verbales, visuelles und/oder audiovisuelles Datenmaterial erhoben und interpretativ ausgewertet, um auf dieser Basis den Untersuchungsgegenstand im Kontext detailliert zu beschreiben sowie Hypothesen und Theorien zu entwickeln. Der qualitative Forschungsansatz

folgt primär einer **theorieentdeckenden Forschungslogik**, wobei das induktive, datengestützte Vorgehen besonders wichtig ist.

Wie jede wissenschaftliche Studie beginnt auch eine qualitative Untersuchung mit der Festlegung des **Forschungsthemas** und der Zuspitzung des Forschungsproblems. Im Unterschied zur quantitativen Forschung geht es jedoch nicht darum, im Vorfeld ein elaboriertes theoretisches Modell zu entwickeln und daraus die zu testenden Hypothesen abzuleiten, die sich auf Relationen zwischen einzelnen Variablen beziehen. Vielmehr werden auf der Basis des **Forschungsstandes** und vorliegender Theorien meist einige wenige, offene Forschungsfragen formuliert, die durch theoretische Vorannahmen nicht zu sehr einzuschränken sind (Prinzip der reflektierten theoretischen Offenheit; ► Abschn. 2.3.2).

Anschließend wird auf der Basis einer oder mehrerer qualitativer Methodologien entschieden, mit welchem **Untersuchungsdesign** die Studie zu realisieren ist. Dabei ist z. B. festzulegen, an wie vielen und welchen Orten die Studie durchzuführen ist, wie oft und wie lange Datenerhebungen erfolgen sollen. Auch über die Zusammensetzung der Stichprobe und die Wahl der Datenerhebungsmethoden sind Entscheidungen zu treffen. Eine Operationalisierung einzelner Variablen entfällt in der qualitativen Forschung, da diese darauf verzichtet, vor der Datenerhebung bereits alle interessierenden Merkmale exakt festzulegen (wie dies z. B. bei einem standardisierten Fragebogen mit exakt vorgegebenen Fragen und Antwortalternativen der Fall ist). Stattdessen werden für die nicht-strukturierte Datenerhebung in qualitativen Studien nicht-standardisierte Datenerhebungsinstrumente verwendet (z. B. offene Fragen bzw. Erzählaufforderungen im Interview, frei formulierte Feldnotizen), die reichhaltiges verbales, visuelles und/oder audiovisuelles Rohdatenmaterial generieren. Das **nicht-strukturierte bzw. offene Vorgehen** soll bei der Datenerhebung zu mehr Gegenstandsangemessenheit führen und auch das Aufdecken vorher nicht erwarteter Aspekte ermöglichen.

Da zu jedem Einzelfall umfassendes Material gesammelt wird, beschränkt man sich in der qualitativen Forschung auf **kleine Stichproben**, die sich aus bewusst – gemäß ihrem Informationsgehalt – ausgewählten Fällen zusammensetzen (zu Typen qualitativer Stichproben ► Abschn. 9.2).

Idealtypisch für qualitative Studien ist ein **zirkuläres Vorgehen** (► Abschn. 2.3.3): Die jeweils erhobenen Daten werden unmittelbar einer ersten **Datenaufbereitung** und **Datenanalyse** unterzogen, auf der Basis der so gewonnenen Zwischenergebnisse wird schrittweise die **weitere Datenerhebung** gesteuert, d. h. der Aufmerksamkeitsfokus kann bei der Datenerhebung auf neue Aspekte gerichtet werden, zusätzliche Fälle können in das Sample Eingang finden. Idealerweise werden Datenerhebung und vorläufige Datenauswertung so lange fortgesetzt, bis keine substanziell neuen Erkenntnisse mehr entstehen (theoretische Sättigung).

In der abschließenden Phase der Datenaufbereitung und detaillierten interpretativen Datenanalyse wird das reichhaltige qualitative Rohdatenmaterial zu **neuen Hypothesen oder Theorien verdichtet**. In der qualitativen Forschung findet somit ein Großteil der Arbeit – insbesondere der Theoriearbeit – erst während bzw. nach der Datenerhebung statt. Sind Datenanalyse und Theoriebildung abgeschlossen, so sind noch die Grenzen der Aussagekraft der Studie zu verdeutlichen und die Konsequenzen für die zukünftige Forschung sowie für die Praxis aufzuzeigen. Im Zuge der **Ergebnispräsentation** wird die Studie der wissenschaftlichen Fachcommunity und zuweilen auch der breiten Öffentlichkeit vorgestellt.

Der große Vorteil des nicht-strukturierten und zirkulären Vorgehens besteht darin, dass der Forschungsprozess im Verlauf einer Studie gleichzeitig einen inhaltlichen Lernprozess für die Forschenden darstellt und Zwischenergebnisse das weitere methodische Vorgehen steuern können. Im Idealfall kann somit das Verständnis des Untersuchungsgegenstandes im Zuge einer Studie schrittweise verbessert und immer größere Gegenstandsnähe hergestellt werden. Der Nachteil dieses offenen Vorgehens besteht darin, dass der Forschungsprozess nicht vorab hinsichtlich seines Zeit-, Personal- und Kostenaufwandes präzise planbar ist. Eine solche **Planbarkeit** ist aber in der Grundlagen- wie Anwendungsforschung in der Regel notwendig. Häufig werden deswegen auch in der qualitativen Forschung Stichprobengrößen vorab festgelegt. Zirkulär wird dann lediglich im Zuge der Datenanalyse verfahren, indem kein vorgefertigtes Analyseraster angelegt wird, sondern die inhaltlichen Kategorien zur Interpretation des qualitativen Datenmaterials schrittweise erarbeitet und mit zunehmender Erschließung des Datenmaterials auch immer wieder modifiziert und präzisiert werden. Der sequenzielle quantitative und der zirkuläre qualitative Forschungsprozess sind in ◘ Abb. 1.1 einander gegenübergestellt.

1.3.3 Forschungsprozesse in der Mixed-Methods-Sozialforschung

Nachdem quantitative und qualitative empirische Sozialforschung lange in einem Konkurrenzverhältnis diskutiert wurden und jeder Ansatz für sich generelle Überlegenheit beanspruchte, wird heute überwiegend von einem **Ergänzungsverhältnis** ausgegangen: Beide Ansätze werden als legitime wissenschaftliche Forschungsstrategien

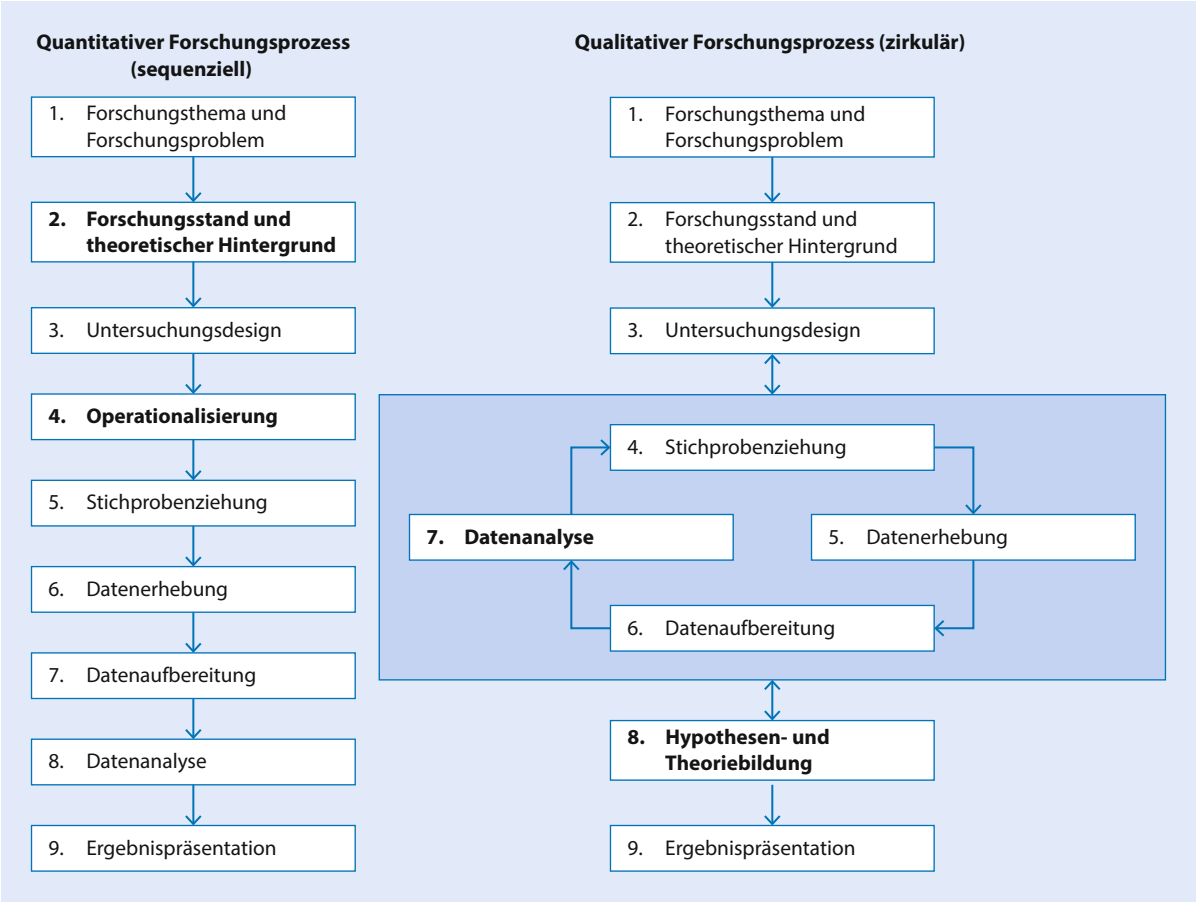


Abb. 1.1 Schematische Darstellung des quantitativen und qualitativen Forschungsprozesses. Mit Hervorhebung (Fettdruck) der meist besonders arbeitsintensiven Phasen. (Mod. nach Witt, 2001. Mit freundlicher Genehmigung des Forums Qualitativer Sozialforschung)

in den Human- und Sozialwissenschaften betrachtet. Der in den letzten Jahren sich herauskristallisierende **Mixed-Methods-Ansatz** („mixed methods research“) geht noch einen Schritt über die „friedliche Koexistenz“ qualitativer und quantitativer Studien hinaus: Er befasst sich damit, wie qualitative und quantitative Forschungsmethoden im Rahmen einer einzelnen Studie bzw. ihrer Teilstudien sinnvoll miteinander zu verknüpfen sind, um maximalen Erkenntnisgewinn zu erzielen (Schreier & Odağ, 2010; Teddlie & Tashakkori, 2009; ▶ Abschn. 2.4; ▶ Abschn. 7.1).

Die Verknüpfung unterschiedlicher Methoden zur Erhöhung der Aussagekraft einer Studie ist dabei nicht neu: Innerhalb der quantitativen Forschung wird sie z. B. durch den **Multitrait-Multimethod-Ansatz** (Campbell & Fiske, 1959; ▶ Abschn. 10.4) und innerhalb der qualitativen Forschung z. B. durch die sog. **Triangulation** (Flick, 2008) realisiert. Kennzeichnend für den Mixed-Methods-Ansatz ist jedoch eine Methodenkombination, die qualitative und quantitative Methoden sowie Methodologien und wissenschaftstheoretische Paradigmen miteinander koppelt.

Will man eine Studie gemäß dem Mixed-Methods-Ansatz durchführen, so ist damit zu rechnen, dass sich der **Forschungsprozess verlängert** und/oder dass der **Forschungsprozess komplexer** wird, weil verschiedene qualitative und quantitative Arbeitsschritte sequenziell und/oder parallel abzuarbeiten sind. Mixed-Methods-Studien variieren in ihrer Komplexität sehr stark.

Die **einfachste Form von Mixed-Methods-Studien** besteht darin, dass zur Bearbeitung des ausgewählten Forschungsproblems sequenziell sowohl eine qualitative als auch eine quantitative Teilstudie durchgeführt und die Befunde direkt aufeinander bezogen werden (sequenzielles Mehrmethoden-Design):

- **Vorstudienmodell:** Relativ häufig wird eine qualitative Studie vorgeschaltet, welche z. B. der Generierung und Präzisierung von Hypothesen dient (z. B. quali-

tative Interviewstudie), die dann in der zweiten Teilstudie quantitativ getestet werden (z. B. quantitative Fragebogenstudie).

- **Vertiefungsmodell:** Auch die umgekehrte Reihenfolge ist möglich und sinnvoll: Zu einer gut etablierten Theorie wird eine hypothesenprüfende quantitative Studie durchgeführt (z. B. quantitative Dokumentenanalyse). Statistisch nachgewiesene Effekte werden in einer zweiten qualitativen Teilstudie detaillierter beschrieben und an Fallbeispielen illustriert (z. B. qualitative Dokumentenanalyse).

Bei sequenziellen Mehrmethodenstudien gemäß dem Vorstudien- oder Vertiefungsmodell würde sich der Forschungsprozess verlängern, weil sowohl ein qualitativer als auch ein quantitativer Forschungsprozess komplett zu durchlaufen sind. Bei anderen Mixed-Method-Designs sind über die Verlängerung des Forschungsprozesses hinaus auch parallele und zirkuläre Arbeitsschritte einzuplanen (► Abschn. 7.1).

Lernquiz

1. Welche Wege der Produktion und Begründung von Wissen existieren außerhalb der wissenschaftlichen Forschung?
2. Was versteht man unter Wissenschaftsbildung und Methodenkompetenz?
3. Warum spricht man von „empirischer" Sozialforschung? Wie ordnen sich Sozialwissenschaften in das System wissenschaftlicher Disziplinen ein?
4. Worin unterscheiden sich quantitative und qualitative Sozialforschung?
5. Worin bestehen die Gemeinsamkeiten quantitativer und qualitativer Sozialforschung?
6. Was versteht man unter interdisziplinärer Forschung? Welche Voraussetzungen müssen dafür gegeben sein?
7. Was ist mit Mixed-Methods-Research im Kontext der empirischen Sozialforschung gemeint?
8. Grenzen Sie Grundlagen- und Anwendungsforschung hinsichtlich ihrer Fragestellungen, Methoden und Ergebnisse voneinander ab.
9. Was bedeutet es, wenn ein Zeitschriftenartikel sehr oft zitiert wird? Wie kommt dies üblicherweise zustande?
10. Wie lässt sich der Forschungsprozess in der quantitativen Sozialforschung beschreiben? Worin unterscheidet er sich vom qualitativen Forschungsprozess?

Übungsaufgaben

1. Die Deutsche Gesellschaft für Psychologie (DGPs) verfügt über eine Fachgruppe „Methoden und Evaluation" und die Deutsche Gesellschaft für Soziologie (DGS) über eine Sektion „Methoden der empirischen Sozialforschung" sowie eine Sektion „Methoden der qualitativen Sozialforschung". Die Gesellschaft für Evaluation (DeGEval) hat einen Arbeitskreis „Methoden in der Evaluation" gebildet. Lesen Sie sich auf den jeweiligen Internetpräsenzen die Selbstbeschreibungen der genannten Methodengruppen durch. Charakterisieren Sie auf dieser Basis die Psychologie und die Soziologie sowie die Evaluationsforschung hinsichtlich ihrer Nähe zum quantitativen, qualitativen und/oder Mixed-Methods-Ansatz der empirischen Sozialforschung.
2. Betrachten Sie drei sozialwissenschaftliche Fachzeitschriften zum Jugendalter, das *Journal of Youth and Adolescence* (Springer), das *Journal of Youth Studies* (Taylor Francis) und *Young. Nordic Journal of Youth Research* (Sage) anhand ihrer Webpräsenzen (beachten Sie besonders die Rubriken: „Aims & Scope" sowie „Manuscript Submission") dahingehend: a) ob sie laut Selbstbeschreibung primär multi-, inter- oder transdisziplinär ausgerichtet sind, b) ob sie eher quantitative und/oder qualitative Beiträge veröffentlichen und c) welchen Impact Factor sie jeweils haben.
3. Die „Marienthal-Studie" ist eine klassische soziologische Studie aus den 1930er-Jahren, in der die Folgen der Langzeitarbeitslosigkeit in einer österreichischen Siedlung nach Schließung der dortigen Textilfabrik untersucht wurden (Jahoda, Lazarsfeld, & Zeisel, 1933; ► Abschn. 10.1.2). Bekannt geworden ist die Studie vor allem durch die Verwendung eines breiten Spektrums reaktiver (z. B. Befragung) und nicht-reaktiver Methoden (z. B. Beobachtung, Dokumentenanalyse), die sich teils als qualitativ und teils als quantitativ einordnen lassen. Informieren Sie sich aus geeigneten Online- und/oder Offline-Quellen über die Studie und nennen Sie mindestens zwei konkrete qualitative und quantitative Datenerhebungsmethoden, die zum Einsatz kamen.

Weiterführende Literatur

- Denzin, N. & Lincoln, Y.S. (Eds.) (2012). *The SAGE handbook of qualitative research.* (4th ed.). Thousand Oaks: Sage.
- Fowler, F. (2008). *Survey research methods.* (4th ed.). Thousand Oaks: Sage.
- Hancock, G. R. & Mueller, R.O. (Eds.) (2010). *The reviewer's guide to quantitative methods in the social sciences*. New York: Routledge.

- Holling, H. & Schmitz, B. (Hrsg.) (2010). *Handbuch Statistik, Methoden und Evaluation*. Göttingen: Hogrefe.
- Kelle, U. (2008). *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung: Theoretische Grundlagen und methodologische Konzepte* (2. Aufl.). Wiesbaden: VS Verlag.
- Kuckartz, U. (2012). *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: VS Verlag.
- Mey, G. & Mruck, K. (Hrsg.) (2010). *Handbuch qualitative Forschung in der Psychologie.* Wiesbaden: VS Verlag.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.
- Tashakkori, A. & Teddlie, C. (Eds.) (2010). *The SAGE handbook of mixed methods in social and behavioral research* (2nd ed.). Thousand Oaks: Sage.

Literatur

Backhouse, R. E. & Fontaine, P. (Eds.). (2010). *The history of the social sciences since 1945*. Cambridge: Cambridge University Press.

Barrell, J. J., Aanstoos, C., Rechards, A. C., & Arons, M. (1987). Human science research methods. *Journal of Humanistic Psychology, 27*(4), 424–457.

Berekoven, L., Eckert, W., & Ellenrieder, P. (2009). *Marktforschung: Methodische Grundlagen und praktische Anwendung* (12. Aufl.). Wiesbaden: Gabler.

Blanckenburg, C. v., Böhm, B., Dienel, H.-L., & Legewie, H. (2005). *Leitfaden für interdisziplinäre Forschergruppen: Projekte initiieren – Zusammenarbeit gestalten.* Stuttgart: Steiner.

Böschen, S. & Schulz-Schäffer, I. (2003). *Wissenschaft in der Wissensgesellschaft*. Wiesbaden: Westdeutscher Verlag.

Bryman, A. (2008). *Social research methods*. Oxford: Oxford University Press.

Buber, R. & Holzmüller, H. (2009). *Qualitative Marktforschung: Konzepte – Methoden – Analysen* (2. Aufl.). Wiesbaden: Gabler.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks: Sage.

Denzin, N. & Lincoln, Y. S. (Eds.). (2012). *The SAGE handbook of qualitative research* (4th ed.). Thousand Oaks: Sage.

Dilthey, W. (1923). *Ideen über eine beschreibende und zergliedernde Psychologie* (Ges. Schrifttum, Bd. 5). Leipzig: Teubner.

Eagly, A. H. & Carlib, L. L. (2003). The female leadership advantage: An evaluation of the evidence. *The Leadership Quarterly, 14*(6), 807–834.

Engelhardt, A. & Kajetzke, L. (Hrsg.). (2010). *Handbuch Wissensgesellschaft. Theorien, Themen und Probleme*. Bielefeld: transcript.

Farr, R. H., Forssell, S. L., & Patterson, C. J. (2010). Parenting and child development in adoptive families: Does parental sexual orientation matter? *Applied Developmental Science, 14*(3), 154–178.

Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest, 13*, 3–66.

Flick, U. (2008). *Triangulation: Eine Einführung* (2.Aufl.). Wiesbaden: VS Verlag.

Freeman, H. E. & Rossi, P. H. (1984). Furthering the applied side of sociology. *American Sociological Review, 49*(4), 571–580.

Grant, E. (1996). *The foundations of modern science in the middle ages: Their religious, institutional and intellectual contexts.* Cambridge: Cambridge University Press.

Gravetter, F. J. & Forzano, L. B. (2009). *Research methods for the behavioral sciences* (3rd ed.). Belmont: Wadsworth.

Hammersley, M. (1992). The paradigm wars: reports from the front. *British Journal of Sociology of Education, 13*(1), 131–143.

Hammersley, M. (2006). *Media bias in reporting social research? The case of reviewing ethnic inequalities in education*. London: Routledge.

Hancock, G. R. & Mueller, R. O. (Eds.). (2010). *The reviewer's guide to quantitative methods in the social sciences*. New York: Routledge.

Hemilä, H., Chalker, E., & Douglas, B. (2007). Vitamin C for preventing and treating the common cold. *Cochrane Library of Systematic Reviews*(3), Art.-Nr. CD000980.

Hirsch Hadorn, G., Hoffmann-Riem, H., Biber-Klemm, S., Grossenbacher-Mansuy, W., Joye, D., Pohl, C., et al. (Eds.). (2008). *Handbook of transdisciplinary research*. Heidelberg: Springer.

Holling, H. & Schmitz, B. (Hrsg.). (2010). *Handbuch Statistik, Methoden und Evaluation.* Göttingen: Hogrefe.

Jahoda, M., Lazarsfeld, P. F., & Zeisel, H. (1933). *Die Arbeitslosen von Marienthal. Ein soziographischer Versuch über die Wirkungen langandauernder Arbeitslosigkeit.* Leipzig: Hirzel.

Jungert, M., Romfeld, E., Sukopp, T., & Voigt, U. (Hrsg.). (2010). *Interdisziplinarität. Theorie, Praxis, Probleme*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Kelle, U. (2008). *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte* (2. Aufl.). Wiesbaden: VS Verlag.

Keuth, H. (1989). *Wissenschaft und Werturteil. Zu Werturteilsdiskussion und Positivismusstreit.* Tübingen: Mohr.

Kidd, S. A. (2002). The role of qualitative research in psychological Journals. *Psychological Methods, 7*(1), 126–138.

Kromrey, H. (2009). *Empirische Sozialforschung. Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung* (12. Aufl.). Stuttgart: UTB.

Kuckartz, U. (2012). *Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: VS Verlag.

Lincoln, Y. S. (2010). „What a long, strange trip it's been"; Twenty-five years of qualitative and new paradigm research. *Qualitative Inquiry, 16*(1), 3–9.

Lincoln, Y. S., Guba, E. G., & Lynham, S. A. (2012). Paradigmatic controversies, contradictions and emerging confluences, revisited. In N. K. Denzin & Y. S. Lincoln (Eds.), *Sage handbook of qualitative research* (4th ed.). Thousand Oaks: Sage.

Malesky, A. & Keim, J. (2001). Mental health professionals' perspectives on sex offender registry web sites. *Sexual Abuse: Journal of Research and Treatment, 13*(1), 53–63.

Mey, G. & Mruck, K. (Hrsg.). (2010). *Handbuch qualitative Forschung in der Psychologie*. Wiesbaden: VS Verlag.

Mitchell, M. L. & Jolley, J. M. (2010). *Research design explained* (7th ed.). Belmont: Wadsworth.

Mruck, K. & Mey, G. (2010). Einleitung. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS Verlag.

Neumann, W. L. (2003). *Social research methods: Qualitative and quantitative methods*. Boston: Allyn & Bacon.

Popper, K. (1934/1989). *Logik der Forschung* (9. Aufl.). Tübingen: Mohr Siebeck.

Porter, T. M. & Ross, D. (Eds.). (2003). *The Cambridge history of science. Volume 7: The modern social sciences*. Cambridge: Cambridge University Press.

Reichertz, J. (2007). Qualitative Sozialforschung – Ansprüche, Prämissen, Probleme. *Erwägen, Wissen, Ethik EWE, 18*(2), 195-208.

Robinson, F. P. (1978). *Effective Study* (6th ed.). New York: Harper & Row.

Salerno, J. M., Najdowski, C. J., Stevenson, M. C., Wiley, T. R. A., Bottoms, B. L., Vaca, R., Jr., et al. (2010). Psychological mechanisms underlying support for juvenile sex offender registry laws: Prototypes, moral outrage, and perceived threat. *Behavioral Sciences and the Law, 28*(1), 58–83.

Schibany, A. & Gassler, H. (2010). *Nutzen und Effekt der Grundlagenforschung*. Graz: JOANNEUM RESEARCH.

Schreier, M. & Odağ, Ö. (2010). Mixed Methods Designs. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS Verlag.

Smith, M. B. (1994). „Human Science"-Really!: A theme for the future of psychology. *Journal of Humanistic Psychology, 34*(2), 111–116.

Snow, C. P. (1959). *The two cultures and the scientific revolution*. New York: The Syndics of the Cambridge University Press.

Sukopp, T. (2010). Interdisziplinarität und Transdisziplinarität. Definitionen und Konzepte. In M. Jungert, E. Romfeld, T. Sukopp & U. Voigt (Hrsg.), *Interdisziplinarität. Theorie, Praxis, Probleme* (S. 13–29). Darmstadt: Wissenschaftliche Buchgesellschaft.

Tashakkori, A. & Teddlie, C. (Eds.). (2010). *Handbook of mixed methods in social and behavioral research.* (2nd ed.). Thousand Oaks: Sage.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating quantitative and qualitative approaches in the social and behavioral sciences.* Thousand Oaks: Sage.

Weber, B. (2010). Challenges of social science literacy – editorial. *Journal of Social Science Education, 9*(4), 2–5.

Wirtz, M. A. (Hrsg.) (2013). *Dorsch – Lexikon der Psychologie* (16. Aufl.). Bern: Huber.

Witt, H. (2001). Forschungsstrategien bei qualitativer und quantitativer Sozialforschung. *Forum Qualitative Sozialforschung, 2*(1), Art. 8. Abgerufen 20. November, 2013, unter http://www.qualitative-research.net/index.php/fqs/article/view/969/2114.

Wissenschaftstheoretische Grundlagen der empirischen Sozialforschung

Nicola Döring, Jürgen Bortz

2.1 Wissenschaftstheorie im Überblick – 34

2.2 Das quantitative Paradigma und der Kritische Rationalismus – 36

2.3 Das qualitative Paradigma und der Sozialkonstruktivismus – 63

2.4 Das Mixed-Methods-Paradigma und der Pragmatismus – 72

Literatur – 78

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_2, © Springer-Verlag Berlin Heidelberg 2016

2

Lernziele

- Wissen, wozu eine wissenschaftstheoretische Fundierung der empirischen Sozialforschung dient.
- Den kritischen Rationalismus als wissenschaftstheoretische Basis des quantitativen Paradigmas der empirischen Sozialforschung darstellen können.
- Neuere wissenschaftstheoretische Ansätze im quantitativen Paradigma kennen.
- Die wissenschaftstheoretischen Grundlagen des qualitativen Paradigmas der empirischen Sozialforschung anhand von fünf Grundprinzipien erläutern können.
- Fünf verschiedene Positionen zu Mixed-Methods als möglichem drittem Paradigma der empirischen Sozialforschung kennen und wissenschaftstheoretische Grundannahmen des Pragmatismus als Basis eines eigenen Mixed-Methods-Paradigmas benennen können.

Zur Wissenschaft gehört nicht nur die Produktion wissenschaftlicher Erkenntnisse, sondern immer auch die **Reflexion des Erkenntnisprozesses** (zur Abgrenzung der Wissenschaft von anderen Erkenntnisformen ▶ Kap. 1). Wir verlangen dementsprechend in der Praxis der empirischen Sozialforschung, dass jede Studie, die Wissenschaftlichkeit beansprucht, ihr methodisches Vorgehen offenlegt, die Ergebnisse mit Vorsicht interpretiert und selbstkritisch auf Einschränkungen des Geltungsbereiches, auf mögliche Lücken, Fehler oder Verzerrungen durch die genutzten theoretischen Vorannahmen und Methoden hinweist und widersprüchliche Befunde sowie unbeantwortete Fragen darlegt (▶ Kap. 3 zu Qualitätskriterien in der empirischen Sozialforschung).

In der **Wissenschaftstheorie** geht der Anspruch der Selbstreflexion des Erkenntnisprozesses noch einen Schritt weiter: Nicht die Stärken und Schwächen einer einzelnen Studie sollen kritisch durchdacht werden, sondern das **Fundament empirisch-wissenschaftlichen Arbeitens** generell soll begründet werden. Diese Aufgabe fällt in den Zuständigkeitsbereich der **Philosophie** als einer Formalwissenschaft (zur Gliederung der wissenschaftlichen Disziplinen ▶ Abschn. 1.1.2). Denn ausschlaggebend für eine stichhaltige Begründung wissenschaftlichen Vorgehens ist z. B. die Logik. In der Wissenschaftstheorie werden also philosophisch auf der Metaebene die Möglichkeiten und Grenzen diskutiert, unter denen Wissenschaft überhaupt sinnvoll betrieben werden kann und Erkenntnis oder gar Erkenntnisfortschritt möglich sind. Derartige Überlegungen wirken notgedrungen oft „abgehoben", und sind durch ihre argumentative Komplexität sowie die zahlreichen Fachbegriffe nicht leicht nachzuvollziehen. Als Reflexion auf der Metaebene muss Wissenschaftstheorie „abstrakt" sein und sich z. B. mit verschiedenen theoretischen Konzepten von „Wirklichkeit" oder „Wahrheit" befassen. Gleichzeitig ist sie aber nicht praxisfern. Denn sie hat den Anspruch, das konkrete wissenschaftliche Arbeiten zu begründen.

Die Wissenschaftstheorie ist **normativ** ausgerichtet und gibt anhand philosophischer Argumente vor, wie Wissenschaft ablaufen sollte, um gültige Erkenntnisse zu liefern. In wissenschaftstheoretischen Kontroversen spielen aber auch **deskriptive** Disziplinen wie die Wissenschaftsgeschichte und die Wissenschaftssoziologie eine wichtige Rolle – nicht zuletzt als Korrektiv und Inspiration für die normative Wissenschaftstheorie: Abweichungen zwischen der Wissenschaftspraxis einerseits und den wissenschaftstheoretischen Vorgaben andererseits können dazu veranlassen, die normativen Vorgaben hinsichtlich ihrer Umsetzbarkeit auf den Prüfstand zu stellen und ggf. zu modifizieren.

In der empirischen Sozialforschung wird heute oft zwischen **drei Paradigmen bzw. Forschungsstrategien**, **Forschungsansätzen** oder **Forschungsmodellen** unterschieden (wir verwenden hier im Zusammenhang mit wissenschaftstheoretischen Grundlagen den auch in der internationalen Fachliteratur etablierten Paradigmen-Begriff, der jedoch abzugrenzen ist von dem wissenschaftshistorischen Paradigmen-Begriff von Thomas Kuhn, 1962, ▶ Abschn. 2.2.11):

- Im **quantitativen Paradigma** wird üblicherweise ein linearer und stark strukturierter Forschungsprozess realisiert, der mit Theoriearbeit und der Ableitung von Hypothesen beginnt, mit standardisierten Erhebungsinstrumenten anhand möglichst repräsentativer Stichproben numerische Daten (Messwerte) erhebt und am Ende in eine statistische Datenanalyse zur Hypothesenprüfung mündet.
- Im **qualitativen Paradigma** wird üblicherweise ein zirkulärer bzw. spiralförmiger und bewusst wenig strukturierter Forschungsprozess realisiert, der mithilfe nicht- oder allenfalls teil-standardisierter Erhebungsverfahren anhand von Einzelfällen oder bewusst ausgewählten kleinen Samples nicht-numerische Daten (meist verbale Daten) erhebt, diese interpretierend auswertet und daraus schrittweise neue Hypothesen und Theorien bildet.
- Im **Mixed-Methods-Paradigma** werden komplexe Forschungsprozesse realisiert, die in ihren verschiedenen Phasen oder Teilstudien jeweils quantitative und qualitative Forschungsstrategien umsetzen und verschiedentlich miteinander verzahnen.

Die drei Paradigmen unterscheiden sich in ihren grundlegenden Vorstellungen darüber, wie sozialwissenschaftliche Forschung überhaupt fundierte Erkenntnisse über die soziale Wirklichkeit gewinnen und begründen kann. Auch über die Beschaffenheit der zu untersuchenden sozialen

Erfahrungswirklichkeit und die Rolle der Forschenden in der Gesellschaft sowie ihr Verhältnis zu den Untersuchungsgegenständen gibt es unterschiedliche Vorannahmen. Eine Studie ist nicht deswegen dem qualitativen Paradigma zuzurechnen, *weil* sie qualitative Daten nutzt, sondern es gilt umgekehrt: Das Vorgehen gemäß qualitativer Forschungslogik führt dazu, dass der Forschungsprozess **absichtlich nicht oder nur wenig strukturiert und standardisiert** wird, so dass *infolgedessen* hauptsächlich mit qualitativen Daten gearbeitet wird. Ebenso ist eine Studie nicht deswegen dem quantitativen Paradigma zuzurechnen, *weil* sie quantitative Daten nutzt, auch hier gilt umgekehrt: Das Vorgehen gemäß quantitativer Forschungslogik führt dazu, dass der Forschungsprozess **hochgradig strukturiert und standardisiert** abläuft und *infolgedessen* hauptsächlich mit quantitativen Daten bzw. Messwerten gearbeitet wird. Die etablierten Bezeichnungen „quantitative" versus „qualitative" Sozialforschung werden deswegen auch kritisiert, und es wird alternativ der Aspekt der Strukturierung/Standardisierung des Forschungsprozesses inklusive des dahinter stehenden wissenschaftstheoretischen Verständnisses hervorgehoben (Kromrey, 2009, S. 24f.).

Der entscheidende Unterschied zwischen quantitativem und qualitativem Paradigma in der empirischen Sozialforschung liegt nicht auf der Ebene der Daten (bzw. der Datenerhebungs- und Datenauswertungsmethoden). Tatsächlich liegt das zentrale Unterscheidungskriterium auf der Ebene der Forschungslogik bzw. der wissenschaftstheoretischen Begründung des Vorgehens. Aus der jeweiligen Forschungslogik ergibt es sich dann, dass im sog. quantitativen Ansatz primär mit numerischem Datenmaterial und im sog. qualitativen Ansatz primär mit verbalem Datenmaterial gearbeitet wird.

Das vorliegende Buch stellt den empirischen Forschungsprozess von der Wahl des Forschungsproblems über Entscheidungen zu Untersuchungsdesign, Stichprobenziehung, Datenerhebungsmethoden und Datenanalyseverfahren bis zur Ergebnispräsentation in den Mittelpunkt. Die Darstellung differenziert dabei an entsprechenden Stellen zwischen quantitativem und qualitativem Ansatz. Über Vorgehensweisen in den einzelnen Arbeitsphasen des Forschungsprozesses wird ausführlich berichtet. Ein mechanisches Umsetzen einzelner Methoden quasi nach **Rezeptbuch** entspricht dabei nicht den Anforderungen der Wissenschaftlichkeit. Vielmehr wird von Forschenden auch eine wissenschaftstheoretische Reflexion der Ziele, Möglichkeiten und Grenzen der eigenen Forschungstätigkeit sowie der eigenen Rolle verlangt, um zu tragfähigen Erkenntnissen zu gelangen und verantwortungsvoll im Sinne der Wissenschaftsethik (► Abschn. 4.2) handeln zu können.

Die fundamentale Bedeutung der Wissenschaftstheorie wird oft zum Anlass genommen, wissenschaftstheoretische Grundlagen am Beginn des Studiums zu vermitteln. Ohne Kenntnis und Erfahrung der Forschungspraxis stellen sich die wissenschaftstheoretischen Debatten aber oft als wenig greifbar dar. Wir gehen davon aus, dass sich ein Verständnis von Wissenschaftstheorie am besten entwickelt, wenn mit wachsender forschungspraktischer Erfahrung immer wieder der Forschungsprozess auf der Metaebene reflektiert und mit den wissenschaftstheoretischen Vorgaben des jeweils gewählten Paradigmas abgeglichen wird.

Die folgende Darstellung umreißt wissenschaftstheoretische Grundlagen und konzentriert sich dabei auf den Bezug zur Forschungspraxis. Es wird versucht, die wichtigsten Prämissen herauszuarbeiten, die heute in den drei Paradigmen der empirischen Sozialforschung weitgehend konsensfähig sind. Wir orientieren unsere Darstellung an der Einteilung der drei Paradigmen der empirischen Sozialforschung, da diese Strukturierung sowohl in der wissenschaftstheoretischen Diskussion vertreten ist, als auch offenbar das Selbstverständnis vieler Forschender kennzeichnet. Auf historisch überholte wissenschaftstheoretische Modelle und auf Kontroversen innerhalb der jeweiligen Ansätze wird verwiesen, allerdings werden diese nicht detailliert behandelt.

Das Kapitel beginnt mit einigen **Begriffsklärungen** (► Abschn. 2.1). Dann widmet es sich dem **quantitativen Paradigma** und der dort vorherrschenden Wissenschaftstheorie, dem **Kritischen Rationalismus** (► Abschn. 2.2). Anschließend geht es um das **qualitative Paradigma**, das nicht wie das quantitative Paradigma auf einer zentralen Wissenschaftstheorie basiert, sondern auf unterschiedlichen Erkenntnistheorien, die jedoch in zentralen Prinzipien übereinstimmen. Eine besonders wichtige Erkenntnistheorie des in sich ausdifferenzierten qualitativen Paradigmas ist der **Sozialkonstruktivismus**, der in der angloamerikanischen Literatur auch als Interpretationism bezeichnet wird (► Abschn. 2.3). Zu der Frage, ob und wie ein eigenständiges drittes **Mixed-Methods-Paradigma** sinnvoll abzugrenzen und wissenschaftstheoretisch zu begründen ist, existieren unterschiedliche Positionen, die sich vor allem auch darüber streiten, inwiefern es möglich ist, grundverschiedene Wissenschaftstheorien zu verbinden. Aktuell wird innerhalb des Mixed-Methods-Paradigmas vor allem der **Pragmatismus** als wissenschaftstheoretische Position diskutiert (► Abschn. 2.4).

Das vorliegende Lehrbuch steht in der Tradition des quantitativen Paradigmas und des Kritischen Rationalismus. Daraus ist abzuleiten, dass wissenschaftstheoretische Positionen, die quantitativ-statistische Forschung

grundsätzlich ablehnen, von uns nicht geteilt – wohl aber dargestellt – werden. Umgekehrt impliziert das Bekenntnis zum Kritischen Rationalismus nicht die Ablehnung qualitativer Verfahren, die im Entdeckungszusammenhang seit jeher ihren Platz haben.

Umstrittener ist die Frage, inwiefern eine qualitative Forschungsstrategie in der Lage ist, nicht nur relevante und interessante neue Hypothesen und Theorien hervorzubringen (dies bezweifelt niemand), sondern diese auch anhand nachvollziehbarer Kriterien als vorläufig bestätigt oder unbestätigt auszuweisen und anzugeben, auf welche nicht-untersuchten Kontexte die Ergebnisse generalisierbar sind. Ein häufiger Streitpunkt sind dabei **Anzahl und Auswahl der untersuchten Fälle**, auf denen qualitative Studien basieren – die empirische Basis erscheint aus Sicht des quantitativen Paradigmas oft zu klein und beliebig für verallgemeinerbare Erkenntnisse. Was Verallgemeinerbarkeit im qualitativen Paradigma bedeutet und wie sie sicherzustellen ist, wird kontrovers diskutiert (► Abschn. 2.3; ► Abschn. 3.3). Ein weiterer Streitpunkt ist der **Umgang mit Vorwissen**. Die Skepsis des qualitativen Ansatzes gegenüber der Arbeit mit ausgewählten, theoretisch vordefinierten Konzepten (Variablen) sowie das Plädoyer für eine möglichst offene Gegenstandsbetrachtung widersprechen der Vorstellung des Kritischen Rationalismus, dass Forschungsprozesse am besten durch theoretische Vorgaben klar strukturiert werden. Das zuweilen propagierte Ideal eines theoriefreien Vorgehens ist aber auch im qualitativen Paradigma zunehmend ersetzt worden durch Vorschläge dazu, wie mit Vorwissen – seien es die Alltagsüberzeugungen der Forschenden oder vorliegende wissenschaftliche Theorien – im Rahmen qualitativ-empirischer Forschungsprozesse konstruktiv umzugehen ist (► Abschn. 2.3.2; zum Verhältnis von Theorie und Empirie in der qualitativen soziologischen Forschung s. Kalthoff, Hirschauer, & Lindemann, 2008).

Wir sehen den wissenschaftstheoretischen Paradigmenstreit aus der Perspektive des Kritischen Rationalismus als **fruchtbar** an, da er – wenn er auf den Austausch von Argumenten und nicht auf wechselseitige Abwertung hinausläuft – hilft, eigene Positionen zu klären und zu hinterfragen. Wir sind überzeugt, dass jenseits paradigmatischer Differenzen auch viele **Gemeinsamkeiten** zwischen Forschenden bestehen und Grundprinzipien der Wissenschaftlichkeit paradigmenübergreifend geteilt werden (► Abschn. 3.1). Nicht zuletzt ist daran zu erinnern, dass insbesondere in der Anwendungsforschung (wie z. B. der Markt- oder der Evaluationsforschung) qualitative und quantitative Forschungsstrategien seit langem friedlicher und fruchtbarer koexistieren als es so mancher philosophische Text zum wissenschaftstheoretischen „Paradigmen-Krieg“ vielleicht vermuten lässt.

2.1 Wissenschaftstheorie im Überblick

In wissenschaftstheoretischen Diskussionen werden zahlreiche Fachbegriffe verwendet, von denen wir einige Grundbegriffe im Folgenden kurz einführen. Anschließend gehen wir auf die Unterscheidung zwischen expliziten und impliziten Wissenschaftstheorien ein.

2.1.1 Wissenschaftstheoretische Grundbegriffe

Die **allgemeine Wissenschaftstheorie** („philosophy of science“) befasst sich als Teilgebiet der Philosophie mit den Möglichkeiten und Grenzen wissenschaftlichen Erkenntnisgewinns. Dabei stehen Vorannahmen über die Beschaffenheit der Untersuchungsgegenstände und über den menschlichen Verstand, die Merkmale von wissenschaftlichen Methoden und Methodologien, Anforderungen an gültige Theorien, sowie gesellschaftliche Funktion und Verantwortung der Wissenschaft im Fokus (zur Einführung in die Wissenschaftstheorie siehe z. B. Chalmers, 2007). Neben der allgemeinen Wissenschaftstheorie existieren auch **fachspezifische Wissenschaftstheorien**. So behandelt z. B. die Wissenschaftstheorie der Psychologie („philosophy of psychology“) Vorannahmen über den Untersuchungsgegenstand im Sinne unterschiedlicher Menschenbilder oder der Beschaffenheit des Seelischen (z. B. Gadenne, 2004; Herzog, 2012).

Innerhalb der allgemeinen wie der fachspezifischen Wissenschaftstheorie wird zwischen den Bereichen der Ontologie, Axiologie und Epistemologie differenziert:

- In der **Ontologie** (Lehre des Seienden; „ontology“) geht es innerhalb der Wissenschaftstheorie um die Frage, wie der Gegenstand einer Wissenschaft beschaffen ist. Eine grundlegende ontologische Frage ist z. B. die, ob wir von einer unabhängig von unserem Bewusstsein existierenden Wirklichkeit ausgehen **(Realismus)** oder ob wir annehmen, dass die Wirklichkeit nur in unserer Vorstellung existiert **(Idealismus)**. Eine weitere wichtige Frage betrifft Merkmale der Wirklichkeit: Ist sie geordnet und funktioniert nach bestimmten Ursache-Wirkungs-Prinzipien **(Kausalität)**, so dass Vorhersagen möglich sind, oder ist sie ungeordnet und unterliegt keinen vorhersagbaren Prozessen **(Chaos)?** Speziell für die Sozialwissenschaften ist ontologisch die Frage zu klären, wie die soziale Wirklichkeit zu erklären ist, d. h. ob ihre Regelhaftigkeit ebenso wie in den Naturwissenschaften auf objektivierbare Ursachen zurückführbar ist **(kausale Erklärung)** und/oder ob sie vor allem durch subjektive Motive der Handelnden zustande kommt **(intentionale Erklärung)**.

- In der **Axiologie** (Lehre von den Werten; „axiology“) beschäftigt man sich innerhalb der Wissenschaftstheorie mit Werten und ihrer Bedeutung im Kontext der Wissenschaft. In den Sozialwissenschaften interessiert hierbei z. B. die Frage, welchen Einfluss die **Wertvorstellungen** der Forschenden (insbesondere ihre Haltung zu sozialen und gesellschaftlichen Problemen) auf ihre Forschungsaktivitäten haben können und sollen. Wertfragen betreffen nicht nur die individuellen Forschenden, sondern auch das Wissenschaftssystem und dessen Normen und Regeln (zur **Wissenschaftsethik** ▶ Abschn. 4.2).
- In der **Epistemologie** (Lehre von der Erkenntnis, „epistemology“) befasst man sich im Kontext der Wissenschaftstheorie mit den Voraussetzungen, Möglichkeiten und Grenzen des wissenschaftlichen Erkenntnisgewinns. Dabei geht es im Kern darum, wie gültige wissenschaftliche Aussagen bzw. spezifische Aussagensysteme – **wissenschaftliche Theorien** – begründet und von falschen Aussagen unterschieden werden können. Damit ist letztlich auch die Frage nach der Definition von **Wahrheit** berührt. Die Epistemologie widmet sich vordringlich der Frage, auf welchen Wegen bzw. mit welchen Methoden wissenschaftliche Erkenntnisse gewonnen werden können. Die Reflexion der wissenschaftlichen Methoden und des Forschungsprozesses allgemein bezeichnet man als **Methodologie**. Darüber hinaus existieren **spezielle Methodologien** als wissenschaftliche Grundlagen unterschiedlicher Forschungsstrategien (z. B. Methodologie der Experimentalforschung versus Methodologie der Umfrageforschung innerhalb des quantitativen Paradigmas; ▶ Abschn. 3.1.1). Ein wichtiges Instrumentarium allgemeiner wissenschaftlicher Methodologie ist die **Logik**. Denn wir verlangen von wissenschaftlichen Aussagen in jedem Fall, dass sie logisch widerspruchsfrei sind. Ohne die Verpflichtung auf Logik wären eine stringente Argumentation bzw. ein wissenschaftlicher Diskurs kaum möglich. Darüber hinaus spielt in den Erfahrungswissenschaften auf erkenntnistheoretischer Ebene die Verknüpfung von Theorien mit der Erfahrungswirklichkeit bzw. den **Daten** eine zentrale Rolle. Drei zentrale **Wege des wissenschaftlichen Schlussfolgerns zwischen Daten und Theorien** sind dabei die **Induktion**, die **Deduktion** und die **Abduktion**.

Induktion – Die Induktion („induction“) ist eine Schlussfolgerung **vom Speziellen auf das Allgemeine**, in der empirischen Sozialforschung eine Schlussfolgerung von empirischen Daten über einzelne beobachtete Fälle auf übergeordnete wissenschaftliche Theorien, die über eine Reihe von Fällen verallgemeinerbar sind. Beim induktiven Schließen beginnt der Erkenntnisprozess mit Daten, und es werden aus den Daten schrittweise Muster herausgearbeitet. Mittels Induktion sollen **neue Theorien gebildet**, aber auch bestätigt werden. Die Induktion ist heute vor allem im **qualitativen Paradigma** der empirischen Sozialforschung etabliert.

Deduktion – Die Deduktion („deduction“) ist eine Schlussfolgerung **vom Allgemeinen auf das Spezielle**, in der empirischen Sozialforschung eine Schlussfolgerung von Theorien auf empirische Daten, die von der Theorie vorhergesagt werden. Beim deduktiven Schließen beginnt der Erkenntnisprozess mit einer Theorie, aus der man empirisch prüfbare Hypothesen ableitet und im Falle von deren Widerlegung anhand von Daten die Theorie kritisiert bzw. im Falle ihrer Nicht-Widerlegung die Theorie als vorläufig bestätigt ansieht. Das sog. deduktiv-nomologische Erklärungsmodell dient der **Theorieprüfung** im **quantitativen Paradigma** der empirischen Sozialforschung (▶ Abschn. 2.2.8).

Abduktion – Bei der Abduktion („abduction“) beginnt der Erkenntnisprozess wie bei der Induktion mit den Daten, allerdings werden im Unterschied zur Induktion nicht die in den Daten erkennbaren Muster schrittweise systematisch herausgearbeitet, sondern es werden gerade die unverständlichen Merkmalskombinationen betrachtet und durch einen plötzlichen gedanklichen Sprung wird eine neue erklärende Hypothese gebildet. Die Abduktion ist also ein **kreativer Prozess der Generierung neuer Hypothesen aus Daten**, wobei vor allem die geistige Haltung der Forschenden entscheidend ist (zur Einführung in die Abduktion s. Reichertz, 2003). In der qualitativen Sozialforschung wird neben der Induktion auch die Abduktion zum Erkenntnisgewinn herangezogen.

Von der Wissenschaftstheorie als Teilgebiet der Philosophie, die präskriptiv vorgibt, wie wissenschaftliche Erkenntnisse zu gewinnen und zu begründen sind, hebt sich die **empirische Wissenschaftsforschung** ab, die beschreibt und erklärt, wie Wissenschaft funktioniert (vgl. Felt, Nowotny, & Taschwer, 1995). Die **Wissenschaftsgeschichte** rekonstruiert die historische Entwicklung wissenschaftlicher Erkenntnisse, etwa die Bildung und Verbreitung neuer Theorien und die Abwendung von alten Theorien. Besonders gut untersucht ist die Wissenschaftsgeschichte der Physik, auf die z. B. sowohl Vertreter des Kritischen Rationalismus (z. B. Popper, 1934/1989) als auch deren Kritiker (z. B. Kuhn, 1962) Bezug nehmen. Die **Wissenschaftssoziologie** betrachtet Wissenschaft als soziales System innerhalb der Gesellschaft und interessiert sich unter anderem dafür, wer an Wissenschaft teilnimmt und wer nicht, wie Wissenschaft finanziert wird, welchen sozialen Normen das Wissenschaftssystem folgt und wie die praktische wissenschaftliche Arbeit abläuft (vgl. Buchinger & Felt, 2006; Knorr, 1991).

Wissenschaftsgeschichte und Wissenschaftssoziologie liefern Hinweise darauf, ob und inwiefern die **Wissenschaftspraxis** damals und heute den von der Wissenschaftstheorie vorgegebenen Regeln folgt oder nicht. Diskrepanzen zwischen Wissenschaftspraxis und Wissenschaftstheorie haben immer wieder auch der Wissenschaftstheorie neue Impulse gegeben. Zwar kann aus der Praxis (dem Sein) nicht die Norm (das Sollen) abgeleitet werden, aber wissenschaftstheoretische Vorgaben, die offenbar gar nicht eingehalten werden oder die gerade bei wichtigen wissenschaftlichen Entdeckungen verletzt

wurden, verlieren an Überzeugungskraft und bedürfen möglicherweise einer Anpassung.

Allerdings kann das Offenlegen von Diskrepanzen zwischen wissenschaftstheoretischen Vorgaben und Forschungspraxis auch dazu veranlassen, Maßnahmen zu ergreifen, um die Wissenschaftspraxis zu verändern. So postuliert z. B. die wissenschaftstheoretische Position des Kritischen Rationalismus (► Abschn. 2.2), dass Erkenntnisfortschritt durch das Aussondern empirisch widerlegter Theorien zustande kommt und somit theoriekonträre Befunde besonders hohen Erkenntniswert besitzen. Gleichzeitig zeigt sich jedoch in der Wissenschaftspraxis die klare Tendenz, vor allem theoriebestätigende Ergebnisse als relevant und publikationswürdig zu erachten. Dieser Widerspruch wird als Problem der Wissenschaftspraxis interpretiert und immer wieder zum Anlass genommen, die vollständige Publikation von Studien inklusive aller theoriekonträren (d. h. auch: statistisch nicht-signifikanten) Ergebnisse zu fordern. So wurden schon Fachzeitschriften gegründet, die bewusst nur Studien publizieren, bei denen die Daten der Forschungshypothese widersprechen bzw. bei denen sie die Nullhypothese, dergemäß kein Effekt vorliegt, stützen (z. B. *Journal of Articles in Support of the Null Hypothesis*: http://www.jasnh.com). Probleme in der Wissenschaftspraxis können auch daraus resultieren, dass die Wissenschaft nicht unabhängig genug operieren kann und sich faktisch politischen und ökonomischen Zwängen unterwerfen muss (zu Fehlfunktionen der Wissenschaft siehe z. B. Fischer, 2007).

2.1.2 Explizite und implizite Wissenschaftstheorien

Bei **expliziten Wissenschaftstheorien** handelt es sich um ausformulierte philosophische Theorien darüber, wie Wissenschaft betrieben werden sollte (z. B. Kritischer Rationalismus, Sozialkonstruktivismus, Pragmatismus). Im Unterschied dazu existieren auch **implizite Wissenschaftstheorien**. Das sind die meist nicht ausformulierten epistemologischen, ontologischen und axiologischen Vorstellungen der Forschenden und der Wissenschaftsgemeinschaft, die sich in der wissenschaftlichen Praxis widerspiegeln. Sie lassen sich mithilfe wissenschaftshistorischer und wissenschaftssoziologischer Analysen rekonstruieren. Zudem können alle Forschenden ihre impliziten wissenschaftstheoretischen Positionen herausarbeiten und hinterfragen, indem sie ihre eigene Forschungstätigkeit reflektieren und darüber nachdenken, warum sie manche Vorgehensweisen und Methodenentscheidungen für wissenschaftlich, andere dagegen für unwissenschaftlich halten.

Wenn wir z. B. von akademischen Abschlussarbeiten eine gründliche Aufarbeitung des bisherigen Forschungs- und Theoriestandes zum untersuchten Problem fordern, dann impliziert dies eine epistemologische Position, der gemäß Wissenschaft kumulativ funktioniert und auf bestehenden Theorien und Erkenntnissen aufzubauen ist. Wenn Zeitschriften die Veröffentlichung von Replikationsstudien – d. h. von Wiederholungsstudien – ablehnen, weil diese nicht „originell" genug seien, dann spricht daraus eine erkenntnistheoretische Position, dergemäß die strenge Prüfung vorhandener Theorien weniger bedeutsam ist als das Generieren neuer Theorien. Wenn bei der Beurteilung von Forschungsanträgen oder Doktorarbeiten methodische Strenge und die Orientierung an etablierten wissenschaftlichen Methoden und Methodologien gefordert wird, steht dahinter offenbar die wissenschaftstheoretische Position, dass Erkenntnisgewinn durch die Nutzung eines etablieren Methodeninstrumentariums gefördert und nicht behindert wird.

2.2 Das quantitative Paradigma und der Kritische Rationalismus

In ► Kap. 1 wurde bereits beschrieben, dass erfahrungswissenschaftlicher Erkenntnisgewinn in den Natur- wie in den Sozial- und Humanwissenschaften auf der systematischen (d. h. methodisch angeleiteten und methodologisch begründeten) Sammlung, Aufbereitung und Analyse empirischer **Daten** und gleichzeitig immer auch auf **Theorien** basiert. Ohne Theorien sind Daten nicht interpretierbar und ohne Daten sind erfahrungswissenschaftliche Theorien nicht zu prüfen und weiterzuentwickeln.

Als führender wissenschaftstheoretischer Ansatz der quantitativ-empirischen Sozialforschung gibt der **Kritische Rationalismus** begründet vor, wie erfahrungswissenschaftliche Theorien zu formulieren und zu prüfen sind. Der Kritische Rationalismus geht auf den österreichisch-britischen Philosophen **Karl Raimund Popper** (1902–1994) und seine bereits 1934 in erster Auflage erschienene Abhandlung **„Logik der Forschung"** zurück (Popper, 1934/1989). Er bietet ein wissenschaftstheoretisches Fundament für alle Erfahrungswissenschaften, wird aber oft an Beispielen aus der Physik erläutert. Die Anwendung des Kritischen Rationalismus speziell auf die Sozialwissenschaften wurde bereits von Popper (1962) skizziert und maßgeblich von dem deutschen Philosophen und Soziologen **Hans Albert** vorangetrieben, insbesondere durch sein 1968 in erster Auflage erschienenes **„Traktat über kritische Vernunft"** (Albert, 1968/1991). Zahlreiche weitere Publikationen zum Kritischen Rationalismus von Albert liegen vor (z. B. Albert, 2000, 2011).

Ein weiterer führender zeitgenössischer Vertreter des Kritischen Rationalismus ist der britische Philosoph und Popper-Schüler **David W. Miller**, dessen Werke über Popper und den Kritischen Rationalismus in vielen Sprachen erschienen sind (Miller, 1994, 2006).

Im Folgenden wird der Kritische Rationalismus mit seinen wichtigsten Annahmen und Begriffen vorgestellt und damit in wissenschaftstheoretisches Denken eingeführt. Dabei wird anhand von Studienbeispielen die Brücke zwischen wissenschaftstheoretischer Metareflexion einerseits und Forschungspraxis andererseits geschlagen. Auf Missverständnisse sowie Grenzen des Ansatzes wird ausdrücklich hingewiesen. Ebenso werden abschließend einige neuere wissenschaftstheoretische Ansätze im quantitativen Paradigma angesprochen.

2.2.1 Rationalismus statt Empirismus

Der Kritische Rationalismus wurde von Karl Popper (1934/1989) als ausdrückliches **Gegenmodell zu Empirismus bzw. Positivismus** entwickelt. Der Empirismus bzw. Positivismus geht davon aus, dass im Zuge der Sammlung empirischer Daten bzw. positiver Evidenzen durch **Induktionsschluss** gesicherte allgemeingültige Theorien abgeleitet und bestätigt werden können. Beispiel: Man beobachtet an einigen Gewässern die Schwäne und sieht nur weiße Schwäne. Aus diesen empirischen Beobachtungen wird die Theorie abgeleitet und für gültig erklärt: „Alle Schwäne sind weiß". (In der erkenntnistheoretischen Diskussion wird zur Veranschaulichung der Argumente oft mit sehr einfachen „Mini-Theorien" operiert, sozialwissenschaftliche Theorien sind in der Regel wesentlich komplexer und bestehen aus einem ganzen Aussagensystem; zu Merkmalen sozialwissenschaftlicher Theorien ▶ Abschn. 2.2.10 und zu typischen Forschungshypothesen in der empirischen Sozialforschung ▶ Abschn. 5.2).

Die empiristische bzw. positivistische **Absicherung von Wissen** durch Induktionsschluss ist laut Kritischem Rationalismus indessen logisch gar nicht möglich. Die Theorie (bzw. Gesetzesaussage oder orts- und zeitungebundene All-Aussage) „Alle Schwäne sind weiß" lässt sich durch den Nachweis von einem weißen Schwan – oder auch von mehreren oder gar sehr vielen weißen Schwänen – niemals eindeutig bestätigen bzw. verifizieren. Denn auch wenn man massenhaft weiße Schwäne findet, bleibt immer offen, ob es nicht doch schwarze oder bunte Schwäne gibt, die bislang nur der eigenen Beobachtung entgangen sind. Gesicherte wissenschaftliche Erkenntnis über All-Aussagen kann somit nie durch **Verifikation** begründet werden. Die All-Aussage „Alle Schwäne sind weiß" lässt sich jedoch durch Nachweis eines einzigen nicht-weißen (z. B. schwarzen) Schwans eindeutig widerlegen (falsifizieren).

Während im Empirismus die empirischen Daten sowohl Ausgangspunkt als auch Gültigkeitskriterium der Erkenntnis bilden, geht der Kritische Rationalismus davon aus, dass immer die durch den menschlichen Verstand (**Ratio**; deswegen „Rationalismus") gebildeten **Theorien der Startpunkt wissenschaftlichen Erkenntnisgewinns** sind: Theorien werden als Vermutungen über die Realität formuliert, können niemals zweifelsfrei bestätigt, aber – unter bestimmten Bedingungen – durch einen **Deduktionsschluss** von der Theorie auf eine überprüfbare empirische Hypothese und deren Konfrontation mit Daten widerlegt werden. Erkenntnisgewinn besteht also darin, durch **Falsifikation** die ungültigen Theorien auszusondern. Eine Theorie, die einen Falsifikationsversuch übersteht, gilt als **vorläufig bestätigt**. Eine Theorie, die viele Falsifikationsversuche überstanden hat, bezeichnen wir als **bewährt**.

Um den zentralen Stellenwert des Falsifikationsprinzips zu betonen, wird der Kritische Rationalismus auch als **Falsifikationismus** bezeichnet. Dabei handelt es sich jedoch nicht um einen naiven Falsifikationismus, der jedes – wie auch immer zustande gekommene – abweichende Einzelergebnis zum Anlass nimmt, ganze Theoriegebäude zum Einsturz zu bringen. Vielmehr ist der Kritische Rationalismus als **methodologischer Falsifikationismus** zu kennzeichnen, der die methodischen Voraussetzungen der Theorieprüfung durch Daten reflektiert. Denn wenn Daten einer Theorie widersprechen, kann das sowohl an der Fehlerhaftigkeit der Theorie liegen als auch an der Fehlerhaftigkeit der Daten bzw. der verwendeten Messinstrumente (z. B. Operationalisierungs- oder Messfehler). Die Daten selbst sind also immer auch theoriehaltig (basieren z. B. auf einer bestimmten Operationalisierung und Messtheorie) und müssen ihrerseits hinterfragt werden, bevor man sie zum Anlass nehmen kann, eine inhaltliche Theorie als falsifiziert oder als vorläufig bestätigt zu betrachten ▶ Abschn. 2.2.8). Um zum Ausdruck zu bringen, dass das kritische Hinterfragen aller Behauptungen und Theorien der Motor des Erkenntnisfortschritts ist, kennzeichnet Popper (1934/1989) seine Position auch als **Kritizismus**.

Kritischer Rationalismus – Gemäß der wissenschaftstheoretischen Position des Kritischen Rationalismus („critical rationalism") kommt Erkenntnis dadurch zustande, dass der menschliche Verstand (Ratio) zunächst Theorien als Vermutungen über die Realität aufstellt. Aus diesen Theorien werden deduktiv empirisch prüfbare Hypothesen abgeleitet und anhand von Daten kritisch in der Wissenschaftsgemeinschaft geprüft. Die kritische Prüfung von Hypothesen umfasst darüber hinaus immer auch eine kritische Betrachtung der Voraussetzungen der Datengewinnung (d. h. eine Methodenkritik) sowie eine kritische Auseinandersetzung mit konkurrierenden Hypothesen und Theorien.

Verifikation und Falsifikation – Sichere Erkenntnis durch Bestätigung von Theorien anhand von Daten (Verifikation; „verification") ist im Verständnis des Kritischen Rationalismus nicht möglich, da ein Induktionsschluss nicht logisch zwingend ist. Lediglich die Widerlegung von Theorien durch Daten (Falsifikation; „falsification") ist – unter bestimmten Bedingungen – zu rechtfertigen auf der Basis des Deduktionsschlusses.

Falsifikationismus bzw. Kritizismus – Der Kritische Rationalismus beschreibt Erkenntnisfortschritt als Aussondern nicht-bestätigter Theorien durch Falsifikation bzw. umgekehrt als Zurückbehalten von nicht-falsifizierten – d. h. vorläufig bestätigten bzw. bewährten – Theorien. Er wird deswegen auch als Falsifikationismus („falsificationism") sowie als Kritizismus („criticism") bezeichnet und stellt ein ausdrückliches Gegenmodell zu dem auf Verifikation basierenden Empirismus bzw. Positivismus dar.

Immer wieder wurde und wird die wissenschaftstheoretische Position des Kritischen Rationalismus in der Fachliteratur als „**Positivismus**" oder „**Post-/Neopositivismus**" etikettiert, und zwar insbesondere von Kritikern (vor allem von Vertretern der Kritischen Theorie der Frankfurter Schule, die den Kritischen Rationalismus im sog. Positivismusstreit in der Soziologie als wissenschaftstheoretisch verfehlten Ansatz angriffen; Adorno et al., 1969; ▶ Abschn. 2.2.11) sowie auch in der Mixed-Methods-Literatur, die eine Integration quantitativer und qualitativer Forschungsstrategien anstrebt (z. B. Teddlie & Tashakkori, 2009; ▶ Abschn. 2.4). Es sei deswegen noch einmal betont: Der Kritische Rationalismus ist im eigenen Verständnis *keine* Variante oder Weiterentwicklung des Positivismus, sondern vertritt ausdrücklich ein geradezu entgegengesetztes Wissenschaftsverständnis: Nicht auf der Basis einer Anhäufung empirischer Daten lassen sich allgemeingültige Theorien begründen, sondern durch das Aufstellen von Theorien (Vermutungen) und deren Falsifikationen (Widerlegungen) wird wissenschaftlicher Erkenntnisfortschritt vorangetrieben (Popper, 1963/2009, S. 236):

> » Unsere Versuche, Wissen über unsere Welt zu erlangen, enthalten nur ein einziges rationales Element: die kritische Prüfung unserer Theorien. Die Theorien selbst sind Versuche, die Lösung eines Problems zu erraten: bestenfalls eine Vermutung. Wir wissen nicht, sondern wir raten. Wenn mich jemand fragt: ‚Woher weißt Du?', so antworte ich: ‚Ich weiß nicht, ich rate nur. Und wenn Du an meinem Problem interessiert bist, bitte kritisiere meine Vermutung; und wenn Du einen Gegenvorschlag machst, dann laß mich versuchen, ihn meinerseits zu kritisieren.' Das ist, glaube ich, die wahre Theorie der Erkenntnis (die ich Ihnen hiermit zur Kritik vorlege) [...] Unsere Theorie sagt: der Fortschritt des Wissens besteht aus *Vermutungen und Widerlegungen*.

Zur erkenntnistheoretischen Abgrenzung zwischen Kritischem Rationalismus mit seinem deduktiven Falsifikationsprinzip einerseits und Empirismus bzw. Positivismus mit dem Induktionsprinzip andererseits hat Popper (1984) zudem die Metaphern von Kübel und Scheinwerfer benutzt: Gemäß **induktiv-positivistischer Kübeltheorie der Erkenntnis** ist unser Verstand wie ein leerer Kübel, der nach und nach mit dem Wasser der Erkenntnis gefüllt wird, d. h. durch Sinneswahrnehmungen sammeln wir schrittweise gesichertes Wissen über die Realität an. Demgegenüber basiert laut **kritisch-rationalistischer Scheinwerfertheorie der Erkenntnis** der Wissensgewinn nicht auf passiver Wahrnehmung, sondern auf aktiver Forschung: Eine bestimmte von uns aufgestellte Theorie bzw. Hypothese veranlasst uns in der Wissenschaft, gezielt bestimmte Ausschnitte der Wirklichkeit zu beobachten. Die gewonnenen Daten werden zur kritischen Prüfung der Hypothese genutzt. Der menschliche Verstand formuliert Hypothesen und diese fungieren als Scheinwerfer, indem sie den Untersuchungsgegenstand in bestimmter Weise beleuchten. Der Untersuchungsgegenstand seinerseits reflektiert das Scheinwerferlicht in spezifischer Weise, die den Vorannahmen entsprechen oder widersprechen kann. Die Scheinwerfertheorie erklärt wissenschaftlichen **Erkenntnisgewinn als aktiven Prozess**, der von den Theorien/Hypothesen bzw. Scheinwerfern der Forschenden ausgeht, wobei jedoch das vom Untersuchungsgegenstand reflektierte Licht nicht nur vom Scheinwerfer, sondern auch maßgeblich von den Merkmalen des Untersuchungsgegenstands abhängt (s. zu Poppers Kübel- und Scheinwerfertheorie der Erkenntnis auch Wurzer, 1994).

2.2.2 Wahrheitssuche statt Wahrheitsbehauptung

Im Alltag gehen wir in der Regel von der ontologischen Annahme aus, dass die von uns wahrgenommene Erfahrungswirklichkeit nicht ausschließlich in unserem Bewusstsein als Vorstellung oder Idee existiert (das wäre die Position des **Idealismus**), sondern unabhängig von unserem Bewusstsein besteht (das ist die Position des **Realismus**). Aussagen, die auf die Realität zutreffen, mit ihr übereinstimmen bzw. mit ihr korrespondieren, betrachten wir als wahr (das ist die **Korrespondenztheorie der Wahrheit**). Von der Realität abweichende Aussagen dagegen gelten als unwahr und zwar je nach Hintergrund als Täuschung, Irrtum, Lüge etc.

Lassen sich diese Alltagsvorstellungen über Realität und Wahrheit auch auf die Wissenschaft übertragen? Ist es laut wissenschaftstheoretischer Position des Kritischen Rationalismus Ziel der empirischen Sozialforschung, Theorien zu finden, die in dem Sinne wahr sind, dass sie der sozialen Wirklichkeit entsprechen? Prinzipi-

ell ja, denn der Kritische Rationalismus geht davon aus, dass erfahrungswissenschaftliche Theorien der Wirklichkeit entsprechen oder widersprechen können. Allerdings ist es gemäß Kritischem Rationalismus niemals möglich, den Wahrheitsgehalt einer Theorie unzweifelhaft zu belegen. Alles Wissen ist grundsätzlich **Vermutungswissen** und immer von Irrtum bedroht (das ist die Position des **Fallibilismus**). Der Kritische Rationalismus wendet sich damit entschieden gegen alle dogmatischen Heilslehren und Pseudowissenschaften (► Abschn. 3.1.1), die einen absoluten und letztgültigen Wahrheitsanspruch erheben und keine rationale Kritik mehr zulassen. Gemäß Fallibilismus ist Erkenntnisgewinn ein fortwährender Prozess der kritischen Diskussion konkurrierender Theorien, der nie für beendet erklärt werden kann. Dass eine zweifelsfreie Letztbegründung von Aussagen nicht möglich ist, wird vor allem mit dem **Münchhausen-Trilemma** verdeutlicht (► Das Münchhausen-Trilemma der Wahrheitsbegründung).

Obwohl der Kritische Rationalismus betont, dass niemand begründet den Anspruch erheben kann, die letztgültige Wahrheit gefunden zu haben, besteht er doch darauf, dass wissenschaftliche Forschung **Wahrheitssuche** bedeutet. Eine pessimistische Haltung, der gemäß letztlich keinerlei Erkenntnis über soziale Wirklichkeit möglich ist, ist dem Kritischen Rationalismus ebenso fremd wie eine relativistische Haltung, der gemäß widersprüchliche Wahrheiten über exakt denselben Sachverhalt gleichberechtigt nebeneinander bestehen können. Vielmehr sieht der Kritische Rationalismus die Erkenntnissuche in der kritischen Auseinandersetzung mit konkurrierenden Theorien und ihrer Bewährung an der Erfahrungswirklichkeit. Auf diese Weise kommt es durch **Versuch und Irrtum** zu einer schrittweisen Annäherung an die Wahrheit (Popper, 1934/1989, S. 225):

> » Der Ehrgeiz, recht zu behalten, verrät ein Mißverständnis: nicht der *Besitz* von Wissen, von unumstößlichen Wahrheiten macht den Wissenschaftler, sondern das rücksichtslos kritische, das unablässige *Suchen* nach Wahrheit.

Mit Theorien oder Wahrheitsbehauptungen, die nicht prinzipiell falsifizierbar sind, befindet man sich gemäß Kritischen Rationalismus nicht mehr im Bereich der Erfahrungswissenschaft. Im Kritischen Rationalismus ist die **Falsifizierbarkeit** das **Abgrenzungs- oder Demarkationskriterium**, anhand dessen wissenschaftliche Theorien von religiösen Lehren, politischen Ideologien oder sonstigen Weltanschauungen abgegrenzt werden.

Das Münchhausen-Trilemma der Wahrheitsbegründung

Das Münchhausen-Trilemma wird bezeichnet nach dem als „Lügenbaron“ bekannt gewordenen Münchhausen, einem im 18. Jahrhundert lebenden deutschen Adligen, dem historisch wie literarisch diverse Lügengeschichten zugeschrieben werden, unter anderem die, er habe einmal sich selbst samt Pferd am eigenen Schopf aus dem Sumpf gezogen. In der wissenschaftstheoretischen Diskussion ist mit dem Münchhausen-Trilemma gemeint, dass der Versuch, eine Aussage unfehlbar als wahr zu begründen, letztlich immer scheitert, denn keine der drei möglichen argumentativen Strategien führt zum Erfolg (Albert, 1968/1991, S. 15f.):

- **Infiniter Regress**: Um die Wahrheit einer Aussage zu begründen wird auf Aussagen bzw. Argumente zurückgegriffen, die ihrerseits begründet werden müssen. Die dafür angeführten Argumente müssen ihrerseits begründet werden. Die Begründung der Begründung der Begründung etc. findet somit niemals ein definitives Ende und somit ist keine letztgültige Wahrheitsbehauptung möglich.
- **Logischer Zirkel**: Hier wird eine scheinbar schlüssige Begründungskette aufgebaut, wobei jedoch angeführte Argumente bereits in den vorgelagerten Argumenten oder in der Definition des zu begründenden Sachverhalts enthalten waren. Die scheinbar gelieferte Letztbegründung steckte bereits in den Prämissen, so dass man sich argumentativ im Kreis dreht.
- **Abbruch des Verfahrens**: Um nicht ewig im infiniten Regress gefangen zu sein, kann in der Begründungskette an irgendeinem Punkt innegehalten werden mit dem Argument, diese Begründung sei nun ausreichend, weil sie doch offensichtlich, evident, intuitiv erfassbar – jedenfalls nicht weiter zu hinterfragen – sei. Doch dieser Abbruch des Begründungsverfahrens erfordert die Festsetzung von und den Glauben an eine bestimmte Letztbegründung, die nicht mehr kritisiert werden darf (das entspricht dann einem Dogma). Ein Wahrheitsanspruch, der auf dem Abbruch des Begründungsverfahrens basiert, ist aus wissenschaftstheoretischer Sicht des Kritischen Rationalismus willkürlich und läuft auf Dogmatismus hinaus. Da wir in der Forschungspraxis jeweils mit bestimmten Vorannahmen arbeiten (auf die wir uns nur durch Abbruch eines potenziell unendlichen Begründungsprozesses stützen können), kann darauf basierendes Wissen immer nur als vorläufige und weiterhin kritisierbare Erkenntnis (nicht als unumstößliche Wahrheit) anerkannt werden.

Das Münchhausen-Trilemma begründet die Position des Fallibilismus, dass die Hoffnung, sich in der Wissenschaft sozusagen argumentativ am eigenen Schopf aus dem Sumpf der Unsicherheit ziehen und zu unumstößlichen Wahrheiten gelangen zu können, immer zum Scheitern verurteilt ist. (Es ist zu beachten, dass Probleme der Wahrheitsbegründung und logischen Beweisführung natürlich nicht nur im Kritischen Rationalismus diskutiert werden, sondern eine lange Tradition in der Philosophie haben.)

Falsifizierbarkeit bedeutet, dass aus den in einer Theorie enthaltenen orts- und zeitübergreifenden allgemeinen Gesetzen konkrete Vorhersagen über die Verhältnisse in der sozialen Wirklichkeit zu bestimmten Zeiten an bestimmten Orten logisch-deduktiv ableitbar sind, die sich anhand von empirischen Daten prinzipiell widerlegen lassen. Beispiel: Die Theorie „Rauchen verursacht Lungenkrebs" ist falsifizierbar, denn aus ihr lässt sich die empirisch prüfbare Hypothese ableiten „Lungenkrebs tritt bei einer Stichprobe von Rauchern überzufällig häufiger auf als bei einer vergleichbaren Stichprobe von Nichtrauchern". Zu beachten ist, dass wir in den Sozial- und Humanwissenschaften aufgrund der Komplexität des Menschen nicht von deterministischen, sondern von probabilistischen (wahrscheinlichkeitstheoretischen) Theorien ausgehen: Die Theorie „Rauchen verursacht Lungenkrebs" prognostiziert also nicht, dass jeder einzelne Raucher zwingend an Lungenkrebs erkrankt, sondern dass Lungenkrebs – in Abhängigkeit von weiteren Einflussfaktoren – bei Rauchern mit höherer Wahrscheinlichkeit als bei Nichtrauchern auftritt (► Abschn. 2.2.8).

Demgegenüber ist die Aussage „Die Seelen der Heiligen sind bei Gott und können sich bei Gott für die Gläubigen einsetzen." empirisch nicht falsifizierbar, da die angesprochenen Sachverhalte nicht zur Beobachtungswirklichkeit gehören. Auch Existenzaussagen bzw. Es-gibt-Sätze („Es gibt Kinder, die niemals weinen."), Kann-Sätze („Bei Zigarettenkonsum kann es zu Lungenkrebs kommen.") und normative Aussagen („Widdergeborene sollten im Januar auf ihre Finanzen achten.") erfüllen nicht das Kriterium der Falsifizierbarkeit. Ebenso wenig sind tautologische Aussagen, die immer wahr sind und keine Falsifikatoren haben („Wenn der Hahn kräht auf dem Mist, ändert sich das Wetter oder es bleibt wie es ist."), oder kontradiktorische Aussagen, die immer unwahr sind und keine Konfirmatoren besitzen („Wenn eine Person keinen Wein trinkt, dann trinkt sie Chardonnay."), als wissenschaftliche Theorien oder Hypothesen mangels Falsifizierbarkeit unbrauchbar (vgl. dazu auch Kromrey, 2009).

Fallibilismus – Gemäß Kritischem Rationalismus ist jedes Wissen fehlbar und somit nur hypothetisches Wissen bzw. Vermutungswissen. Ein letztgültiger Wahrheitsanspruch kann, wie das Münchhausen-Trilemma verdeutlicht, nicht begründet werden (Fallibilismus; „fallibilism").

Wahrheitssuche – Wissenschaftliche Forschung besteht im Verständnis des Kritischen Rationalismus bzw. Fallibilismus in einer nie endenden Wahrheitssuche („search for truth") durch kritische Prüfung von Theorien, wobei man sich durch Versuch und Irrtum der Wahrheit zumindest annähert. Dies setzt voraus, dass wissenschaftliche Theorien so zu formulieren sind, dass man sie prinzipiell empirisch prüfen und falsifizieren kann (Falsifizierbarkeit). Nicht-falsifizierbare Behauptungen sind laut Kritischem Rationalismus nicht Gegenstand der Erfahrungswissenschaft.

2.2.3 Kritischer statt naiver Realismus

In Abgrenzung zum Alltagsverständnis, demgemäß die äußere Realität so ist, wie wir sie wahrnehmen (sog. **naiver Realismus**), wird im Kritischen Rationalismus überwiegend ein **kritischer Realismus** vertreten, demgemäß a) eine vom menschlichen Bewusstsein unabhängige, bestimmten Gesetzmäßigkeiten folgende Wirklichkeit existiert und diese b) zumindest teilweise für den Menschen erkennbar ist, wobei jedoch in Abhängigkeit vom menschlichen Wahrnehmungs- und Denkapparat immer mit mehr oder minder starken Verzerrungen zu rechnen ist (vgl. Albert, 1987).

Wenn der Kritische Rationalismus ontologisch also den Idealismus (demgemäß die Wirklichkeit nur im menschlichen Bewusstsein als „Idee" existiert) verwirft und den Realismus (demgemäß die Wirklichkeit unabhängig vom menschlichen Bewusstsein existiert) bevorzugt, ist dies eine Vorannahme, die ihrerseits nicht prüfbar ist. Popper argumentiert, dass der Realismus im Vergleich zum Idealismus aus verschiedenen Gründen deutlich plausibler ist (vgl. Popper, 1973, S. 39ff.), betont aber, dass das Bekenntnis zum Realismus letztlich eine Glaubensfrage darstellt: Wissenschaftliche Forschung „ist geleitet von dem unwissenschaftlichen, metaphysischen (aber biologisch erklärbaren) Glauben, dass es Gesetzmäßigkeiten gibt, die wir entschleiern, entdecken können." (Popper, 1934/1989, S. 223).

Kritischer Realismus – Im Kritischen Rationalismus wird die Existenz einer außerhalb des menschlichen Bewusstseins existierenden, bestimmten Gesetzmäßigkeiten folgenden Wirklichkeit angenommen, über die der Mensch prinzipiell Erkenntnisse gewinnen kann, wenn auch das Wissen immer unsicher bleibt (kritischer Realismus; „critical realism").

Basissatz- und Korrespondenzproblem – Gemäß kritischem Realismus sind empirische Daten keine reinen Abbilder der Realität, sondern im Zuge der Forschung erzeugte, stets von Theorien durchsetzte Beschreibungen, deren Übereinstimmung mit der beobachtbaren Wirklichkeit (sog. Basissatzproblem) sowie deren Entsprechung mit den jeweiligen Begriffen/Konstrukten innerhalb der zu prüfenden Theorie (sog. Korrespondenzproblem) ihrerseits theoretisch zu fassen und kritisch-empirisch zu prüfen sind anhand sog. Hilfs- oder Instrumententheorien.

Im deduktiv-falsifizierenden Wissenschaftsverständnis des Kritischen Rationalismus geht es darum, Theorien über die Erfahrungswirklichkeit anhand empirischer Hypothesenprüfung kritisch zu beurteilen. Wenn die Daten aber nicht per se als wahre Abbilder der Realität bzw. als Tatsachen zu betrachten sind (das wäre die Position des naiven Realismus), sondern gemäß kritischem Realismus stets ihrerseits zu hinterfragen sind, ergeben sich epistemologisch zwei Probleme des Kritischen Rationalismus, das sog. Basissatz- und das Korrespondenzproblem:

1. **Basissatzproblem**: In der empirischen Sozialforschung sollen die Theorien bzw. die aus ihnen abgeleiteten Hypothesen an der sozialen Wirklichkeit auf ihren Wahrheitsgehalt geprüft werden. Zu diesem Zweck werden Daten über die Wirklichkeit erhoben, diese Messwerte bzw. Beobachtungsaussagen (sog. Protokollsätze) werden auch als **Basissätze** bezeichnet. Das Basissatzproblem besteht laut Kritischem Rationalismus nun darin, dass unsere Basissätze bzw. empirischen Daten verzerrt und falsch sein können, d. h. die realen Verhältnisse nicht korrekt widergeben. So gehen wir z. B. davon aus, dass die in einer Studie erhobenen Intelligenztestergebnisse den gezeigten Intelligenzleistungen der Untersuchungspersonen tatsächlich entsprechen. Basissätze können jedoch gemäß kritischem Realismus nicht beanspruchen theoriefrei Tatsachen zu beschreiben, denn sie sind im Zuge der Forschung auf der Basis theoretischer Vorannahmen produziert worden (▶ Abschn. 2.2.1 zur Scheinwerfertheorie der Erkenntnis): „Es gibt keine reinen Beobachtungen: sie sind von Theorien durchsetzt und werden von Problemen und von Theorien geleitet“ (Popper, 1934/1989, S. 76). Basissätze sind im Verständnis des Kritischen Rationalismus auch nicht verifizierbar. Stattdessen können sie nur einer kritischen Prüfung unterzogen und **durch Beschluss der Scientific Community als vorläufig gültig anerkannt** werden.
   Popper (1934/1989, S. 74f.) spricht in diesem Zusammenhang metaphorisch auch von einem **Gerichtsverfahren**, in dem über die vorläufige Anerkennung von Basissätzen entschieden wird. In diesem Gerichtsverfahren werden Entstehung und Begleitumstände der Basissätze anhand von Kriterien geprüft, die ihrerseits wiederum konsensbedürftig sind. So wird man in der Forschungspraxis eine **Aussage über das Intelligenztestergebnis einer Untersuchungsgruppe**, das auf den gemessenen Testwerten der einzelnen Gruppenmitglieder basiert, als Basissatz anerkennen, wenn bekannt ist, dass ein etablierter Intelligenztest (▶ Abschn. 10.4.3) genutzt und regelkonform durchgeführt und ausgewertet wurde. Der Basissatz würde dagegen z. B. angezweifelt werden, wenn ein unpublizierter oder veralteter Test eingesetzt wurde, dessen testtheoretische Eigenschaften unbekannt oder unbefriedigend sind, wenn die Testdurchführung und Testauswertung nicht durch qualifiziertes Personal erfolgten oder wenn die Bearbeitungszeit und die Bearbeitungsbedingungen nicht für alle Testpersonen gleichermaßen festgelegt wurden. Die Basissätze sind die Beobachtungssätze, mit denen die Realität beschrieben wird. Dass diese Realitätsbeschreibung unsicher ist – kein „Felsengrund“, sondern eher **„Sumpfland“** – gibt Popper (1934/1989, S. 75f.) zu. Hier besteht die Möglichkeit, dass die Forschenden sich irren, dass sie – um die Metapher der Scheinwerfertheorie (▶ Abschn. 2.2.1) aufzugreifen – einen defekten Scheinwerfer verwenden und somit zu Basissätzen kommen, die von der Realität abweichen. Da jedoch die Unsicherheit von Basissätzen bekannt ist und sie somit jederzeit hinterfragt werden können, sollten Fehler und Irrtümer aufzudecken sein, insbesondere durch **Offenlegung des methodischen Vorgehens** (inklusive der Erhebungsinstrumente und Rohdaten) sowie durch **Wiederholungs- bzw. Replikationsstudien** anderer Forschungsteams. Mit dem Hinweis, dass die Wissenschaftsgemeinschaft Basissätze durch Konsens anerkennt, werden argumentativ gemäß Münchhausen-Trilemma (▶ Abschn. 2.2.2) der infinite Regress und der Zirkelschluss vermieden, stattdessen wird das Argumentationsverfahren abgebrochen. Aber eben nicht mit dem Dogma der letztgültigen Wahrheitsbehauptung, sondern mit einer kritisierbaren Vermutung.

Basissatzproblem – Das Basissatzproblem („problem of basic statements“) bezieht sich auf die Problematik des Nachweises, dass Beobachtungsprotokolle bzw. die erhobenen empirischen Daten tatsächlich mit der Wirklichkeit übereinstimmen. Die Übereinstimmung kann nicht bewiesen, sondern nur in einem – ständiger Kritik unterzogenen – Konsens der Wissenschaftsgemeinschaft festgelegt werden. Damit handelt es sich bei empirischen Daten im Verständnis des Kritischen Rationalismus nicht einfach um reine Fakten, sondern immer um theoriegeladene Aussagen.

2. **Korrespondenzproblem**: Dass die Daten mit der Realität übereinstimmen, reicht nicht aus, um sie begründet zur Theorieprüfung nutzen zu können. Die gemessenen Variablen (Indikatoren) müssen inhaltlich auch den **abstrakten Begriffen** bzw. **theoretischen Konstrukten** entsprechen, die in der zu prüfenden Theorie vorkommen. Man spricht vom Korrespondenzproblem, um die Problematik anzusprechen, dass wir von konkreten empirischen Indikatoren bzw. einer **Beobachtungssprache** (also einer Sprache, die sich auf beobachtete Indikatoren bezieht: z. B. Punktwert in einem Intelligenztest) die Brücke schlagen müssen zu abstrakten theoretischen Begriffen bzw. einer **theoretischen Sprache** (also einer Sprache, die sich auf theoretische Konzepte bezieht: z. B. Bedeutung des Konstrukts der Intelligenz). Selbst ein vorläufig als gültig anerkannter Basissatz kann zur Prüfung von Theorien untauglich sein, wenn der Messwert des Indikators nicht genau dem inhaltlichen Konstrukt in der Theorie entspricht. Dies kann z. B. der Fall sein, wenn die zu prüfende Theorie auf eine andere Intelligenzkomponente oder ein anderes Intelligenzverständnis abzielt als der verwendete Intelligenztest. Die Ent-

sprechung zwischen Aussagen über empirische Indikatoren und Aussagen über theoretische Konstrukte (sog. Zweisprachenproblem) wird im kritischen Rationalismus ebenfalls nicht als gesichert vorausgesetzt. Stattdessen müssen die **Korrespondenzregeln** als kritisch zu diskutierende und **empirisch zu prüfende Hypothesen** behandelt werden (Schnell, Hill, & Esser, 2008, S. 79).

Korrespondenzproblem – Das Korrespondenzproblem („problem of correspondence") bezieht sich auf die Problematik des Nachweises, dass die in einer empirischen Untersuchung eingesetzten Indikatoren tatsächlich das erfassen, was mit den in der zu prüfenden Theorie verwendeten abstrakten Begriffen bzw. theoretischen Konstrukten gemeint ist. Diese Übereinstimmung kann nicht bewiesen, sondern nur in einem – ständiger Kritik unterzogenen – Konsens der Wissenschaftsgemeinschaft festgelegt werden.

Gemäß Realismusverständnis des Kritischen Rationalismus sind empirische Daten keine Abbilder der Realität, sondern durch die Forschenden und ihre Theorien und Methoden („Scheinwerfer") erzeugte **hypothetische Aussagen**, deren Übereinstimmung mit der Realität (Basissatzproblem) sowie deren Entsprechung mit den jeweiligen Konstrukten innerhalb der Theorie (Korrespondenzproblem) ihrerseits theoretisch zu fassen und kritisch-empirisch zu prüfen sind. Man spricht in Abgrenzung zu der jeweils interessierenden inhaltlichen Theorie **(Kerntheorie)** hierbei von den im Zuge empirischer Forschung ergänzend verwendeten **Instrumenten- oder Hilfstheorien** (vgl. Lakatos, 1977). Beispiel: Um eine entwicklungspsychologische Kerntheorie zur Intelligenzentwicklung über die Lebensspanne empirisch mittels Intelligenztest zu prüfen, wird u. a. immer auch eine Instrumententheorie im Sinne einer Intelligenztesttheorie benötigt, die begründet, dass der verwendete Intelligenztest tatsächlich das Konstrukt „Intelligenz" misst und wie er korrekt anzuwenden ist, um gültige Messwerte bei den Untersuchungspersonen zu erzeugen. Neben der Instrumententheorie im engeren Sinne, die sich auf die verwendeten Messinstrumente bezieht, treffen Hilfstheorien auch Aussagen über diverse untersuchungs- oder personengebundene Einflüsse, die als Störungen interpretiert werden, welche die Ergebnisse verfälschen können.

Während man erkenntnistheoretisch von Hilfs- oder Instrumententheorien spricht, die Aussagen darüber treffen, inwiefern die in einer Studie erhobenen Daten etwas über das Verhältnis von Theorie und Realität aussagen, spricht man in der empirisch-quantitativen Forschungspraxis von **Gütekriterien von Messinstrumenten** bzw. genauer: von **Gütekriterien von Schlussfolgerungen, die aus den gewonnenen Messwerten gezogen werden**. Zentrales Gütekriterium ist dabei die Konstruktvalidität. Zudem sind auch die theoretisch begründete **Güte von Forschungsdesigns** bzw. genauer: die Güte der aus den Studien gezogenen Schlussfolgerungen (interne und externe Validität) sowie die **Güte von Stichproben** (Repräsentativität) wichtig für die Interpretation von Daten im Hinblick auf ihre Aussagekraft über die Realität einerseits und die zu prüfende Kerntheorie andererseits (zu Qualitäts- bzw. Gütekriterien der quantitativen Forschung ► Abschn. 3.2 und ► Abschn. 10.4.1).

Die Anerkennung von Basissätzen und die Korrespondenz zwischen konkreten empirischen Indikatoren und abstrakten theoretischen Konstrukten erfolgt im Konsens der **Scientific Community** auf der Grundlage von expliziten und impliziten Hilfs- bzw. Instrumententheorien. Einer Kerntheorie widersprechende Daten werden also nicht automatisch als Widerlegung der Theorie aufgefasst (das wäre **naiver Falsifikationismus** mit **naivem Realismus**), sondern sie werden unter Berücksichtigung der Hilfs- und Instrumententheorien kritisch betrachtet, was auch dazu führen kann, dass die widersprechenden Daten als fehlerhaft oder ungeeignet zur Theorieprüfung zurückgewiesen werden können (**methodologischer Falsifikationismus** mit **kritischem Realismus**). Gemäß Kritischem Rationalismus ist all dies im Diskurs der Scientific Community zu klären, der jeweils die theoretischen Vorannahmen sowie die methodischen Details einer Studie detailliert vorzulegen sind, so dass die Interpretationsspielräume nicht beliebig ausdehnbar sind: Widersprechende, aber solide erhobene Daten können nicht einfach als „messfehlerbehaftet" oder „nicht repräsentativ" wegdiskutiert werden, um die Theorie vor Kritik zu bewahren (dies würde einer Immunisierung gegen Kritik und somit unwissenschaftlichem Vorgehen entsprechen). Theorien können aber auch nicht für ungültig erklärt werden, wenn man nur sehr anfechtbares widersprechendes Datenmaterial vorzuweisen hat.

Methodologischer Falsifikationismus – Die kritische Prüfung von inhaltlichen Kerntheorien muss im Verständnis des Kritischen Rationalismus wegen des Basissatz- und Korrespondenzproblems immer auch unter Berücksichtigung der für die Studie relevanten Hilfs- oder Instrumententheorien erfolgen (methodologischer Falsifikationismus; „methodological falsificationism"). Dabei sind nicht zuletzt Replikationsstudien (Wiederholungsstudien, „replication studies") besonders wichtig, die nämlich nicht nur auf derselben inhaltlichen Kerntheorie, sondern auch auf denselben Hilfs- und Instrumententheorien basieren wie die jeweiligen Vorläuferstudien.

Ein wichtiges Maß für den **Bewährungsgrad einer Theorie** sind schließlich **Anzahl und Strenge der Falsifikationsversuche**, denen sie ausgesetzt wurde. Insbesondere Replikationsstudien, die Vorläuferstudien identisch oder mit geringen Modifikationen wiederholen, stellen eine strenge Bewährungsprobe bislang nicht-falsifizierter inhaltlicher Kerntheorien sowie vor allem auch der

zugehörigen Instrumenten- und Hilfstheorien dar. Dieser wissenschaftliche Regulationsmechanismus greift jedoch nur insofern, wie tatsächlich Replikationsstudien in nennenswertem Umfang durchgeführt und publiziert werden.

Bewährungsgrad einer Theorie – Der Bewährungsgrad („corroboration") einer erfahrungswissenschaftlichen Theorie zu einem bestimmten Sachverhalt ist im Verständnis des Kritischen Rationalismus von der Anzahl und Strenge der überstandenen Replikations- bzw. Falsifikationsversuche abhängig.

2.2.4 Grad der Falsifizierbarkeit und Informationsgehalt von Theorien

Theorien unterscheiden sich in ihrem **Informationsgehalt** (empirischen Gehalt; „empirical content") bzw. im **Grad ihrer Falsifizierbarkeit** („degree of falsifiability") oder anders gesagt im **Grad ihrer Allgemeingültigkeit** (Schnell et al., 2008, S. 62ff.). Eine Theorie, die einen weiten Gültigkeitsbereich beansprucht und sehr präzise Vorhersagen macht, hat einen höheren Informationsgehalt und kann potenziell durch mehr widersprechende Beobachtungsbeispiele (d. h. **Falsifikatoren**) widerlegt werden als eine Theorie mit einem sehr engen Gültigkeitsbereich und sehr vagen Vorhersagen. Gehen wir im einfachsten Fall von einer Kausaltheorie aus, die auf einer einzigen Annahme in Form eines **Wenn-Dann-Satzes** basiert: Im Wenn-Teil wird die Ursache (bzw. die Ausprägung der unabhängigen Variable) genannt und im Dann-Teil die Wirkung (bzw. die Ausprägung der abhängigen Variable): „Wenn gewalthaltige digitale Spiele genutzt werden, dann zeigen sich verstärkt aggressive Gedanken".

Der Informationsgehalt (d. h. der Grad der Allgemeingültigkeit) sowie die Zahl der möglichen Falsifikatoren *sinken*, wenn man den **Wenn-Teil der Hypothese** (bei unveränderter Dann-Komponente) durch **Und-Komponenten** (Konjunktion) stärker eingrenzt: „Wenn im Jugendalter sehr viele gewalthaltige digitale Spiele genutzt werden **und** eine Vernachlässigung durch die Eltern vorliegt **und** schlechte Bildungs- und Berufschancen bestehen **und** das Selbstwertgefühl gering ist, dann zeigen sich verstärkt aggressive Gedanken". Würde man weitere Und-Komponenten hinzufügen (z. B. … **und** keine Geschwister im Haushalt leben **und** das Geburtsdatum der 13.4.2000 ist), so könnte durch den Wenn-Teil im Extrem ein Einzelfall spezifiziert sein. Die Theorie hätte dann nur einen Einzelfall als potenziellen Falsifikator. Sie wäre angesichts der im Kritischen Rationalismus geforderten Allgemeingültigkeit wissenschaftlicher Theorien (nahezu) informationslos, weil sie nur auf eine einzige Person anwendbar ist. Demgegenüber führt eine Erweiterung des Wenn-Teils um **Oder-Komponenten** (Disjunktion) zu einer **Erhöhung des Informationsgehaltes**: „Wenn gewalthaltige digitale Spiele **oder** gewalthaltige Fernsehprogramme genutzt werden, dann zeigen sich verstärkt aggressive Gedanken". Potenzielle Falsifikatoren dieser Theorie sind hier nicht nur die Nutzergruppen gewalthaltiger Spiele, sondern auch gewalthaltiger TV-Programme. Die Theorie hat einen höheren Informationsgehalt bzw. einen höheren Verallgemeinerungsgrad, denn sie adressiert nicht nur Computerspiele, sondern auch Fernsehsendungen und somit eine größere Gruppe von Medienangeboten bzw. Mediennutzenden.

Während die Theorie durch logische Und-Erweiterung des Wenn-Teils (und damit Eingrenzung des Anwendungsbereichs) an Informationsgehalt verliert, führt die **Und-Erweiterung des Dann-Teils einer Hypothese** zu einer Steigerung des Informationsgehalts, d. h. sie lässt sich dann prinzipiell besser falsifizieren, weil die Wirkungen präzisiert sind. „Wenn gewalthaltige digitale Spiele genutzt werden, dann zeigen sich verstärkt aggressive Gedanken **und** aggressive Verhaltensweisen". Personen, die auf die Nutzung gewalthaltiger digitaler Spiele nicht mit aggressiven Verhaltensweisen reagieren, können nun zur Falsifikation der Theorie beitragen, während das bei der einfachen Dann-Komponente, die nur auf aggressive Gedanken abzielt, nicht der Fall ist. Bei einer logischen **Oder-Erweiterung des Dann-Teils** („Wenn gewalthaltige digitale Spiele genutzt werden, dann zeigen sich verstärkt aggressive Gedanken *oder* aggressive Verhaltensweisen") sinkt dagegen wiederum der Informationsgehalt der Theorie, da die vorhergesagten Wirkungen weniger stark präzisiert sind.

Informationsgehalt einer Theorie – Falsifizierbare Theorien unterscheiden sich in ihrem Informationsgehalt (empirischen Gehalt; „empirical content") bzw. im Grad der Falsifizierbarkeit („degree of falsifiability"), d. h. gleichzeitig im Grad ihrer Allgemeingültigkeit und Präzision. Eine Theorie mit hohem Informationsgehalt zeichnet sich durch einen breiten Gültigkeitsbereich (Oder-Erweiterung der Wenn-Komponente: höhere Allgemeingültigkeit) und sehr präzise Vorhersagen aus (Und-Erweiterung der Dann-Komponente: höhere Präzision). Für sie existieren mehr Falsifikatoren als für eine Theorie mit geringem Informationsgehalt. Kann sich die Theorie mit hohem Informationsgehalt trotz vielfältiger Falsifikationsmöglichkeiten empirisch dennoch bewähren, trägt dies im Verständnis des Kritischen Rationalismus stärker zum Erkenntnisgewinn bei als die Prüfung informationsärmerer Theorien.

2.2.5 Raffinierter statt einfacher methodologischer Falsifikationismus

Gemäß **methodologischem Falsifikationismus** werden Hypothesen auf der Basis widersprechender Daten nur dann verworfen, wenn die der Datenerhebung zugrunde liegenden Methoden und Theorien (v. a. auch Hilfs- und Instrumententheorien) die Wissenschaftsge-

meinschaft überzeugen (▶ Abschn. 2.2.3). Im Falle einer Falsifikation wird dann aber typischerweise nicht die gesamte theoretische Vorarbeit verworfen und ein ganz neues wissenschaftliches „Raten" über denselben Sachverhalt beginnen. Vielmehr wird man eher versuchen, die offenbar **fehlerhafte Theorie zu modifizieren**.

Eine typische Theoriemodifikation im Licht hypothesenkonträrer Daten ist die **Exhaustion**, bei welcher der Geltungsbereich der Theorie eingeschränkt wird, um die theoriekonträren Fälle auszuschließen (Holzkamp, 1968, S. 159ff.; 1972). Beispiel: Der Gesetzesaussage „Alle Schwäne sind weiß" widersprechend werden in einer bestimmten Region X durch zuverlässige Beobachtung schwarze Schwäne entdeckt. Statt die Theorie komplett zu verwerfen wird sie durch Exhaustion angepasst: „Alle Schwäne sind weiß außer den Schwänen in Region X" bzw. in der für wissenschaftliche Theorien gebräuchlichen Wenn-Dann-Formulierung: „Wenn es sich bei den Untersuchungsobjekten um Schwäne handelt *und* wenn sie nicht in Region X leben, dann sind sie weiß". Die modifizierte Theorie muss dann noch einmal geprüft werden. Dabei könnte man erneut widersprechende Daten finden, z. B. graue Schwanküken. Statt die Theorie an dieser Stelle zu verwerfen, wäre sie weiter zu exhaurieren: „Wenn es sich bei den Untersuchungsobjekten um Schwäne handelt **und** wenn sie nicht in Region X leben **und** wenn sie ausgewachsen sind, dann sind sie weiß". Die Exhaustion von Theorien führt zwar meist dazu, dass sich in Folgestudien keine oder kaum noch widersprechende Daten zeigen und somit der empirische Bestätigungsgrad steigt, gleichzeitig sinkt aber der Informationsgehalt der Theorie, weil ihr Anwendungsbereich immer enger eingegrenzt wird (Und-Erweiterung der Wenn-Komponente ▶ Abschn. 2.2.4).

Exhaustion – Theoriekonträre empirische Ergebnisse können zum Anlass genommen werden, die betreffende Theorie nicht komplett zu verwerfen, sondern zu modifizieren. Eine typische Theoriemodifikation besteht in der Exhaustion („exhaustion"), bei der der Geltungsbereich der Theorie eingeschränkt wird, um theoriekonträre Fälle auszuschießen. Diese Einschränkung des Geltungsbereiches läuft auf eine Erweiterung des Wenn-Teils der Theorie durch eine oder mehrere logische Und-Komponenten hinaus. Exhaustion kann den empirischen Bestätigungsgrad einer Theorie steigern, allerdings auf Kosten ihres Informationsgehaltes bzw. Allgemeingültigkeitsanspruchs.

Gibt es neben der Exhaustion eine andere Möglichkeit, wissenschaftstheoretisch begründet mit Theorien weiterzuarbeiten, wenn methodisch solide theoriekonträre Daten vorliegen und die Theorie somit eigentlich verworfen werden muss? Der sog. **raffinierte methodologische Falsifikationismus** bejaht diese Frage. Er wurde als Weiterentwicklung von Poppers klassischem methodologischen Falsifikationismus (▶ Abschn. 2.2.3) durch den ungarischen Philosophen **Imre Lakatos** (1977) vorgeschlagen. Die „Raffinesse" besteht dabei darin, dass man neben der vorläufigen empirischen Bestätigung einer Theorie ihre Qualität auch im Hinblick auf ihre **heuristische Erklärungskraft im Vergleich zu konkurrierenden Theorien** zum selben Sachverhalt bewertet. Gute Theorien sollen nicht nur empirisch bewährt sein, sondern auch einen möglichst hohen heuristischen Erklärungswert haben, sollen also die Wissenschaft zu interessanten weiteren Studien animieren, die Ableitung neuer Hypothesen erlauben und innovative Forschungsfragen anregen. Legt man auch dieses zweite Kriterium an, kann es durchaus rational im Sinne des Erkenntnisfortschritts sein, eine heuristisch fruchtbare Theorie trotz widersprechender Daten zunächst beizubehalten und weiter mit ihr zu arbeiten. Insbesondere wenn die Alternative nur darin besteht, eine Theorie zu akzeptieren, die zwar (z. B. in Folge von mehrfachen Exhaustionen) inzwischen besser bestätigt, aber gleichzeitig in ihrem Informationsgehalt (▶ Abschn. 2.2.4) deutlich reduziert ist. Um noch einmal Poppers Scheinwerfer-Metapher (▶ Abschn. 2.2.1) zu bemühen: Wenn mit einer vorliegenden Theorie ein leistungsstarker Scheinwerfer zur Verfügung steht, mit dem ein großer Gegenstandsbereich ausgeleuchtet werden kann, lohnt es in der Logik des raffinierten methodologischen Falsifikationismus, vorerst mit diesem weiterzuarbeiten und widersprüchliche empirische Befunde hinzunehmen, weil dies am Ende zu besseren Theorien führen kann als das Arbeiten mit einer kleinen Taschenlampe, die einen viel engeren Realitätsausschnitt beleuchtet.

Nicht nur präskriptiv scheint der raffinierte Falsifikationismus überzeugender, er ist auch deskriptiv eher mit bisherigem Erkenntnisfortschritt vereinbar: Nach Lakatos (1977) erklärt der raffinierte methodologische Falsifikationismus den wissenschaftlichen Erkenntnisprozess besser als der einfache Falsifikationismus, weil aus der Wissenschaftsgeschichte zahlreiche Beispiele dafür vorliegen, dass eine Theorie trotz vieler erfolgreicher Falsifikationen dennoch nicht verworfen, sondern aufrecht erhalten wurde und sich **im Laufe der Zeit durch Abwandlungen** in ihrem Informationsgehalt gesteigert und gleichzeitig auch empirisch bewährt hat. Es wird dabei der Kern der Theorie („hard core of theoretical assumptions") in einer Weise verändert, die nicht nur auf Exhaustion hinausläuft, sondern die Erklärungskraft der Theorie steigert. Gleichzeitig wird auch an den Hilfs- und Instrumententheorien, dem Schutzgürtel („protective belt of auxiliary hypotheses") der Kerntheorie, gearbeitet (z. B. Verbesserung von Operationalisierungen).

Es ist zu beachten, dass aus diesen und anderen in diesem Kapitel vorgestellten wissenschaftstheoretischen Überlegungen meist keine präzisen Anweisungen für die Forschungspraxis abzuleiten sind: So lässt sich eben nicht als genaue Entscheidungsregel angeben, für welche zusätzliche Menge an Erklärungskraft bei einer Theorie eine wie große Menge an widersprechenden Daten vorläufig noch akzeptabel ist. Wichtig ist vielmehr der Grundgedanke, Theorien nicht isoliert hinsichtlich ihrer empirischen Bewährung zu betrachten, sondern auch ihre Erklärungskraft einzubeziehen und zwar jeweils im Vergleich zu konkurrierenden Theorien – damit wird ein **Theorienpluralismus** als normal und nützlich für den Erkenntnisgewinn über einen bestimmten Sachverhalt angenommen. Dieser Theorienpluralismus ist in der empirischen Sozialforschung auch deswegen üblich, weil verschiedene Theorien über einen Sachverhalt bzw. ein Forschungsproblem teilweise unterschiedliche Ursachenfaktoren berücksichtigen und sich somit durchaus ergänzen können.

Zudem betont Lakatos (1977), dass Kerntheorien und die zugehörigen Hilfstheorien nicht singulär geprüft, sondern im Zuge von wissenschaftlichen **Forschungsprogrammen** über längere Zeiträume hinweg kontinuierlich bearbeitet und weiterentwickelt werden. Diese **dynamische Perspektive** ist bei der Bewertung von Theorien einzubeziehen, weil aus dieser Sicht aktuelle Falsifikationen wiederum zu relativieren sind: Einem innovativen neuen Theorieentwurf sollte man nicht sofort abverlangen, einem etablierten Forschungsprogramm hinsichtlich empirischer Bewährung überlegen zu sein. Vielmehr ist zu bewerten, ob und wie die neue Theorie die Forschung zu inspirieren verspricht und ob sie sich zukünftig zu einem erfolgreichen Forschungsprogramm entwickeln kann, das seine empirische Bewährung sowie seinen Informationsgehalt schrittweise steigert. Mit Lakatos (1977) kann also das Weiterarbeiten mit einer vorliegenden Theorie rational verteidigt werden, auch wenn falsifizierende Daten vorliegen. Freilich funktioniert auch dies nicht beliebig im Sinne einer Immunisierung von Theorien vor jeglicher Kritik, sondern muss die Scientific Community im konkreten Fall mit Argumenten überzeugen.

Raffinierter methodologischer Falsifikationismus – Während der klassische methodologische Falsifikationismus Theorien primär anhand ihrer empirischen Bewährung beurteilt, bezieht der raffinierte methodologische Falsifikationismus („refined methodolocial falsificationism“) als zweites Qualitätskriterium von Theorien ihre Erklärungskraft („explanatory power“) ein. Eine Theorie mit großer Erklärungskraft zeichnet sich dadurch aus, dass aus ihr zahlreiche interessante Hypothesen abzuleiten und innovative Ideen für die Forschung zu gewinnen sind. Sie kann den Erkenntnisgewinn stärker fördern als eine gut bewährte Theorie, die jedoch nur wenige Effekte erklären kann. Das Kriterium der Erklärungskraft einer Theorie ist relativ zu anderen Theorien zum selben Sachverhalt zu bewerten.

2.2.6 Entdeckungs-, Begründungs- und Verwendungszusammenhänge von Theorien

Der Erkenntnisprozess wird in die Bereiche der Entdeckung (Bildung von Theorien), der Begründung/Rechtfertigung (Prüfung von Theorien) und der Verwendung/Verwertung (praktische Anwendung von Theorien) eingeteilt (Alisch, 2005; Popper, 1934/1989, S. 6f.; Reichenbach, 1938). Das Augenmerk des Kritischen Rationalismus liegt auf der Frage, wie erfahrungswissenschaftliche Theorien anhand empirischer Daten zu widerlegen bzw. bei ausbleibender Widerlegung als vorläufig bestätigt bzw. bewährt anzunehmen sind. Diese Fragen betreffen den **Begründungszusammenhang** wissenschaftlicher Erkenntnisse. Das methodologische Prinzip ist dabei das deduktive Falsifikationsprinzip. Startpunkt sind bereits vorhandene Theorien über den Untersuchungsgegenstand.

Begründungszusammenhang von Theorien – Der Kritische Rationalismus konzentriert sich im Zusammenhang mit wissenschaftlichem Erkenntnisgewinn auf den Begründungszusammenhang. Mit dem Begründungs- oder Rechtfertigungszusammenhang („context of justification“) ist die Frage angesprochen, ob und wie eine Theorien intersubjektiv nachvollziehbar als ungültig (falsifiziert) oder als vorläufig bestätigt bzw. bewährt (bislang nicht falsifiziert) ausgewiesen werden kann. Die Prüfung von wissenschaftlichen Theorien soll laut Kritischem Rationalismus in der Forschung gemäß dem deduktiven Falsifikationsprinzip methodisch angeleitet in nachvollziehbarer Weise anhand von empirischen Daten erfolgen. Wobei gemäß raffiniertem methodologischem Falsifikationismus immer auch die Erklärungskraft der Theorie im Vergleich zu anderen Theorien ein Beurteilungskriterium ist.

Wie man zu den zu prüfenden Theorien kommen kann oder soll, wird im Kritischen Rationalismus nicht genau vorgegeben. Der **Entdeckungszusammenhang** ist laut Popper gekennzeichnet durch das „Erraten“ von Gesetzmäßigkeiten in der materialen und sozialen Realität; entsprechende Einfälle von Forschenden seien allenfalls psychologisch, nicht jedoch erkenntnislogisch erklärbar (Popper, 1934/1989, S. 7). Doch betrachtet man die wissenschaftstheoretischen Annahmen des Kritischen Rationalismus zum Erkenntnisfortschritt, so lassen sich verschiedene Empfehlungen für den Entdeckungszusammenhang ableiten: Da von einem kumulativen Erkenntnisprozess ausgegangen wird (schrittweise Annäherung an die Wahrheit durch wiederholten Versuch und Irrtum) ist es aus Sicht des Kritischen Rationalismus sinnvoll, mit bereits vorhandenen und vorläufig bestätigten Theorien weiter zu arbeiten. Sei es, dass man anstelle eigener Hypothesengenerierung ganz bewusst **etablierte Theorien** (also mutmaßlich gute „Scheinwerfer“; zur Scheinwerfertheorie der Erkenntnis ► Abschn. 2.2.1) nutzt, um

durch **Replikation** ihren Bewährungsgrad zu prüfen und zu sichern. Sei es, dass man vorliegende Theorien kritisch durchdenkt, modifiziert und die **überarbeitete Theorie selbst prüft** oder sie zunächst als **reine Theoriearbeit** (▶ Abschn. 6.3.1) der Wissenschaftsgemeinschaft zur Prüfung vorlegt. Mit dieser Vorgehensweise bewegt man sich innerhalb eines bestimmten Forschungsprogramms im Sinne von Lakatos (1977) und hilft bei der Weiterentwicklung einer Theorie.

Schließlich sind aber auch im Kritischen Rationalismus **gänzlich neue Theoriebildungen** wünschenswert, sei es, dass diese durch systematisches empirisches Arbeiten mittels Abduktion oder Induktion gewonnen werden (▶ Abschn. 2.2.1) oder durch anekdotische Alltagserfahrungen und Intuition zustande kommen. Hier sind der Kreativität prinzipiell keine Grenzen gesetzt. Allerdings müssen Theorien, die der Wissenschaftsgemeinschaft vorgeschlagen werden, unter anderem relevant, widerspruchsfrei sowie falsifizierbar sein (zu Merkmalen von Theorien ▶ Abschn. 2.2.10). Andernfalls würde die gemäß Kritischem Rationalismus sich notwendigerweise anschließende Arbeit der Theorieprüfung nicht lohnenswert erscheinen. Die Relevanz von Forschungsproblemen kann anhand unterschiedlicher wissenschaftsimmanenter sowie praxisbezogener und ethischer Kriterien begründet werden (▶ Abschn. 3.2).

Entdeckungszusammenhang von Theorien – Im Entdeckungszusammenhang („context of discovery") geht es darum, wie neue Theorien und Hypothesen von den Forschenden aufgestellt werden. Im Verständnis des Kritischen Rationalismus ist dies keine wissenschaftstheoretische, sondern eine psychologische Frage der Kreativität und Intuition. Eine logisch stringente Methode zur Produktion neuer wissenschaftlicher Ideen wird nicht für möglich gehalten, da gute Einfälle immer auch ein irrationales Moment haben. In der Forschungspraxis haben sich dennoch bestimmte Strategien und Techniken der Bildung neuer Theorien etabliert (z. B. Theoriearbeit, empirische Vorstudien).

Mit dem **Verwendungszusammenhang** schließlich ist zweierlei gemeint: zum einen die Nutzung wissenschaftlicher Theorien (allgemeiner: wissenschaftlicher Forschungsergebnisse) **im Kontext der Wissenschaft** selbst (z. B. als Anregung für weitere empirische Studien), zum anderen die Nutzung wissenschaftlicher Theorien und Befunde zur **Lösung praktischer Probleme** (Alisch, 2005, S. 58f.). Insbesondere anwendungswissenschaftliche bzw. technologischen Theorien, wie sie z. B. in der Evaluationsforschung eine wichtige Rolle spielen (▶ Abschn. 18.1.1) sind hier angesprochen, aber auch grundlagenwissenschaftliche Theorien können die Basis für praktische Entscheidungen oder die Konzeption praktischer Maßnahmen bilden. Aus erkenntnistheoretischer Perspektive kann der Verwendungszusammenhang nach dem Begründungszusammenhang eine weitere Bewährungsprobe für Theorien darstellen: Theorien, die **praktische Anwendbarkeit** („practicality") zeigen, auf die in der Praxis oft und offenbar erfolgreich zurückgegriffen wird, kann ein höherer Bewährungsgrad zugeschrieben werden (zur Verwertbarkeit als Qualitätskriterium erziehungswissenschaftlicher Theorien: Heid & Harteis, 2005; zur Kluft zwischen Theorie und Praxis: Sandelands, 1990).

Dabei ist jedoch zu beachten, dass im praktischen Verwendungszusammenhang außerhalb der Wissenschaft nicht die Forschenden über den Umgang mit der Theorie bestimmen, sondern vor allem die Praktiker, so dass Nicht-Anwendung kein Beleg für mangelnde Gültigkeit sein muss und erfolgreiche Anwendung oft gar nicht genügend empirisch dokumentiert wird. Im Verwendungszusammenhang stellen sich vor allem auch axiologische Fragen der ethischen Verantwortung empirischer Sozialforschung: Werden die wissenschaftlichen Theorien in der Praxis möglicherweise genutzt, um Ziele zu verfolgen, die unseren Wertvorstellungen widersprechen?

Verwendungszusammenhang von Theorien – Mit dem Verwendungszusammenhang („context of application") ist die Nutzung wissenschaftlicher Theorien und Forschungsergebnisse gemeint. Diese erfolgt teils innerhalb der Wissenschaft (z. B. vorliegende Theorien und Befunde werden als Grundlage weiterer wissenschaftlicher Studien herangezogen), teils aber auch außerhalb der Wissenschaft in diversen Praxisfeldern (z. B. eine psychologische Theorie wird zur Entwicklung einer psychotherapeutischen Intervention herangezogen; eine erziehungswissenschaftliche Theorie wird zur Gestaltung einer Schulreform herangezogen). Eine Theorie, die in der Praxis oft und offenbar erfolgreich genutzt wird, kann einen hohen praktischen Bewährungsgrad beanspruchen. Allerdings sind die Verhältnisse in der Praxis von vielen Faktoren beeinflusst, so dass praktische Verwertbarkeit keine umfassende, systematische Theorieprüfung darstellt (diese ist im Verständnis des Kritischen Rationalismus im Begründungszusammenhang zu leisten).

2.2.7 Ethische Verantwortung und Objektivität in der Forschung

Gemäß Kritischem Rationalismus dient Wissenschaft der Wahrheitssuche. Ziel der empirischen Sozialforschung ist es dementsprechend, einen Bestand an bewährten und informationshaltigen Theorien aufzubauen, mit denen die Gesetzmäßigkeiten der sozialen Wirklichkeit möglichst gut erklärt und weitere Forschungsaktivitäten angeregt werden können. Dabei wird für die wissenschaftliche Theorieprüfung **Objektivität im Begründungszusammenhang** des Erkenntnisgewinns verlangt. Damit ist gemeint, dass die Falsifikation einer Theorie in intersubjektiv nachvollziehbarer Weise anhand von Daten und Argumenten erfolgen soll, so dass das Ergebnis der Theorieprüfung von den Einstellungen, Werten und Vorurteilen der einzelnen Forschenden gegenüber dem Forschungsgegenstand unabhängig ist. Es wird – etwas missverständ-

lich – auch vom Prinzip der **Wertfreiheit** gesprochen, damit ist gemeint, dass die Bewertung des untersuchten Problems durch die Forschenden nicht ausschlaggebend für das Ergebnis sein darf (zum sog. **Werturteilsstreit** ▶ Abschn. 2.2.11). Das Prinzip der wertneutralen Forschung ist in der quantitativen empirischen Sozialforschung unter anderem festgelegt als Gütekriterium der Objektivität von Messinstrumenten sowie als Anforderung der Replizierbarkeit ganzer Studien und ihrer Befunde. Es zu beachten, dass auch eine so definierte „wertfreie" oder wertneutrale Forschung sich immer an den **Werten der Forschungs- und Wissenschaftsethik** orientiert (▶ Kap. 4).

Beispiel: Das Ergebnis der statistischen Überprüfung der Hypothese „Die Ausländerfeindlichkeit hat in den letzten zehn Jahren in Deutschland zugenommen" an einem bevölkerungsrepräsentativen Datensatz unter Verwendung eines standardisierten Messinstruments zur Ausländerfeindlichkeit muss unabhängig davon sein, wie die jeweiligen Forschenden eine Zunahme der Ausländerfeindlichkeit finden würden oder ob sie selbst als In- oder Ausländer zu klassifizieren wären. Dass die Theorieprüfung an einem Datensatz wertneutral immer zum selben Ergebnis führt, wird durch die Standardisierung und Transparenz des Vorgehens und die dadurch mögliche Überprüfung innerhalb der Scientific Community sichergestellt. Denn das Falsifikationsprinzip meint nicht nur, dass alle Forschenden selbstkritisch und ergebnisoffen ihre eigenen Theorien prüfen, sondern dass generell der Theoriebestand auch von anderen Forschungsgruppen kritisch geprüft wird. Die Veranlassung dazu ist vor allem dann gegeben, wenn im Sinne des Theorienpluralismus mehrere konkurrierende Theorien zum selben Sachverhalt existieren und vergleichende Theorieprüfungen durchführbar sind.

Dieser ausdrückliche Anspruch der **Objektivität** oder **Wertfreiheit bei der Prüfung von Hypothesen** gemäß wissenschaftstheoretischem Verständnis des Kritischen Rationalismus wird sehr häufig falsch verstanden. Nämlich immer dann, wenn nicht berücksichtigt wird, dass er sich speziell auf den Begründungszusammenhang bezieht, nicht auf den Entdeckungs- und auch nicht auf den Verwendungszusammenhang wissenschaftlicher Theorien.

Der Kritische Rationalismus geht selbstverständlich davon aus, dass alle Forschenden ein eigenes Weltbild, Urteile und Vorurteile, subjektive Erfahrungen, Werte und Glaubenssätze etc. mitbringen und diese nicht im Forschungslabor einfach ablegen können, um zu „objektiven Forschern" zu werden. Auch geht der Kritische Rationalismus in axiologischer Hinsicht nicht davon aus, dass Forschende im quantitativen Ansatz gesellschaftlichen Problemen grundsätzlich „wertfrei" gegenüberstehen oder die bestehenden gesellschaftlichen Verhältnisse kritiklos als „objektive Realität" hinnehmen (zur Kritik am Objektivitätsverständnis des Kritischen Rationalismus ▶ Abschn. 2.2.11). Tatsächlich spielen **Wertfragen** gemäß Kritischem Rationalismus im **Entdeckungszusammenhang** von Theorien eine wichtige Rolle: Sozialkritische Positionen können die Wahl bestimmter Forschungsthemen und auch die Formulierung konkreter Theorien mitbeeinflussen, in denen dann z. B. Mechanismen der Ausgrenzung, Diskriminierung oder Unterdrückung erklärt werden. Wissenschaftstheoretisch ist der Vorwurf, der Kritische Rationalismus fordere oder fördere durch seine normativen Vorgaben sozial verantwortungslose oder gleichgültige „neutrale Forschende" nicht zutreffend.

Axiologische Fragen nach der Bedeutung von Werten in der Wissenschaft und der Verantwortung der sozialwissenschaftlich Forschenden im Hinblick auf soziale Probleme sind – genau wie andere wissenschaftstheoretische Probleme – einerseits normativ zu diskutieren, andererseits aber auch auf die Wissenschaftspraxis zu beziehen. Hier zeigt die Forschungspraxis, dass quantitative sozialwissenschaftliche Forschung sich häufig kritisch mit bestehenden sozialen Verhältnissen und mit sozialen Problemen auseinandersetzt. In Bereichen der Anwendungsforschung, in denen es buchstäblich um Leben und Tod geht – etwa in der medizinischen Forschung oder der Evaluationsforschung zu Projekten der Armutsbekämpfung – gilt strengste empirisch-quantitative Forschung zum Nachweis von Kausalität heute als „Goldstandard" (Clemens & Demombynes, 2011). Denn niemandem ist geholfen, wenn wohlmeinende Forschende gemeinsam mit Betroffenen einseitig ein gut gemeintes Projekt zur Armutsbekämpfung in Afrika loben, das sich in der Praxis dann aber gar nicht objektiv als wirksam und nützlich für die Dorfgemeinschaften erweist. Wissenschaftliche Objektivität gemäß kritisch-rationalistischem Wissenschaftsverständnis ist bei der Theorieprüfung als Verpflichtung zu **methodisch strengem, intersubjektiv nachvollziehbarem wissenschaftlichen Arbeiten gemäß definierten Qualitätskriterien** (▶ Kap. 3), in der Anwendungsforschung, in der nicht selten Interessensgruppen bestimmte Wunschergebnisse erwarten, zudem noch als **wissenschaftliche Unabhängigkeit** oder **Überparteilichkeit** zu verstehen (▶ Abschn. 18.1.4). Sich z. B. humanistischen Werten der Gerechtigkeit verbunden zu fühlen, darf und soll die **Auswahl von sozial- und humanwissenschaftlichen Forschungsprojekten** bestimmen (z. B. Evaluation von Projekten der Entwicklungszusammenarbeit versus Evaluation von Konsumprodukten). Es darf durch die eigenen Wertvorstellungen oder durch Parteilichkeit für Betroffene aber nicht vorbestimmt sein, ob man eine bestimmte Theorie falsifiziert oder nicht (z. B. Aussage über

die Wirksamkeit oder Nicht-Wirksamkeit des evaluierten Armutsbekämpfungsprojekts hinsichtlich der damit verfolgten Ziele). Denn damit wäre der Erkenntnisgewinn sabotiert.

In welcher Weise wissenschaftliche Erkenntnisse in der Praxis umgesetzt werden, hängt von vielen, vor allem auch wirtschaftlichen und politischen Faktoren ab. Auf den **Verwendungs- bzw. Verwertungszusammenhang** ihrer wissenschaftlichen Erkenntnisse haben Forschende deswegen in der Regel wenig Einfluss. Eine empirisch vorerst bewährte Theorie, welche die Entstehung von Ausländerfeindlichkeit erklärt, kann prinzipiell genutzt werden für Praxismaßmaßnahmen, mit denen Ausländerfeindlichkeit geschürt oder abgebaut wird. Hier können Forschende insofern gemäß ihren Wertvorstellungen aktiv werden, dass sie intendierte Nutzungen ihrer Forschungsergebnisse aktiv fördern (z. B. durch Öffentlichkeitsarbeit; ▶ Abschn. 13.4). Derartige Aktivitäten, mit denen sich Forschende als ethisch und politisch verantwortliche Bürgerinnen und Bürger engagieren, finden jedoch aus Sicht des Kritischen Rationalismus jenseits der eigentlichen wissenschaftlichen Tätigkeit (d. h. Wahrheitssuche durch kritische Prüfung von Theorien) statt.

Es sind vor allem **humanistische Werte** der Freiheit, Gleichheit, Demokratie und Menschenrechte, die in der Wissenschaftsgemeinschaft als konsensfähig gelten und im Entdeckungs- und Verwendungszusammenhang diskutiert werden. Während die **Forschungsethik** sich auf die Rechte der Untersuchungsteilnehmenden und die Vermeidung von Beeinträchtigungen für sie durch den Forschungsprozess selbst konzentriert (▶ Abschn. 4.1), betreffen axiologische Fragen der **ethischen Verantwortung von Wissenschaft** generell vor allem die Wahl der Forschungsthemen und somit den Aspekt ihrer Relevanz (zur Wissenschaftsethik ▶ Abschn. 4.2). So werden im Allgemeinen Forschungsprojekte als relevanter eingeschätzt, deren Erkenntnisse dazu genutzt werden können, humanistische Werte zu verwirklichen. Dabei ist zu beachten, dass gerade in der Anwendungsforschung (wie z. B. der Evaluationsforschung) die ethische Bewertung von Maßnahmen auch sehr kontrovers sein kann, wenn Anspruchsgruppen mit unterschiedlichen Interessen, Werten oder kulturellen Hintergründen involviert sind (▶ Abschn. 18.1.4).

Nicht zuletzt geht der Kritische Rationalismus auch davon aus, dass ein Wissenschaftsverständnis, das jeden absoluten Wahrheitsanspruch verneint und sich dafür der fortwährenden gemeinsamen Wahrheitssuche und dem kritischen Diskurs mit rationalen Argumenten verschreibt, den **Werten einer freiheitlich-demokratischen Gesellschaft** entspricht und vor Denkverboten sowie der Etablierung dogmatischer Heilslehren schützen kann.

Wertneutralität im Begründungszusammenhang – Die Prüfung von wissenschaftlichen Theorien soll gemäß Kritischem Rationalismus methodisch angeleitet anhand von Daten erfolgen und zwar im Wissenschaftssystem ergebnisoffen und wertfrei/wertneutral bzw. objektiv hinsichtlich des inhaltlichen Ergebnisses. Dies wird vor allem durch die Strukturierung, Standardisierung und detaillierte Offenlegung des Vorgehens sowie die kritische Prüfung durch Peers (d. h. Forschende im jeweiligen inhaltlichen Gebiet) sichergestellt. Mit Objektivität ist aber nicht gemeint, dass Forschende der sozialen Wirklichkeit gegenüber generell eine „neutrale" Haltung einnehmen sollen oder können oder dass sie gegenüber ihren eigenen Theorien eine „neutrale" Haltung haben sollen oder können. Vielmehr gibt der Kritische Rationalismus vor, dass die Wertvorstellungen der Forschenden hinsichtlich des Forschungsproblems im Begründungszusammenhang von Theorien keine Rolle spielen und das Ergebnis der Theorieprüfung nicht vorbestimmten dürfen. Als wichtigstes Korrektiv gilt dabei die kritische Prüfung und Replikation durch Fachkollegen.

Werte im Entdeckungs- und Verwendungszusammenhang – Wertvorstellungen der Forschenden dürfen und sollen darüber mitentscheiden, welchen Forschungsproblemen sie sich zwecks Theoriebildung zuwenden oder nicht zuwenden (Entdeckungszusammenhang) und inwiefern sie eine bestimmte Anwendung ihrer Forschungsergebnisse und ihrer Theorien aktiv fördern oder nicht fördern (Verwendungszusammenhang).

2.2.8 Deduktiv-nomologische Erklärungen der Realität

Wissenschaftlicher Erkenntnisgewinn vollzieht sich gemäß Kritischem Rationalismus in den Natur- wie den Sozialwissenschaften durch das Überprüfen von Theorien, die sich auf Ursache-Wirkungs-Relationen (Wenn-Dann-Relationen) in der Wirklichkeit beziehen. Speziell wird hierbei das so genannte **deduktiv-nomologische Erklärungsmodell** (kurz: **DN-Modell**) angewendet, das auch als **Hempel-Oppenheim-Schema** (kurz: **HO-Schema**) bezeichnet wird (Hempel & Oppenheim, 1948).

Das DN-Modell ist ein **aussagenlogisches Modell**. Es verknüpft eine allgemeine Gesetzesaussage **(nomologische Aussage)** mit einer logisch daraus abgeleiteten **(deduzierten)** empirisch prüfbaren Hypothese, die sich auf einen konkreten (raum-zeitlich spezifizierten) sozialen Kontext bezieht. Je nachdem, wie die Prüfung der empirischen Hypothese anhand der erhobenen Daten ausgeht, wird auf die vorläufige Gültigkeit oder Nicht-Gültigkeit des Gesetzes (bzw. der wissenschaftlichen Kerntheorie, zu welcher das Gesetz gehört) zurückgeschlossen. Sofern die Forschungshypothese logisch korrekt aus dem Gesetz abgeleitet wurde, gilt: Wenn die Forschungshypothese empirisch *nicht* zutrifft, muss (sofern die Randbedingungen und die Hilfstheorien zutreffen) im Rückschluss die Korrektheit des zu prüfenden Gesetzes bzw. der Theorie in Zweifel gezogen werden. Dieser Schluss wird in der Logik allgemein auch als **modus tollens** bezeichnet: Wenn aus A logisch immer B

Tabelle 2.1 Die Theorieprüfung im Kritischen Rationalismus basiert auf dem deduktiv-nomologischen Erklärungsmodell bestehend aus Explanans und Explanandum bzw. zu prüfender Gesetzesaussage und daraus logisch abgeleiteter empirischer Hypothese (hier illustriert anhand einer Studie zur symbolischen Selbstergänzung)

Konzept	Erläuterung	Beispiel einer Studie zur symbolischen Selbstergänzung
Explanans (lat. das Erklärende)	Eine **allgemeine Gesetzesaussage** (nomologische Aussage, meist kausale Aussage) liegt als Teil einer wissenschaftlichen Kerntheorie vor, etwa ein kausales Gesetz in der Form „Wenn A und B, dann C“.	**Theorie der symbolischen Selbstergänzung** („self-completion theory“; Wicklund & Gollwitzer, 1982): Wenn Menschen eine bestimmte Identität (z. B. akademische, sportliche, politische Identität) wichtig ist und wenn sie hinsichtlich dieser Identität Defizite bei sich wahrnehmen, dann kompensieren sie dies durch das Präsentieren von identitätsbekräftigenden Symbolen (z. B. beschönigende Selbstbeschreibungen, Besitz von Statusobjekten).
	Aus dem allgemeingültigen Gesetz wird eine **empirische Hypothese** für einen konkreten sozialen Kontext abgeleitet. Die in der **Wenn-Komponente** genannten Randbedingungen A und B (Ursachen, d. h. bestimmte Ausprägungen der unabhängigen Variablen) sind in dem betrachteten Kontext erfüllt (dies muss **empirisch beobachtbar** sein).	**Wenn** Forschenden ihre akademische Identität wichtig ist und wenn sie bislang nur vergleichsweise wenige akademische Leistungen vorzuweisen haben (z. B. wenige wissenschaftliche Publikationen) …
Explanandum (lat. das zu Erklärende)	Das mit der **Dann-Komponente** zu erklärende Phänomen C (Wirkung, d. h. bestimmte Ausprägungen der abhängigen Variablen) liegt in dem betrachteten Kontext vor (dies muss **empirisch beobachtbar** sein).	… **dann** greifen sie stärker auf selbstergänzende Symbole zurück (z. B. Nennung ihrer akademischen Titel und Positionen in der Signatur ihrer E-Mails oder auf ihrer Webseite (vgl. Harmon-Jones, Schmeichel, & Harmon-Jones, 2009).

folgt und wenn B nicht zutrifft, dann trifft auch A nicht zu (Popper, 1934/1989, S. 2, 16).

Zusammengefasst: Um eine allgemeine Gesetzesaussage bzw. eine Theorie gemäß Kritischem Rationalismus einem Falsifikationsversuch zu unterziehen, muss man aus ihr eine konkrete, empirisch prüfbare Hypothese logisch stringent ableiten und anhand von Daten zeigen, ob die Hypothese als vorläufig bestätigt oder widerlegt gelten kann. Struktur und Elemente des deduktiv-nomologischen Erklärungsmodells sind in Tab. 2.1 an einem Beispiel aus der psychologischen Fachliteratur dargestellt. Die zu prüfende theoretische Gesetzesaussage (typischerweise Wenn-Dann-Aussage) und die Randbedingungen werden zusammen auch als **Explanans** bezeichnet, das zu erklärende Phänomen dagegen ist das **Explanandum** (Kromrey, 2009, S. 80; Popper, 1962; Schnell et al., 2008, S. 58).

Aus jeder Gesetzesaussage lassen sich prinzipiell sehr **viele Hypothesen ableiten**, in dem hier betrachteten Beispiel zur Theorie der symbolischen Selbstergänzung (Tab. 2.1) etwa Hypothesen zu unterschiedlichen Typen von Identitäten bei unterschiedlichen Personengruppen in unterschiedlichen Kulturkreisen. Die Überzeugungskraft eines Falsifikationsversuchs hängt also auch davon ab, wie plausibel und relevant die abgeleitete Hypothese ist und inwiefern davon auszugehen ist, dass ihre empirische Prüfung tatsächlich auch zu hypothesenkonträren Ergebnissen führen kann.

Bei der Anwendung des deduktiv-nomologischen Erklärungsmodells in der empirischen Forschung ist zu beachten, dass der logische Schluss des modus tollens **deterministische Gesetze und Hypothesen** voraussetzt (die durch einzelne Gegenevidenzen klar widerlegbar wären). In der empirischen Forschung – das betrifft die Natur- und Sozialwissenschaften gleichermaßen – haben wir es jedoch mit **probabilistischen (wahrscheinlichkeitstheoretischen) Gesetzen bzw. Theorien** zu tun. Probabilistische Erklärungen erheben nicht den Anspruch, zwingend für jeden Einzelfall zu gelten, sondern lediglich mit bestimmter Wahrscheinlichkeit zuzutreffen. Wahrscheinlichkeitsaussagen sind somit durch hypothesenkonträre Fälle nicht zu widerlegen (bzw. zu falsifizieren), da Ausnahmen von der Regel ja zulässig sind. Trotzdem lassen sich laut Popper logische Beziehungen herstellen, da eine theoretisch vorhergesagte Wirkung unterschiedlichen empirischen Wahrscheinlichkeiten (Auftretenshäufigkeiten) mehr oder weniger gut entspricht (Popper, 1934/1989, S. 146). Letztlich argumentiert Popper, dass sich die Wissenschaftsgemeinschaft sinnvollerweise darauf einigen kann, **Falsifikationskriterien einzuführen** in der Weise, dass Wahrscheinlichkeitsgrenzen festgelegt werden, so dass man z. B. extrem geringe Auftretenshäufigkeiten faktisch als Falsifikation einer Hypothese auffassen kann, welche das Auftreten des Effekts postuliert hatte (Popper, 1934/1989, S. 152ff.).

Popper erläutert die Hypothesenprüfung am Beispiel von probabilistischen Hypothesen in der Physik. Dass wir auch in den Sozialwissenschaften mit probabilistischen und nicht mit deterministischen Erklärungen arbeiten, ist

umso plausibler und mit der Individualität der Einzelfälle zu begründen: Eine sozialwissenschaftliche Theorie kann in der Regel nicht alle Einflussfaktoren auf menschliches Erleben und Verhalten im Zusammenhang mit dem untersuchten sozialen Phänomen vollständig aufklären. Indem sie einige wichtige Ursachenfaktoren betrachtet, andere dagegen vernachlässigt, kann sie das Phänomen nur zum Teil erklären und muss abweichende Fälle hinnehmen. Zudem wird auch argumentiert, dass menschliches Erleben und Verhalten möglicherweise gar nicht vollständig erklärbar ist, sondern allein schon die Wirksamkeit des freien Willens zu individuellen Abweichungen von deterministischen Gesetzmäßigkeiten führen würde. Dagegen wird wiederum eingewandt, dass auch vermeintlich freie Willensentscheidungen möglicherweise nach bestimmten Gesetzmäßigkeiten entstehen.

Festzuhalten bleibt, dass die Theorien und daraus abgeleiteten Hypothesen in den Sozial- und Humanwissenschaften nicht deterministisch (d. h. ausnahmslos für jeden Einzelfall gültig), sondern probabilistisch (d. h. für einen großen Teil der Fälle gültig) sind. Wenn aber die Theorie „Rauchen verursacht Lungenkrebs“ nicht deterministisch („Alle Raucher entwickeln Lungenkrebs“ oder „Wenn eine Person raucht, dann bekommt sie immer Lungenkrebs“), sondern probabilistisch gemeint ist („Raucher entwickeln mit höherer Wahrscheinlichkeit Lungenkrebs als vergleichbare Nichtraucher“), lässt sie sich auch nicht durch ein einzelnes Gegenbeispiel einer rauchenden Person ohne Lungenkrebs widerlegen. Vielmehr ist die **probabilistische Hypothese** zu prüfen, dass Raucher mit höherer Wahrscheinlichkeit an Lungenkrebs erkranken als vergleichbare Nichtraucher. Bei der Prüfung probabilistischer Hypothesen wird die Übereinstimmung zwischen Daten und Hypothese in einem wahrscheinlichkeitstheoretischen Modell bestimmt und über einen statistischen Signifikanztest geprüft, wobei man als Falsifikationskriterium das per Konvention festgelegte Signifikanzniveau verwendet (▶ Abschn. 2.2.9, ▶ Abschn. 12.5.1).

Zusammengefasst: Im Kritischen Rationalismus erfolgt die Hypothesenprüfung mittels Falsifikation anhand des deduktiv-nomologischen Erklärungsmodells. Da wir mit wahrscheinlichkeitstheoretischen Hypothesen arbeiten, die durch hypothesenkonforme Einzelfälle weder verifiziert, noch durch hypothesenkonträre Einzelfälle falsifiziert werden können, muss zur Entscheidung ein sinnvolles **wahrscheinlichkeitstheoretisches Falsifikationskriterium** in der Weise eingeführt werden, dass hohe Auftretenswahrscheinlichkeiten für den Effekt und sehr geringe Auftretenswahrscheinlichkeiten gegen den Effekt sprechend ausgelegt werden (dies ist eine methodologische Entscheidung, die im **statistischen Signifikanztest** formalisiert ist). Ohnehin wird kein naiver Falsifikationismus verfolgt, sondern im Sinne des methodologischen Falsifikationismus das Ergebnis der Hypothesenprüfung stets kritisch in Hinsicht auf die Aussagekraft der Daten und verwendeten Methoden hinterfragt. Insbesondere muss die Berechtigung des **Rückschlusses vom Ergebnis der statistischen Hypothesenprüfung auf die zugrundeliegende Theorie** diskutiert werden. Dieser Rückschluss ist nur gerechtfertigt, wenn wir die Hilfs- und Instrumententheorien als gültig akzeptieren, auf denen die Methodenentscheidungen in den einzelnen Phasen des empirischen Forschungsprozesses basieren (◘ Tab. 2.2).

Darüber hinaus ist bei der Prüfung einer **Theorie mit kausalem Erklärungsanspruch** bei der Ergebnisinterpretation immer auch kritisch zu hinterfragen, ob die erhobenen Daten tatsächlich etwas über den Verursachungsmechanismus (Kausalität) aussagen können oder ob sie nicht vielmehr nur Zusammenhänge belegen, deren ursächliches Zustandekommen offen bleibt (dies ist vor allem eine Frage des Untersuchungsdesigns einer hypothesenprüfenden Studie, wobei experimentelle Designs Kausalität am besten nachweisen können; ▶ Abschn. 7.6.1).

Probabilistische Theorien und Hypothesen – In der empirischen Sozial- und Humanforschung arbeitet man nicht mit deterministischen Theorien und Hypothesen, die den Anspruch erheben, ausnahmslos jeden Einzelfall zu erklären („Alle Raucher entwickeln Lungenkrebs“). Stattdessen setzt man eine (bislang nicht erklärte oder grundsätzlich nicht erklärbare) Individualität der Einzelfälle voraus und erhebt mit probabilistischen (wahrscheinlichkeitstheoretischen) Theorien und Hypothesen („probabilistic theories/hypotheses“) lediglich den Anspruch, die untersuchten Phänomene der Tendenz nach bzw. auf Aggregatebene zu erklären („Raucher entwickeln mit höherer Wahrscheinlichkeit Lungenkrebs als Nichtraucher“).

Prüfung probabilistischer Hypothesen durch statistische Signifikanztests – Da probabilistische Theorien und Hypothesen nicht durch widersprechende Einzelfälle falsifizierbar sind, müssen begründete Falsifikationskriterien eingeführt werden. Zur Prüfung von Forschungshypothesen durch Daten dient im quantitativen Paradigma meist der statistische Signifikanztest, der Stichprobenergebnisse wahrscheinlichkeitstheoretisch bewertet und konventionell mit einem Signifikanzniveau von $\alpha = 5\,\%$ (oder $1\,\%$) als Falsifikationskriterium operiert. Das Ergebnis eines statistischen Hypothesentests ist deswegen immer mit Unsicherheit behaftet: Statistisch liefert bei einem Signifikanzniveau von $5\,\%$ dementsprechend von 20 Signifikanztests ein Signifikanztest ein falsch-positives Ergebnis (d. h. ein statistisch signifikantes Ergebnis, obwohl die Forschungshypothese nicht zutrifft).

Interpretation der Ergebnisse von statistischen Signifikanztests – Der Rückschluss vom Ergebnis eines statistischen Hypothesentests (signifikantes versus nicht-signifikantes Ergebnis) auf die zu prüfende Theorie ist definitionsgemäß mit einem statistischen Fehler behaftet. Er ist darüber hinaus auf theoretischer Ebene unsicher und muss kritisch diskutiert werden, da die Aussagekraft der Daten von der Gültigkeit diverser Hilfs- und Instrumententheorien abhängt, auf denen die Methodenentscheidungen im Forschungsprozess (z. B. Operationalisierung, Stichprobenauswahl) basieren.

Zur Verdeutlichung des deduktiv-nomologischen Erklärungsmodells wurde als Studienbeispiel die Überprüfung

Tabelle 2.2 Empirischer Forschungsprozess zur Gewinnung der Daten und damit verbundene explizite und implizite Hilfs- und Instrumententheorien in einer Studie zur symbolischen Selbstergänzung. (Beispiel in Anlehnung an die Studien von Harmon-Jones et al., 2009. Mit freundlicher Genehmigung von John Wiley and Sons)

Phase im Forschungsprozess	Methodische Entscheidungen in einer Studie zur symbolischen Selbstergänzung	Hilfs- bzw. Instrumententheorien in einer Studie zur symbolischen Selbstergänzung
Untersuchungsdesign (► Kap. 7)	Das symbolische Selbstergänzungsverhalten von Forschenden mit Professorenstatus wird im Rahmen einer nicht-experimentellen Querschnittstudie mittels Dokumentenanalyse untersucht.	Es wird angenommen, dass bei Forschenden mit Professorenstatus generell eine hohe Relevanz der akademischen Identität vorliegt und somit die Kerntheorie der symbolischen Selbstergänzung anwendbar ist.
Operationalisierung und Messung (► Kap. 8)	Die Messung der wissenschaftlichen Produktivität (unabhängige Variable bzw. Prädiktor) erfolgt anhand der Anzahl der Publikationen sowie der Anzahl der Zitationen der jeweiligen Untersuchungsperson in der Datenbank „Web of Science“. Dabei werden die Publikations- und die Zitationszahl jeweils an der Anzahl der nach der Promotion verstrichenen Jahre relativiert. Beide Werte werden z-standardisiert und additiv zu einem Index der akademischen Produktivität zusammengefasst. Die Messung des symbolischen Selbstergänzungsverhaltens (abhängige Variable bzw. Kriterium) erfolgt anhand der E-Mail-Signaturen der Untersuchungspersonen: Für die Nennung von 1.) Doktortitel, 2.) akademischer Position (z. B. Professorenstatus) und 3.) weiteren akademischen Meriten (z. B. Wissenschaftspreis) wird jeweils ein Punkt vergeben, so dass der Grad der symbolischen Selbstergänzung in der E-Mail-Signatur in dem gebildeten Index zwischen 0 und 3 variiert.	Es wird inhaltlich angenommen, dass Publikations- und Zitationszahl tatsächlich Indikatoren des Konstrukts wissenschaftliche Produktivität sind. Es wird messtheoretisch angenommen, dass der gebildete Index die Variabilität der akademischen Produktivität auf Kardinalskalen-Niveau abbildet (zum Skalenniveau ► Abschn. 8.4). Es wird inhaltlich angenommen, dass die Titel-Nennung in der E-Mail-Signatur tatsächlich Ausdruck von symbolischem Selbstergänzungsverhalten ist. Es wird messtheoretisch angenommen, dass der gebildete Index die Variabilität im symbolischen Selbstergänzungsverhalten auf Kardinalskalen-Niveau abbildet (zum Skalenniveau ► Abschn. 8.4). All diese Annahmen betreffen das Korrespondenzproblem (► Abschn. 2.2.3).
Stichprobenziehung (► Kap. 9)	Es wird anhand der teilöffentlichen E-Mail-Kommunikation in einem wissenschaftlichen Institut sowie zu einer Konferenz eine Gelegenheitsstichprobe von $n = 44$ E-Mails von Hochschullehrenden in den USA gezogen.	Es wird davon ausgegangen, dass der Stichprobentyp der Gelegenheitsstichprobe (► Abschn. 9.3.1) zwar nicht für die Population aller E-Mail-schreibenden Hochschullehrenden in den USA repräsentativ ist, aber die Variabilität der Merkmale ausreichend widerspiegelt, um sinnvoll ihre Kovariation bzw. Korrelation zu bestimmen. Zudem wird theoretisch angenommen, dass der Stichprobenumfang genügend Teststärke (► Abschn. 14.1) für eine aussagekräftige inferenzstatistische Datenanalyse (► Kap. 12) liefert.
Datenerhebung (► Kap. 10)	Es erfolgt eine non-reaktive Datenerhebung durch Dokumentenanalyse (► Abschn. 10.6): Die gesammelten E-Mails werden quantitativ-inhaltsanalytisch hinsichtlich der Nennung der akademischen Titel und Positionen in der Signatur ausgewertet und die wissenschaftliche Produktivität der Hochschullehrenden wird über die Literaturdatenbank „Web of Science“ recherchiert. Messwerte werden gemäß der vorher festgelegten Operationalisierung zugewiesen.	Es wird theoretisch angenommen, dass die inhaltsanalytische Auswertung der E-Mails und die Datenbankabfragen unverzerrt und messfehlerfrei erfolgen. Besondere Vorkehrungen (z. B. Einsatz von „blinden“ Kodierern, welche die Forschungshypothese der Studie nicht kennen) oder Kontrollen (Bestimmung der Inter-Kodierer-Übereinstimmung; ► Abschn. 10.6.3) finden nicht statt. Diese Annahmen betreffen das Basissatzproblem (► Abschn. 2.2.3).
Datenbereinigung (► Kap. 11)	Aufgrund der non-reaktiven Datenerhebung mittels Dokumentenanalyse treten keine unvollständigen Datensätze (z. B. durch Antwortverweigerung) oder unplausiblen Werte (z. B. durch Scherzantworten) auf, so dass in dieser Hinsicht keine Datenbereinigung notwendig ist.	Theoretische Annahmen darüber, welche der untersuchten Fälle von der Datenauswertung ausgeschlossen werden müssen (z. B. aufgrund einer zu hohen Zahl an fehlenden oder unplausiblen Werten), entfallen, da diesbzgl. keine Datenbereinigung notwendig ist. Entsprechende Annahmen würden das Basissatzproblem (► Abschn. 2.2.3) betreffen.

der Theorie der symbolischen Selbstergänzung herangezogen (Tab. 2.1). Dazu wurde aus der Theorie der symbolischen Selbstergänzung eine empirisch prüfbare Hypothese abgeleitet, die sich auf eine spezielle Form der symbolischen Selbstergänzung im Zusammenhang mit der akademischen Identität von Wissenschaftler/innen bezieht (Harmon-Jones, Schmeichel, & Harmon-Jones, 2009). Im Folgenden soll nun genauer betrachtet werden, wie die zur Hypothesenprüfung notwendigen Daten gewonnen werden, auf welche Hilfs- und Instrumententheorien dabei im Zuge des Forschungsprozesses zurückgegriffen wird, und wie anhand der gewonnenen quantitative Daten dann der statistische Hypothesentest vollzogen und vor dem Hintergrund der bis hierher dargestellten wissenschaftstheoretischen Annahmen des Kritischen Rationalismus korrekt zu interpretieren ist.

2.2.9 Datengewinnung und statistische Hypothesenprüfung

Um eine Hypothese anhand von Daten prüfen zu können, müssen im Zuge eines geordneten und dokumentierten **empirischen Forschungsprozesses** zunächst Daten gewonnen werden, die geeignet sind, um sinnvoll etwas über die infrage stehende Hypothese (und auf dieser Basis dann auch über die Theorie) aussagen zu können. Wir haben schon festgehalten, dass der Forschungsprozess in der empirisch-quantitativen Sozialforschung in einzelne Phasen eingeteilt werden kann (▶ Abschn. 1.3), in denen jeweils bestimmte **methodische Entscheidungen** zu treffen sind. Diese Entscheidungen beeinflussen den **Informationsgehalt der gewonnenen Daten**. Daten sind im Verständnis des Kritischen Rationalismus keine reinen Faktenbeobachtungen, sondern gemäß Basissatz- und Korrespondenzproblem (▶ Abschn. 2.2.3) theoriegeladene Aussagen, die vom Vorverständnis abhängen und von der Wissenschaftsgemeinschaft kritisch zu bewerten sind. Die Theoriehaltigkeit der Daten bezieht sich dabei nicht nur auf die inhaltliche Kerntheorie (hier im Beispiel: Theorie der symbolischen Selbstergänzung), sondern vor allem auch auf die methodischen **Hilfs- bzw. Instrumententheorien**. Das sind die theoretischen Annahmen, die im Forschungsprozess zugrunde gelegt werden, um bestimmte Methodenentscheidungen zu rechtfertigen. Es ist wichtig, sich diese oft impliziten (d. h. nicht ausdrücklich ausformulierten) Hilfs- und Instrumententheorien bewusst zu machen (Tab. 2.2).

Laut Hypothese in unserem Beispiel zur symbolischen Selbstergänzung (Tab. 2.1) soll geprüft werden, ob akademisch unproduktivere Forschende in stärkerem Maße und akademisch produktivere Forschende in geringerem Maße im Sinne symbolischen Selbstergänzungsverhaltens in ihren E-Mail-Signaturen auf ihre akademischen Titel hinweisen, sofern ihnen ihre akademische Identität wichtig ist. Als **Untersuchungsdesign** (zu unterschiedlichen Untersuchungsdesigns ▶ Kap. 7) wird eine **nicht-experimentelle Querschnittstudie** gewählt. Symbolisches Selbstergänzungsverhalten wird also nicht experimentell manipuliert, sondern im natürlichen Umfeld beobachtet: Wie stark tritt es bei mehr oder minder produktiven Forschenden auf? Dabei beschränkt man sich auf einen einzigen Messzeitpunkt. Zudem wird die Studie als **non-reaktive Dokumentenanalyse** angelegt, d. h. die Datenerhebung erfolgt ohne direkten Kontakt zu den Untersuchungsobjekten (so dass z. B. Verzerrungen durch sozial erwünschtes Antworten oder Gedächtniseffekte entfallen). Stattdessen wird das symbolische Selbstergänzungsverhalten der Untersuchungsobjekte direkt ihren E-Mail-Texten und ihre akademische Produktivität der wissenschaftlichen Literaturdatenbank „Web of Science“ (zu wissenschaftlichen Literaturdatenbanken ▶ Abschn. 6.1.2) entnommen. Die Untersuchung beschränkt sich aus forschungsökonomischen Gründen auf die USA als Heimatland des Untersuchungsteams.

Alle methodischen Entscheidungen im Forschungsprozess sind mit spezifischen Hilfs- und Instrumententheorien verknüpft und beeinflussen damit den **Theoriegehalt und die Interpretierbarkeit der Daten**. Die Methodenentscheidungen müssen deswegen in der Publikation der Studie im Sinne der Präsentationsqualität als Gütekriterium (▶ Abschn. 3.1.2) detailliert offen gelegt und von der Wissenschaftsgemeinschaft für inhaltlich gerechtfertigt gehalten werden. Nur wenn die Hilfs- und Instrumententheorien anerkannt werden, sind die resultierenden Daten aussagekräftig, um über die Hypothese (und auf dieser Basis dann auch über die Theorie) entscheiden zu können.

Die Entscheidung über die Hypothese findet in der quantitativen Sozialforschung typischerweise anhand eines **statistischen Hypothesentests** statt, meist eines klassischen Nullhypothesen-Signifikanztests (zur Logik des statistischen Signifikanztests ▶ Abschn. 12.5.1). Der grobe Ablauf der statistischen Hypothesenprüfung und die Ergebnisinterpretation sind Tab. 2.3 zu entnehmen.

Sozialwissenschaftliche Hypothesen sind – wie bereits ausgeführt – **Wahrscheinlichkeitsaussagen** (probabilistische Aussagen), die sich durch konträre Einzelfälle prinzipiell nicht widerlegen (falsifizieren) lassen. Sozialwissenschaftliche Hypothesen machen zudem verallgemeinernde Aussagen über Populationen, die in der Regel nicht vollständig, sondern nur ausschnitthaft (anhand von Stichproben) untersucht werden können, so dass auch eine Verifikation der Hypothese nicht möglich ist. Man steht also – wie bereits angesprochen – vor dem Dilemma, anhand von empirischen Daten probabilistische Hypothesen

Tabelle 2.3 Konfrontation der empirischen Hypothese mit den Daten durch statistische Hypothesenprüfung mittels klassischem Signifikanztest bei einer Studie zum symbolischen Selbstergänzungsverhalten. (Beispiel in Anlehnung an Harmon-Jones et al., 2009. Mit freundlicher Genehmigung von John Wiley and Sons)

Struktur des statistischen Hypothesentests	Beispiel: Überprüfung einer Hypothese zum symbolischen Selbstergänzungsverhalten
Inhaltliche Forschungshypothese auf der Basis der Kerntheorie	Wenn Forschenden ihre akademische Identität wichtig ist und wenn sie im Vergleich zu anderen Forschenden weniger akademische Leistungen erreicht haben (z. B. weniger Publikationen), dann greifen sie stärker auf selbstergänzende Symbole zurück (z. B. Nennung ihrer akademischen Titel und Positionen in der Signatur ihrer E-Mails).
Statistisches Hypothesenpaar	**Forschungs-/Alternativhypothese**: Je weniger Publikationen Forschende haben, umso stärker greifen sie in ihren E-Mail-Signaturen auf die Nennung von akademischen Titeln und Positionen zurück. Oder anders formuliert: Es besteht bei Forschenden ein negativer Zusammenhang zwischen der Anzahl der Publikationen einerseits und der Anzahl der genannten Titel und Positionen andererseits. Der Korrelationskoeffizient ρ (rho) in der Population hat ein negatives Vorzeichen: $\boldsymbol{H_1: \rho < 0}$ **Nullhypothese**, die das Gegenteil der Forschungshypothese behauptet bzw. den postulierten Effekt negiert: Es besteht bei Forschenden kein Zusammenhang oder ein positiver Zusammenhang zwischen der Anzahl der Publikationen und der Anzahl der genannten Titel und Positionen: $\boldsymbol{H_0: \rho \geq 0}$
Ergebnis der statistischen Hypothesenprüfung	Stichprobenumfang $n = 44$ Bivariater Korrelationskoeffizient in der Stichprobe: $\boldsymbol{r = -.29}$ Per bivariatem Korrelationstest (► Abschn. 12.5.2) berechnete Irrtumswahrscheinlichkeit p (einseitig): $\boldsymbol{p = .03}$
Interpretation des statistischen Ergebnisses im Hinblick auf die **Falsifikation oder vorläufige Bestätigung (Nicht-Falsifikation) der Hypothese**	Es zeigte sich hypothesenkonform ein negativer Zusammenhang zwischen Publikationsleistung und Nennung akademischer Titel in der E-Mail-Signatur in der Stichprobe ($r = -.29$). Dieser negative Zusammenhang erwies sich im Signifikanztest als überzufällig bzw. statistisch signifikant (die Irrtumswahrscheinlichkeit war mit $p = .03$ kleiner als das konventionelle Signifikanzniveau von $\alpha = .05$). Somit entscheiden wir uns für die **Annahme der Forschungshypothese**. Sie gilt als vorläufig bestätigt. Bei einer umfassenden Ergebnisinterpretation sind drei weitere Aspekte zu beachten: So ist die **Effektgröße** zu diskutieren: Offenbar existiert ein überzufälliger negativer Zusammenhang, allerdings handelt es sich um eine numerisch eher schwache Korrelation, deren **praktische Bedeutung** im Kontext anderer Studien zu bewerten ist (zur Effektgröße ► Abschn. 14.2). Zudem ist die **Generalisierbarkeit des Befundes** aus einer Gelegenheitsstichprobe von $n = 44$ E-Mails zu hinterfragen (Problem der **externen Validität**; ► Abschn. 3.2.2). Dass die Studie einen hypothesenkonformen überzufälligen Effekt gezeigt hat, spricht für die Theorie, lässt aber nicht den Rückschluss zu, dass dieser Effekt bei allen Forschenden in den USA vorliegt. Dazu wären Studien mit weiteren und ggf. auch höherwertigen Stichproben notwendig, welche die Population der Forschenden umfassender abdecken (► Abschn. 9.3). Nicht zuletzt stellt sich die Frage nach der **kausalen Interpretierbarkeit** des Befundes (Problem der **internen Validität**, ► Abschn. 3.2.2). Statistisch nachweisbar war in der Studie ein überzufälliger negativer linearer Zusammenhang gemäß einem Produkt-Moment-Korrelationskoeffizienten von $r = -.29$. Ob dieser Zusammenhang dadurch zustande kommt, dass die geringere Produktivität das beobachtete symbolische Selbstergänzungsverhalten verursacht, oder ob hier andere Kausalmechanismen wirken, muss diskutiert werden.

prüfen zu wollen, die sich der Form nach weder verifizieren noch falsifizieren lassen. Der Ausweg aus dieser Situation besteht in der Festlegung von Prüfkriterien, die zur Falsifikation herangezogen werden. Eines der wichtigsten Prüfkriterien bei der Hypothesenprüfung ist die **statistische Signifikanz**. Ein Ergebnis ist statistisch signifikant, wenn die Irrtumswahrscheinlichkeit p (d. h. die Auftretenswahrscheinlichkeit des Stichprobenergebnisses unter der Annahme, dass in der Population kein Effekt vorliegt) kleiner als das als Falsifikationskriterium festgelegte Signifikanzniveau von üblicherweise $\alpha = 5\,\%$ ausfällt (► Abschn. 12.5.1).

2.2.10 Prüfung und Weiterentwicklung von Theorien

Da die aus der Kerntheorie abgeleitete Forschungshypothese sich empirisch nicht als falsch erwiesen hat (d. h. anhand des statistischen Signifikanztests nicht falsifiziert wurde), ist auch die Theorie nicht als falsch zurückzuweisen. Vielmehr wird (unter Annahme der Gültigkeit der Hilfs- und Instrumententheorien gemäß ◘ Tab. 2.2 und des statistischen Signifikanztests gemäß ◘ Tab. 2.3) angesichts des gescheiterten Falsifikationsversuchs der Hypothese auch die Theorie als vorläufig bestätigt betrachtet. Dass bei der Ergebnisinterpretation eines statistischen Hypothesentests die Effektgröße, die externe und die interne Validität zu berücksichtigen sind, wurde in ◘ Tab. 2.3 bereits angesprochen. Diese Punkte seien noch einmal aufgegriffen und vertieft.

Dass die Theorie der symbolischen Selbstergänzung auf akademische Identität und Titel-Nennung in der Online-Kommunikation anwendbar ist, wird zusätzlich dadurch gestützt, dass Harmon-Jones, Schmeichel und Harmon-Jones (2009) in ihrer Publikation nicht nur eine, sondern **drei empirische Prüfungen der Theorie an unterschiedlichen Stichproben** berichten. Neben der in ◘ Tab. 2.3 dargestellten Studie mit $n = 44$ E-Mails ($r = -.29$, $p = .03$) wurde in zwei weiteren Untersuchungen mit $n = 29$ E-Mails ($r = -.35$, $p = .04$) sowie mit $n = 793$ Webseiten ($r = -.12$, $p = .001$) jeweils dasselbe Ergebnismuster im Sinne eines negativen Zusammenhangs zwischen akademischer Leistung und symbolischem Selbstergänzungsverhalten gefunden. Durch den Nachweis des Effekts in unterschiedlichen Stichproben und Kontexten ist seine Gültigkeit und Verallgemeinerbarkeit **(externe Validität)** besser abgesichert. Die **Effektgröße** (▶ Abschn. 14.2) wird bei Korrelationskoeffizienten über den Determinationskoeffizienten r^2 ausgedrückt und liegt im Beispiel also zwischen $r^2 = (-.12)^2 = 1.4\,\%$ und $r^2 = (-.35)^2 = 12.2\,\%$. Inwiefern diese statistischen Effektgrößen von **theoretischer und praktischer Bedeutung** sind – muss inhaltlich diskutiert werden (vorzugsweise durch Vergleiche mit Ergebnissen anderer Studien aus dem entsprechenden Forschungsfeld).

Es ist zu beachten, dass alle drei Studien nicht-experimentelle querschnittliche Korrelationsstudien darstellen und somit zwar einen systematischen Merkmalszusammenhang, nicht jedoch **Kausalität** belegen. Denn zum gültigen Nachweis einer Ursache-Wirkungs-Relation müsste diese empirisch aktiv hergestellt werden im Rahmen eines experimentellen oder quasi-experimentellen Designs oder müsste zumindest über die Zeit hinweg in einer Längsschnittstudie beobachtet werden. Da wir bei Kausalität davon ausgehen, dass die Ursache der Wirkung vorausgeht, ist es im vorliegenden Beispiel unplausibel, dass der Merkmalszusammenhang durch eine umgekehrte Kausalrichtung zustande kommt (die Titel-Nennung kann nicht die Publikationsleistung verursachen). Allerdings könnte der Merkmalszusammenhang durch Drittvariablen erzeugt werden, die sowohl die Publikationsleistung als auch die Titelnennung beeinflussen. Der stringente empirische Nachweis einer bestimmten Kausalrelation ist wissenschaftlich sehr anspruchsvoll, da es darum geht, alle Alternativerklärungen theoretisch und empirisch auszuschließen. Diese Problematik ist vor allem mit dem Untersuchungsdesign und dem Gütekriterium der **internen Validität** verknüpft (▶ Abschn. 3.2.2). Vorschnelle kausale Interpretationen von Korrelationen sind ein sehr verbreiteter Fehler.

Eine **monokausale Hypothese** benennt im Wenn-Teil nur *eine* Ursache, eine multikausale Hypothese führt im Wenn-Teil *mehrere* Ursachen an. In den Sozial- und Humanwissenschaften gehen wir in der Regel davon aus, dass soziale Sachverhalte durch mehrere verschiedene Ursachen und oft auch deren Zusammenspiel hervorgebracht werden und somit multikausale Erklärungen angemessen sind. **Multikausale Hypothesen** mit vielen konjunktiven Und-Komponenten im Wenn-Teil haben jedoch einen geringeren Informationsgehalt (▶ Abschn. 2.2.4). Insofern besteht hier bei der Theoriebildung ein Spannungsverhältnis zwischen dem Bestreben, einerseits die Komplexität der sozialen Wirklichkeit abzubilden und andererseits möglichst allgemeingültige Erklärungen zu finden, die unter vielen Umständen zutreffen, wenn auch vielleicht in unterschiedlichem Maße. Eine einfache Wenn-Dann-Hypothese lässt auch multikausale Erklärungen zu, denn sie behauptet nicht, dass die mit dem Wenn-Teil verbundene unabhängige Variable die einzige Erklärung für die abhängige Variable ist. Generell werden **Theorien mittlerer Reichweite** („middle range theories“) in den Sozialwissenschaften für erstrebenswert gehalten (Merton, 1949), da sie sowohl informationsreich als auch falsifizierbar sind. Sie grenzen sich einerseits von hochkomplexen allumfassenden Theorien zur Erklärung „der Gesellschaft“ oder „des Menschen“ ab, die eine große Fülle von Einflussfaktoren (inklusive historischer und kultureller) berücksichtigen, aber forschungspraktisch kaum empirisch prüfbar sind. Und sie grenzen sich andererseits von Ad-hoc-Theorien ab, die sich auf einen sehr begrenzten Sachverhalt beziehen und somit zwar mit geringem Aufwand zu prüfen sind, aber keinen großen Erklärungswert besitzen.

Theoretische Erklärungen, bei denen einzelne Ursachen betrachtet werden, postulieren den jeweiligen Kausaleffekt unter der sog. **Ceteris-Paribus-Klausel**, d. h. unter der Annahme, dass alle anderen Bedingungen gleich

sind. Dies muss dann auch im Untersuchungsdesign realisiert werden, etwa in der Weise, dass in einer Studie (z. B. zur Erklärung von Lungenkrebs) zwei Gruppen miteinander verglichen werden (Raucher und Nichtraucher), die sich ausschließlich im betrachteten Ursachenfaktor (Rauchen) und nicht hinsichtlich anderer Faktoren (z. B. genetische Disposition, Schadstoff-Exposition, Lebensstil) systematisch unterscheiden. Die vorläufige Bestätigung der Theorie, dass Rauchen Lungenkrebs erzeugt, entbindet dabei nicht davon, weitere Ursachen für Lungenkrebs ausfindig zu machen. Im Zuge der Wahrheitssuche können und sollen letztlich verschiedene Verursachungstheorien gegeneinander abgewogen oder auch integriert werden (z. B. können besondere Effekte durch das gleichzeitige Zusammenwirken unterschiedlicher Ursachenfaktoren wie z. B. Rauchen, Ernährung und Bewegung entstehen). Die Wahrheitssuche gemäß Versuch und Irrtum ist im Verständnis des Kritischen Rationalismus über die präzise Bearbeitung einzelner Fragestellungen zu bewerkstelligen. Damit wird ein **analytisches Vorgehen** präferiert, das die Komplexität der Erfahrungswirklichkeit auf einzelne Merkmale und Merkmalszusammenhänge herunterbricht (im Unterschied zu einem holistischen Vorgehen, das den Anspruch verfolgt, die Wirklichkeit in ihrer Komplexität möglichst vollständig und ganzheitlich abzubilden).

Monokausale Erklärung – Ein Sachverhalt wird durch **einen einzigen Ursachenfaktor** erklärt. Theorien mit einem einzigen Erklärungsfaktor werden oft als zu stark vereinfachend betrachtet. Sie können durch weitere Theorien ergänzt werden, die andere Ursachen des fraglichen Sachverhalts adressieren.

Multikausale Erklärung – Ein Sachverhalt wird durch **mehrere Ursachenfaktoren** erklärt. Es wird davon ausgegangen, dass viele soziale Sachverhalte jeweils durch mehrere Ursachen und deren Zusammenwirken hervorgerufen werden, weshalb multikausale Erklärungen den monokausalen meist überlegen sind. Der Kritische Rationalismus strebt danach, die zentralen Ursachenfaktoren analytisch herauszuarbeiten. Eine ganzheitliche (holistische) Herangehensweise, dergemäß alle erdenklichen Einflussfaktoren gleichzeitig betrachtet werden, wird abgelehnt, da hier die Gefahr gesehen wird, dass dogmatische Weltbilder Einzug halten auf Kosten differenzierter empirischer Analyse und theoriekritischer Diskussion.

Wenn empirische Daten zur Theorieprüfung und Theorieentwicklung verwendet werden, sind verschiedene **Arten der unvollständigen Erklärung** zu vermeiden (vgl. Schnell et al., 2008, S. 68ff.):

- **Nachträgliche Ad-hoc-Erklärung**: Ein theoriekonträrer Befund wird im Nachhinein durch eine neue alternative Theorie bzw. durch **neu eingeführte Wirkfaktoren** erklärt. Diese neue Theorie wurde induktiv aus den Daten gewonnen, ist somit unbestätigtes Vermutungswissen und muss als solches im Sinne des Kritischen Rationalismus der Wissenschaftsgemeinschaft zur eingehenden Prüfung vorgelegt werden. Solange entsprechende inhaltliche Ad-hoc-Erklärungen jedoch ungeprüft bleiben (was meist der Fall ist), haben sie noch keinen besonderen Erkenntniswert. Nachträgliche Ad-hoc-Erklärungen können sich auch auf die **Hilfs- und Instrumententheorien** beziehen, so etwa wenn ein der inhaltlichen Theorie widersprechendes Ergebnis im Nachhinein darauf zurückgeführt wird, dass eine der verwendeten Hilfstheorien ungültig sei (z. B. dass in der Befragungssituation nicht ehrlich, sondern weitgehend unehrlich geantwortet wurde). Auch derartige Ad-hoc-Erklärungen durch modifizierte Hilfstheorien sind zunächst reine Vermutungen und bedürfen der empirischen Prüfung.
- **Partielle Erklärung**: Eine Theorie sagt relativ allgemeine Wirkungen voraus (z. B. gesteigertes symbolisches Selbstergänzungsverhalten), empirisch werden jedoch sehr konkrete Wirkungen (z. B. Nennung akademischer Titel in E-Mail-Signaturen) geprüft. Daneben sind viele andere Formen des symbolischen Selbstergänzungsverhaltens möglich. Dass und warum es gerade zu dem E-Mail-Selbstdarstellungsverhalten kommt, wird somit durch die Theorie nur partiell erklärt.
- **Erklärung mit impliziten Gesetzen**: Hierbei handelt es sich um Gesetzesaussagen, bei denen die Wenn-Komponente bzw. die **unabhängigen Variablen** sog. **Globalvariablen** darstellen, also sehr grobe Kategorien (z. B. Lebensalter, Ethnizität oder Geschlecht), die im Einzelnen mit einer Vielzahl von Wirkfaktoren verknüpft sind (z. B. biografische Erfahrungen, Rollenerwartungen, sozioökonomischer Status, Einstellungen, körperliche Merkmale etc.). Gesetzesaussagen wie „Ältere sind weniger technikinteressiert als Jüngere“ oder „Frauen sind weniger technikinteressiert als Männer“ sind Beispiele für Erklärungen mit impliziten Gesetzen, denn hier werden bestimmte Wirkmechanismen jenseits des puren kalendarischen Lebensalters oder des biologischen Geschlechts unterstellt (z. B. Bedeutung des biografischen Zeitpunkts, an dem man mit neuen Technologien in Berührung kommt; Geschlechtsrollenerwartungen; Einfluss von Familie und Peer-Group). Der wissenschaftliche Erkenntniswert ist höher, wenn anstelle des Operierens mit Globalvariablen die impliziten Annahmen expliziert und die eigentlichen Wirkfaktoren untersucht werden. Umgekehrt: Die vorläufige Bestätigung einer Gesetzesaussage über Globalvariablen sagt nichts über die dahinter liegenden Ursache-Wirkungs-Mechanismen aus.

Es bleibt festzuhalten, dass die Frage, ob und inwiefern inhaltliche Theorien und zugehörige Hilfstheorien

2

Anforderungen des Kritischen Rationalismus an wissenschaftliche Theorien

Eine sozialwissenschaftliche **Theorie** ist ein in sich geschlossenes (d. h. von anderen Theorien abgrenzbares) Aussagensystem zur Beschreibung, Erklärung und Vorhersage eines bestimmten sozialen Phänomens bzw. Bereichs der sozialen Wirklichkeit. Sie besteht aus Definitionen, Axiomen und Propositionen, die sich mit ihren theoretischen Konstrukten auf einzelne Merkmale der sozialen Wirklichkeit beziehen, die im quantitativen Paradigma als Variablen zu messen sind (vgl. Schnell et al., 2008, S. 54):

- Mit **Definitionen** werden jeweils die Bedeutungen aller verwendeten Begriffe bzw. theoretischen Konstrukte eindeutig festgelegt und damit auch ihre Relationen (z. B. was sind Ober- und Unterbegriffe).
- **Axiome** sind Aussagen über Zusammenhänge zwischen mindestens zwei theoretischen Konstrukten, deren Gültigkeit im Rahmen der Theorie unterstellt und nicht mehr geprüft wird. Die Axiome einer Theorie müssen voneinander unabhängig sein und dürfen nicht im Widerspruch zueinander stehen. Axiome beziehen sich in sozialwissenschaftlichen Theorien oft auf grundlegende Eigenschaften des Menschen (z. B. „Menschen streben nach Belohnungen“ oder „Menschen sind selbstreflexiv“) und werden als Menschenbilder zugrunde gelegt.
- Die **Theoreme** bzw. **Propositionen** sind die eigentlichen inhaltlichen Aussagen der spezifischen Theorie, sie leiten sich aus den Axiomen und Definitionen ab. Eine Proposition ist eine Verknüpfung zwischen mindestens zwei Konstrukten (z. B. Wenn-Dann-, Je-Desto-Aussage), die einen möglichst **großen Geltungsbereich** (wenig beschränkter Raum-Zeit-Bezug) sowie einen **festgelegten Objektbereich** (z. B. Jugendliche im Alter zwischen 14 und 19 Jahren) hat, für den sie gilt (Allaussage; vgl. Schnell et al., 2008, S. 62). Von besonderer Bedeutung sind kausale Propositionen, die Ursache-Wirkungs-Relationen ausdrücken und somit Erklärungen und Prognosen erlauben. Je nach empirischem Bewährungsgrad werden die inhaltlichen Aussagen einer Theorie auch als **Gesetze** (sehr hoher Bewährungsgrad) oder als **Hypothesen** (noch sehr geringer Bewährungsgrad) bezeichnet. Theorien enthalten aber auch beschreibende Propositionen. Manche Theorien bestehen im Kern nur aus ein oder zwei Theoremen, Propositionen, Gesetzen oder Hypothesen. So besteht die Theorie der symbolischen Selbstergänzung im Wesentlichen aus einem Theorem (Tab. 2.1).

Die Annahmen einer Theorie müssen eigentlich vollständig und systematisch **explizit** dargestellt werden (eben mit allen Definitionen, Axiomen und Theoremen und deren Relationen untereinander). In der sozialwissenschaftlichen Forschungspraxis geschieht dies jedoch nur selten. Stattdessen wird der Grundgedanke der Theorie mit einigen zentralen Theoremen und Definitionen anschaulich umschrieben (Axiome werden oft nicht expliziert). Eine vollständige Explizierung und Formalisierung von Theorien erfolgt in der empirischen Sozialforschung allenfalls in computergestützten Simulationsstudien (► Abschn. 6.3.3). Ist eine Theorie nicht vollständig und strukturiert expliziert, kann eine Einschätzung ihrer Widerspruchsfreiheit (siehe unten) erschwert sein.

Nicht mehr direkter Bestandteil der Theorie sind die **empirischen Hypothesen**, die logisch stringent aus ihr abgeleitet (deduziert) werden. Im Unterschied zum Theorem bzw. zur **theoretischen Hypothese** als Bestandteil der Theorie mit jeweils breitem Geltungs- und Objektbereich, wird bei der empirischen Hypothese der raum-zeitliche Geltungsbereich und oft auch der Objektbereich eingeschränkt auf solche Arten von Untersuchungsobjekten und raum-zeitlichen Gegebenheiten, die empirisch im Rahmen einer konkreten Studie untersuchbar sind. Auch sollte zu jeder empirischen Hypothese angebbar sein, wie die theoretischen Konstrukte als empirische Indikatoren bzw. Variablen zu messen sind.

Zur Beurteilung von Theorien in der empirischen Sozialforschung sind folgende Kriterien heranzuziehen (Asendorpf & Neyer, 2012, S. 5; Pähler, 1986):

1. Theorien müssen **innere Widerspruchsfreiheit** („internal consistency“) aufweisen. Mangels systematischer Darstellung der meisten sozialwissenschaftlichen Theorien (z. B. nicht alle Definitionen und Axiome der Theorie sind vollzählig ausformuliert, sondern manche bleiben implizit) kann es schwierig sein, die logische Konsistenz eines in der Literatur gefundenen Theoriegebäudes im Detail zu bewerten.
2. Theorien müssen zudem **äußere Widerspruchsfreiheit** („external consistency“) aufweisen, das heißt, sie dürfen nicht im Widerspruch zu gut bewährten und allgemein anerkannten anderen Theorien stehen. Wohl aber dürfen sie im Widerspruch stehen zu konkurrierenden Theorien mit ähnlichem Bewährungsgrad oder geringerem Informationsgehalt.
3. Theorien müssen dem Kriterium der **Falsifizierbarkeit** („falsifiability“) genügen. Das heißt, die in der Theorie verwendeten theoretischen Konstrukte, die sich auf Merkmale der sozialen Wirklichkeit beziehen, müssen so eindeutig definiert sein, dass sie prinzipiell empirisch erfassbar bzw. messbar sind. Aus der Theorie müssen sich empirisch prüfbare Hypothesen eines untersuchbaren raum-zeitlichen Geltungsbereichs und Objektbereichs logisch stringent ableiten lassen. Falsifizierbarkeit setzt nicht nur voraus, dass die theoretischen Konstrukte als Variablen operationalisierbar und messbar sind, sondern auch, dass Untersuchungsbedingungen zu schaffen sind, in denen das gesamte Variationsspektrum der Ausprägungen der Variablen vorkommt, denn anhand eines eingeschränkten Variationsspektrums können die behaupteten Ursache-Wirkungs-Relationen nicht geprüft werden. Umgekehrt: Eine Theorie ist falsifizierbar, wenn man aus ihr Hypothesen ableiten kann, für die genau angegeben werden kann, welche empirischen Ergebnisse als hypothesenkonträr oder als hypothesenkonform zu betrachten sind. Anhand der empirischen Annahme oder Ablehnung der Hypothese ist dann die Falsifikation oder Weiterentwicklung der Theorie zu diskutieren.
4. Theorien sollen einen möglichst **hohen Informationsgehalt** („empirical content“) haben. Das heißt, ihr Geltungsbereich sollte möglichst breit (hohe **Allgemeinheit**) und die **Präzision** ihrer Vorhersagen möglichst groß sein. Somit existieren für die Theorie dann viele potenzielle Falsifikatoren (theoriewiderlegende Untersuchungseinheiten, auf die die Ursachen, aber nicht die Wirkungen zutreffen). Dagegen hat eine Theorie zum selben Sachverhalt einen geringeren Informationsgehalt, wenn sie einen engen Geltungsbereich aufweist und Wirkungen nur vage vorhersagt, so dass in der Beobachtungswirklichkeit nur eine kleine Zahl von Fällen bzw. potenziellen Falsifikatoren existieren, an denen die Theorie geprüft werden kann (geringerer Grad an Falsifizierbarkeit; „degree of falsifiability“).

5. Theorien sollen eine möglichst **große Erklärungskraft** („explanatory power“) besitzen. Sie sollen sich nicht nur auf deskriptive Aussagen beschränken, sondern Ursache-Wirkungs-Relationen angeben, die das Zustandekommen bestimmter Phänomene erklären sowie Vorhersagen über das zukünftige Auftreten des Phänomens erlauben. Zudem hat im Theorienvergleich diejenige Theorie die höhere Erklärungskraft, aus der mehr neue Hypothesen ableitbar sind bzw. mit deren Hilfe mehr Effekte kausal erklärt werden können.
6. Insbesondere bei technologischen Theorien (► Abschn. 18.1.1) ist auch die **praktische Anwendbarkeit** („practicality“) ein Gütekriterium, also die Frage, inwiefern die Theorie Vorhersagen macht, die für die Gestaltung von Praxismaßnahmen handlungsleitend sein können.
7. Theorien sollen möglichst **große Einfachheit bzw. Sparsamkeit** („parsimony“, „succintness“) aufweisen. Das heißt, sie sollten mit möglichst wenigen Konstrukten, Axiomen und Theoremen auskommen und nur so komplex wie unbedingt notwendig sein. Von zwei konkurrierenden Theorien ist (bei ansonsten vergleichbarer Qualität der Theorien) die einfachere zu bevorzugen. Das Sparsamkeitsprinzip ist auch als **Ockhams Rasiermesser** („Ockham’s razor“) bekannt, benannt nach dem Philosophen William Ockham (1285–1345), der dafür plädierte, unnötig komplizierte theoretische Erklärungen sozusagen abzurasieren.
8. Eine Theorie sollte einen möglichst **hohen empirischen Bewährungsgrad** („degree of corroboration“) aufweisen, das heißt möglichst viele und möglichst strenge Falsifikationsversuche überstanden haben. Besonders gut erkennbar ist ein hoher Bewährungsgrad an einer großen Zahl theorieprüfender Studien, die übereinstimmend die postulierten Effekte zeigen, was vor allem mittels Metaanalysen nachgewiesen wird (► Kap. 16). Generell besteht jedoch die Gefahr, den empirischen Bewährungsgrad einer Theorie zu überschätzen, da im Sinne des sog. Publication Bias theoriekonträre Befunde seltener publiziert werden als theoriekonforme (► Abschn. 16.1.3).

In der Forschungspraxis arbeiten Studien oft mit einzelnen theoretischen Hypothesen bzw. Theoremen, die wissenschaftstheoretisch nicht den Status einer vollwertigen Theorie haben, etwa weil sie kein in sich abgeschlossenes, konsistentes Aussagensystem bilden, eine geringe kausale Erklärungskraft aufweisen oder einen sehr geringen empirischen Bewährungsgrad besitzen. In diesem Fall wird oft von einem **theoretischen Modell**, einem **Theorierahmen** bzw. einem **theoretischen Analyserahmen** gesprochen, um den erkenntnistheoretisch begrenzten Status der theoretischen Aussagen zu kennzeichnen. Die einmalige Prüfung von Hypothesen, die aus einem selbst entwickelten Theorierahmen abgeleitet wurden, trägt im Vergleich zur wiederholten Prüfung einer etablierten Theorie aus Sicht des Kritischen Rationalismus in geringerem Maße zum theoretischen Erkenntnisgewinn bei. Denn durch jeweils einmalige Prüfung immer wieder neu konstruierter Modelle fehlt es an ausreichenden Falsifikationsversuchen, um den Bewährungsgrad bewerten zu können.

anhand von empirischen Hypothesenprüfungen vorläufig zu bestätigen, zu widerlegen oder zu modifizieren sind, aus Sicht des Kritischen Rationalismus einer **gründlichen Diskussion in der Wissenschaftsgemeinschaft** bedarf und in der quantitativen Sozialforschung mit dem Ausgang des statistischen Signifikanztests (oder eines anderen hypothesenprüfenden inferenzstatistischen Verfahrens, ► Abschn. 12.2) keineswegs abschließend beantwortet ist.

Wissenschaftlicher Erkenntnisfortschritt ist im Verständnis des Kritischen Rationalismus maßgeblich theoriegetrieben. Dementsprechend ist eine gründliche Theoriearbeit notwendig, sowohl was die interessierende inhaltliche Kerntheorie als auch mögliche Konkurrenztheorien und die jeweils verwendeten Hilfs- und Instrumententheorien betrifft. Nicht zuletzt ist eine wissenschaftstheoretische Reflexion notwendig, um zu begründen, ob und wie mit den Ergebnissen der statistischen Hypothesenprüfung im Hinblick auf die Theoriebewertung umzugehen ist (► Anforderungen des Kritischen Rationalismus an wissenschaftliche Theorien).

Neben dem Typus der hypothesenprüfenden **explanativen Studie**, die auf Theorieprüfung ausgerichtet ist, werden in der quantitativen empirischen Sozialforschung oft populationsbeschreibende **deskriptive Studien** durchgeführt, die der Gegenstandsbeschreibung dienen und in denen es z. B. um die Erfassung der Verbreitung oder Ausprägung bestimmter Merkmale in der Bevölkerung geht, ohne dass damit ein kausaler Erklärungsanspruch verbunden wird. Hier gelten prinzipiell dieselben wissenschaftstheoretischen Überlegungen wie bei explanativen Studien. Insbesondere müssen Basissatz- und Korrespondenzproblem entsprechend gelöst werden (d. h. Auswahl eines geeigneten Messinstruments für das interessierende Konstrukt und Realisierung einer möglichst unverzerrten Datenerhebung). **Explorative Studien** sind dagegen oft qualitativ angelegt.

2.2.11 Grenzen des Kritischen Rationalismus

Das Wissenschaftsverständnis des Kritischen Rationalismus ist in der empirisch-quantitativen Sozialforschung weithin akzeptiert (► Abschn. 2.2.11). Dennoch werden Grenzen des Ansatzes sowohl von Seiten der Forschungspraxis als auch der Wissenschaftstheorie aufgezeigt. Während Popper (1934/1989) den Induktionsschluss strikt ablehnt und das deduktiv-nomologische Modell propagiert, argumentieren Kritiker, der sog. Deduktionsschluss

sei in Wirklichkeit seinerseits induktiv (Kelle, 2008). Infrage gestellt wird auch, inwiefern der Kritische Rationalismus ein schlüssiges Wahrheitskonzept bietet. Denn er geht ja gemäß Basissatzproblem davon aus, dass die Daten nicht direkt an der Wirklichkeit abgeglichen, sondern auf der Basis etablierter Methoden und durch Beschluss der Wissenschaftsgemeinschaft anerkannt werden (► Abschn. 2.2.3). Ein solcher konventionalistischer Wahrheitsbegriff passe nicht zum korrespondenztheoretischen Wahrheitsverständnis, dass man sich im Laufe der Zeit mit den Theorien der Wirklichkeit immer besser annähere (Thornton, 2013).

Kritikpunkte betreffen also mit dem Wahrheitsverständnis und dem deduktiv-nomologischen Erklärungsmodell einerseits das argumentative Fundament des Ansatzes und nehmen auf traditionsreiche philosophische Diskurse Bezug. Andererseits betreffen sie korrigierbare Schwächen wie die mangelnde Umsetzung bestimmter Prinzipien des Ansatzes (z. B. ungenügende Würdigung von nicht-signifikanten Befunden sowie von Replikationsstudien in der bisherigen Forschungspraxis).

Vernachlässigung des Entdeckungszusammenhangs

Der Kritische Rationalismus konzentriert sich auf den Begründungszusammenhang bzw. die Überprüfung von Theorien durch Falsifikation (► Abschn. 2.2.6). Über den Entdeckungszusammenhang – also die Generierung von neuen Theorien – wird nichts ausgesagt. Denn hierfür gibt es laut Kritischem Rationalismus keine logischen Regeln. Zuweilen wird es jedoch als Defizit des Ansatzes aufgefasst, dass er den Ausgangspunkt des wissenschaftlichen Erkenntnisgewinns in den Theorien sieht, deren Bildung jedoch nicht genauer betrachtet und systematisiert. Popper (1934/1989, S. 7) spricht von kreativen „Einfällen“ und von „Intuition“, wenn es um das Aufstellen neuer Theorien geht. Doch willkürliche Behauptungen sind damit nicht gemeint. Insbesondere der raffinierte methodologische Falsifikationismus weist darauf hin, dass gute Theorien im Rahmen von Forschungsprogrammen durch Weiterentwicklung vorhandener Theorien entstehen können.

Wir werden in ► Abschn. 6.3 genauer darauf eingehen, wie man sinnvollerweise zu einem theoretischen Rahmen für eine eigene empirische Studie kommt: Theorien können durch Weiterentwicklung vorliegender theoretischer Modelle entstehen, im Zuge explorativer empirischer Studien gewonnen, durch Variation von Methoden oder auch im Zuge von Alltagsbeobachtungen und Introspektion kreiert werden. Da Theoriebildung durch Induktion und Abduktion im qualitativen Paradigma im Zentrum steht, werden für die Theorieentwicklung im quantitativen Ansatz oft qualitative Vorstudien durchgeführt.

Zweifel an der Existenz sozialer Gesetzmäßigkeiten

Der Kritische Rationalismus geht davon aus, dass wissenschaftlicher Erkenntnisgewinn in den Natur- und in den Sozialwissenschaften denselben Regeln unterliegt **(Prinzip der Einheitswissenschaft)**, da sowohl die Natur als auch das soziale Leben kausalen Gesetzmäßigkeiten folgen. Dass solche Gesetzmäßigkeiten existieren, wird ontologisch vorausgesetzt.

Kritiker bezweifeln jedoch, dass die soziale Wirklichkeit anhand **allgemeingültiger Gesetzmäßigkeiten** erklärbar ist. Denn charakteristisch für Menschen sei ihr Bewusstsein und die Tatsache, dass sie nicht einfach mechanisch auf äußere Reize reagieren. Das sei ein verfehltes mechanistisches Menschenbild, welches vernachlässigt, dass Menschen sozialen Situationen Bedeutungen verleihen und planvoll handeln. Deswegen treten die Anhängerinnen und Anhänger einer interpretativen bzw. qualitativen Sozialwissenschaft für ein Bild des selbstbestimmt handelnden Menschen ein, dessen Erleben und Verhalten man nicht durch Benennen äußerer, objektiv beobachtbarer Wirkfaktoren „erklären“, sondern nur durch kommunikatives Nachvollziehen der subjektiven Weltsicht und inneren Gründe der Akteurinnen und Akteure „verstehen“ könne. „Die Natur erklären wir, das Seelenleben verstehen wir“, so hatte schon der Philosoph und Psychologe Wilhelm Dilthey (1883/2006) programmatisch eine Trennung natur- und geisteswissenschaftlicher Forschung postuliert (zur Abgrenzung der verschiedenen Wissenschaftsdisziplinen ► Abschn. 1.1.2).

Diese griffige polarisierte Gegenüberstellung von **Verstehen** und **Erklären** ist jedoch zu hinterfragen. So werden auch im quantitativen Ansatz die Sichtweisen und Sinngebungen der Beteiligten berücksichtigt (z. B. Erhebung von Motiven, Werten oder Zielen zur Erklärung von Handlungen) und es werden auch im qualitativen Ansatz Erklärungen genutzt (z. B. intentionale Erklärungen). Auch wird im quantitativen Paradigma kein deterministischer oder universalistischer Erklärungsanspruch erhoben: Jede sozialwissenschaftliche Theorie kann menschliches Verhalten und Erleben im Verständnis des Kritischen Rationalismus nur der Tendenz nach vorhersagen und lässt individuelle Ausnahmen zu, und jede Theorien hat jeweils eine limitierte Reichweite und ist z. B. nur auf bestimmte kulturelle oder organisationale Kontexte anwendbar. Das ontologische Verständnis des Kritischen Rationalismus, der von einer von sozialen Gesetzmäßigkeiten geprägten sozialen Wirklichkeit ausgeht, unterscheidet sich von dem Wirklichkeitsverständnis des Sozialkonstruktivismus im qualitativen Paradigma ► Abschn. 2.3), wobei diese Differenz aber nur ungenügend über das Begriffspaar Verstehen-Erklären zu fassen

und Gegenstand anspruchsvoller philosophischer Auseinandersetzungen ist.

Zweifel an Möglichkeit und Nutzen wissenschaftlicher Objektivität

Der vom Kritischen Rationalismus vertretene Anspruch **objektiver Theorieprüfung** wird vor allem von gesellschaftskritischen Positionen grundlegend infrage gestellt mit dem Argument, dass bei sozialen Sachverhalten eine **Trennung zwischen Erkenntnissubjekt und Erkenntnisobjekt** gar nicht möglich sei: Das gesamte Wissenschaftssystem sowie die einzelnen Forschenden sind Teil der Gesellschaft und partizipieren an bestehenden Macht- und Herrschaftsverhältnissen, was sich dann auch in einer durch den privilegierten Standpunkt der Forschenden entsprechend verzerrten (und eben nicht objektiven) Beschreibung und Erklärung der sozialen Wirklichkeit widerspiegele. Zudem stelle sich die Wissenschaft, wenn sie selbst nicht politisch Position beziehe, letztlich in den Dienst der gesellschaftlichen Herrschaftsverhältnisse und liefere Erkenntnisse, die zu beliebigen Zwecken missbraucht werden können.

Als Gegenmodell wird eine **emanzipatorische Wissenschaft** verlangt, die ausdrücklich Partei ergreift gegen Unterdrückung und für marginalisierte Bevölkerungsgruppen und darauf ausgerichtet ist, Erkenntnisse zu produzieren, die dazu dienen, Herrschaftsverhältnisse abzubauen und Gerechtigkeit und Gleichberechtigung zu fördern (z. B. **Prinzip der Subjektivität und Perspektivität** sowie Parteilichkeit statt Objektivität; ► Abschn. 2.3.5). Dementsprechend sei auch das zentrale Gültigkeitskriterium für Theorien nicht deren objektiv feststellbare Übereinstimmung mit der Wirklichkeit, sondern vor allem ihr Nutzen für gesellschaftliche Verbesserungen. Dies wiederum setzt voraus, dass man verbindlich festlegt, was als emanzipatorisch gelten und wirken kann und was nicht. Dieses Vorgehen wird im Kritischen Rationalismus als Rückfall in Dogmatismus abgelehnt. Der Kritische Rationalismus hält daran fest, dass es Aufgabe der Wissenschaft ist, herauszufinden, wie die soziale Wirklichkeit beschaffen ist und funktioniert und dass es Aufgabe der Politik ist, die soziale Wirklichkeit zu gestalten und Entscheidungen darüber zu treffen, wie sie beschaffen sein soll.

Unterschiedliche Auffassungen darüber, ob der Anspruch wissenschaftliche Unabhängigkeit bzw. Objektivität bei der Prüfung von Theorien a) überhaupt erreichbar und b) gesellschaftlich nützlich bzw. ethisch verantwortbar ist, sind in Kontroversen zwischen quantitativem und qualitativem Paradigma bis heute zentral. Dass man vorzugsweise mit strukturierten deduktiven Verfahren zu intersubjektiv nachvollziehbaren und auch replizierbaren Erkenntnissen über die soziale Wirklichkeit gelangt, ist die Grundannahme des quantitativen Paradigmas, während das qualitative Paradigma auf eine unstrukturierte und induktive Vorgehensweise setzt. Auf der Meta-Ebene bestehen trotzdem **Gemeinsamkeiten** sowohl in **epistemologischen Fragen** (z. B. unumgehbare Notwendigkeit der intersubjektiven Einigung hinsichtlich Gültigkeit oder Glaubwürdigkeit von Erkenntnissen, da absoluter Wahrheitsanspruch nicht erhoben werden kann; Bewusstsein darüber, dass Forschung immer ein aktiver Konstruktionsprozess ist, dessen Vorannahmen jeweils transparent zu machen und zu diskutieren sind) als auch in **axiologischen Fragen** (z. B. Bekenntnis zu humanitären Werten und gesellschaftlicher Verantwortung sozialwissenschaftlicher Forschung). Die Kontroverse um den Objektivitäts-Anspruch hat eine lange Geschichte in der empirischen Sozialforschung (► Werturteilsstreit und Positivismusstreit in der deutschen Soziologie).

Kluft zwischen Wissenschaftstheorie und Wissenschaftspraxis

Der Kritische Rationalismus schreibt mit dem Falsifikationismus normativ vor, wie wissenschaftlich zu forschen ist, um logisch begründet zu wissenschaftlichen Erkenntnissen über die Wirklichkeit zu gelangen. Das Grundlagenwerk von Popper (1934/1989) heißt „Logik der Forschung“ und nicht „Praxis der Forschung“, was verdeutlicht, dass es um die logische Begründung von Vorschriften für wissenschaftliches Forschen geht.

Kritik am Falsifikationismus kann nun daher rühren, dass man die zugrunde liegende logische Argumentation infrage stellt, etwa die Anwendbarkeit des deduktiv-nomologischen Modells auf probabilistische Theorien und Hypothesen. Eine andere Stoßrichtung der Kritik fragt nach der Bewährung der vom Kritischen Rationalismus formulierten Vorgaben in der alltäglichen Forschungspraxis: Gibt es Hinweise darauf, dass der logisch begründete Falsifikationismus in der Praxis tatsächlich funktioniert? Wurden falsche Theorien in der Vergangenheit durch Falsifikation widerlegt und dann auch aufgegeben? Oder muss man sich eingestehen, dass in vielen Fällen wissenschaftlicher Erkenntnisfortschritt in einer Weise erfolgt, die mit den Vorgaben des Kritischen Rationalismus gar nicht vereinbar ist?

Letztere Position wurde besonders namhaft von dem US-amerikanischen Wissenschaftshistoriker **Thomas S. Kuhn** (1922–1996) in seinem in erster Auflage 1962 erschienen Buch **Die Struktur wissenschaftlicher Revolutionen** (Kuhn, 1962) vertreten. Kuhn zeigte auf, dass in der Wissenschaftsdisziplin der Physik in einer bestimmten historischen Phase ein zentrales Theoriemodell – z. B. die Epizykeltheorie von Ptolemäus – vorherrschend ist und die Sichtweisen der Forschenden prägt. Diese zentrale Theorie geht mit einem speziellen Weltbild

2

Werturteilsstreit und Positivismusstreit in der deutschen Soziologie

Historisch wurde die Kontroverse um Objektivität versus Parteilichkeit bzw. politischer Einflussnahme vor allem in wissenschaftstheoretischen Kontroversen in der deutschen Soziologie ausgetragen, erst als Werturteilsstreit und dann als Positivismusstreit.

Der Werturteilsstreit

„Eine empirische Wissenschaft vermag niemanden zu lehren, was er soll, sondern nur was er kann und – unter Umständen – was er will." Diese von dem Soziologen Max Weber Anfang des 20. Jahrhunderts vertretene Position (Weber, 1904/1951, S. 151) führte zum sog. Werturteilsstreit. Webers Position, dass eine empirische Wissenschaft nicht normativ vorgeben könne, welche Werte („Zwecke", Ziele) gesellschaftlich richtig oder falsch sind, sondern dass sie nur „Zwecke" (Ziele), „Mittel" (Maßnahmen, um Zwecke/Ziele zu erreichen) und deren Relationen hinsichtlich ihrer Merkmale untersuchen könne, wurde von Soziologen im „Verein für Socialpolitik" scharf kritisiert, die der Auffassung waren, dass Wissenschaft nur dann sinnvoll sei, wenn ihre Ergebnisse auch moralisch fundiert sind und dementsprechend zu moralisch vertretbaren praktischen Ergebnissen und Konsequenzen führen. Wäre die Bewertung wissenschaftlicher Erkenntnisse einfach allen Menschen freigestellt, führe das zu Beliebigkeit und würde damit gerade den wissenschaftlichen Anspruch auf gesicherte Erkenntnis untergraben. Webers Plädoyer für „Wertfreiheit" bedeutete, dass er der Soziologie nicht die Rolle zuschrieb, über Werte zu entscheiden und die richtigen Werte der Gesellschaft vorzugeben, sondern diese Aufgabe der Politik zuwies.

Der Positivismusstreit

Unter dem Stichwort Positivismusstreit erlebte die deutsche Soziologie in den 1960er-Jahren eine zweite vehemente axiologische Auseinandersetzung über die Frage nach Werten und gesellschaftlicher Verantwortung im Wissenschaftsbetrieb (Keuth, 1989). Der Positivismusstreit fand maßgeblich zwischen Vertretern des Kritischen Rationalismus, vor allem Karl Popper und Hans Albert, und Vertretern der Frankfurter Schule („Kritische Theorie"), vor allem Theodor W. Adorno und Jürgen Habermas statt (zusammenfassend Adorno et al., 1969). In ähnlichem Sinne wie Max Weber spricht sich der Kritische Rationalismus für „Wertfreiheit" in dem Sinne aus, dass Wertfragen und Fragen der Gültigkeit von Hypothesen voneinander zu trennen sind.

Habermas (1983) als Vertreter der Frankfurter Schule, die in der Tradition der dialektischen Sozialphilosophie die Gesellschaft als Ganzes (Totalität) theoretisch fassen und verändern wollte, argumentierte, dass die angeblich „wertfreie", „reine" Forschung auf ihrer Suche nach „Wahrheit" letztlich nur kritiklos die bestehenden ungerechten Verhältnisse bestätige und aufrechterhalte. Dieser „positivistisch halbierte Rationalismus" sei auf reine Zweckrationalität verpflichtet, d. h., es würden Theorien und Techniken entwickelt, mit denen das soziale Leben weitreichend beeinflusst werden kann, ohne dass darüber reflektiert werde, welchen Sinn und Wert all dies habe.

Als Gegenmodell zur **Zweckrationalität**, die nur angibt, mit welchen Methoden welcher Zweck zu erreichen ist und damit leicht missbraucht werden kann, steht das dialektische Konzept der unteilbaren Vernunft, die Wissen nicht loslöst von Werten und praktischen Entscheidungen (**Wertrationalität**). Diese Vernunft hofften die Vertreter der Frankfurter Schule durch die Methode der **Dialektik** zu erreichen (Dialektik: Verfahren der Erkenntnisgewinnung, das durch den Wechsel von Argument/These und Gegenargument/Antithese die Begrenztheit einer theoretischen Idee zu erkennen und in einer Synthese zu überwinden sucht; Simon-Schäfer, 1993). Diese Methode soll es ermöglichen, verborgene Widersprüche und Erkenntnisinteressen zu erkennen: „Wir bringen z. B. zu Bewusstsein, daß empirisch-analytische Forschungen technisch verwertbares Wissen hervorbringen, aber kein Wissen, das zur hermeneutischen Klärung des Selbstverständnisses handelnder Subjekte verhilft" (Habermas, 1969, S. 161).

Kennzeichnend für den Positivismusstreit waren weitreichende Missverständnisse und persönliche Animositäten zwischen den Disputanten. So hatten z. B. die Anhänger des kritischen Rationalismus nie behauptet, dass Wissenschaftler wertfrei und unvoreingenommen gegenüber gesellschaftlichen Problemen arbeiten könnten oder sollten.

> » Es ist gänzlich verfehlt anzunehmen, daß die Objektivität der Wissenschaft von der Objektivität des Wissenschaftlers abhängt. Und es ist gänzlich verfehlt zu glauben, daß der Naturwissenschaftler objektiver ist als der Sozialwissenschaftler. Der Naturwissenschaftler ist ebenso parteiisch wie alle anderen Menschen, und er ist leider – wenn er nicht zu den wenigen gehört, die dauernd neue Ideen produzieren – gewöhnlich äußerst einseitig und parteiisch für seine eigenen Ideen eingenommen … Was man als wissenschaftliche Objektivität bezeichnen kann, liegt einzig und allein in der kritischen Tradition, die es trotz aller Widerstände so oft ermöglicht, ein herrschendes Dogma zu kritisieren. Anders ausgedrückt, die Objektivität der Wissenschaft ist nicht eine individuelle Angelegenheit der verschiedenen Wissenschaftler, sondern eine soziale Angelegenheit ihrer gegenseitigen Kritik … Sie hängt daher zum Teil von einer ganzen Reihe von gesellschaftlichen und politischen Verhältnissen ab, die diese Kritik ermöglichen. (Popper, 1969/1972, S. 112)

Objektivität im Verständnis des Kritischen Rationalismus meint intersubjektive Nachvollziehbarkeit und kritische Auseinandersetzung mit Theorien und Befunden innerhalb der Wissenschaftsgemeinschaft, ohne dass eine vorbestimmte moralische oder politische Position über den Wahrheitsgehalt einer Theorie entscheidet. Dies setzt die implizite Annahme voraus, dass das Wissenschaftssystem genügend Pluralität beinhaltet, um grundlegende Kritik sowie Vielfalt der gewählten Forschungsprobleme und entwickelten Theorien zu ermöglichen und nicht einer allseits geteilten impliziten Ideologie folgt.

einher (z. B. geozentrisches Weltbild, demgemäß die Erde im Zentrum des Universums steht und von den anderen Planeten umkreist wird) und bildet gemeinsam mit ihm das **Paradigma einer Fachdisziplin** (mit dem Paradigma ist bei Kuhn eine fachlich dominierende inhaltliche Theorie mitsamt Weltbild gemeint, dieser Paradigmen-Begriff ist nicht auf das quantitative und qualitative Paradigma im Sinne sozialwissenschaftlicher Forschungsstrategien anwendbar). Trotz falsifizierender Hinweise wird laut Kuhn in Zeiten der Normalwissenschaft gemeinschaftlich an diesem Paradigma festgehalten. Eine grundlegend neue Theorie setze sich nicht durch einzelne Falsifikationen durch, denn auch das zugehörige Weltbild müsste im Zuge eines Paradigmenwechsels geändert werden. Demnach bedarf es für das Verwerfen eines Paradigmas eine regelrechte **wissenschaftliche Revolution** in dem jeweiligen Fach (z. B. Ablösung des geozentrischen Weltbilds durch das heliozentrische Weltbild, demgemäß die Erde und die anderen Planeten des Sonnensystems um die Sonne kreisen). Oft müsse erst eine **neue Generation von Forschenden** antreten, weil die Vertreter des alten Modells und Weltbildes eben in der Regel nicht so grundlegend umdenken können oder wollen. Zudem lassen sich unterschiedliche Paradigmen auch nicht direkt vergleichen, da sie durch neue Grundannahmen geprägt sind **(Inkommensurabilität)**.

Kuhn stellte also das Idealbild des Kritischen Rationalismus, nämlich dass die Forschenden sich der ständig selbstkritischen, ergebnisoffenen Wahrheitssuche verschrieben haben, infrage und betonte, dass sie eben oft ihrem angestammten Theoriemodell und Weltbild treu bleiben. Dementsprechend sei wissenschaftlicher Fortschritt auch keine schrittweise Annäherung an die Wahrheit, sondern geprägt durch historische Umbrüche. Auf die Details der Auseinandersetzung zwischen Kuhn und Popper kann hier nicht genauer eingegangen werden. Im Hinblick auf die Sozialwissenschaften ist jedoch zu betonen, dass grundlegende Weltbilder wie in der Physik eigentlich kaum vorhanden sind, die Theorienlandschaft sehr viel kleinteiliger ist und somit Theorienstreit nicht unbedingt gleich „Paradigmenstreit" bedeutet.

Mit dem **raffinierten methodologischen Falsifikationismus** (▶ Abschn. 2.2.5) hat **Imre Lakatos** (1977) versucht, eine alternative Betrachtung des Wissenschaftsfortschritts zu liefern. Er widerspricht Poppers Idee, dass einzelne Theorien regelmäßig durch Falsifikation verworfen werden, was sehr kurze Theorielebenszyklen nach sich ziehen würde. Und er widerspricht Kuhns Idee, dass Theorien im Zusammenhang mit Weltbildern sehr lange überleben bis sie schließlich durch Generationenwechsel abgelöst werden. Stattdessen betont er, dass Forschung in Form von **Forschungsprogrammen** stattfindet, also über einige Jahre aufeinander aufbauenden Theoriereihen, die teils mit bestätigenden, aber immer auch mit widersprechenden Daten konfrontiert sind. Manche dieser Forschungsprogramme sind **degenerativ**, schränken etwa ihren Geltungsbereich zunehmend ein und verlieren an Wert. Andere dagegen sind **progressiv** und liefern schließlich unerwartete, innovative Vorhersagen. Der raffinierte methodologische Falsifikationismus und das Konzept der Forschungsprogramme haben u. a. in der Psychologie große Akzeptanz gefunden (Herrmann, 1976).

Aufbauend auf Kuhns (1962) Analyse der Wissenschaftsgeschichte, geht der Wissenschaftstheoretiker **Paul Feyerabend** (1924–1994) noch einen Schritt weiter. Nach seiner Auffassung zeigt die Wissenschaftsgeschichte, dass wichtige neue Erkenntnisse immer wieder durch Vorgehensweisen entstanden sind, die methodologische Regeln verletzen. In der Konsequenz empfiehlt Feyerabend somit provokativ einen **methodologischen Anarchismus**, demgemäß sich Forschende gerade nicht bestimmten logisch begründeten methodischen Vorschriften unterwerfen, sondern methodische Vorgaben flexibel handhaben und durchaus auch verletzen sollten nach dem viel zitierten Motto „anything goes" (Feyerabend, 1979). Nach Feyerabend stellen methodologische Regeln also keinen Rahmen dar, der Wissenschaftsfortschritt anleiten kann, sondern bilden eher ein einengendes Korsett für die Kreativität der Forschenden. Feyerabends radikale Thesen gegen methodische Zwänge wurden viel diskutiert, spielen aber heute in der Forschungspraxis keine entscheidende Rolle. Denn Feyerabend vertrat letztlich eine relativistische Position, dergemäß nicht nur nicht genau vorgegeben werden kann, welches methodische Vorgehen wissenschaftlich einzuschlagen ist, sondern auch keine sichere Theorieprüfung möglich ist. In der Konsequenz würden damit dann Theorien und Hypothesen, die einander widersprechende Aussagen über dasselbe Phänomen treffen (z. B. Ausländerfeindlichkeit nimmt in einem bestimmten Zeitraum in einer bestimmten Bevölkerungsgruppe zu versus nimmt nicht zu), nebeneinander stehen, ohne dass über ihre Gültigkeit entschieden werden kann.

Eine solche relativistische Haltung (dergemäß beliebig viele widersprüchliche Wahrheiten über denselben Sachverhalt gleichberechtigt nebeneinander existieren) wird im Kritischen Rationalismus ebenso wie jeder Dogmatismus (demgemäß man beansprucht, die einzige Wahrheit unfehlbar gefunden zu haben) abgelehnt zugunsten einer fortwährenden gemeinsamen Wahrheitssuche. Auf der Ebene einzelner Forschungsfelder innerhalb der quantitativen Sozialforschung lässt sich durchaus zeigen, dass bestimmte Theorien und Befunde im Laufe der Zeit durch widersprechende Empirie zurückgewiesen, modifiziert oder durch andere Theorien ersetzt werden. Konkrete Hinweise auf derartige Prozesse sind vor allem jenen wissenschaftlichen Beiträgen zu entnehmen, die sich der

Forschungssynthese widmen und z. B. im Rahmen systematischer Reviews den Theorie- und Forschungsstand zusammenfassend bewerten (▶ Kap. 16).

2.2.12 Neuere wissenschaftstheoretische Ansätze im quantitativen Paradigma

Der klassische Kritische Rationalismus, der in diesem Kapitel ausführlich dargestellt wurde, ist im quantitativen Paradigma der empirischen Sozial- und Humanforschung bis heute als **Grundverständnis** etabliert (Gadenne, 2006). Empirische Studien dazu, welche wissenschaftstheoretischen Positionen von Forschenden unterschiedlicher sozial- und humanwissenschaftlicher Disziplinen und Forschungsfelder im Einzelnen vertreten werden, fehlen jedoch weitgehend.

Wissenschaftstheoretische Fragen werden in der empirischen Sozialforschung anders diskutiert als in der Philosophie. So hat innerhalb der Philosophie der Kritische Rationalismus heute eine sehr umstrittene Stellung; zahlreiche konkurrierende wissenschaftstheoretische Modelle werden in der aktuellen **philosophischen Fachliteratur** diskutiert (z. B. Fachzeitschriften wie *International Studies in the Philosophy of Science*, Taylor & Francis, oder *Philosophy of Science*, University of Chicago Press). Aus philosophischer Sicht ist der Kritische Rationalismus trotzdem zu verteidigen, bedarf aber vor dem Hintergrund aktueller philosophischer Debatten einer kritischen Überarbeitung (Rowbottom, 2011).

Aus Sicht der empirischen Sozial- und Humanforschung besteht das Problem, dass eine Metareflexion des eigenen wissenschaftlichen Vorgehens im Sinne der Wissenschaftlichkeit notwendig ist (▶ Abschn. 3.1), dass gleichzeitig aber die philosophische Fachliteratur zur Wissenschaftstheorie ohne umfassende Vorkenntnisse in Philosophie kaum erschlossen werden kann. Eine wichtige Rolle in der Auseinandersetzung mit Wissenschaftstheorie spielen deswegen **disziplinspezifische Aufarbeitungen** und Weiterentwicklungen wissenschaftstheoretischer Modelle. So hat z. B. der Psychologe Theo Herrmann für die Psychologie in der Tradition von Imre Lakatos und seinem raffinierten methodologischen Falsifikationismus (▶ Abschn. 2.2.5) das Konzept der **Forschungsprogramme** adaptiert (Herrmann, 1976) und in der Tradition von Popper (▶ Abschn. 2.2.1) die Erkenntnissuche in der psychologischen Forschung als fortwährenden **Problemlösungsprozess** rekonstruiert (Herrmann, 1979).

Eine im Vergleich zu dem seit den 1930er-Jahren diskutierten Kritischen Rationalismus deutlich jüngere Wissenschaftstheorie ist der in den 1970er-Jahren entwickelte **wissenschaftstheoretische Strukturalismus** (auch als **Non-Statement-View** bezeichnet), der auf Natur- und Sozialwissenschaften angewendet wird (Balzer, Moulines, & Sneed, 1987; Sneed, 1971; Stegmüller, 1979). Kennzeichnend und namensgebend für diesen wissenschaftstheoretischen Ansatz ist, dass Theorien, Befunde und Forschungsprozesse formalisiert mit **mathematischen Strukturen** (Mengenlehre) dargestellt werden. Darüber hinaus unterscheidet sich der Strukturalismus maßgeblich vom Kritischen Rationalismus, indem er Theorien nicht als rein sprachliche Aussagensysteme betrachtet, sondern den **Anwendungs- und Empirie-Bezug jeweils als Teil der Theorie** begreift. Dementsprechend enthält eine Theorie gemäß Strukturalismus als Strukturelemente neben ihren Kernannahmen auch Aussagen über Anwendungen, Daten und Hilfstheorien (z. B. Messtheorien). Wie sich die psychologische Experimentalforschung mithilfe des Strukturalismus wissenschaftstheoretisch begründen lässt, wurde von dem Psychologen Rainer Westermann (2000) dargestellt.

Wissenschaftstheoretische Debatten werden nicht zuletzt auch durch **Probleme in der Wissenschaftspraxis** befruchtet. Spektakuläre Fälschungen und Manipulationen im Umgang mit Daten in der Psychologie (▶ Abschn. 4.2.2) haben z. B. eine Debatte rund um die Verbreitung von **theoretisch und statistisch falsch-positiven Befunden** angestoßen. Kritiker monieren, dass es viel zu leicht möglich sei, Daten frei zu erfinden oder erhobene Daten so zu behandeln, dass am Ende Wunschergebnisse zustande kommen. Offenbar werden die Modalitäten der Datenerhebung und Datenauswertung oft nicht transparent genug berichtet und Einzelbefunde in der Praxis viel zu selten wiederholten Falsifikationsversuchen ausgesetzt. Detailliertere Offenlegung des methodischen Vorgehens sowie eine groß angelegte Replikations-Initiative sollen diesen Missständen entgegnen wirken (▶ Abschn. 4.2.2). Gleichzeitig wird aber auch darauf hingewiesen, dass in der Psychologie viele **statistisch falsch-negative Befunde** verbreitet sind, weil ein Großteil von quantitativen Studien aufgrund zu kleiner Stichprobenumfänge (oder anderer Probleme wie z. B. mangelnder Kontrolle von Störvariablen) eine zu geringe Teststärke aufweisen und somit von vorne herein keine statistisch signifikanten Ergebnisse liefern können (▶ Abschn. 14.1). Gegen diese statistisch falsch-negativen Befunde, die im Zusammenhang mit der Hypothesenprüfung im Begründungszusammenhang (▶ Abschn. 2.2.6) auftreten, wird z. B. die Planung von Studien mit optimalen Stichprobenumfängen empfohlen (▶ Abschn. 14.3).

Daneben existieren **theoretisch falsch-negative Befunde**, die dadurch zustande kommen, dass gültige theoretische Alternativerklärungen zur eigenen Forschungs-

hypothese (die das Phänomen womöglich viel besser und umfassender erklären) gar nicht erst in Erwägung gezogen werden (Fiedler, Kutzner, & Krueger, 2012, S. 663). So erklärt die Terror Management Theory, dass die Konfrontation mit Symbolen des Todes (z. B. Bestattungsunternehmen, Leichenwagen) dazu führt, dass Menschen konservativer denken (sich z. B. auf kulturelle Traditionen und Familienwerte besinnen), um die aktivierte Sterblichkeitsangst zu bewältigen. Aber vielleicht wird durch diese sehr enge Theorie eine umfassendere Theorie fälschlich negiert. Vielleicht würden Symbole des Lebens (Geburt, Naturwunder) dieselben Effekte zeigen. Vielleicht geht es bei den Todessymbolen gar nicht um Sterblichkeitsangst, sondern allgemeiner um die Frage nach dem Lebenssinn (vgl. Fiedler et al., 2012, S. 664f.)? Eine derartige Alternativtheorie mit breiterem Anwendungsbereich, die dann eben dazu führen würde, andere Typen von Stimuli in Untersuchungen einzubeziehen, muss im Vorfeld der Studienplanung im Zuge der Theoriearbeit entwickelt werden.

Die Problematik theoretisch falsch-negativer Befunde im Sinne des Übersehens der besseren theoretischen Erklärung betrifft also den Entdeckungszusammenhang (► Abschn. 2.2.6) und sollte dazu veranlassen, im Kritischen Rationalismus die Theoriebildung stärker wissenschaftstheoretisch zu reflektieren und z. B. systematische Theoriearbeit in der Lehre dem akademischen Nachwuchs umfassender zu vermitteln.

Wissenschaftstheoretische Überlegungen auf der Metaebene sind notwendig, wenn es darum geht, Maßnahmen gegen falsch-positive sowie gegen falsch-negative Befunde auszubalancieren und die Folgen für den wissenschaftlichen Erkenntnisgewinn insgesamt im Blick zu behalten. In dieser Debatte wird in der Psychologie u. a. die Position vertreten, dass gute Wissenschaft sich vor allem dadurch auszeichnet, dass zu einem interessierenden Forschungsproblem im Entdeckungszusammenhang konkurrierende Theorien entwickelt, unter Beachtung von Alternativerklärungen kritisch diskutiert und schließlich exakt ausformuliert werden. Im Begründungszusammenhang sind diese Theorien in einem **experimentum crucis** auf den Prüfstand zu stellen, indem ihre unterschiedlichen Prognosen des Studienergebnisses mit den Daten konfrontiert werden (Fiedler et al., 2012, S. 667). Hier zeigt sich in rezenten wissenschaftstheoretischen Debatten das grundlegende Bekenntnis zum Kritischen Rationalismus und seinem Falsifikationsprinzip ebenso wie die Anforderung, die klassischen wissenschaftstheoretischen Grundlagen immer wieder neu im Licht aktueller Beispiele und Probleme aus der Forschungspraxis zu diskutieren und zu akzentuieren (hier z. B. hinsichtlich der vorrangigen Bedeutung von Theoriearbeit im Entdeckungszusammenhang).

2.3 Das qualitative Paradigma und der Sozialkonstruktivismus

Während das quantitative Paradigma der empirischen Sozialforschung sich am naturwissenschaftlich-analytischen Vorgehen orientiert und z. B. eine exakte Definition und Messung all derjenigen Merkmale der sozialen Wirklichkeit verlangt, deren Ausprägungen und Verknüpfungen (v. a. Ursache-Wirkungs-Relationen) untersucht werden sollen, steht das qualitative Paradigma in der **Tradition der Geisteswissenschaften**. Es zielt primär auf eine verstehend-interpretative Rekonstruktion sozialer Phänomene in ihrem jeweiligen Kontext ab, wobei es vor allem auf die Sichtweisen und Sinngebungen der Beteiligten ankommt, also darauf, was ihnen wichtig ist, welche Lebenserfahrungen sie mitbringen und welche Ziele sie verfolgen etc. Beide Ansätze unterscheiden sich somit in ihrem Verständnis von Sozialwissenschaft, d. h. in ihren epistemologischen, ontologischen und axiologischen Annahmen zum wissenschaftlichen Vorgehen, was sich nicht zuletzt in einem anderen Rollenverständnis der Forschenden sowie in einem anderen Ablauf des Forschungsprozesses niederschlägt. Dementsprechend arbeiten beide Paradigmen mit unterschiedlichen Untersuchungsdesigns, Stichprobentypen, Datenerhebungs- und Datenauswertungsverfahren. Der statistischen Analyse von Messwerten im quantitativen Paradigma steht im qualitativen Paradigma die Interpretation von nicht-numerischem, meist sprachlich kodiertem (dabei häufig narrativ verfasstem, aber auch durch Beobachtung konstituiertem) Datenmaterial gegenüber.

Bevor wir Grundzüge des wissenschaftstheoretischen Fundaments qualitativer empirischer Sozialforschung skizzieren, sei noch einmal betont, dass es sich bei dem so genannten qualitativen Paradigma um einen **Sammelbegriff** für eine ausgesprochen heterogene Menge unterschiedlicher expliziter und impliziter Wissenschaftstheorien, Forschungsstrategien und Denkstile handelt, die verschiedenen Fachdisziplinen, historischen Perioden und nationalen Wissenschaftstraditionen entstammen (vgl. Breuer, 2010, S. 37; Lincoln, 2010; Reichertz, 2007). Im Unterschied zum quantitativen Paradigma, in dem der Kritische Rationalismus zentraler Bezugspunkt ist, lässt sich ein so einheitlicher Bezugsrahmen im qualitativen Paradigma nicht benennen. Da es aber auch zu umfassend wäre, die unterschiedlichen wissenschaftstheoretischen Positionen innerhalb des qualitativen Paradigmas (z. B. Hermeneutik, Dialektik, Phänomenologie, Praxeologie, Poststrukturalismus, Postmodernismus, Grounded-Theory-Methodologie) jeweils einzeln auszuführen und in ihren Gemeinsamkeiten und Differenzen einander ge-

genüber zu stellen, arbeitet man in der Darstellung der wissenschaftstheoretischen Grundlagen qualitativer Sozialforschung üblicherweise mit abstrahierten Grundannahmen und Grundprinzipien, die von den verschiedenen qualitativen Richtungen weitgehend geteilt werden und die wir im Folgenden aufgreifen werden.

Ein Vorschlag, bei aller Heterogenität das **Grundverständnis des qualitativen Paradigmas** pointiert zusammenzufassen, stammt von Ernst von Kardorff (1995, S. 4):

> » Der kleinste gemeinsame Nenner der qualitativen Forschungstraditionen lässt sich vielleicht wie folgt bestimmen: Qualitative Forschung hat ihren Ausgangspunkt im Versuch eines vorrangig deutenden und sinnverstehenden Zugangs zu der interaktiv „hergestellt" und in sprachlichen wie nicht-sprachlichen Symbolen repräsentiert gedachten sozialen Wirklichkeit. Sie bemüht sich dabei, ein möglichst detailliertes Bild der zu erschließenden Wirklichkeitsausschnitte zu liefern. Dabei vermeidet sie so weit wie möglich, bereits durch rein methodische Vorentscheidungen den Bereich möglicher Erfahrung einzuschränken oder rationalistisch zu „halbieren". Die bewusste Wahrnehmung und Einbeziehung des Forschers und der Kommunikation mit den „Beforschten" als konstitutives Element des Erkenntnisprozesses ist eine zusätzliche, allen qualitativen Ansätzen gemeinsame Eigenschaft: Die Interaktion des Forschers mit seinen „Gegenständen" wird systematisch als Moment der „Herstellung" des „Gegenstandes" selbst reflektiert.

Die Übereinstimmungen zwischen den verschiedenen qualitativen Forschungsstrategien und Wissenschaftsverständnissen werden angesichts ihrer starken Ausdifferenzierung („Vielfalt ohne rechte Einheit") teilweise noch vorsichtiger eingeschätzt (Reichertz, 2007, S. 197):

> » Allenfalls kann man sagen, dass bei fast allen qualitativen Ansätzen „Sinn" (Gegenstand der Forschung) und „Rekonstruktion" (Anspruch der Forschung) als „Grundtöne" immer wieder aufklingen. Weil das so ist, macht es aus meiner Sicht keinen Sinn, von der Qualitativen Sozialforschung zu sprechen, sondern, wenn überhaupt, sollte man das Ganze das Feld der qualitativen Methoden nennen, in dem die Hinwendung zum menschlich erzeugten „Sinn" immer wieder zu hören ist und in dem diese Orientierung als wesentliche Währung gehandelt wird. Zusätzlich soll gelten, dass sie empirisch arbeiten und ihre Theoriebildung auf die Daten beziehen, sie möglicherweise sogar aus ihnen emergieren lassen wollen, dass sie also etwas „Sinnhaftes" (re-)konstruieren wollen.

Die folgende Darstellung von **fünf wissenschaftstheoretischen Grundprinzipien** des qualitativen Paradigmas orientiert sich an entsprechenden Strukturierungen einschlägiger Lehr- und Handbuchbeiträge (vgl. Breuer, 2010, S. 37ff.; Kruse, 2013b; Lamnek, 2010, S. 19ff.). Für differenzierte Übersichten unterschiedlicher Wissenschaftstheorien im qualitativen Paradigma mit zahlreichen Verweisen auf die entsprechenden Primärquellen siehe z. B. Denzin und Lincoln (2005) oder Creswell (2013) sowie für eine kritische Bestandsaufnahme der qualitativen Forschung in Deutschland Reichertz (2007) und die zusammen mit seinem Beitrag veröffentlichten Kommentare. Reflexionen und Kontroversen über die wissenschaftstheoretischen Grundlagen des qualitativen Paradigmas sind u. a. auch der Open-Access-Zeitschrift *Forum Qualitative Sozialforschung FQS* (http://www.qualitative-research.net) zu entnehmen.

2.3.1 Erstes Prinzip: Ganzheitliche und rekonstruktive Untersuchung lebensweltlicher Phänomene

Wenn die Sozial- und Humanwissenschaften die soziale Erfahrungswirklichkeit untersuchen wollen, stellt sich auf wissenschaftstheoretischer Ebene die ontologische Frage nach der Beschaffenheit dieser sozialen Wirklichkeit. Auf der Basis entsprechender Annahmen zum Wesen des Menschen sowie des zwischenmenschlichen Zusammenlebens ist dann eine geeignete Forschungsstrategie zu wählen. Im qualitativen Paradigma wird davon ausgegangen, dass Menschen sich und ihre jeweilige Umwelt reflektieren, im Austausch miteinander Sinnzuschreibungen vornehmen, sich selbst und ihr Umfeld durch ihre Handlungen zielgerichtet gestalten und verändern. Als Zentrum menschlichen Erlebens und Handelns wird die **alltägliche Lebenswelt** angesehen, die durch ihre jeweiligen Akteurinnen und Akteure, durch soziale Regeln und Normen sowie kulturelle und historische Einflüsse geprägt ist. Das Konzept der Lebenswelt wurde ursprünglich von dem Begründer der Phänomenologie Edmund Husserl (1859–1938) eingeführt und für die Sozialwissenschaften maßgeblich von dem österreichischen Soziologen Alfred Schütz (1899–1959) ausgearbeitet (Schütz, 1932; Schütz & Luckmann, 1975/2003); diverse Weiterentwicklungen des Konzeptes folgten.

> » Die Wissenschaften, die menschliches Handeln und Denken deuten und erklären wollen, müssen mit einer Beschreibung der Grundstrukturen der vorwissenschaftlichen, für den – in der natürlichen Einstellung verharrenden – Menschen selbstverständlichen Wirk-

> lichkeit beginnen. Diese Wirklichkeit ist die alltägliche Lebenswelt. Sie ist der Wirklichkeitsbereich, an der der Mensch in unausweichlicher, regelmäßiger Wiederkehr teilnimmt. Die alltägliche Lebenswelt ist die Wirklichkeitsregion, in die der Mensch eingreifen und die er verändern kann, indem er in ihr durch die Vermittlung seines Leibes wirkt. Zugleich beschränken die in diesem Bereich vorfindlichen Gegenständlichkeiten und Ereignisse, einschließlich des Handelns und der Handlungsergebnisse anderer Menschen, seine freien Handlungsmöglichkeiten. Sie setzen ihm zu überwindende Widerstände wie auch unüberwindliche Schranken entgegen. Ferner kann sich der Mensch nur innerhalb dieses Bereichs mit seinen Mitmenschen verständigen, und nur in ihm kann er mit ihnen zusammenwirken. Nur in der alltäglichen Lebenswelt kann sich eine gemeinsame kommunikative Umwelt konstituieren. Die Lebenswelt des Alltags ist folglich die vornehmliche und ausgezeichnete Wirklichkeit des Menschen. Unter alltäglicher Lebenswelt soll jener Wirklichkeitsbereich verstanden werden, den der wache und normale Erwachsene in der Einstellung des gesunden Menschenverstandes als schlicht gegeben vorfindet. Mit schlicht gegeben bezeichnen wir alles, was wir als fraglos erleben, jeden Sachverhalt, der uns bis auf weiteres unproblematisch ist. (Schütz & Luckmann, 1975/2003, S. 29).

Die sozialen Phänomene, die es zu untersuchen gilt, zeichnen sich nach diesen ontologischen Vorannahmen dadurch aus, dass sie durch Kommunikation und Interaktion hervorgebracht werden (zum Sozialkonstruktivismus ► Abschn. 2.3.6) und dabei von den Besonderheiten der beteiligten Individuen und ihrer jeweiligen Lebenswelt geprägt sind. Daraus wiederum ergibt sich die Anforderung an sozialwissenschaftliche Forschung, eine holistische bzw. **ganzheitliche Untersuchung lebensweltlicher Phänomene** anzustreben und dabei die Sicht- und Handlungsweisen der Beteiligten detailliert zu beschreiben und nachzuvollziehen. Dies kann vor allem dann gelingen, wenn die Forschenden mit den Beforschten in Kommunikation treten und deren Lebenswelten genau kennenlernen, also z. B. ihren Alltag teilnehmend beobachten oder sich ihre Lebensgeschichten erzählen lassen, anstatt z. B. mit künstlichen Laborsituationen und/oder standardisierten Fragebögen zu arbeiten und jeweils auf der Basis vorformulierter Theorien nur einzelne Aspekte des sozialen Geschehens zu analysieren, wie das im quantitativen Paradigma getan wird. Die ganzheitliche Herangehensweise im qualitativen Paradigma wird als Gegenmodell zu der als reduktionistisch abgelehnten analytisch-variablenzentrierten Vorgehensweise des quantitativen Paradigmas verstanden.

Die ganzheitliche Untersuchung auf der Ebene der Lebenswelten führt im Ergebnis zu **zeit- und orts-, (sub-)kultur- sowie personengebundenen Erkenntnissen**. Denn Lebenswelten werden durch die interagierenden Individuen kontextspezifisch gestaltet und befinden sich in ständigem Wandel, so dass ihre wissenschaftliche Rekonstruktion jeweils für sich steht und nicht oder nur beschränkt verallgemeinerbar ist. Aus dem **Menschenbild** von reflektierten und handlungsmächtigen Individuen und der durch sie geprägten vielfältigen, komplexen und dynamischen Lebenswelten wird die Schlussfolgerung gezogen, dass einzelne Aspekte menschlichen Erlebens und Verhaltens grundsätzlich nicht durch allgemeine Gesetzmäßigkeiten erklärbar sind, wie das im quantitativen Ansatz angenommen wird. Eine solche Vorstellung wird als mechanistisches Menschenbild abgelehnt.

In Bezug auf ihre Lebenswelten wird den Akteurinnen und Akteuren mitunter ein **Expertenstatus** zugeschrieben. Dieses Expertenwissen der Beteiligten gilt es im Verständnis qualitativer Sozialforschung zu erschließen, so dass Forschende vor allem auch die Rolle von Lernenden und Kommunikationspartnern einnehmen anstatt sozusagen über die Köpfe der Beteiligten hinweg zu agieren und mit vorgefertigten Theorien und Methoden eine vermeintlich objektive, wissenschaftliche Sichtweise den sozialen Phänomenen überzustülpen.

Prinzip der ganzheitlichen und rekonstruktiven Untersuchung lebensweltlicher Phänomene – Ontologisch wird die im qualitativen Paradigma empirischer Sozialforschung untersuchte soziale Wirklichkeit auf der Ebene der alltäglichen Lebenswelt der Individuen angesiedelt. Denn menschliches Erleben und Handeln lässt sich im Verständnis des qualitativen Paradigmas nicht durch allgemeine Gesetzmäßigkeiten kausal erklären, sondern nur anhand der jeweils individuellen Weltsichten und Sinngebungen der Beteiligten in ihren jeweiligen Alltagszusammenhängen verstehen und intentional erklären. Durch diese Gegenstandsbestimmung befinden sich nicht-lebensweltliche Makro- und Mikro-Phänomene außerhalb des Fokus der qualitativen Sozialforschung.

Mit dem Fokus auf die Lebenswelt und deren Rekonstruktion wird in der qualitativen Sozialforschung der Gegenstandsbereich bzw. das **Spektrum der untersuchbaren Forschungsprobleme deutlich eingeschränkt**. Denn all jene Strukturen oder Einflussgrößen, die in der sozialen Wirklichkeit eine Rolle spielen, aber nicht das direkte Handeln und Wahrnehmen der Beteiligten betreffen und/oder ihnen nicht bewusst sind, bleiben bei einer Rekonstruktion ihrer Handlungs- und Denkweisen definitionsgemäß ausgeklammert. Das betrifft z. B. auf der Makroebene ökologische Parameter des Klimawandels sowie auf der Mikroebene neurologische Prozesse. Qualitativer Forschung zugänglich sind lediglich Diskurse über derartige Phänomene, zu deren Erfassung Messungen und entsprechende Messgeräte notwendig sind.

Die detaillierte Rekonstruktion von Lebenswelten unter Berücksichtigung der Besonderheiten der jeweils beteiligten Individuen wirft die Frage nach der Verallgemeinerbarkeit der Befunde auf. **Generalisierbares Wissen** zu erzeugen wird auch im qualitativen Paradigma als Ziel betrachtet (Lincoln, 2010). Hierbei wird Generalisierbarkeit aber nicht wie im quantitativen Paradigma z. B. über eine ausschnittweise Betrachtung sozialer Phänomene anhand repräsentativer Stichproben angezielt, sondern durch eine Abstraktion der Befunde einzelner qualitativer lebensweltbezogener Studien. Möglich ist es z. B. durch Typenbildung über Individuen zu generalisieren oder bei der Theoriebildung Kontextfaktoren ausdrücklich zu berücksichtigen, so dass Einschätzungen darüber möglich sind, auf welche bislang nicht untersuchten Lebenswelten sich die Befunde übertragen lassen oder nicht.

Beispiel: Anhand von qualitativer Forschung in drei verschiedenen Unternehmen werden Prozesse der Diskriminierung gegenüber Erwerbstätigen mit Migrationshintergrund am jeweiligen Arbeitsplatz und im jeweiligen Kollegenkreis rekonstruiert (z. B. unterschiedliche Arten von Misstrauensausdruck, rassistische Witze). Die qualitative Forschung kann anhand der Fallstudien bestimmte Diskriminierungsmuster herausarbeiten und auch Kontextfaktoren benennen, so dass Anhaltspunkte vorliegen, auf welche nicht-untersuchten Arbeitsplatzbedingungen die Muster übertragbar sein sollten. Was diese Form der Generalisierung qualitativer Befunde jedoch nicht leisten kann, sind Aussagen darüber, wie verbreitet welche Form der Diskriminierung in unterschiedlichen Branchen unterschiedlicher Nationen ist. Hierfür wären dann quantitative Studien indiziert, die aus der Population der Unternehmen verschiedener Branchen eines oder mehrerer Länder möglichst repräsentative Stichproben ziehen, Indikatoren von Diskriminierung erfassen und statistisch im Überblick darstellen.

2.3.2 Zweites Prinzip: Reflektierte theoretische Offenheit zwecks Bildung neuer Theorien

Die verschiedenen wissenschaftstheoretischen Ansätze innerhalb der qualitativen Sozialforschung stimmen weitgehend darin überein, dass wissenschaftlicher Erkenntnisgewinn vor allem auf die **Bildung genuin neuer Hypothesen und Theorien** über den interessierenden Untersuchungsgegenstand ausgerichtet sein sollte. Mit dem Fokus auf Theoriebildung (statt Theorieprüfung) wird somit das Hauptarbeitsfeld der qualitativen Sozialforschung im **Entdeckungszusammenhang** (und nicht wie im Kritischen Rationalismus im Begründungszusammenhang) verortet (► Abschn. 2.2.6).

Als Quelle neuen Wissens wird in der qualitativen Sozialforschung vor allem die detaillierte und möglichst unverzerrte Beschreibung und Rekonstruktion des interessierenden sozialen Phänomens angesehen. Die A-priori-Festlegung auf eine einzelne Theorien gilt als hinderlich: Denn entweder ist diese vorformulierte Theorie zutreffend, dann brächte ihre Bestätigung kein neues, sondern nur redundantes Wissen hervor. Oder sie ist unzutreffend, dann würde sie die Forschung in die falsche Richtung lenken. Um also der Zielsetzung der innovativen Theoriebildung nachzukommen wird im qualitativen Paradigma eine durch das **Prinzip der theoretischen Offenheit** geprägte Annäherung an den Gegenstand – und das heißt in der qualitativen empirischen Sozialforschung eben meist: an die Untersuchungspersonen in ihren alltäglichen Lebenswelten – verlangt. Eine gänzlich theoriefreie Vorgehensweise im Sinne **absoluter Offenheit** ist dabei unmöglich. Denn jeder Wahrnehmungsakt bei einer teilnehmenden Feldbeobachtung oder jeder Kommunikationsakt bei einem Leitfaden-Interview setzt erkenntnistheoretisch auf Seiten der Forschenden bestimmte theoretische Vorannahmen voraus. Andernfalls könnten sie das relevante Feld gar nicht identifizieren, sich dort gar nicht orientieren und gar nicht sinnverstehend mit den Feldmitgliedern kommunizieren. Das Prinzip der Offenheit kann also nur eine **relative** bzw. **reflektierte Offenheit** fordern (vgl. Breuer, 2010). Das bedeutet: Eine dem Forschungsprozess vorgelagerte Festlegung auf eine bestimmte, den weiteren Forschungsprozess dann vorstrukturierende und steuernde wissenschaftliche Theorie über den Gegenstand soll ausdrücklich vermieden werden.

Wohl aber darf und soll man sich über den zu untersuchenden Sachverhalt vorinformieren und sich auch die eigenen, im Zusammenhang mit der Studie relevanten Vorerfahrungen und Vorannahmen bewusst machen. Diese fließen in die Forschung ein und können z. B. ein erstes Verständnis des Gegenstandes ermöglichen. Gemäß dem Prinzip der Offenheit müssen die Forschenden sich aber dann aktiv darum bemühen, Datenmaterial zu sammeln, das über ihre Vorannahmen hinausgeht oder diese infrage stellt und somit neue Perspektiven für die angestrebte Theoriebildung eröffnet. Sofern mit theoretischen Konstrukten gearbeitet wird, dürfen diese nicht vorab definitorisch festgelegt, sondern in Anlehnung an Blumer (1954) lediglich als **sensibilisierende Konzepte** („sensitizing concepts“) verstanden werden (vgl. Kruse, 2013c). Für eine Bedeutungserweiterung oder Bedeutungsveränderung dieser sensibilisierenden Konzepte im Zuge der Forschung muss man offen bleiben (zu theoretischen Vorannahmen in der Grounded-Theory-Methodologie ► Abschn. 10.6.2).

Prinzip der reflektierten theoretischen Offenheit zwecks Bildung neuer Theorien – Erkenntnisgewinn zielt im Paradigma der qualitativen Sozialforschung vor allem auf die Bildung neuer Hypothesen und Theorien ab, die passgenau auf den jeweiligen Untersuchungsgegenstand (d. h. ein lebensweltliches soziales Phänomen) zugeschnitten sind. Deswegen sollen sich die Forschenden eben nicht mit einer bereits im Vorfeld ausgewählten fertigen Theorie und vordefinierten theoretischen Konstrukten mit einem durchstrukturierten Forschungsprozess dem Gegenstand nähern. Stattdessen sollen sie ihr (immer in gewisser Weise vorhandenes und auch notwendiges) Vorverständnis kritisch hinterfragen und ausdrücklich offen sein für Fälle, Situationen und Daten, die ihren Vorannahmen widersprechen und somit über die bisherigen theoretischen Vorstellungen hinaus neue Kenntnisse über den Untersuchungsgegenstand vermitteln.

Innerhalb des Feldes qualitativer Methoden wird das Prinzip der theoretischen Offenheit unterschiedlich ausgelegt und umgesetzt. Das Spektrum reicht von Haltungen, die auf möglichst große Unvoreingenommenheit durch weitgehende Theorieabstinenz abzielen, um neue Hypothesen direkt aus den Daten zu schöpfen, bis zu Ansätzen, die mit der Explizierung umfangreicher theoretischer Überlegungen in den Forschungsprozess starten, um diese im Zuge der Datenerhebung infrage zu stellen und zu modifizieren.

2.3.3 Drittes Prinzip: Zirkularität und Flexibilität des Forschungsprozesses zwecks Annäherung an den Gegenstand

In der qualitativen Sozialforschung sollen maßgeschneiderte neue Theorien über soziale Phänomene in alltäglichen Lebenswelten generiert werden, die ein möglichst ganzheitliches Verständnis des Geschehens ermöglichen. Dabei werden nicht nur theoretische Festlegungen im Vorfeld der Datenerhebung abgelehnt (▶ Abschn. 2.3.2), sondern auch vorstrukturierte methodische Abläufe und standardisierte Datenerhebungsinstrumente, wie sie im quantitativen Paradigma zur Theorieprüfung genutzt werden, als Hindernisse beim Erkenntnisgewinn betrachtet. Denn im qualitativen Ansatz steht eben nicht die Theorieprüfung, sondern die Theoriebildung im Mittelpunkt. Und diese soll durch eine theoretische und methodische Strukturierung – die letztlich auf eine Engführung beim Erkunden und Entdecken sozialer Phänomene in der untersuchten Lebenswelt hinauslaufen würde – nicht behindert werden.

Analog der theoretischen Offenheit wird somit **methodische Flexibilität** verlangt. Der Forschungsprozess als solcher wird nicht vorab detailliert durchgeplant und dann linear abgearbeitet, sondern idealerweise flexibel an den sich entfaltenden Erkenntnisprozess angepasst: In Abhängigkeit von den Zwischenergebnissen einer Studie können z. B. gezielt zusätzliche Fälle in die Stichprobe aufgenommen (theoretische Stichprobenziehung ▶ Abschn. 9.2.1) oder neue Fragen in einen Interview-Leitfaden integriert werden. Indem der Forschungsprozess einer einzelnen qualitativen Studie in **mehreren Zyklen** die Phasen der Untersuchungsplanung, Datenerhebung und Datenanalyse durchläuft und die Vorgehensweise jeweils nachkorrigiert, können mittlerweile entdeckte Irrtümer, Ungenauigkeiten oder Lücken im Verständnis des Gegenstandes auf Seiten der Forschenden schrittweise behoben werden.

Das hier angesprochene Prinzip wird in der Literatur teilweise als „**Prinzip der Prozessualität**“ bezeichnet oder es wird auf den „Prozesscharakter“ qualitativer Forschung verwiesen (vgl. Kruse, 2013a; Lamnek, 2010, S. 21f.). Da jedoch empirische Forschung auch im quantitativen Paradigma prozesshaft verläuft, ist der Verweis auf die Zirkularität und Flexibilität des qualitativen Forschungsprozesses präziser. Dabei ist jedoch zu beachten, dass auch der oft verwendete Begriff der **Zirkularität** noch unscharf ist: Denn beim Durchlaufen eines qualitativen Forschungsprozesses werden nicht dieselben Zyklen wiederholt, sondern durch flexible Adjustierung des Vorgehens soll bei jedem Zyklus die Gegenstandsnähe erhöht werden, so dass man sich idealerweise **spiralförmig** dem Gegenstand annähert. Dass sich trotzdem nicht die Spiralförmigkeit, sondern die Zirkularität als Kennzeichnung eingebürgert hat, liegt daran, dass sie auf den begrifflich etablierten (auf den Philosophen Martin Heidegger, 1927/2006, zurückgehenden) Zirkel des Verstehens – den **hermeneutischen Zirkel** („hermeneutic circle“, „hermeneutical circle“) – als schrittweise Annäherung einer Textinterpretation an die Textbedeutung verweist:

> » Wer einen Text verstehen will, vollzieht immer ein Entwerfen. Er wirft sich einen Sinn des Ganzen voraus, sobald sich ein erster Sinn im Text zeigt. Ein solcher zeigt sich wiederum nur, weil man den Text schon mit gewissen Erwartungen auf einen bestimmten Sinn hin liest. Im Ausarbeiten eines solchen Vorentwurfs, der freilich beständig von dem revidiert wird, was sich beim weiteren Eindringen in den Sinn ergibt, besteht das Verstehen dessen, was da steht. (Gadamer, 1960/1990, S. 271)

Gemäß hermeneutischem Zirkel wird die Bedeutung eines Textes erschlossen, indem man auf der Basis von Vorkenntnissen und Erwartungen beim ersten Lesen ein Grobverständnis entwickelt und dieses im Zuge des wiederholten Durcharbeitens des Textes immer weiter verfeinert, wobei man die Textteile im Kontext des Textganzen und das Textganze im Licht der Textteile deutet, insofern ist auch der hermeneutische Zirkel eigentlich eine **hermeneutische Spirale**. Der hermeneutische Zirkel wird

oft als spezifisches Verfahren des Erkenntnisgewinns in den Geisteswissenschaften und der qualitativen Sozialforschung betrachtet.

Prinzip der Zirkularität und Flexibilität des Forschungsprozesses zwecks Annäherung an den Gegenstand – Der qualitative Forschungsprozess wird nicht vorab vollständig durchgeplant, festgelegt und dann linear abgearbeitet, sondern idealerweise flexibel gestaltet: In mehreren Untersuchungszyklen innerhalb einer Studie sollen anhand der Zwischenergebnisse jeweils das Untersuchungsdesign, die Auswahl der Fälle und/oder die Datenerhebungs- und Datenanalyse-Hilfsmittel überarbeitet (Flexibilität des Vorgehens) und dem untersuchten Gegenstand somit schrittweise immer besser angepasst werden (Zirkularität bzw. Spiralförmigkeit der verstehenden Annäherung an den Untersuchungsgegenstand).

In der qualitativen Forschungspraxis wird das Prinzip der Zirkularität und Flexibilität primär im Kontext der Datenanalyse umgesetzt, indem im Zuge der interpretierenden Auswertung des qualitativen Datenmaterials nach und nach Codes gebildet und verfeinert werden (zur qualitativen Datenanalyse ▶ Abschn. 12.1). Bei vielen qualitativen Interviewstudien werden in der Praxis dagegen Fallauswahl und Interview-Leitfaden nicht zirkulär entwickelt, sondern im Vorfeld festgelegt (zur Kluft zwischen dem Anspruch theoretischer Stichprobenbildung und der Forschungspraxis ▶ Abschn. 9.2.1). Entsprechende **Vorplanungen und lineare Abläufe** haben dabei oft organisatorische und forschungsökonomische Gründe: So müssen z. B. für die Zeit- und Budgetplanung bei drittmittelfinanzierten Studien Anzahl und Rekrutierungswege der untersuchten Fälle vorab kalkuliert werden. Sofern die Datenerhebung parallel durch verschiedene Interviewende erfolgt, sind die Möglichkeiten einer schrittweisen Verbesserung des Leitfadens ebenfalls limitiert. Der Verzicht auf eine verbesserte Annäherung an den Gegenstand durch forschungsökonomisch bedingte Einschränkungen der methodischen Flexibilität stellt im Verständnis des qualitativen Paradigmas eine Qualitätseinbuße dar, deren Konsequenzen für die jeweilige Studie zu bedenken und zu diskutieren sind.

2.3.4 Viertes Prinzip: Forschung als Kommunikation und Kooperation zwischen Forschenden und Beforschten

In der qualitativen Sozialforschung wird Forschungsaktivität primär als Kommunikation zwischen Forschenden und Beforschten verstanden, d. h. die Daten werden meist im interaktiven Kontakt mit dem Untersuchungsfeld hervorgebracht: Sei es, dass Forschende mit Untersuchungspersonen in einen direkten oder medienvermittelten (telefonischen, computervermittelten) Dialog eintreten (z. B. im Zuge von Feldgesprächen, die Feldbeobachtungen ergänzen, ▶ Abschn. 10.1.2, oder durch unterschiedliche Typen von qualitativen Interviews, ▶ Abschn. 10.2.2). Das **direkte Gespräch von Angesicht zu Angesicht** wird meist als eine Form der qualitativen Datenerhebung betrachtet, die den Prozess des Fremdverstehens besonders fördert, da Forschende einen ganzheitlichen Eindruck von den Interviewpersonen erhalten, verständnissichernde nonverbale Signale verfügbar, spontane Rückfragen möglich und oft Hintergrundinformationen (z. B. über das Wohnumfeld oder den Arbeitsplatz, wo das Interview stattfindet) zugänglich sind. **Medienvermittelte qualitative Interviews** können jedoch unter bestimmten Umständen vorteilhafter sein (z. B. weil online anderweitig nicht erreichbare Zielgruppen kontaktierbar sind und mediale Enthemmung eine offenere Kommunikation über schambesetzte Inhalte ermöglicht, ▶ Abschn. 10.2.2).

Daneben werden in der qualitativen Sozialforschung aber auch **vorgefundene Dokumente** analysiert, die nicht der Kommunikation zwischen Forschenden und Beforschten entstammen und gar nicht für Forschungszwecke generiert wurden, sondern ohnehin in den alltäglichen Lebenswelten entstehen (z. B. Schulaufsätze von Kindern; Beiträge in Online-Selbsthilfegruppen; persönliche Tagebücher; Urlaubsfotos; Briefkorrespondenz; ▶ Abschn. 10.6.2). Diese Datenquellen werden im qualitativen Paradigma ebenfalls besonders geschätzt, da es sich um genuine (nicht durch den Forschungsprozess erzeugte) Äußerungen in der Lebenswelt handelt. Um die Interpretation dieser Quellen im Sinne der Beteiligten sicherzustellen, wird ergänzend zur qualitativen Dokumentenanalyse teilweise der Kontakt zu den jeweiligen Autorinnen und Autoren hergestellt, so dass diese z. B. die Interpretationen der Forschenden prüfen können (kommunikative Validierung; ▶ Abschn. 3.3). Zudem kann eine sinnrekonstruierende Analyse entsprechender Dokumente teilweise auf weitere Hintergrundinformationen aus der Lebenswelt zurückgreifen (z. B. Schulakten der Schülerinnen und Schüler, deren Schulaufsätze analysiert werden – sofern dies datenschutzrechtlich und forschungsethisch abgesichert ist).

Eine auf Verstehen ausgerichtete Kommunikation zwischen Forschenden und Beforschten verlangt eine **kooperative Haltung**. Von den Beforschten wird erwartet, dass sie sich Zeit nehmen für umfassende Interviews, dass sie Zugang zu ihren Lebenswelten gewähren, ggf. Dokumente aushändigen und den Forschenden bei Verständnisproblemen helfen. Im Unterschied zu standardisierten Datenerhebungen, wie sie im quantitativen Paradigma üblich sind, erfordert die nicht-strukturierte qualitative Vorgehensweise in der Regel einen höheren Grad an aktiver Beteiligung und Vertrauen (z. B. zweistündiges biografisches

Interview im eigenen Wohnzimmer versus zehnminütiges anonymes Ausfüllen eines standardisierten Online-Fragebogens). Das **Verhältnis zwischen Forschenden und Beforschten** ist deswegen im qualitativen Paradigma besonders wichtig: Ein distanziertes, durch Misstrauen oder Machtgefälle geprägtes Verhältnis, das in eine gestörte Kommunikation mündet, würde die wissenschaftlichen Erkenntnismöglichkeiten beschränken. Es würde außerdem dem Menschenbild des qualitativen Ansatzes widersprechen, demgemäß die Beforschten als reflektierte Forschungspartner ernst genommen und nicht von den Forschenden vereinnahmt werden sollen.

Prinzip der Kommunikation und Kooperation zwischen Forschenden und Beforschten – Da der qualitative Ansatz in der Regel auf die Rekonstruktion der Sichtweisen der am untersuchten sozialen Phänomen Beteiligten abzielt, ist die direkte (zuweilen auch medienvermittelte) Kommunikation zwischen Forschenden und Beforschten das zentrale Erkenntnismittel. Damit diese Kommunikation ein möglichst weitreichendes Fremdverstehen ermöglicht, ist eine kooperative Haltung auf beiden Seiten notwendig. Die Beforschten müssen sich vertrauensvoll öffnen und die Forschenden müssen sich persönlich in den Kontakt einbringen und ein dem jeweiligen Gegenüber und dem Forschungsziel angemessenes Arbeitsbündnis aufbauen, dieses reflektieren und dokumentieren.

Wie ein konstruktives und gleichberechtigtes Verhältnis zwischen Forschenden und Beforschten im Einzelnen theoretisch zu fassen und praktisch zu fördern ist, wird kontrovers diskutiert. Häufig wird das dem Forschungsprozess zugrundeliegende **Arbeitsbündnis** als „Rapport" konzeptualisiert (► Abschn. 10.2.2). Auch Konzepte wie Partnerschaftlichkeit, Anwaltschaftlichkeit, Parteilichkeit, Freundschaftlichkeit und auf jeden Fall Respekt werden angeführt. Welches Verhältnis im konkreten Fall im Sinne des Erkenntnisgewinns und der Forschungsethik anzustreben und zu realisieren ist, hängt von vielen Faktoren ab. Da in der qualitativen Forschung aufgrund ihrer sozialkritischen Tradition (Denzin & Lincoln, 2005) oft besonderes Augenmerk auf der Untersuchung marginalisierter Bevölkerungsgruppen und ihrer Lebensbedingungen liegt, zielt der Appell zur Partnerschaftlichkeit in erster Linie darauf ab, dass die durch Bildung und sozialen Status meist privilegierten Forschenden keine machtasymmetrische Situation konstruieren und dem Gegenüber nicht ihr Weltbild überstülpen (**„research down"**-Konstellation mit mehr Macht auf Seiten der Forschenden), sondern wirklich die Sichtweisen der Beteiligten zur Geltung kommen lassen. Dies setzt ehrliches Interesse, Einfühlungsvermögen und Offenheit voraus. Wachsende Bedeutung zur Überwindung einseitiger Definitionsmacht der Forschenden gewinnt auch **partizipative Forschung**, bei der die Beforschten gleichberechtigt in alle Phasen des Forschungsprozesses einbezogen und ggf. auch auf Projektstellen beschäftigt werden (Unger, 2012; zu partizipativen Ansätzen in der Evaluationsforschung ► Abschn. 18.2.4).

Doch nicht nur zu große Distanz, auch **übergroße emotionale Nähe** kann erkenntnistheoretisch problematisch sein, etwa weil sie ihrerseits Kommunikationsblockaden erzeugt. In der ethnografischen Feldforschung wird unreflektierte Identifikation der Forschenden mit dem Untersuchungsfeld als **„going native"** problematisiert (► Abschn. 10.1.2). Zudem kann freundschaftliche Nähe in manchen Forschungskontexten auch gar nicht hergestellt werden: Wie weit kann und soll z. B. das partnerschaftliche Arbeitsbündnis bei Studien in extremistischen und gewaltbereiten Szenen aussehen? Nicht zuletzt ist zu beachten, dass auch in der qualitativen Forschung zahlreiche Forschungskontexte existieren (z. B. in der Wirtschafts- und Organisationsforschung), in denen eher eine **Machtasymmetrie zu Lasten der Forschenden** vorliegt (z. B. durch Alter, Geschlecht, Nationalität oder berufliche Position), so dass sie Gefahr laufen, wiederum von den Beforschten vereinnahmt zu werden, wenn sie nicht genügend Distanz und Widerständigkeit aufbauen (**„research up"**-Konstellation mit weniger Macht auf Seiten der Forschenden; z. B. junge Doktorandin oder junger Doktorand untersucht Geschlechterdiskriminierung in der Arbeitswelt anhand von Interviews mit Führungskräften). Machtasymmetrien zwischen Forschenden und Beforschten im Sinne von „research down"- oder „research up"-Konstellationen können sich nicht nur in den Interaktionen bei der Datenerhebung niederschlagen, sondern auch in Kontroversen rund um die Ergebnisinterpretation und Veröffentlichung von Studienergebnissen (Breuer, 2011a).

Das individuelle Verhältnis zwischen den einzelnen Forschenden zu den Beforschten wird als genuiner Bestandteil des Erkenntnisprozesses betrachtet und soll gerade nicht durch Standardisierung nivelliert werden. Stattdessen stellt es eine notwendige Voraussetzung für das Fremdverstehen dar, muss dafür jedoch gründlich reflektiert und dokumentiert werden, wofür dann wiederum Anleitungen gefragt sind (z. B. Helfferich, 2011). Die **subjektiven Merkmale der Forschenden** wie z. B. Vorerfahrungen, Persönlichkeitseigenschaften und Auftreten, die allesamt das Arbeitsbündnis beeinflussen, werden im qualitativen Paradigma nicht als „Störfaktoren" betrachtet, sondern als Ressourcen, die jeweils spezifische Zugänge zur Bedeutungswelt des Gegenübers eröffnen. Ein von dieser Subjektivität und persönlichen Bezogenheit losgelöster Kommunikations- und Verständigungsprozess wird im qualitativen Paradigma nicht für sinnvoll gehalten.

2.3.5 Fünftes Prinzip: Selbstreflexion der Subjektivität und Perspektivität der Forschenden

Dem Kritischen Rationalismus liegt eine Erkenntnistheorie zugrunde, dergemäß der Wahrheitsgehalt von Theorien anhand des nachprüfbaren und kritisierbaren Vergleichs zwischen einer Theorie und den zugehörigen Beobachtungsdaten sowie zwischen der Theorie und konkurrierenden Theorien erfolgt. Dabei sollte eine **Theorieprüfung** anhand von Daten unabhängig von der Person der Forschenden und ihren Einstellungen zu denselben Ergebnissen führen (objektive, d. h. intersubjektiv nachvollziehbare und replizierbare Theorieprüfung in der Forschungsgemeinschaft).

Demgegenüber wird im qualitativen Paradigma eine Erkenntnistheorie favorisiert, die davon ausgeht, dass wissenschaftlicher Erkenntnisgewinn nicht von der Person der Forschenden und ihren sozialen, kulturellen und individuellen Eigenheiten, Vorerfahrungen und Vorkenntnissen abzulösen ist. Denn der Forschungsprozess im qualitativen Paradigma, der sich auf die **theoriebildende Rekonstruktion** lebensweltlicher sozialer Phänomene bezieht, nicht selten ein temporäres „Eintauchen" in die Lebenswelten der Untersuchten verlangt (z. B. bei ethnografischer Feldbeobachtung) sowie eine intensive persönliche Kommunikation mit den Untersuchungsteilnehmenden beinhaltet (meist in Form mündlicher Forschungsgespräche im privaten oder beruflichen Umfeld der Untersuchungsteilnehmenden), ist immer an die **Subjektivität der Forschenden** geknüpft. Während die Eigenheiten der Forschenden im quantitativen Paradigma bei der Theorieprüfung durch Strukturierung des Vorgehens und Standardisierung der Instrumente sozusagen neutralisiert werden sollen, wird im qualitativen Paradigma gerade eine solche Standardisierung abgelehnt, stattdessen dürfen und sollen sich die Forschenden bei ihrer Theoriebildung individuell jeweils auf den Untersuchungsgegenstand einlassen, da gerade in der Subjektivität die Voraussetzung für Erkenntnisgewinn gesehen wird (Subjektivität als Erkenntnisquelle aus psychoanalytischer Sicht: Devereux, 1976; reflektierte Subjektivität im Kontext der Grounded-Theory-Methodologie: Breuer, 2011b; Breuer, Mruck, & Mey, 2011).

Gefordert wird im Sinne der Wissenschaftlichkeit jedoch eine **genaue Reflexion und Dokumentation der subjektiven Perspektiven der Forschenden** und ihrer Einflüsse auf den Forschungsprozess sowie die Forschungsergebnisse. Diese kritische Selbstreflexion betrifft alle Phasen des Forschungsprozesses, vor allem die Themenwahl, das Untersuchungsdesign, die Datenerhebung, die Datenanalyse und die Ergebnispräsentation. Qualitativ Forschende müssen sich und der Wissenschaftsgemeinschaft darüber Rechenschaft ablegen, welchen eigenen Bezug sie zum Forschungsthema haben, welche positiven wie negativen Vorerfahrungen sie z. B. mitbringen oder welche soziale Stellung sie im Vergleich zu den Untersuchungspersonen einnehmen (vgl. Breuer, 2010, S. 44). Die Subjektivität und Perspektivität der Forschenden kommt z. B. darin zum Tragen, dass sie in unterschiedlicher Weise einen Rapport zu Interviewpartnern aufbauen und dementsprechend unterschiedliche Gespräche führen (Helfferich, 2011) oder dass sie Textdokumente unterschiedlich deuten. Bei der Selbstreflexion der eigenen Subjektivität sollen im qualitativen Paradigma verschiedene Arbeitstechniken helfen: Ein wichtiges Korrektiv sind die Mitforschenden im selben Projekt (mit denen man sich über Vorgehensweisen, Daten und Zwischenergebnisse austauscht) sowie die Untersuchungspersonen, denen Zwischenergebnisse zur Überprüfung vorgelegt werden können (z. B. im Zuge einer kommunikativen Validierung) oder die im Sinne partizipativer Ansätze in alle Forschungsphasen aktiv eingebunden werden können (z. B. Mitentscheidung über Forschungsdesigns), sowie außenstehende Fachkolleginnen und -kollegen, mit denen Abläufe und Ergebnisse eines Projekts diskutiert werden (z. B. in Deutungsgruppen im Rahmen der Grounded-Theory-Methodologie).

Die Selbstreflexion der Subjektivität und Perspektivität der Forschenden zielt nicht nur auf idiosynkratische Vorstellungen und Vorerfahrungen, sondern im Sinne eines im qualitativen Paradigma favorisierten ausdrücklich emanzipatorischen Wissenschaftsverständnisses (Denzin, Lincoln, & Smith, 2008) auch auf die **Machtverhältnisse** zwischen Forschenden und Beforschten. Eine subjektivistische Erkenntnistheorie ist somit auch mit axiologischen Fragen der von den Forschenden vertretenden **Werte** sowie der Verantwortung der Wissenschaft eng verbunden.

Prinzip der Selbstreflexion der Subjektivität und Perspektivität der Forschenden – Das qualitative Paradigma empirischer Sozialforschung vertritt eine subjektivistische Erkenntnistheorie, dergemäß der Erkenntnisprozess immer untrennbar mit der Person der Forschenden – ihrer Subjektivität und Perspektivität – verknüpft ist. Ihr persönliches Verhältnis zum jeweiligen Untersuchungsgegenstand und zu den Untersuchungspersonen müssen Forschende deswegen kritisch reflektieren. Die Reflexion der Subjektivität und Perspektivität gilt als zentrale Erkenntnisquelle und muss entsprechend auch dokumentiert werden. Das betrifft nicht zuletzt die Position und Perspektive der Forschenden im Kontext sozialer Hierarchien und gesellschaftlicher Machtverhältnisse.

Wenn Erkenntnisgewinn nicht als objektivierbare (d. h. durch strukturierte Abläufe und standardisierte Instrumente von der Person der Forschenden abstrahierbare) Annäherung an die Wahrheit, sondern immer als subjektgebundener Verstehens- und Rekonstruktionsprozess

begriffen wird, stellt sich die Frage, aus welcher Perspektive und für welche Interessen Wissen generiert wird bzw. generiert werden sollte. Aus gesellschafts- und wissenschaftskritischem Blickwinkel wird im qualitativen Paradigma hinterfragt, inwiefern sich in den Perspektiven und Standpunkten der Forschenden z. B. gesellschaftlich privilegierte und machterhaltende Tendenzen widerspiegeln und eben die Sichtweisen von gesellschaftlich unterprivilegierten und marginalisierten Gruppen zu kurz kommen. Zwar bemüht sich der qualitative Ansatz um eine Rekonstruktion der Perspektiven der am untersuchten Phänomen Beteiligten und kooperiert mit ihnen (Prinzip der Kommunikation und Kooperation ▶ Abschn. 2.3.4), letztlich werden der Forschungsprozess und die Ergebnispräsentation aber primär von den Forschenden gesteuert. Perspektiven einzunehmen, die nicht dem gesellschaftlichen Mainstream entsprechen und dabei Forschungsmethoden zu nutzen, die nicht dem akademischen Mainstream folgen (sog. **kritische Methodologie**; Denzin, Lincoln, & Smith, 2008), wird in manchen Feldern der qualitativen Sozialforschung als ideale oder notwendige Ausrichtung begriffen.

Die **epistemologische Standpunkt-Theorie** („standpoint theory“) als ein spezieller Ansatz innerhalb der Wissenschaftsverständnisse im Feld der qualitativen Methoden fordert deswegen nicht nur eine Selbstreflexion der Perspektivität der Forschenden, sondern auch die Einnahme einer bestimmten **werthaltigen Perspektive**, etwa Parteilichkeit für Benachteiligte. Darüber hinaus wird eine verstärkte direkte **Partizipation von marginalisierten Gruppen** am Wissenschaftssystem gefordert. Dahinter steht die Vorstellung, dass z. B. Fragen der gesellschaftlichen und sozialen Benachteiligung wie sie u. a. in der feministischen Forschung, den Queer Studies oder den Postcolonial Studies untersucht werden, am besten auch von unmittelbar betroffenen Forschenden untersucht werden können, da diese eine entsprechende Problemperspektive bereits mitbringen (z. B. Denzin et al., 2008). Reflexivität der eigenen Perspektivität bedeutet also für qualitativ Forschende in diesem Wissenschaftsverständnis, sich die Bedeutung der eigenen Position und Perspektive für den Erkenntnisprozess bewusst zu machen (z. B. mögliche „blinde Flecken“, die daraus resultieren, dass man selbst von bestimmten Privilegien profitiert).

Die Frage, inwiefern sich eine bestimmte ideologische Perspektive der Forschenden bei konkreten Untersuchungsgegenständen auf den Forschungsprozess und dessen Ergebnisse auswirkt, ist komplex und wird nicht nur umfassend wissenschaftstheoretisch diskutiert, sondern verstärkt auch empirisch untersucht: Welche sozialwissenschaftlichen Studien liefern überzeugendere und praktisch hilfreichere Erklärungsmodelle für die Unterrepräsentanz von Frauen in der akademischen Forschung? Quantitative oder qualitative Studien? Studien, die mit oder Studien, die ohne ausdrücklich feministische Perspektive arbeiten? (vgl. Rolin, 2006; Wylie, 2009; für einen Überblick **feministischer Erkenntnistheorien** s. Anderson, 2012).

Auch wenn über die große Bedeutung der Perspektivität der Forschenden in der qualitativen Forschung Einigkeit besteht, so ist im Detail durchaus strittig, wie am besten mit ihr umzugehen ist. Die zugrunde liegenden wissenschaftstheoretischen Modelle sind pluralistisch und zum Teil kontrovers (z. B. Standpunkt-Theorie versus postmoderne Theorie) und unterliegen fortlaufendem Wandel.

2.3.6 Der Sozialkonstruktivismus als wissenschaftstheoretische Basis des qualitativen Paradigmas

Um dem wissenschaftstheoretischen Hintergrund des qualitativen Paradigmas trotz der angeführten Heterogenität der Ansätze einen Namen zu geben, wird oft vom **Interpretativen Paradigma** („interpretationism“, „interpretivism“) gesprochen und/oder auf den **Sozialkonstruktivismus** („social constructivism“) als Basis verwiesen. So konstatiert z. B. Reichertz (2007, S. 197) dass im Feld der qualitativen Methoden „fast alle Ansätze mit (sozial-)konstruktivistischen Prämissen“ operieren. Dabei ist jedoch zu beachten, dass ontologische, epistemologische und axiologische Positionen in unterschiedlichen Strömungen des Konstruktivismus wiederum stark ausdifferenziert sind. Zentrale Prämissen des wissenschaftstheoretischen Sozialkonstruktivismus werden im Folgenden skizziert (für sozialkonstruktivistische Wissenschaftstheorie allgemein: Kukla, 2000, und für sozialkonstruktivistische Grundlagen der qualitativen Sozialforschung: Creswell, 2013; Kruse, 2013d; Soini, Kronkquist, & Huber, 2011).

- **Epistemologische Prämissen:** Um Erkenntnisse über die durch soziale Konstruktionsprozesse hervorgebrachten sozialen Phänomene (sog. **Konstruktionen erster Ordnung**) zu gewinnen, müssen im Verständnis des Sozialkonstruktivismus eben die sozialen Konstruktionsprozesse rekonstruiert, d. h. die Sicht- und Handlungsweisen der Beteiligten verstanden werden. Der wissenschaftliche Erkenntnisprozess wird dabei seinerseits als Konstruktionsprozess aufgefasst und zwar im Sinne einer **Konstruktion zweiter Ordnung**: Die Forschenden erstellen in ihren wissenschaftlichen Arbeiten Konstruktionen, mit denen sie verstehend nachvollziehen, wie die Beforschten im

Alltag die interessierenden sozialen Phänomene herstellen und verstehen. Das Wahrheits- oder Gültigkeitskriterium für qualitative Forschungsergebnisse ist die Übereinstimmung der Konstruktionen der Forschenden mit den Konstruktionen der Beforschten, die u. a. als Glaubwürdigkeit bezeichnet wird (zu Gütekriterien qualitativer Sozialforschung ▶ Abschn. 3.3). Sofern an einem bestimmten sozialen Sachverhalt unterschiedliche Personengruppen beteiligt sind, die ihrerseits unterschiedliche Verständnisse von dem Phänomen haben, führt die wissenschaftliche Rekonstruktion dann auch zu unterschiedlichen Konstruktionen zweiter Ordnung, die nebeneinander bestehen (Beispiel: der Nutzen einer neuen Unterrichtsmethode wird von Lehrenden, Lernenden und deren Angehörigen unterschiedlich wahrgenommen und bewertet).

- **Ontologische Prämissen:** Soziale Phänomene werden als Ergebnisse sozialer Konstruktionsprozesse verstanden, d. h. sie werden durch die in ihren Lebenswelten handelnden Individuen gemeinsam hervorgebracht, sind dabei vom jeweiligen situativen und kulturellen Kontext abhängig und unterliegen historischem Wandel. Ein essentialistisches Verständnis sozialer Phänomene als objektiv vorgefundener Sachverhalte wird abgelehnt, stattdessen wird **anti-essentialistisch** die kollektive und individuelle **Gemachtheit und Veränderbarkeit des Sozialen** betont. Beispiel: Arbeitslosigkeit ist im sozialkonstruktivistischen Verständnis keine objektive Tatsache, sondern ein sozial in einem bestimmten gesellschaftlichen Zusammenhang hergestelltes Phänomen. Es setzt u. a. ein bestimmtes System der Erwerbsarbeit voraus, die unterschiedliche Verteilung von Beschäftigungschancen, die unterschiedliche Bewertung von Erwerbs- und Reproduktionsarbeit sowie die spezifische ressourcengenerierende und sinngebende Funktion von Erwerbsarbeit etc. Wenn man nun von „den Arbeitslosen" spricht, so könnte dies essentialistisch so verstanden werden, als ob Arbeitslosigkeit als „natürliches" und ggf. unveränderliches Phänomen verstanden wird. Um im Gegensatz dazu das sozialkonstruktivistische Gegenstandsverständnis zu betonen, würde man z. B. von Personen sprechen, die „als Arbeitslose bezeichnet werden" oder „denen Arbeitslosigkeit zugeschrieben" wird. Da Sprache ein zentrales Element sozialer Konstruktion von Wirklichkeit ist, wird bei der wissenschaftlichen Gegenstandsbeschreibung aus sozialkonstruktivistischer Sicht immer auch Sprachkritik virulent (d. h. das Hinterfragen vorherrschender Bezeichnungen).
- **Axiologische Prämissen:** Das sozialkonstruktivistische Wirklichkeits- und Wissenschaftsverständnis betont die Handlungsmacht aller Individuen und damit auch ihre ethische Verantwortung. Für die qualitativ Forschenden leitet sich daraus der Anspruch ab, soziale Probleme aus Sicht der Betroffenen und Beteiligten zu rekonstruieren, um zum Abbau von Missständen beizutragen bzw. um die Beforschten zu ermächtigen, besser für ihre Interessen einzutreten. Dies soll von Seiten der Forschenden nicht nur geschehen durch das Produzieren nützlicher Forschungsergebnisse, sondern in manchen qualitativen Ansätzen bereits durch das aktive Einbeziehen der Beforschten in den Forschungsprozess, wodurch sie z. B. ihre Identität stärken, ihr Reflexionsvermögen schulen oder praktische Veränderungsmaßnahmen einleiten können (z. B. in der Aktionsforschung oder in der Empowerment-Evaluation, ▶ Abschn. 18.2.4).

2.4 Das Mixed-Methods-Paradigma und der Pragmatismus

Dass es im Zuge empirischer Sozialforschung oft sinnvoll ist, qualitative und quantitative Daten – und somit auch qualitative und quantitative Datenerhebungsmethoden (z. B. Leitfaden-Interview sowie standardisierte Fragebogenerhebung) – zu nutzen, ist völlig unstrittig. In der Grundlagen-, vor allem aber in der Anwendungsforschung ist ein entsprechender **Methoden-Mix** gängig. Im vorliegenden Lehrbuch werden deswegen qualitative und quantitative Datenerhebungsmethoden umfassend behandelt.

Unterschiedliche Auffassungen existieren jedoch dahingehend, wie qualitative und quantitative Datenerhebungen sinnvoll **in einen Gesamtforschungsprozess zu integrieren** und daraus gültige wissenschaftliche Aussagen zu einem bestimmten Forschungsproblem abzuleiten sind. Damit sind Methodologie und Wissenschaftstheorie angesprochen. Denn im Zuge einer quantitativen Studie und im Verständnis des Kritischen Rationalismus lassen sich qualitative Datenerhebungen ohne weiteres integrieren – typischerweise im Entdeckungszusammenhang zur Generierung von neuen Theorien, die anschließend im Begründungszusammenhang quantitativ-statistisch geprüft werden. Auch im Zuge einer qualitativen Studie und im Verständnis des Sozialkonstruktivismus lassen sich quantitative Datenerhebungen einbeziehen – typischerweise indem deskriptiv-statistische Ergebnisse zur Häufigkeit oder Ausprägung bestimmter Merkmale oder Phänomene in der qualitativen Gesamtinterpretation Berücksichtigung finden. Die Nutzung unterschiedlicher Datentypen kann hier als spezielle Form der Triangulation – nämlich der **Daten-Triangulation** – verstanden werden (vgl. Flick, 2008).

Während ein Methoden-Mix innerhalb des quantitativen oder innerhalb des qualitativen Paradigmas dem jeweils „anderen" Datentyp eine ergänzende Funktion zuschreibt, plädiert das Mixed-Methods-Paradigma für eine **neue Methodologie**, d. h. für Forschungsprozesse, die nicht nur qualitative und quantitative Daten (und somit Datenerhebungsmethoden), sondern qualitative und quantitative Forschungsstrategien (also auch z. B. Untersuchungsdesigns, Stichprobentypen und Datenanalyseverfahren) miteinander kombinieren. Damit stellt sich die Frage, ob und wie es zu rechtfertigen ist, Erkenntnisprozesse, die auf ganz unterschiedlichen wissenschaftstheoretischen Vorannahmen über die Beschaffenheit der sozialen Wirklichkeit, die Rolle der Forschenden und die ideale Gestaltung des Forschungsprozesses basieren, überhaupt miteinander zu verknüpfen. Die Mixed-Methods-Literatur ist deswegen auch eine gute Anlaufstelle, um sich über wissenschaftstheoretische Grundlagen und Kontroversen der empirischen Sozialforschung zu informieren (z. B. Kelle, 2008). Wir gehen im Folgenden zunächst auf verschiedene Positionen hinsichtlich der Legitimierung und Fundierung eines Mixed-Methods-Paradigmas ein und beschreiben anschließend die wissenschaftstheoretische Position des Pragmatismus, der von einem Teil der Mixed-Methods-Community als wissenschaftstheoretische Grundlage herangezogen wird.

Neben der wissenschaftstheoretischen Frage, unter welchen Umständen und wie ein Mixed-Methods-Ansatz den sozialwissenschaftlichen Erkenntnisgewinn fördern kann, hat die aktuelle Diskussion um Mixed-Methods-Forschung auch eine **wissenschaftspolitische Ebene**: So läuft das Propagieren eines Mixed-Methods-Ansatzes darauf hinaus, sowohl quantitative als auch qualitative Forschungsstrategien als legitim und gleichberechtigt anzuerkennen und den in einer Reihe von sozialwissenschaftlichen Disziplinen und Forschungsfeldern bislang randständigeren qualitativen Ansatz stärker in den wissenschaftlichen Mainstream einzubeziehen. Dies wird einerseits als Chance, andererseits aber auch als Gefahr diskutiert: Kann die teilweise zum „Paradigmen-Krieg" eskalierte Polarisierung zwischen quantitativem und qualitativem „Lager" durch Mixed-Methods konstruktiv überwunden und eine **Demarginalisierung qualitativer Forschung** erreicht werden? Würden verstärkt Mixed-Methods-Designs verlangt und z. B. bei der Begutachtung von Zeitschriftenartikeln oder Drittmittelanträgen honoriert, würde dies auf eine stärkere Berücksichtigung von qualitativen Forschungsstrategien im Wissenschaftssystem hinauslaufen. Andererseits wird die Gefahr gesehen, dass im Mixed-Methods-Ansatz möglicherweise wichtige Grundprinzipien qualitativer Sozialforschung aufgegeben oder verwässert würden. Eine Integration in den Mainstream könnte schließlich im Widerspruch stehen zu der gerade aus der Distanz zum Mainstream erwachsenden dezidiert **wissenschaftskritischen Position** zumindest von Teilen des qualitativen Paradigmas. Diese Bemerkungen sollen dafür sensibilisieren, dass theoretische und politische Kontroversen rund um die „richtige" Form des sozialwissenschaftlichen Forschens innerhalb der Mixed-Methods-Bewegung fortleben und nicht einfach beigelegt sind.

2.4.1 Fünf Positionen zu Mixed-Methods als drittem Paradigma

Wie kann und soll Mixed-Methods-Forschung als drittes Forschungsparadigma der empirischen Sozialforschung neben dem quantitativen und dem qualitativen Paradigma bestehen? Fünf Hauptpositionen lassen sich differenzieren (vgl. Greene, 2008; Teddlie & Tashakkori, 2009):

1. Inkommensurabilitäts-These

Die **Inkommensurabilitäts-These**, die auch als **puristische Position** bezeichnet wird, geht davon aus, dass quantitatives und qualitatives Paradigma wissenschaftstheoretisch nicht vergleichbar und auch nicht miteinander vereinbar sind. Dementsprechend sollte eine Studie entweder eindeutig bzw. „pur" im qualitativen oder im quantitativen Paradigma verankert werden, je nachdem, welches Paradigma bevorzugt bzw. für das grundsätzlich überlegene gehalten wird. **Ein eigenes Mixed-Methods-Paradigma wird als sinnlos abgelehnt**.

Der Inkommensurabilitäts-These widerspricht die wachsende Zahl an Mixed-Methods-Studien sowie die beginnende Institutionalisierung des Mixed-Methods-Paradigmas (Teddlie & Johnson, 2009, S. 76) mit eigenen internationalen Konferenzreihen, Lehrbüchern und referierten Fachzeitschriften (z. B. Journal of Mixed Methods Research), die allesamt von der Integrierbarkeit qualitativer und quantitativer Forschungsstrategien einschließlich deren wissenschaftstheoretischer Grundlagen ausgehen.

2. Aparadigmatischer Standpunkt

Der **aparadigmatische Standpunkt** betrachtet Datenerhebungsmethoden und Studiendesigns einerseits sowie wissenschaftstheoretische Paradigmen andererseits als weitgehend unabhängig voneinander. Dementsprechend seien ein philosophischer Paradigmenstreit oder eine **Paradigmensuche für Mixed-Methods-Forschung unnötig**. Das gelte vor allem für die Anwendungsforschung, die mit ihren Erkenntnissen zur Lösung von Praxisproblemen beitragen will und dazu einfach alle verfügbaren Datenquellen und Forschungsstrategien nutzen sollte, die zur Beantwortung der Forschungsfrage beitragen. Ohnehin orientierten sich Forschende in der Praxis kaum an

philosophischen Meta-Überlegungen, so dass erfolgreiche Mixed-Methods-Studien de facto ohne paradigmatische bzw. wissenschaftstheoretische Festlegungen möglich seien.

Der aparadigmatische Standpunkt umgeht die in Grundlagen- wie Anwendungsforschung zentrale Frage nach der genauen Begründung des Gültigkeitsanspruchs wissenschaftlicher Aussagen. Er ist innerhalb der Mixed-Methods-Community nicht besonders verbreitet, vielmehr bemüht man sich hier gerade um die wissenschaftstheoretische Reflexion von Integrationsmöglichkeiten qualitativer und quantitativer Forschungsstrategien einschließlich entsprechender Gütekriterien für alle Phasen von Mixed-Methods-Forschungsprozessen (► Abschn. 3.4).

3. Komplementaritäts-These

Gemäß **Komplementaritäts-These** sind quantitatives und qualitatives Paradigma wissenschaftstheoretisch grundlegend verschieden, wobei jedoch kein Ansatz generell über- oder unterlegen sei. Vielmehr haben beide jeweils spezifische Stärken und Schwächen. Es ist deswegen sinnvoll, beide Forschungsstrategien zu nutzen, wobei sie jedoch nicht miteinander vermischt werden sollten. Durch Vermischung würden die Stärken des jeweiligen Ansatzes verwässert. Eine Mixed-Methods-Studie sollte somit **jeweils vollständige qualitative und vollständige quantitative Teilstudien** enthalten.

Einzuwenden ist hierbei, dass ein Nebeneinander qualitativer und quantitativer Studien ja ohnehin existiert und es dann nur eine Frage der Bezeichnung ist, ob von zwei Studien oder zwei Teilstudien die Rede ist. Für Mixed-Methods-Forschung wird im Allgemeinen eine stärkere Verknüpfung quantitativer und qualitativer Forschungsstrategien gefordert, die über ein Neben- oder Nacheinander separater Teilstudien hinausgeht.

4. Dialektischer Standpunkt

Der **dialektische Standpunkt** geht ebenso wie die Komplementaritäts-These von spezifischen Stärken und Schwächen sowohl des qualitativen als auch des quantitativen Paradigmas aus. Er setzt sich das Ziel, die Widersprüche beider Paradigmen nicht durch Separierung in einzelne Teilstudien auszuschalten (wie die Komplementaritäts-These vorschlägt), sondern sie im Zuge **dialektischen Denkens** im Rahmen einer integrierten Mixed-Methods-Studie fruchtbar zu machen. Gerade die zu Tage tretenden Gegensätze und Probleme böten die Chance verbesserten Erkenntnisgewinns.

Eine Umsetzungsmöglichkeit des dialektischen Mixed-Methods-Standpunkts besteht darin, dasselbe Forschungsproblem parallel von einem im quantitativen und einem im qualitativen Paradigma arbeitenden Team untersuchen zu lassen und über die jeweiligen Vorgehensweisen, Ergebnisse und wissenschaftstheoretischen Vorannahmen fortlaufend in einen intensiven Austausch zu treten, der letztlich die Forschungsstrategien und Befunde beider Teams modifiziert und dadurch zu Erkenntnissen führt, die bei der Beschränkung auf ein Paradigma nicht entstanden wären. Aufgrund des großen Aufwandes wird eine solche dialektische Vorgehensweise selten umgesetzt.

5. Standpunkt eines neuen Mixed-Methods-Paradigmas

Der Standpunkt eines neuen Mixed-Methods-Paradigmas schließlich postuliert die Eigenständigkeit von Mixed-Methods-Forschung, die über Kombination oder Dialog von quantitativem und qualitativem Ansatz hinausgeht und auch eine **eigene wissenschaftstheoretische Grundlage** beinhaltet. Hierfür wird meist als erkenntnistheoretische Basis der philosophische **Pragmatismus** vorgeschlagen.

Der paradigmatische Standpunkt, der Mixed-Methods-Forschung mit dem Pragmatismus als eigener erkenntnistheoretischer Basis begründet, besagt nicht, dass einfach beliebig qualitative und quantitative Vorgehensweisen zu kombinieren seien, sofern das „praktisch" erscheint (dies entspricht eher dem aparadigmatischen Standpunkt). Vielmehr handelt es sich beim philosophischen Pragmatismus um eine wissenschaftstheoretische Position, welche wissenschaftstheoretische Grundannahmen von quantitativem und qualitativem Paradigma nicht als unüberwindbare Gegensätze versteht, sondern z. B. als Extrempole auf verschiedenen Dimensionen, auf denen jeweils auch Mittelpositionen begründet eingenommen werden können. Der Grad der Ausarbeitung und die Überzeugungskraft der erkenntnistheoretisch-pragmatischen Position werden – außerhalb und innerhalb des Mixed-Methods-Paradigmas – kontrovers diskutiert.

2.4.2 Der Pragmatismus als wissenschaftstheoretische Basis eines Mixed-Methods-Paradigmas

Der philosophische Pragmatismus ist eine Denkrichtung, der gemäß Erkenntnisgewinn weder induktiv durch die Wahrnehmung der Realität (also empiristisch-positivistisch) noch deduktiv anhand von verstandesmäßig aufgestellten Theorien (also rationalistisch) erfolgt, sondern vor allem im Zusammenhang mit dem zielgerichteten Handeln in der Lebenswelt (also pragmatisch). Bei der Beurteilung von Begriffen und Theorien geht es nicht um deren Übereinstimmung mit der Erfahrungswirklichkeit

oder um deren Passung in ein analytisch-theoretisches Begriffssystem, sondern vor allem um die praktischen Konsequenzen. Insbesondere solche Theorien werden als gültig akzeptiert, die sich als brauchbar und nützlich erweisen, um durch praktisches Handeln bestimmte Ziele zu erreichen.

Begründet wurde der philosophische Pragmatismus in der zweiten Hälfte des 19. Jahrhunderts in den USA maßgeblich durch **Charles Sanders Pierce** (1839–1914), **William James** (1842–1910) und **John Dewey** (1859–1952). Dieser klassische nordamerikanische Pragmatismus wurde in der zweiten Hälfte des 20. Jahrhunderts als Neo-Pragmatismus verstärkt wieder aufgegriffen und weiterentwickelt, etwa durch **Richard Rorty** (1931–2007) und **Hilary Putnam** (1926–).

Der Pragmatismus ist in sich stark ausdifferenziert und es liegen unterschiedliche Beiträge aus der Mixed-Methods-Literatur vor, die versuchen, die Grundzüge philosophisch-pragmatischen Denkens zu rekonstruieren und auf empirisch-sozialwissenschaftliche Studien in einem Mixed-Methods-Paradigma anzuwenden (z. B. Cornish & Gillespie, 2009; Greene, 2008; Johnson & Onwuegbuzie, 2004). Die wichtigsten epistemologischen, ontologischen und axiologischen Prämissen werden im Folgenden grob skizziert.

Epistemologische Prämissen

Aus der Perspektive einer pragmatischen Wissenschaftstheorie geht es – sehr vereinfacht gesagt – nicht darum, Wissen zu generieren, das wahr ist in dem Sinne, dass es die äußere Realität korrekt abbildet oder die Sichtweisen der Beteiligten korrekt rekonstruiert. Stattdessen wird Wissen als Werkzeug betrachtet, das dabei helfen soll, erfolgreich zielgerichtet zu handeln. Theorien, die das **Kriterium der Nützlichkeit** erfüllen, werden als gültig akzeptiert. Theorien, die nicht erfolgreich zur Lösung praktischer Probleme beitragen können, werden zurückgewiesen. Theorien, die keine praktischen Anwendungsmöglichkeiten erkennen lassen, werden als irrelevant betrachtet. Das pragmatische Gültigkeitskriterium für Theorien ist definitionsgemäß **interessensbezogen**. Denn eine Theorie, die der Lösung bestimmter Probleme oder dem Erreichen bestimmter Ziele dient, stellt sich in den Dienst konkreter Interessen. Aus pragmatischer Sicht folgt daraus, dass die verfolgten Interessen jeweils transparent gemacht und diskutiert werden sollten, d. h. axiologischen Fragen nach Werten wird große Bedeutung beigemessen.

Der Pragmatismus ist insofern **pluralistisch**, als er von einer großen Vielfalt menschlicher Interessen ausgeht, die in der wissenschaftlichen Forschung adressiert werden können. Er ist aber nicht relativistisch in dem Sinne, dass es zu jeder konkreten Fragestellung „viele Wahrheiten" gibt. Stattdessen entscheidet jeweils die praktische Nützlichkeit, ob eine Theorie als gültig anerkannt wird. Eine „Hierarchie" der empirischen Erkenntnis, etwa in dem Sinne, dass immer der Kausalitätsnachweis mit einer randomisierten Kontrollgruppenstudie als „Goldstandard" zu betrachten ist, wird abgelehnt. Denn bestimmte praktische Fragestellungen lassen sich mit dieser Methodologie nicht bearbeiten. Beispiel: Wenn das Ziel verfolgt wird, die Lebenssituation von Patienten mit Diabetes zu verbessern, ist medizinisches Wissen nützlich wie etwa gesicherte Erkenntnisse über eine effektive Insulinersatztherapie, die mittels randomisierter Kontrollgruppenstudien gewonnen wird. Gleichzeitig ist es zur Krankheitsbewältigung für die Betroffenen auch nützlich zu erfahren, wie andere Patientinnen und Patienten ihren Alltag gestalten, wie sie die mit Diabetes verbundenen körperlichen, sozialen und mentalen Probleme bewältigen und eine positive Identität als chronisch Kranke aufbauen. Solches Wissen kann besonders gut mithilfe qualitativ-rekonstruktiver Verfahren gewonnen werden, etwa durch biografische qualitative Interviews (► Abschn. 10.2.2). Sofern sich die Erkenntnisse als hilfreich für Betroffene erweisen, haben sie erkenntnistheoretisch denselben Status wie Ergebnisse quantitativer Studien (vgl. Cornish & Gillespie, 2009). Diese epistemologische Perspektive erlaubt die Vereinbarkeit quantitativer und qualitativer Forschungsstrategien zu Mixed-Methods-Strategien. Die detaillierte Ausarbeitung pragmatischer Gültigkeitsnachweise für empirische Befunde aus Mixed-Methods-Studien steht allerdings noch weitgehend aus. Methodologisch werden im Pragmatismus Abduktion, Induktion und Deduktion akzeptiert und kombiniert.

Ontologische Prämissen

In ontologischen Fragen nimmt der Pragmatismus meist eine vermittelnde Position zwischen den Wirklichkeitsverständnissen von quantitativem und qualitativem Paradigma ein: Er geht von einem **transaktionalen Mensch-Umwelt-Verhältnis** aus, in dem weder objektive soziale Gesetzmäßigkeiten noch subjektive Wirklichkeitskonstruktionen vorherrschen, sondern beide im menschlichen Erleben und Handeln in einem Wechselverhältnis zueinander stehen: Die Konstruktion sozialer Phänomene wird als Prozess verstanden, der zwar von den Beteiligten aktiv gestaltet, aber immer auch im Rahmen bestimmter objektiver sozialer Gegebenheiten stattfindet. Darüber hinaus werden Grundsatzdebatten über die „wahre Natur" des Menschen und der sozialen Wirklichkeit oft für überflüssig gehalten, sofern aus ihnen keine eindeutigen Konsequenzen für die Forschungspraxis folgen.

Axiologische Prämissen

Die pragmatische Wissenschaftstheorie geht davon aus, dass die Wahl eines bestimmten Forschungsproblems in ethischer Hinsicht stets ausdrücklich begründungspflichtig ist. Generell sollte empirische Sozialforschung aus philosophisch-pragmatischer Sicht bei der **Lösung sozialer Probleme** auf individueller, kollektiver und gesellschaftlicher Ebene helfen. Forschenden wird somit unter anderem geraten, sich benachteiligten Bevölkerungsgruppen und deren Problemen zuzuwenden. Nicht selten sind soziale Probleme mit ethischen Dilemmata und Interessenkonflikten verbunden. In solchen Fällen wird eine öffentliche Debatte über die gesellschaftlich favorisierte Zielrichtung der Forschung vorgeschlagen. In Evaluationsforschungsprojekten sind die Interessen unterschiedlicher Anspruchsgruppen zu beachten (▶ Abschn. 18.1.4). Schließlich empfiehlt es sich aus pragmatisch-wissenschaftstheoretischer Perspektive, jeden Forschungsprozess dahingehend zu hinterfragen, wessen Interessen das gewonnene Wissen (das pragmatisch immer als Werkzeug zur Lösung von Problemen verstanden wird) dient oder nicht dient.

Lernquiz

1. Welche Aussagen stimmen?
   a) Fragen nach dem Wesen wissenschaftlicher Untersuchungsgegenstände fallen in den Bereich der Ontologie.
   b) Die Erkennbarkeit der Wirklichkeit ist eine epistemologische Frage.
   c) Die Axiologie ist kein Teilbereich der Wissenschaftstheorie.
   d) Die Wissenschaftstheorie ist normativ ausgerichtet.
   e) Wenn die Wissenschaftspraxis von den Vorgaben der Wissenschaftstheorie abweicht, muss die Wissenschaftstheorie verworfen werden.
2. Erläutern Sie das Falsifikationsprinzip des Kritischen Rationalismus!
3. Welche Rolle spielt die Wissenschaftsgemeinschaft im Kritischen Rationalismus, vor allem vor dem Hintergrund, dass sowohl ein naiver Realismus als auch ein naiver Falsifikationismus erkenntnistheoretisch abgelehnt werden?
4. Welche Kriterien werden an Theorien gemäß Kritischem Rationalismus angelegt?
5. Welche Positionen vertritt der Kritische Rationalismus?
   a) Wissenschaftliche Erkenntnis entsteht durch die Sammlung von Daten über die Realität.
   b) Wenn Daten einer wissenschaftlichen Theorie widersprechen, muss diese verworfen werden.
   c) Forschende sollten gegenüber sozialen Sachverhalten eine wertfreie, neutrale Haltung einnehmen.
   d) Für den wissenschaftlichen Erkenntnisfortschritt sind die statistisch signifikanten Ergebnisse ausschlaggebend.
   e) Eine wissenschaftliche Theorie sollte die Komplexität der Wirklichkeit abbilden und möglichst alle Ursachenfaktoren eines Phänomens einbeziehen.
6. Nennen Sie die fünf zentralen wissenschaftstheoretischen Grundprinzipien qualitativer Sozialforschung!
7. Erläutern Sie das Prinzip der Zirkularität und Flexibilität des methodischen Vorgehens im qualitativen Paradigma! Wo stößt es an Grenzen?
8. Was trifft nicht zu?
   a) Im DN-Modell entspricht die abhängige Variable dem Explanandum.
   b) Durch Exhaustion steigt der Bewährungsgrad und sinkt der Informationsgehalt einer Theorie.
   c) Während im quantitativen Paradigma die theoretisch interessierenden Merkmale der Erfahrungswirklichkeit *vor* der Datenerhebung präzise zu definieren und zu operationalisieren sind, werden im qualitativen Paradigma die für das Verstehen der interessierenden Phänomene notwendigen theoretischen Konzepte erst *im Zuge* der Datenerhebung und Datenanalyse ausgearbeitet.
   d) In der qualitativen Forschung werden die subjektiven Sichtweisen der untersuchten Personen durch Kommunikation ergründet, in der quantitativen Forschung wird dagegen nur das Verhalten objektiv erfasst.
   e) Im qualitativen wie im quantitativen Paradigma wird die Forschungstätigkeit als aktiver Konstruktionsprozess verstanden.
   f) Die qualitative Sozialforschung konzentriert sich auf die Untersuchung lebensweltlicher Phänomene. Die quantitative Sozialforschung untersucht diese ebenfalls und analysiert darüber hinaus auch Sachverhalte auf der Mikroebene (z. B. Hirnaktivitäten auf der Basis physiologischer Messungen) sowie auf der Makroebene (z. B. Unterschiede in der Lebenserwartung unterschiedlicher Bevölkerungsgruppen im historischen Wandel oder interkulturellen Vergleich auf der Basis bevölkerungsstatistischer Erhebungen).
9. Welche Positionen werden hinsichtlich der wissenschaftstheoretischen Vereinbarkeit quantitati-

ver und qualitativer Forschungsstrategien zu einer Mixed-Methods-Forschungsstrategie vertreten?

10. Welche Position(en) vertritt die philosophisch-pragmatische Wissenschaftstheorie?
    a) Anwendungswissen ist immer interessenbezogen, Grundlagenwissen dagegen nicht.
    b) Forschungsaktivitäten in den Sozialwissenschaften sollten sich gemäß pragmatischem Wissenschaftsverständnis bevorzugt der Lösung sozialer Probleme widmen, die von der Allgemeinheit als vordringlich angesehen werden.
    c) Gemäß transaktionalem Mensch-Umwelt-Verhältnis geht der Pragmatismus davon aus, dass sich das Erleben und Verhalten von Menschen am besten durch das Zusammenspiel objektiver sozialer Gesetzmäßigkeiten und subjektiver Sinnkonstruktionen beschreiben, erklären und vorhersagen lässt.
    d) Wissen im Allgemeinen und wissenschaftliche Erkenntnisse im Besonderen sind als Hilfsmittel zur praktischen Problemlösung zu betrachten.
    e) Wenn eine Theorie nützliche Beiträge zur Lösung praktischer Probleme liefert, ist sie als gültig anzuerkennen.

Übungsaufgaben

Lesen Sie sich folgenden Artikel durch, in dem Popper seine wissenschaftstheoretische Position mit Blick auf die Sozialwissenschaften (speziell die Soziologie) in 27 Thesen pointiert zusammenfasst:

Popper, K. (1962). Die Logik der Sozialwissenschaften. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 14*, 233–248.

Beantworten Sie auf dieser Basis folgende Fragen:

1. Laut Popper „beginnt die Erkenntnis nicht mit Wahrnehmungen oder Beobachtungen oder der Sammlung von Daten oder von Tatsachen, sondern …“ womit? Teilen Sie diese Auffassung? (Begründung)
2. Popper beschreibt in seinem Beitrag eine Begebenheit auf einer interdisziplinären wissenschaftlichen Konferenz, auf der über „Wissenschaft und Humanismus“ diskutiert wurde. Ein anwesender Sozialanthropologe habe sich dort nicht selbst an den Diskussionen beteiligt, sondern sich stattdessen wie ein „Naturwissenschaftler“ in die Rolle des „objektiven Beobachters“ begeben und den Diskussionsverlauf anhand der verschiedenen Formen des gezeigten verbalen und nonverbalen Verhaltens der Wissenschaftler analysiert, etwa dahingehend wer wen beeinflusst oder das Gespräch dominiert. Um nicht seine „Objektivität“ zu verlieren, habe der Sozialanthropologe laut eigener Aussage seinen eigenen (subjektiven) Standpunkt zu den verhandelten wissenschaftstheoretischen Fragen erst gar nicht geäußert und sich auch nicht in die inhaltliche Debatte verstricken lassen. Popper lehnt eine solche Haltung vehement als „pseudo-naturwissenschaftlich“ und „verfehlten Naturalismus“ ab. Laut Popper entsteht wissenschaftliche Objektivität nämlich nicht durch eine vermeintlich unbeteiligte Beobachterrolle, sondern … wodurch? Welche Bedingungen sind laut Popper notwendig, damit z. B. die betreffende Konferenz möglichst objektive Ergebnisse erbringen kann? Teilen Sie diese Auffassung? (Begründung)
3. Im Hinblick auf Wertfragen differenziert Popper zwischen rein wissenschaftlichen Werten (Wahrheit, Relevanz, Einfachheit, Genauigkeit, Fruchtbarkeit von wissenschaftlichen Aussagen) sowie außerwissenschaftlichen Werten und Interessen (z. B. mögliche Nutzung wissenschaftlicher Forschung für persönliche Bereicherung, industrielle Entwicklung, nationale Verteidigung oder Angriffspolitik) der Beteiligten. In der Wissenschaftspraxis spielen laut Popper sowohl wissenschaftliche als auch außerwissenschaftliche Werte immer eine Rolle. Er fordert deswegen auch keine „Wertfreiheit“, sondern vielmehr … was?

Weiterführende Literatur

- Breuer, F. (1991). *Wissenschaftstheorie für Psychologen*. Münster: Aschendorff. http://www.ssoar.info/ssoar/handle/document/965.
- Chalmers, A. (2007). *Wege der Wissenschaft. Einführung in die Wissenschaftstheorie* (6. Aufl.). Heidelberg: Springer.
- Hecht, H. & Desnizza, W. (2012). *Psychologie als empirische Wissenschaft*. Heidelberg: Springer.
- Herzog, W. (2012). *Wissenschaftstheoretische Grundlagen der Psychologie*. Heidelberg: Springer.
- Kelle, U. (2008). *Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte* (2. Aufl.). Wiesbaden: VS.

- Mey, G. & Mruck, K. (Hrsg.) (2010). *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS.
- Popper, K. (1962). Die Logik der Sozialwissenschaften. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 14*, 233–248.

Literatur

Adorno, T. W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., & Popper, K. (1969). *Der Positivismusstreit in der deutschen Soziologie*. Neuwied: Luchterhand.

Albert, H. (1991). *Traktat über kritische Vernunft* (5. verb. & erw. ed.). Hamburg: Mohr Siebeck. (Original veröffentlicht 1968).

Albert, H. (1987). *Kritik der reinen Erkenntnislehre. Das Erkenntnisproblem in realistischer Perspektive*. Tübingen: Mohr.

Albert, H. (2000). *Kritischer Rationalismus*. Tübingen: Mohr Siebeck.

Albert, H. (2011). *Kritische Vernunft und rationale Praxis*. Tübingen: Mohr Siebeck.

Alisch, L.-M. (2005). Über die (Wechsel-)Beziehung zwischen Entdeckungs-, Begründungs- und Verwendungszusammenhang wissenschaftlicher Sätze. In H. Heid & C. Harteis (Eds.), *Verwertbarkeit – ein Qualitätskriterium (erziehungs-)wissenschaftlichen Wissens?* (S. 55–78). Wiesbaden: VS.

Anderson, E. (2012). Feminist epistemology and philosophy of science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved December 9, 2013, from http://plato.stanford.edu/archives/fall2012/entries/feminism-epistemology/.

Asendorpf, J. B. & Neyer, F. J. (2012). *Psychologie der Persönlichkeit* (5. Aufl.). Heidelberg: Springer.

Balzer, W., Moulines, C. U., & Sneed, J. D. (1987). *An architectonic for science: The structuralist program*. Dordrecht: Reidel.

Blumer, H. (1954). What's wrong with social theory? *American Sociological Review, 18*, 3–10.

Breuer, F. (2010). Wissenschaftstheoretische Grundlagen qualitativer Methodik in der Psychologie. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 35–49). Wiesbaden: VS.

Breuer, F. (2011a). The other talks back. Auslösung von Feldreaktanzen durch sozialwissenschaftliche Re-/Präsentationen. *Forum Qualitative Sozialforschung, 12*(2), Art. 23. Retrieved January 14, 2012, from http://nbn-resolving.de/urn:nbn:de:0114-fqs1102233.

Breuer, F. (2011b). *Reflexive grounded theory. Eine Einführung für die Forschungspraxis*. (2. Aufl.). Heidelberg: Springer.

Breuer, F., Mruck, K., & Mey, G. (2011). Subjektivität und Selbst-/Reflexivität in der Grounded-Theory-Methodologie. In G. Mey & K. Mruck (Hrsg.), *Grounded Theory Reader* (2. überarb. u. erweiterte Aufl., S. 427–448). Wiesbaden: VS.

Buchinger, E. & Felt, U. (Hrsg.). (2006). *Technik- und Wissenschaftssoziologie in Österreich. Stand und Perspektiven*. Wiesbaden: VS.

Chalmers, A. F. (2007). *Wege der Wissenschaft: Einführung in die Wissenschaftstheorie* (5., völlig überarbeitete Aufl.). Heidelberg: Springer.

Clemens, M. & Demombynes, G. (2011). When does rigorous impact evaluation make a difference? The case of the Millenium Villages. *The Journal of Development Effectiveness, 3*(3), 305–339.

Cornish, F. & Gillespie, A. (2009). A pragmatist approach to the problem of knowledge in health psychology. *Journal of Health Psychology, 14*(6), 800–809.

Creswell, J. W. (2013). *Qualitative inquiry & research design – Choosing among five approaches* (3rd. ed.). Los Angeles, CA: Sage.

Denzin, N. K. & Lincoln, Y. S. (Eds.). (2005). *The Sage handbook of qualitative research* (3rd. ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., Lincoln, Y. S., & Smith, L. T. (Eds.). (2008). *Handbook of critical and indigenous methodologies*. Thousand Oaks, CA: Sage.

Devereux, G. (1976). *Angst und Methode in den Verhaltenswissenschaften*. Frankfurt a. M.: Ullstein.

Dilthey, W. (2006). *Einleitung in die Geisteswissenschaften. Versuch einer Grundlegung für das Studium der Gesellschaft und Geschichte*. Göttingen: Vandenhoeck & Ruprecht. (Original veröffentlicht 1883).

Felt, U., Nowotny, H., & Taschwer, K. (1995). *Wissenschaftsforschung. Eine Einführung*. Frankfurt/Main: Campus.

Feyerabend, P. (1979). *Erkenntnis für freie Menschen*. Frankfurt: Suhrkamp.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*(6), 661–669.

Fischer, K. (2007). Fehlfunktionen der Wissenschaft. *Erwägen, Wissen, Ethik EWE, 18*(1), 3–16.

Flick, U. (2008). *Triangulation: Eine Einführung* (2. Aufl.). Wiesbaden: VS.

Gadamer, H.-G. (1990). *Gesammelte Werke Band 1. Hermeneutik I: Wahrheit und Methode* (6.Aufl.). Tübingen: Mohr. (Original veröffentlicht 1960).

Gadenne, V. (2004). *Philosopie der Psychologie*. Bern: Huber.

Gadenne, V. (2006). Empirische Forschung und normative Wissenschaftstheorie. Was bleibt von der Methodologie des kritischen Rationalismus? In A. Diekmann (Hrsg.), *Methoden der Sozialforschung* (S. 33-50). Wiesbaden: VS.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7–22.

Habermas, J. (1969). Analytische Wissenschaftstheorie und Dialektik. In T. W. Adorno, H. Albert, R. Dahrendorf, J. Habermas, H. Pilot & K. Popper (Eds.), *Der Positivismusstreit in der deutschen Soziologie* (pp. 155-192). Neuwied: Luchterhand.

Habermas, J. (1983). *Zur Logik der Sozialwissenschaften*. Frankfurt/Main: Suhrkamp.

Harmon-Jones, C., Schmeichel, B. J., & Harmon-Jones, E. (2009). Symbolic self-completion in academia: evidence from department web pages and email signature files. *European Journal of Social Psychology, 39*(2), 311–316.

Heid, H. & Harteis, C. (Hrsg.). (2005). *Verwertbarkeit. Ein Qualitätskriterium (erziehungs-)wissenschaftlichen Wissens?* Wiesbaden: VS.

Heidegger, M. (2006). *Sein und Zeit* (19. Aufl.). Tübingen: Niemeyer. (Original veröffentlicht 1927).

Helfferich, C. (2011). *Die Qualität qualitativer Daten: Manual für die Durchführung qualitativer Interviews* (4. ed.). Wiesbaden: VS.

Hempel, C. G. & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135–175.

Herrmann, T. (1976). *Die Psychologie und ihre Forschungsprogramme*. Göttingen: Hogrefe.

Herrmann, T. (1979). *Psychologie als Problem*. Stuttgart: Klett-Cotta.

Herzog, W. (2012). *Wissenschaftstheoretische Grundlagen der Psychologie*. Wiesbaden: Springer VS.

Holzkamp, K. (1968). *Wissenschaft als Handlung*. Berlin: de Gruyter.

Holzkamp, K. (1972). *Kritische Psychologie*. Frankfurt/Main: Fischer.

Johnson, R. B. & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Kalthoff, H., Hirschauer, S., & Lindemann, G. (Hrsg.). (2008). *Theoretische Empirie. Zur Relevanz qualitativer Forschung*. Frankfur/Main: Suhrkamp.

Kelle, U. (2008). *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung* (2. Aufl.). Wiesbaden: VS.

Keuth, H. (1989). *Wissenschaft und Werturteil: zu Werturteilsdiskussion und Positivismusstreit*. Tübingen: Mohr Siebeck.

Knorr, C. (1991). *Die Fabrikation von Erkenntnis: Zur Anthropologie der Naturwissenschaft* (3. Aufl.). Frankfurt/Main: Suhrkamp.

Kromrey, H. (2009). *Empirische Sozialforschung* (12. Aufl.). Stuttgart: Lucius & Lucius.

Kruse, J. (2013a). Prinzip der Prozessualität. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (16. Aufl., S. 1220). Bern: Huber.

Kruse, J. (2013b). Qualitative Sozialforschung. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (16. Aufl., S. 1279–1280). Bern: Huber.

Kruse, J. (2013c). Sensitizing concepts. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (16. Aufl., S. 1414). Bern: Huber.

Kruse, J. (2013d). Sozialkonstruktivismus. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (16. Aufl., S. 1455). Bern: Huber.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kukla, A. (2000). *Social constructivism and the philosophy of science*. London: Routledge.

Lakatos, I. (1977). *The methodology of scientific research programmes: Philosophical papers volume I*. Cambridge: Cambridge University Press.

Lamnek, S. (2010). *Qualitative Sozialforschung* (5., vollständig überarb. Aufl.). Weinheim: Beltz.

Lincoln, Y. S. (2010). „What a long, strange trip it's been . . .": Twenty-five years of qualitative and new paradigm research. *Qualitative Inquiry, 16*(1), 3–9.

Merton, R. K. (1949). *Social theory and social structure: Toward the codification of theory and research*. Glencoe: Free Press.

Miller, D. (1994). *Critical rationalism: A restatement and defence*. Chigago, Illinois: Open Court.

Miller, D. (2006). *Out of error: Further essays on critical rationalism*. Burlington, VT: Ashgate.

Pähler, K. (1986). *Qualitätsmerkmale wissenschaftlicher Theorien*. Hamburg: Mohr Siebeck.

Popper, K. (1989). *Logik der Forschung* (9. Aufl.). Tübingen: Mohr Siebeck. (Original veröffentlicht 1934).

Popper, K. (1962). Die Logik der Sozialwissenschaften. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 14*, 233–248.

Popper, K. (2009). *Vermutungen und Widerlegungen*. Tübingen: Mohr Siebeck. (Original veröffentlicht 1963).

Popper, K. (1972). Die Logik der Sozialwissenschaften. In T. W. Adorno, R. Dahrendorf, H. Pilot, H. Albert, J. Habermas & K. Popper (Hrsg.), *Der Positivismusstreit in der deutschen Soziologie* (S. 103–123). Neuwied: Luchterhand. (Original veröffentlicht 1969).

Popper, K. (1973). *Objektive Erkenntnis. Ein evolutionärer Entwurf*. Hamburg: Hoffmann & Campe.

Popper, K. (1984). *Objektive Erkenntnis. Ein evolutionärer Entwurf* (4. Aufl.). Hamburg: Hoffmann & Campe.

Reichenbach, H. (1938). *Experience and prediction. An analysis of the foundations and the structure of knowledge*. Chicago: University of Chicago Press.

Reichertz, J. (2003). *Die Abduktion in der qualitativen Sozialforschung*. Wiesbaden: VS.

Reichertz, J. (2007). Qualitative Sozialforschung – Ansprüche, Prämissen, Probleme. *Erwägen, Wissen, Ethik EWE, 18*(2), 195–208.

Rolin, K. (2006). The bias paradox in feminist standpoint epistemology. *Episteme, 3*(1-2), 125–137.

Rowbottom, D. P. (2011). *Popper's critical rationalism: A philosophical investigation*. London: Routledge.

Sandelands, L. E. (1990). What is so practical about theory? Lewin revisited. *Journal for the Theory of Social Behaviour, 20*(3), 357–379.

Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der empirischen Sozialforschung*. München: Oldenbourg

Schütz, A. (1932). *Der sinnhafte Aufbau der sozialen Welt. Eine Einleitung in die verstehende Soziologie*. Wien: Springer.

Schütz, A. & Luckmann, T. (2003). *Strukturen der Lebenswelt*. Konstanz: UVK. (Original work published in 1975).

Simon-Schäfer, R. (1993). Dialektik. In H. Seiffert & G. Radnitzky (Hrsg.), *Lexikon der Wissenschaftstheorie* (S. 33–36). München: dtv.

Sneed, J. D. (1971). *The logical structure of mathematical physics*. Dordrecht: Reidel.

Soini, H., Kronqvist, E.-L., Huber, G. L., & (Eds.). (2011). Epistemologies for qualitative research. Retrieved January 14, 2014, from http://www.qualitativepsychology.com/files/qualitative-psychology-nexus-viii.pdf.

Stegmüller, W. (1979). *The structuralist view of theories. A possible analogue of the bourbaki programme in physical science*. Berlin: Springer.

Teddlie, C. & Johnson, R. B. (2009). Methodological thought since the twentieth century. In C. Teddlie & A. Tashakkori (Eds.), *The foundations of mixed methods research: Integrating quantitative and qualitative techniques in the social and behavioral sciences* (pp. 62–82). Thousand Oaks, CA: Sage.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research: integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Thornton, S. (2013). Karl Popper. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy. Retrieved December 9, 2013, from http://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=popper.

Unger, H. (2012). Partizipative Gesundheitsforschung. Wer partizipiert woran? *Forum Qualitative Sozialforschung, 13*(1), Art. 7. Abgerufen unter http://nbn-resolving.de/urn:nbn:de:0114-fqs120176.

Kardorff, E. von (1995). Qualitative Sozialforschung – Versuch einer Standortbestimmung. In U. Flick, E. von Kardorff, H. Keupp, L. von Rosenstiel & S. Wolff (Hrsg.), *Handbuch Qualitative Sozialforschung* (2.Aufl., S. 3–10). Weinheim: Beltz.

Weber, M. (1951). Die „Objektivität" sozialwissenschaftlicher und sozialpolitischer Erkenntnis. In M. Weber (Hrsg.), *Gesammelte Aufsätze zur Wissenschaftslehre* (S. 146–214). Tübingen: Mohr. (Original veröffentlicht 1904).

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik*. Göttingen: Hogrefe.

Wicklund, R. A. & Gollwitzer, P. M. (1982). *Symbolic self-completion*. Hillsdale, N. J.: Lawrence Earlbaum.

Wurzer, J. (1994). Karl Poppers Kübel- und Scheinwerfermodell der Erkenntnis. Abgerufen 26. November, 2013, unter http://www.wurzer.org/Homepage/Publikationen/Eintr%C3%A4ge/1994/5/1_Karl_Poppers_K%C3%BCbel_und_Scheinwerfermodell_der_Erkenntnis_files/kuebel_und_scheinwerfer.pdf.

Wylie, A. (2009). Social constructionist arguments in harding's science and social inequality. *Hypatia, 23*(4), 201–211.

Qualitätskriterien in der empirischen Sozialforschung

Nicola Döring, Jürgen Bortz

3.1 Wissenschaftlichkeit und wissenschaftliche Qualität – 84

3.2 Gütekriterien für quantitative Forschung – 93

3.3 Gütekriterien für qualitative Forschung – 106

3.4 Gütekriterien für Mixed-Methods-Forschung – 114

Literatur – 117

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_3, © Springer-Verlag Berlin Heidelberg 2016

3

Lernziele

- Wissenschaft von Nicht-Wissenschaft, Pseudo- und Parawissenschaft abgrenzen können.
- Wissen, wozu Kriterien der wissenschaftlichen Qualität in der empirischen Sozialforschung dienen und wie man sie strukturieren kann.
- Die wichtigsten Gütekriterien der quantitativen Sozialforschung erläutern können.
- Die wichtigsten Gütekriterien der qualitativen Sozialforschung darstellen können.
- Die wichtigsten Gütekriterien der Mixed-Methods-Sozialforschung kennen.

Ziel der empirischen Sozialforschung ist es, **wissenschaftliche Erkenntnisse** über soziale Sachverhalte zu gewinnen. Dreh- und Angelpunkt empirischer Forschung ist dabei immer der **Forschungsprozess** mit seinen verschiedenen Phasen, in denen jeweils unterschiedliche Forschungsmethoden – z. B. Methoden der Datenerhebung und der Datenanalyse – zum Einsatz kommen (► Abschn. 1.3). Auf ganz grundlegender Ebene werden der Ablauf des Forschungsprozesses, die Eigenschaften der einzelnen Forschungsmethoden sowie die Rolle der Forschenden im Verhältnis zum Forschungsgegenstand durch das jeweilige Wissenschaftsverständnis bzw. das **wissenschaftstheoretische Paradigma** begründet. Wir unterscheiden dabei in der empirischen Sozialforschung vor allem das quantitative, das qualitative sowie das Mixed-Methods-Paradigma (► Kap. 2).

Zur Beurteilung der Qualität wissenschaftlicher Erkenntnisse wird meist der Forschungsprozess hinsichtlich seiner Güte bewertet. Die erste grundlegende Frage lautet dabei, ob sich eine Studie überhaupt auf dem Boden der Wissenschaftlichkeit bewegt oder das Vorgehen als nicht-, pseudo- oder parawissenschaftlich zu kennzeichnen ist. Dies ist anhand vier grundlegender **Standards der Wissenschaftlichkeit** zu prüfen, die für quantitative, qualitative sowie Mixed-Methods-Forschung gleichermaßen gelten und auch disziplinübergreifend in den Erfahrungswissenschaften angelegt werden. Es ist zu beachten, dass mit den Standards der Wissenschaftlichkeit hier allgemeine Richtlinien des wissenschaftlichen Vorgehens gemeint sind (z. B. Formulierung eines wissenschaftlichen Forschungsproblems als Ausgangspunkt jeder wissenschaftlichen Studie). Davon abzugrenzen sind Standards im Sinne von Referenzwerten, die angeben, welchen Wert ein gemessenes Kriterium erreichen soll (z. B. Referenzwerte zur Beurteilung des Kriteriums der Messgenauigkeit von Erhebungsinstrumenten).

Wenn Wissenschaftlichkeit gegeben ist, stellt sich in der Regel noch die Frage, ob man es mit einer eher schwachen oder durchschnittlichen, mit einer besonders guten oder gar exzellenten wissenschaftlichen Studie zu tun hat. Für diese Beurteilung werden analog den vier Standards vor allem vier **Kriterien der wissenschaftlichen Qualität** herangezogen, unter denen die methodische Strenge als wissenschaftsimmanentes Kriterium eine herausgehobene Position einnimmt (► Abschn. 3.1).

Die auf allgemeiner Ebene formulierten Kriterien wissenschaftlicher Qualität (und insbesondere methodischer Strenge) müssen bei der Beurteilung konkreter Studien auf die Details der einzelnen Phasen des wissenschaftlichen Forschungsprozesses heruntergebrochen werden. Dabei handelt es sich in den Sozial- und Humanwissenschaften meist um empirische Studien, aber auch nichtempirische wissenschaftliche Studien (z. B. Methoden- oder Theoriearbeiten) sind hinsichtlich ihrer wissenschaftlichen Qualität einzuordnen. Dabei sind **Gütekriterien für quantitative Forschung**, die u. a. auf die Qualität statistischer Auswertungen abzielen (► Abschn. 3.2), von **Gütekriterien für qualitative Forschung**, in denen z. B. die Qualität von Textinterpretationen angesprochen wird (► Abschn. 3.3), zu unterscheiden. Auch existieren eigene **Gütekriterien für Mixed-Methods-Forschung** (► Abschn. 3.4), die u. a. die Qualität der Integration von quantitativem und qualitativem Datenmaterial berücksichtigen.

Wir werden im Laufe des Buches bei der Behandlung der einzelnen Phasen des Forschungsprozesses sowie bei unterschiedlichen qualitativen und quantitativen Methoden der Datenerhebung und Datenanalyse immer wieder auf deren jeweils spezifische Gütekriterien detailliert zu sprechen kommen. In diesem Kapitel geht es um eine **Gesamtsicht der wissenschaftlichen Qualität.** Denn eine sachgerechte Bewertung vorliegender Studien sowie eine sinnvolle Planung eigener Studien erfordert es, die Gesamtqualität des Forschungsprozesses im Blick zu behalten. Dabei wird es in der Praxis kaum gelingen, eine Idealstudie zu finden oder zu realisieren, die sämtlichen Gütekriterien im Höchstmaß genügt. Denn wissenschaftliche Studien müssen oft mit begrenzten zeitlichen, finanziellen und personellen Ressourcen realisiert werden, was zu Abstrichen bei einzelnen Aspekten der methodischen Strenge oder bei anderen Qualitätskriterien führt. Darüber hinaus stehen einige Gütekriterien auch in einem Spannungsverhältnis zueinander: Das einseitige Optimieren eines Kriteriums würde auf Kosten eines anderen Kriteriums gehen, so dass eine geeignete Balance zu finden ist.

Bei der **Qualitätsbewertung** generell – so auch bei der Qualitätsbewertung wissenschaftlicher Studien – ist zu unterscheiden zwischen Kriterien, Indikatoren und Standards (zu Grundlagen zum Qualitätsbegriff und zum Qualitätsmanagement s. Geiger & Kotte, 2007 sowie die ISO-9000-Normenreihe):

- Wissenschaftliche **Qualitäts- oder Gütekriterien** („scientific quality criteria“) sind Merkmale wissenschaftlicher Studien, an denen auf wissenschaftstheoretischer Basis ihre Qualität bzw. Güte festmacht werden soll, was letztlich auf die Beurteilung der durch die Studie gewonnenen Erkenntnisse hinausläuft. Ein Beispiel für ein Qualitätskriterium ist die methodische Strenge. In der quantitativen Forschung hat die methodische Strenge u. a. den Unteraspekt der Messgenauigkeit (Reliabilität) eines standardisierten Messinstrumentes (z. B. geringe Verzerrung der Ergebnisse eines Intelligenztestes durch Messfehler). In der qualitativen Forschung hat die methodische Strenge u. a. den Unteraspekt der kommunikativen Validierung einer qualitativen Datenanalyse (z. B. Interviewpartner bestätigen die Interpretationen, die die Forschenden aus dem Interviewmaterial gewonnen haben bzw. geben zu den Ergebnissen solche Rückmeldungen, dass die Forschenden daraus eine Bestätigung ihrer Resultate schlüssig ableiten können).
- Die Qualitätskriterien werden zunächst als theoretische Konzepte definiert. Für eine Qualitätsbeurteilung bzw. Qualitätsmessung müssen sie für konkrete wissenschaftliche Studien als messbare **Qualitätsindikatoren** („scientific quality indicators“) empirisch fassbar sein. In der Regel existieren pro Qualitätskriterium unterschiedliche Indikatoren. So wird z. B. das Gütekriterium der Messgenauigkeit in der quantitativen Forschung über statistische Reliabilitätskoeffizienten als Qualitätsindikatoren messbar, wobei gemäß Testhalbierungs-, Paralleltest-, Testwiederholungs- oder Interner Konsistenz-Methode verschiedene Reliabilitätskoeffizienten berechnet werden können (► Abschn. 10.4.1 „Hauptgütekriterien psychologischer Tests“ und ► Abschn. 10.4.4). Auch die Gültigkeit von Interpretationen im qualitativen Forschungsansatz wird an unterschiedlichen Indikatoren verdeutlicht, etwa der Rückmeldungen der Interviewpartner (kommunikative Validierung), dem Vergleich mit Schlussfolgerungen aus anderem Datenmaterial zum selben Phänomen – z. B. Beobachtungsdaten ergänzend zu den Interviewdaten (Triangulation) – oder durch die überzeugende Angabe von Argumenten zur Schlüssigkeit der abgeleiteten Interpretation (argumentative Validierung; zu Validitätsformen qualitativer Forschung s. Lamnek, 2010, S. 138).
- Wissenschaftliche **Qualitätsstandards** („scientific/methodological quality standards“, „methodological standards of social scientific research“, „standards for qualitative/quantitative research“) schließlich sind normative Vorgaben dazu, welche Ausprägung die Qualitätsindikatoren jeweils mindestens haben müssen, damit von schlechter, ausreichender oder guter wissenschaftlicher Qualität in diesem Aspekt auszugehen ist (**Mindest-, Regel- und Maximalstandards**). Man spricht auch von **Referenzwerten** bzw. **Referenzbereichen** von Indikatoren. (Der Standard im Sinne eines Referenzwertes für ein mithilfe von Indikatoren gemessenes Kriterium ist – dies sei noch einmal betont – nicht zu verwechseln mit dem Standard als allgemeiner Richtlinie des wissenschaftlichen Vorgehens.)
  Beispiel: Es wird oft festgelegt, dass das quantitative Gütekriterium der Messgenauigkeit erfüllt ist, wenn das Messinstrument einen Reliabilitätskoeffizienten von mindestens .80 (auf einer Skala von 0 bis 1) aufweist. Das Problem besteht nun darin, dass zu vielen Qualitätskriterien bzw. Qualitätsindikatoren keine allgemeingültigen Standards bzw. Referenzbereiche angegeben werden können, sondern dass die Beurteilung meist unter Berücksichtigung der Besonderheiten der jeweiligen Studie erfolgen muss. So hängen z. B. auch die Standards für den Qualitätsindikator der Messgenauigkeit von der Art des Messinstrumentes sowie vom Typus des gemessenen Merkmals ab und können nicht allgemeingültig auf einen Mindeststandard von .80 festgelegt werden. Ebenso existiert kein eindeutiger Standard, der angibt, unter welchen Bedingungen in der qualitativen Forschung eine kommunikative oder argumentative Validierung als mangelhaft, ausreichend, gut oder exzellent zu betrachten ist. Denn dazu müsste allgemeingültig festgelegt werden können, wie viele der Interviewpartner in welchem Verfahren (z. B. mündlich, per E-Mail) wie umfassend und mit welchen Argumenten den Deutungen der Forschenden zustimmen (hohe kommunikative Validität) oder wie viele wie stark überzeugende Argumente zur Stützung einer Interpretation notwendig sind (hohe argumentative Validität).

Letztlich läuft die Qualitätsbeurteilung wissenschaftlicher Forschung darauf hinaus, eine Studie hinsichtlich möglichst aller einschlägigen Qualitätskriterien zu betrachten (d. h. entsprechende Qualitätsindikatoren zu erfassen), deren Ausprägung zu berichten und anhand von Beurteilungsmaßstäben argumentativ einzuordnen. Ein Verweis auf allgemeingültige Standards ist dabei oft nicht möglich, vielmehr muss die Ausprägung jedes Qualitätskriteriums mit Blick auf die Zielsetzung und die Ressourcen der Studie sowie im Vergleich zu üblichen Studien des Forschungsfeldes bewertet werden. Welche Qualitätskriterien dabei im Einzelnen zu betrachten sind, wird in vielen Bereichen der Wissenschaft durch **Kriterienkataloge** festgelegt. Solche Kriterienkataloge sind unterschiedlichen Quellen zu entnehmen, z. B. diesen:

- fachwissenschaftliche Methodenliteratur (z. B. Lincoln & Guba, 1985; Shadish, Cook, & Campbell, 2002),
- wissenschaftliche Fachgesellschaften (z. B. Richtlinien zur Präsentation von Studien in Publikationen der American Psychological Association [APA]: Cooper, 2010; Kriterien ethischer Strenge für Wissenschaft und Forschung der APA: Sales & Folkman, 2000),
- Fachzeitschriften (z. B. Mindeststandards für publikationsfähige Manuskripte der *Zeitschrift für Erziehungswissenschaft:* Beiheft 4/2005 zu „Standards und Standardisierungen in der Erziehungswissenschaft"),
- Branchenverbände (z. B. zahlreiche Richtlinien zur Qualität von Marktforschung: http://ratmarktforschung.de/richtlinien/; internationaler Kodex für ethische Marktforschung: ICC/ESOMAR-Codex: http://www.esomar.org/knowledge-and-standards.php),
- ISO-Normen mit Qualitätsanforderungen für bestimmte Forschungsfelder (z. B. Normen für Markt-, Meinungs- und Sozialforschung: DIN ISO 20252:2012),
- Forschungsförderer (z. B. Richtlinien der Deutsche Forschungsgemeinschaft [DFG] zu förderungswürdigen Projekten),
- Hochschulinstitute (z. B. Kriterienlisten zur Bewertung der wissenschaftlichen Qualität studentischer Hausarbeiten und Abschlussarbeiten, die den Studierenden über das institutseigene Webangebote bekannt gemacht werden).

Entsprechende Kriterienkataloge unterscheiden sich u. a. in ihrem Detaillierungsgrad, in der Bezeichnung, Strukturierung sowie Gewichtung der Kriterien, basieren jedoch im Kern meist auf einem konsensfähigen Grundkanon an Kriterien wissenschaftlicher Qualität.

Neben der Kenntnis der einschlägigen Kriterien und Kriterienkataloge – sowie ausreichender Übung in deren korrekter Anwendung (für ein Lehrbuch mit vielen Beispielen zur Bewertung der Qualität von sozialwissenschaftlichen Zeitschriftenartikeln s. Girden & Kabacoff, 2010) – ist es im Rahmen der Entwicklung von Methodenkompetenz (► Abschn. 1.1.1) auch wichtig, auf der Metaebene zu hinterfragen, welche Gültigkeit eine eigene oder fremde Qualitätsbewertung einer wissenschaftlichen Studie für sich beanspruchen kann: Wird die Güte der Studie von den Urteilenden angemessen eingestuft oder vielleicht über- oder unterschätzt? Die ausdrückliche und verbindliche Festlegung von Qualitätskriterien, -indikatoren und -standards sowie das Einholen mehrerer Expertenurteile (z. B. zwei bis drei Gutachten bewerten die wissenschaftliche Qualität einer Doktorarbeit oder eines eingereichten Zeitschriftenartikels) sowie das Verfahren der doppelblinden Begutachtung – bei der Gutachtende und Begutachtete jeweils anonym bleiben – sind Maßnahmen des Wissenschaftssystems, um zu möglichst unverzerrten und fairen Qualitätsbewertungen von Forschungsarbeiten und den aus ihnen abgeleiteten wissenschaftlichen Erkenntnissen zu kommen.

3.1 Wissenschaftlichkeit und wissenschaftliche Qualität

Was zeichnet Wissenschaft aus? Wie kann man eine wissenschaftliche Studie von nicht-wissenschaftlichem oder gar pseudowissenschaftlichem Vorgehen abgrenzen? Was kennzeichnet gute Wissenschaft? Wie kann man hochwertige wissenschaftliche Arbeiten von durchschnittlichen oder schwachen Beiträgen differenzieren?

Die Bewertung wissenschaftlicher Qualität erfolgt innerhalb des Wissenschaftssystems durch unterschiedliche **wissenschaftsinterne Anspruchsgruppen** mit verschiedenen Zielen:

- Fachkollegen bewerten im Zuge des Peer-Review-Verfahrens die Qualität von eingereichten Zeitschriftenartikeln, was die Basis für Publikationsentscheidungen bildet;
- Fachkollegen bewerten im Zuge des Peer-Review-Verfahrens die Qualität von Projektanträgen, was die Basis für Förderentscheidungen darstellt;
- Hochschullehrende bewerten die Qualität studentischer Qualifikationsarbeiten, um Prüfungsnoten zu vergeben;
- Forschende wählen in der Rolle von Lehrenden besonders gute Studien aus, um sie als Beispiele in der Ausbildung der Studierenden zu verwenden;
- Forschende bewerten die Qualität von Vorläuferstudien, auf denen sie ihre eigene Untersuchung aufbauen oder die sie einer Replikation unterziehen wollen;
- Forschende bewerten die Qualität von publizierten Studien, die sie im Zuge systematischer Forschungssynthese (wie z. B. einer Metaanalyse, ► Kap. 16) gewichtet nach ihrer Qualität zusammenfassen wollen.

Darüber hinaus wird wissenschaftliche Qualität aber auch außerhalb des Wissenschaftssystems durch diverse **wissenschaftsexterne Anspruchsgruppen** beurteilt:

- Stipendiengeber bewerten die Qualität von Promotionsvorhaben, um Förderentscheidungen zu treffen;
- Forschungsförderer bewerten die Qualität von Projektanträgen, um Fördergelder zu vergeben;
- Auftraggebende bewerten den Nutzen der von ihnen finanzierten anwendungswissenschaftlichen Studien (z. B. Marktforschungs- oder Evaluationsstudien);

- Politiker bewerten die Qualität und die Ausrichtung der zum großen Teil staatlich finanzierten Grundlagenforschung;
- Journalisten suchen und bewerten wissenschaftliche Studien, die einen hohen Nachrichtenwert aufweisen bzw. von öffentlichem Interesse sind;
- die Bürgerinnen und Bürger bewerten wissenschaftliche Studien, über die in den Massenmedien oder im Internet berichtet wird, hinsichtlich ihrer Interessantheit und ihres Nutzens für die Allgemeinheit;
- gesellschaftliche Minderheiten bewerten wissenschaftliche Studien, die sich mit ihrer Lebenssituation befassen, hinsichtlich möglicher Missverständnisse und Diskriminierungen;
- Untersuchungspersonen bewerten die Ergebnisse von Studien, an denen sie selbst teilgenommen haben, hinsichtlich ihrer Aussagekraft.

Die **Festlegung und Anwendung von Qualitätskriterien** in der empirischen Sozialforschung ist ihrerseits ein Forschungsgegenstand (Barker & Pistrang, 2005; Breuer & Reichertz, 2001; Bryman, Becker, & Sempik, 2008). So wird z. B. untersucht,

- welche Kriterien der wissenschaftlichen Qualität verschiedene Anspruchsgruppen anlegen (z. B. Befragung von Journalisten, Politikern, Bürgern zu ihren Anforderungen an gute sozialwissenschaftliche Studien);
- ob und wie sich die Gewichtung von Qualitätskriterien historisch verändert (steigen oder sinken innerhalb des Wissenschaftssystems und speziell innerhalb der Sozial- und Humanwissenschaften die Anforderungen an methodische Strenge in den letzten Dekaden?);
- ob bestimmte Qualitätskriterien in einem Spannungsverhältnis zueinander stehen (geht die praktische Nützlichkeit einer Studie meist mit geringerer methodischer Strenge einher?);
- wie zuverlässig Kriterienkataloge angewendet werden (z. B. wie hoch ist die Urteilerübereinstimmung, wenn derselbe Kriterienkatalog zur Einschätzung derselben Studien von unterschiedlichen Urteilenden genutzt wird?);
- wie zuverlässig und unverzerrt das Peer-Review-Verfahren ist (z. B. mit welcher Wahrscheinlichkeit werden Qualitätsdefizite wie Fehler in der statistischen Auswertung oder Plagiate tatsächlich aufgedeckt?) oder
- durch welche Qualitätsmerkmale sich besonders erfolgreiche (z. B. vielzitierte) Studien von anderen unterscheiden.

Probleme und Grenzen der Qualitätsbewertung von Wissenschaft aufzuzeigen, Kriterienkataloge und deren sachgerechte Anwendung immer wieder auf den Prüfstand zu stellen, entspricht dem Selbstverständnis der Wissenschaft, die nicht nur Wissen produziert, sondern den Erkenntnisprozess selbst beobachtet und reflektiert (► Kap. 2). Unschärfen und Kontroversen in der Bewertung wissenschaftlicher Qualität sollten nicht zum Anlass genommen werden, Qualitätsbetrachtungen für obsolet zu erklären, sondern die Prozesse der wissenschaftlichen Qualitätssicherung in der Praxis realistisch einzuschätzen und bei Bedarf im eigenen Arbeitsfeld zu verbessern.

3.1.1 Vier Standards der Wissenschaftlichkeit

Um Wissenschaft zu kennzeichnen, sind zunächst vier allgemeine **Standards der Wissenschaftlichkeit** anzulegen, die in allen Erfahrungswissenschaften und somit auch in allen sozial- und humanwissenschaftlichen Fächern zur Anwendung kommen und sowohl mit dem quantitativen, dem qualitativen sowie dem Mixed-Methods-Paradigmen übereinstimmen (► Kap. 2):

1. Formulierung eines **wissenschaftlichen Forschungsproblems**: Bei dem für die Studie ausgewählten Thema und Forschungsproblem (meist ausdifferenziert in Forschungsfragen und/oder Forschungshypothesen) müssen empirisch untersuchbare und auf dem aktuellen wissenschaftlichen Erkenntnisstand theoretisch erklärbare Sachverhalte adressiert werden. Das gewählte Forschungsproblem muss also in einen bestehenden wissenschaftlichen Forschungs- und Publikationskontext einzuordnen sein (z. B. stellt die Untersuchung von „interkultureller Kommunikation“ ein wissenschaftliches Forschungsthema dar, nicht aber die „Kommunikation mit Außerirdischen“, da die Existenz von Außerirdischen beim aktuellen Kenntnisstand wissenschaftlich nicht anerkannt ist – dementsprechend ist die Behandlung entsprechender Fragen Gegenstand sog. wissenschaftlicher Grenzgebiete bzw. der Parawissenschaft).
2. Realisierung eines – auf das Forschungsproblem zugeschnittenen – **wissenschaftlichen Forschungsprozesses** und Einsatz etablierter wissenschaftlicher Forschungsmethoden und Techniken wissenschaftlichen Arbeitens in allen Phasen des Forschungsprozesses: Die Studie muss als Ablauf eines in verschiedene Phasen strukturierten Forschungsprozesses geplant und umgesetzt werden. Die einzelnen Phasen des Forschungsprozesses, ihre Reihenfolge und Ausgestaltung müssen dabei durch gängige wissenschaftstheoretische Paradigmen und **Methodologien** begründet sein. Das Vorgehen sollte nicht zuletzt durch Verweise auf die Methodenliteratur und/oder das Vorgehen

Tabelle 3.1 Abgrenzung der Wissenschaft von Nicht-, Pseudo- und Parawissenschaft anhand der vier Standards der Wissenschaftlichkeit sowie dem selbsterklärten Anspruch der Wissenschaftlichkeit

	Standard 1: Wissenschaftliches Forschungsproblem	Standard 2: Wissenschaftlicher Forschungsprozess	Standard 3: Wissenschafts- und Forschungsethik	Standard 4: Dokumentation des Forschungsprojektes	Selbsterklärter Anspruch der Wissenschaftlichkeit
Wissenschaft	+	+	+	+	+
Nicht-Wissenschaft	Teilweise	–	–	–	–
Pseudowissenschaft	Teilweise	_	Teilweise	Teilweise	+
Parawissenschaft	–	+	+	+	+

in inhaltlich vergleichbaren publizierten Studien legitimiert sein (z. B. Gestaltung des Forschungsprozesses gemäß Methodologie der quantitativen Experimentalforschung oder gemäß Methodologie der qualitativen ethnografischen Feldforschung). Dass wissenschaftliche **Methoden** zum Einsatz kommen, ist daran festzumachen, dass man sich auf Methoden und Techniken stützt, die in der fachspezifischen Methodenliteratur vorgeschlagen werden und/oder in inhaltlich vergleichbaren publizierten Studien zum Einsatz kommen (z. B. Technik der systematischen Literaturrecherche in wissenschaftlichen Literatur-Datenbanken; Nutzung einer wissenschaftlichen Methode der Datenanalyse wie z. B. eines statistischen Signifikanztests oder einer qualitativen Inhaltsanalyse). Die gewählte Methodologie mitsamt der einzelnen Methoden der Datenerhebung und Datenanalyse muss wissenschaftlich anerkannt sein und dabei auch zum jeweiligen Forschungsproblem passen.

3. Orientierung an der **Wissenschafts- und Forschungsethik:** Der Forschungsprozess muss ethischen Regeln folgen. Eine Studie verliert z. B. den Charakter der Wissenschaftlichkeit, wenn Daten manipuliert, Ideen gestohlen, Sponsoren oder Interessenskonflikte der Forschenden verschwiegen werden (Wissenschaftsethik) und/oder wenn die Untersuchungsteilnehmenden ungerechtfertigt beeinträchtigt oder gar geschädigt werden (Forschungsethik).
4. Vollständige schriftliche **Dokumentation des gesamten Forschungsprojekts** inklusive Archivierung des Datenmaterials: Der Forschungsprozess muss in seinen einzelnen Phasen durch eine genaue Dokumentation intersubjektiv nachvollziehbar bzw. replizierbar sein (d. h. detaillierte Beschreibung und Begründung des Forschungsproblems sowie der gewählten Methodologie und Methoden und ihrer ethischen Aspekte inklusive Dokumentation aller Forschungsmaterialien wie z. B. Fragebögen oder Interviewleitfäden und vor allem der erhobenen Daten sowie der Umstände der Datenerhebung). Ohne eine solche Dokumentation können Außenstehende die Studie nicht einschätzen.

Dass Wissenschaftlichkeit daran gebunden ist, sich nachweislich und ausschließlich auf etablierte wissenschaftliche Methodologien und Methoden zu stützen, ist jedoch nicht als Verbot methodischer Innovation zu verstehen, die gerade auch ein Kennzeichen von Wissenschaft ist. Sofern grundlegende Neuerungen im Forschungsprozess oder in einzelnen Forschungsmethoden eingeführt werden sollen, verlangen diese jedoch eine überzeugende Begründung und müssen von anderen Fachvertretern anerkannt werden.

Die Beurteilung der Wissenschaftlichkeit erfolgt typischerweise **nicht** graduell bzw. als Kontinuum („mehr oder weniger wissenschaftliche" Arbeit), sondern kategorial: Sofern die vier Standards der Wissenschaftlichkeit **prinzipiell** erfüllt bzw. keine groben Verstöße erkennbar sind (als Indikatoren fungieren vor allem die Vollständigkeit des Forschungsberichtes sowie die entsprechenden Angaben zur Wahl von Forschungsthema, Methodologie und Methoden sowie zum Umgang mit ethischen Fragen in selbigem), wird die Studie als wissenschaftlich eingeordnet und somit von anderen Varianten des Erkenntnisgewinns abgegrenzt. Unter den von Wissenschaft abweichenden Erkenntnisformen sind Nicht-Wissenschaft, Pseudowissenschaft und Parawissenschaft besonders hervorzuheben; sie verletzen unterschiedliche Standards der Wissenschaftlichkeit (Tab. 3.1).

Nicht-Wissenschaft

Nicht-Wissenschaft umfasst die unterschiedlichsten im Alltag gebräuchlichen Erkenntnisformen: Menschen generieren und begründen Wissen, indem sie sich auf eigene Lebenserfahrungen, Intuition, gesunden Menschenverstand, Logik, Autoritätspersonen, religiöse Lehren etc. stützen (▶ Abschn. 1.1.1). Diese Formen des Wissens er-

weisen sich häufig als brauchbar, haben jedoch Grenzen, die durch wissenschaftlichen Erkenntnisgewinn überwunden werden können und sollen. Nicht-wissenschaftlicher Erkenntnisgewinn erfüllt nicht die Standards der Wissenschaftlichkeit (insbesondere nicht die Standards zu Forschungsprozess und Dokumentation). Nicht-Wissenschaft beschäftigt sich jedoch teilweise mit denselben sozialen Sachverhalten wie die empirische Sozialforschung (z. B. macht man sich auch im Alltag Gedanken über den Entwicklungsstand von Kindern, die Ursachen von politischen Konflikten oder die Gesundheitsschädlichkeit bestimmter Ernährungsgewohnheiten). Allerdings wird bei Nicht-Wissenschaft **kein Anspruch auf Wissenschaftlichkeit** erhoben. Man spricht deswegen im Zusammenhang mit Nicht-Wissenschaft lieber von „vorwissenschaftlichem" Erkenntnisgewinn oder von „Alltagswissen" statt von „unwissenschaftlichem" Vorgehen.

Beispiel: Im vorliegenden Buch wird immer wieder auf nicht-wissenschaftliche Erkenntnisse zurückgegriffen, so etwa, wenn auf „typische Anfängerfehler" bei Studierenden hingewiesen wird. Hierbei handelt es sich um subjektive Einschätzungen auf der Basis der eigenen Lehrpraxis und Studierendenbetreuung, die sich als hoffentlich nützliche Heuristik erweisen, um der Leserschaft Tipps an die Hand zu geben. Eine wissenschaftliche Studie könnte darüber hinaus viel umfassender und zuverlässiger Auskunft darüber geben, welche Methodenfehler oder methodischen Missverständnisse bei Studierenden unterschiedlicher sozialwissenschaftlicher Fächer heutzutage besonders verbreitet sind, die dementsprechend in Lehrbüchern verstärkt aufzugreifen wären.

Es ist zu beachten, dass Fachbegriffe aus der Wissenschaft (z. B. „Experiment", „Test", „Studie", „Umfrage") teilweise auch im nicht-wissenschaftlichen Bereich genutzt werden.

Beispiel: Die Redakteurin einer Frauenzeitschrift lebt eine Woche ohne Handy und schildert die Ergebnisse dieses **„Experiments"** in einer Reportage, wobei in erster Linie der Unterhaltungswert zählt und nicht der Anspruch erhoben wird, allgemeingültige Erkenntnisse zu gewinnen (zu den Anforderungen, die an ein wissenschaftliches Experiment gestellt werden, ► Abschn. 7.6.1).

Dasselbe gilt für nicht-wissenschaftliche Methoden wie sog. **„Psychotests"** in Publikumszeitschriften, die ebenfalls vor allem Unterhaltungswert haben sollen und nicht vergleichbar sind mit psychologischen Testverfahren als wissenschaftlichen Datenerhebungsmethoden, die entsprechende Gütekriterien erfüllen müssen (► Abschn. 10.4.1 „Abgrenzung zwischen wissenschaftlichen psychologischen Tests und nicht-wissenschaftlichen Tests"). Vorsicht geboten ist auch bei nicht-wissenschaftlichen **„Online-Umfragen"** in Form von spontanen Votings, die z. B. auf Facebook-Profilen oder den Webseiten von Online-Magazinen durchgeführt werden und deren Ergebnisse aufgrund der hochgradig selektierten Stichproben keinesfalls als „Meinungsbild" der Bevölkerung interpretierbar sind (zu unterschiedlichen Stichprobenarten und ihrer Aussagekraft ► Kap. 9).

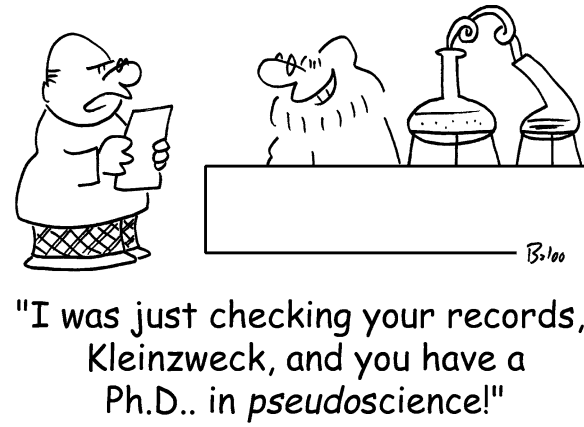
Pseudowissenschaft

Pseudowissenschaft ist dann gegeben, wenn ausdrücklich der Anspruch der Wissenschaftlichkeit erhoben, jedoch faktisch nicht eingelöst wird. Insbesondere die Orientierung an wissenschaftlichen Methodologien und Methoden ist bei Pseudowissenschaft in der Regel verletzt, zudem wird bei der Darstellung des Forschungsproblems meist der aktuelle Forschungsstand ungenügend aufgegriffen. Pseudowissenschaft existiert vor allem in drei unterschiedlichen Varianten, die sich durch die Gründe bzw. Motive der Verletzung der Standards der Wissenschaftlichkeit unterscheiden:

1. **Pseudowissenschaft durch unbeabsichtigte Verletzung von Standards der Wissenschaftlichkeit:** Diese Form der **unbeabsichtigten Pseudowissenschaft** entsteht zum großen Teil aufgrund **mangelnder Methodenkompetenz** (► Abschn. 1.1.1). Die Forschenden glauben, dass sie wissenschaftlich operieren, machen jedoch so gravierende Fehler, dass nicht einmal mehr eine mangelhafte wissenschaftliche Arbeit resultiert, sondern die Wissenschaftlichkeit gänzlich verneint werden muss. Wenn etwa Studierende in ihren Abschlussarbeiten ankündigen, eine wissenschaftliche Studie vorzulegen und wissenschaftliche Gültigkeit für ihre Ergebnisse beanspruchen, dabei aber nicht alle Standards der Wissenschaftlichkeit eingehalten haben, muss leider Pseudowissenschaft konstatiert werden.
   Beispiel: Eine studentische Arbeitsgruppe untersucht Schüchternheit, arbeitet jedoch mangels besseren Wissens primär mit Alltagstheorien (d. h. ignoriert den wissenschaftlichen Forschungsstand) und erfasst Schüchternheit mit einem in einer Publikumszeitschrift gefundenen Psychotest (d. h. nutzt keine etablierte wissenschaftliche Methode zur Messung von Schüchternheit).
   Auch **organisatorische Fehlplanungen** im Forschungsprozess können im Extremfall die Wissenschaftlichkeit kosten, etwa wenn unter übergroßem Zeitdruck kurz vor dem Abgabetermin die Datenanalyse nicht mehr systematisch mit geeigneten wissenschaftlichen Methoden erfolgt, sondern nur selektiv Einzelergebnisse herausgegriffen und als wissenschaftlicher Befund präsentiert werden. Nicht zuletzt führt die **Wahl ungeeigneter (d. h. nicht zum Forschungsproblem passender) wissenschaftlicher**

Methoden bzw. Methodologien zu pseudowissenschaftlichen Studien: So etwa, wenn angegeben wird, man wolle Kausalhypothesen im quantitativen Ansatz prüfen und dann de facto ein nicht-experimentelles Design realisiert wird, das zur Analyse von Kausalität ungeeignet ist. Unbeabsichtigter Pseudowissenschaft bzw. Unwissenschaftlichkeit gilt es durch gute Methodenausbildung vorzubeugen.

2. **Pseudowissenschaft durch vorsätzliche Verletzung von Standards der Wissenschaftlichkeit für persönliche oder wirtschaftliche Vorteile:** Diese Form der **vorsätzlichen Pseudowissenschaft** erfolgt bewusst und absichtlich unter Verletzung von Regeln der Wissenschaftsethik, etwa weil man sich dadurch Arbeit sparen und akademische oder ökonomische Vorteile verschaffen möchte. Ein Beispiel ist die Wissenschaftsfälschung, bei der der Forschungsprozess gar nicht wie vorgegeben durchlaufen wird, sondern die Daten erfunden oder beliebig manipuliert werden, um publikations- und prestigeträchtige Ergebnisse zu erhalten (▶ Abschn. 4.2.2). Vorsätzliche Pseudowissenschaft ist geächtet und wird von verschiedenen Instanzen sanktioniert.
3. **Pseudowissenschaft durch Abweichung von Standards der Wissenschaftlichkeit im Rahmen eines bestimmten Überzeugungssystems und Weltbildes:** Bei dieser Form der **ideologischen Pseudowissenschaft** geht es um Theorien, Phänomene, Produkte oder Verfahren, für die wissenschaftliche Fundierung beansprucht wird, a) ohne dass diese nach gängigen Standards der Wissenschaftlichkeit ausreichend vorhanden ist oder b) obwohl diese nach gängigen Kriterien der Wissenschaftlichkeit als nicht existent einzustufen ist (◘ Abb. 3.1). Ideologische Pseudowissenschaft wird nicht nur von einzelnen Forschenden, sondern von ganzen Bevölkerungsgruppen getragen, die entsprechende Überzeugungen teilen und an diesen festhalten wollen. Widersprechende wissenschaftliche Befunde und kritische theoretische Positionen aus der Mainstream-Wissenschaft werden nicht anerkannt bzw. aktiv bekämpft. Kontroversen an der Grenze zwischen Wissenschaft und ideologischer Pseudowissenschaft werden teilweise sehr heftig geführt und gehen nicht selten mit persönlichen Angriffen einher, da hier konkurrierende Weltbilder aufeinander treffen. Anhänger von pseudowissenschaftlichen Vorstellungen fühlen sich meist von der Wissenschaft unrechtmäßig ausgegrenzt, während Vertreter der Wissenschaft davor warnen, dass ideologische Pseudowissenschaften auf dem Vormarsch sind und durch Fehlinformation der Öffentlichkeit große Gefahren bergen können.



◘ **Abb. 3.1** In der Pseudowissenschaft wird der Anspruch von Wissenschaftlichkeit erhoben, jedoch nicht eingelöst. (© Baloo -Rex May-/Search ID: rman7521, Rights Available from CartoonStock.com)

Beispiele für pseudowissenschaftliche Behauptungen: Die MMR-Kombinationsimpfung (Masern, Mumps und Röteln) bei Kleinkindern kann Autismus auslösen, somit sollten Kinder nicht geimpft werden (eine Position der sog. Impfgegner; Meyer & Reiter, 2004). Ein Klimawandel findet nicht statt bzw. ist nicht durch Menschen verursacht bzw. hat keine dramatischen negativen Konsequenzen, somit sind Klimaschutzmaßnahmen unsinnig (Position der sog. Klimaskeptiker; Boykoff & Boykoff, 2004; Rahmstorf, 2007). Mit Sauerstoff angereichertes Tafelwasser erhöht die Leistungsfähigkeit (Werbebehauptung zur Vermarktung von Erfrischungsgetränken). Homosexualität ist krankhaft und kann durch eine „Konversionstherapie" geheilt werden (Position einiger religiös-christlicher Therapeuten und der sog. Ex-Gay-Bewegung; vgl. Haldeman, 1999; Haldeman, 2002).

Nicht zuletzt in Medizin und Psychologie ist es der wissenschaftlichen Profession wichtig, die praktisch tätigen Ärzte und Psychotherapeuten sowie die Bevölkerung darüber aufzuklären, welche Therapieverfahren als wissenschaftlich nachweisbar wirksam (**evidenzbasiert**) gelten können und bei welchen älteren oder neueren Ansätzen es sich um pseudowissenschaftliche Heilslehren handelt, die wirkungslos oder sogar schädlich sind (zur Abgrenzung evidenzbasierter und pseudowissenschaftlicher psychotherapeutischer Interventionen s. z. B. Lilienfeld, Lynn, & Lohr, 2003; Lilienfeld, Ruscio, & Lynn, 2008; Olatunji, Parker, & Lohr, 2005). Die Aufdeckung von Pseudowissenschaft erfordert eine sachliche Bewertung anhand der vorliegenden wissenschaftlichen Erkenntnisse, d. h. der Befunde von empirischen Studien, die Standards der Wissenschaftlichkeit erfüllen. Dabei ist oft festzustellen, dass sich pseudowissenschaftliche Behauptungen auf Studien stützen,

die die Standards der Wissenschaftlichkeit unzureichend erfüllen und dass umgekehrt Gegenevidenzen aus Studien, die Standards der Wissenschaftlichkeit genügen, dennoch nicht anerkannt werden. Die Abgrenzung zwischen Wissenschaft und ideologischer Pseudowissenschaft kann ohne umfassende Kenntnis des Forschungsstandes im Einzelfall schwierig sein, weil auch pseudowissenschaftliche Positionen zuweilen große öffentliche Aufmerksamkeit genießen. Deswegen sind die wissenschaftlichen Fachgesellschaften gehalten, Stellungnahmen herauszugeben, um populäre pseudowissenschaftliche Positionen als solche zu kennzeichnen.

Parawissenschaft

Parawissenschaft erhebt den Anspruch der Wissenschaftlichkeit und arbeitet auch korrekt auf der Basis wissenschaftlicher Forschungsmethoden. Was Parawissenschaft jedoch von der etablierten Wissenschaft unterscheidet ist ihr Forschungsgegenstand bzw. das gewählte Forschungsproblem. Es bezieht sich auf Phänomene, deren Existenz und Beschaffenheit auf aktuellem wissenschaftlichem Forschungsstand theoretisch nicht erklärbar sind. Man spricht auch von Forschung zu **wissenschaftlichen Grenzgebieten**.

Ein Beispiel für eine Parawissenschaft ist die **Parapsychologie**, die sich mit sog. Psi-Phänomenen befasst, also mit außersinnlicher Wahrnehmung (z. B. Gedankenübertragung, Kontakte mit Verstorbenen) sowie mit Psycho- bzw. Telekinese (z. B. Löffelbiegen mit purer Gedankenkraft). Es gibt eine umfassende Forschung zu Psi-Phänomenen, die Standards der Wissenschaftlichkeit erfüllt: Anomale Bewusstseinsphänomene wie Gedankenübertragung werden u. a. in experimentellen Studien mit modernen psychophysiologischen Messmethoden (z. B. EEG; ► Abschn. 10.5.2) untersucht (zum Überblick s. Ambach, 2012). Führende psychologische Fachzeitschriften wie „Psychological Bulletin“ veröffentlichen Artikel, in denen der Forschungsstand zu Psi-Phänomenen zusammengefasst wird (z. B. Storm, Tressoldi, & Di Risio, 2010). Auch ist eine Fortentwicklung des Theoriestandes in diesem Forschungsfeld zu konstatieren (z. B. Carpenter, 2012).

Ein nicht nur psychologisches, sondern interdisziplinäres Forschungsfeld, das sich mit der wissenschaftlichen Untersuchung aktuell noch unerklärbarer bzw. anomaler Sachverhalte befasst (UFO, Geister etc.), ist die **Anomalistik** (http://www.anomalistik.de). In der Parawissenschaft bzw. Anomalistik wendet man sich den anomalen Sachverhalten ergebnisoffen zu. Weder geht es darum, ihre Existenz unbedingt zu verteidigen (das ist die Haltung der „Gläubigen“ sowie einiger Pseudowissenschaftler), noch darum, ihre Existenz von vorne herein zu leugnen und über die Phänomene zu spotten (das ist die Haltung vieler „Skeptiker“). Vielmehr ist es ein Anliegen zu versuchen, mit etablierten wissenschaftlichen Methoden systematisch Daten zu sammeln und auszuwerten, Theorien zu entwickeln und auf den Prüfstand zu stellen (Truzzi, 1996). So kann eine wissenschaftliche Auseinandersetzung mit dem UFO-Phänomen bedeuten, dass man durch Befragung erhebt, welche unterschiedlichen und übereinstimmenden Vorstellungen über das Phänomen in der UFO-Szene verbreitet sind (Wunder, 2001), dass man die Haltung der Wissenschaft zum UFO-Phänomen politikwissenschaftlich anhand etablierter Theorien diskutiert (Wendt & Duvall, 2008) oder das methodische Vorgehen bei UFO-Falluntersuchungen an wissenschaftlichen Standards misst (Ickinger, 2006).

Wissenschaft – Wissenschaft („science“) erhebt den Anspruch der Wissenschaftlichkeit und erfüllt alle vier Standards der Wissenschaftlichkeit: 1. wissenschaftliches Forschungsproblem, 2. wissenschaftlicher Forschungsprozess, 3. Wissenschafts- und Forschungsethik sowie 4. Dokumentation des Forschungsprojektes.

Nicht-Wissenschaft – Nicht-Wissenschaft („non science“) befasst sich häufig mit ähnlichen Fragestellungen wie wissenschaftliche Forschung (das gilt besonders für die Human- und Sozialwissenschaften mit ihren oft alltagsnahen Untersuchungsgegenständen). Allerdings wird bei Nicht-Wissenschaft kein geordneter Forschungsprozess durchlaufen und auch nicht der Anspruch der Wissenschaftlichkeit erhoben.

Pseudowissenschaft – Pseudowissenschaft („pseudo science“) erhebt den Anspruch der Wissenschaftlichkeit, ohne ihn jedoch nach gängigen Standards der Wissenschaftlichkeit einzulösen. Grob lassen sich unbeabsichtigte, vorsätzliche und ideologische Pseudowissenschaft voneinander abgrenzen. Die Abgrenzung zwischen Wissenschaft und ideologischer Pseudowissenschaft geht oft mit hochkontroversen öffentlichen Debatten einher.

Parawissenschaft – Parawissenschaft („para science“) beschäftigt sich mit anomalen Sachverhalten, deren Existenz und Beschaffenheit auf dem aktuellen wissenschaftlichen Kenntnisstand nicht erklärbar sind. Dabei wird gemäß wissenschaftlicher Methodologie und Methodik vorgegangen und auch der Anspruch der Wissenschaftlichkeit erhoben. Mit zunehmendem Erkenntnisgewinn kann sich eine Parawissenschaft zu einer regulären Wissenschaft entwickeln.

3.1.2 Vier Kriterien der wissenschaftlichen Qualität

Sind die vier Standards der Wissenschaftlichkeit prinzipiell erfüllt, stellt sich die Anschlussfrage, ob es sich um eine gute, durchschnittliche oder eher mangelhafte wissenschaftliche Untersuchung handelt, womit dann auch die Qualität der resultierenden wissenschaftlichen Erkenntnisse angesprochen ist. Für eine entsprechende Beurteilung lassen sich zu allen vier Standards der Wissenschaftlichkeit zusätzliche Qualitätsanforderungen stellen (Tab. 3.2). Auch diese sind wie die Standards der Wissenschaftlichkeit als paradigmen- und disziplin-

Tabelle 3.2 Die vier Standards der Wissenschaftlichkeit und die zugehörigen vier Kriterien der wissenschaftlichen Qualität im Überblick

Standards der Wissenschaftlichkeit *Sie müssen von jeder wissenschaftlichen Studie prinzipiell eingehalten werden.*	Kommentar *Diese Fragen müssen bei einer wissenschaftlichen Studie prinzipiell bejaht werden können.*	Kriterien der wissenschaftlichen Qualität *Sie sind bei wissenschaftlichen Studien graduell sehr unterschiedlich ausgeprägt und differenzieren herausragende, gute, durchschnittliche und schwache Studien.*	Kommentar *Bei diesen Fragen ist anhand von Vergleichsstudien, Referenzwerten aus der Methodenlehre und inhaltlichen Argumenten der Grad der Ausprägung in der jeweiligen wissenschaftlichen Studie abzuschätzen, um ihre Qualität einzustufen.*
1. Wissenschaftliches Forschungsproblem	Bearbeitet die Studie ein Forschungsproblem, das sich in einen anerkannten wissenschaftlichen Forschungs- und Publikationskontext einordnet?	**1. Inhaltliche Relevanz**	In welchem Maße trägt die Studie mit ihren Ergebnissen a) zum wissenschaftlichen Erkenntnisfortschritt (wissenschaftliche/theoretische Relevanz) und/oder b) zur Lösung praktischer Probleme (praktische Relevanz) bei?
2. Wissenschaftlicher Forschungsprozess	Orientiert sich die Studie an etablierten wissenschaftlichen Methodologien und Methoden, die zum Forschungsproblem passen?	**2. Methodische Strenge**	Wie anspruchsvoll sind die gewählten Methodologien und Methoden, wie gut sind sie zur Bearbeitung des Forschungsproblems geeignet, und wie regelkonform werden sie umgesetzt?
3. Wissenschafts- und Forschungsethik	Folgt die Studie den Prinzipien der Wissenschafts- und Forschungsethik?	**3. Ethische Strenge**	Wie konsequent und umfassend werden einzelne Standards der Wissenschafts- und Forschungsethik erfüllt?
4. Dokumentation des Forschungsprojektes	Sind Vorgehen und Ergebnisse der wissenschaftlichen Studie im Detail nachvollziehbar dokumentiert?	**4. Präsentationsqualität**	Wie vollständig, wohlstrukturiert und gut lesbar wird die Studie in ihrem Ablauf und mit ihren Befunden präsentiert und wie umfassend werden die Standards der Berichterstattung des jeweiligen Faches und Publikationsorgans eingehalten?

übergreifend zu betrachten, müssen jedoch disziplin- und paradigmenspezifisch konkretisiert werden. Neben der inhaltlichen Relevanz und der methodischen Strenge – im Englischen als Begriffspaar **„relevance and rigor“** geläufig – werden als weitere Kriterien der wissenschaftlichen Qualität die ethische Strenge und die Präsentationsqualität in jüngster Zeit stärker in der Methodendiskussion hervorgehoben.

1. **Inhaltliche Relevanz** („relevance“): Die Wissenschaftlichkeit verlangt die Wahl eines wissenschaftlichen Forschungsproblems (► Abschn. 5.1). Eine **gute** wissenschaftliche Studie zeichnet sich dadurch aus, dass sie ein inhaltlich besonders relevantes wissenschaftliches Forschungsproblem bearbeitet, wobei in der Grundlagenforschung die wissenschaftliche Relevanz und in der Anwendungsforschung die praktische Relevanz im Vordergrund stehen. Die Relevanz ist primär von der Wahl des Forschungsproblems abhängig, teilweise aber auch von den empirischen Befunden und daraus ableitbaren Schlussfolgerungen. So ist eine Studie mit relevanter Fragestellung letztlich doch von geringer Bedeutung, wenn sie sehr schwer interpretierbare Ergebnisse liefert im Unterschied zu einer Studie, deren Befunde eindeutig sind und sich unmittelbar für die weitere Forschung und/oder Praxis fruchtbar machen lassen.
   - **Wissenschaftliche bzw. theoretische Relevanz** („theoretical relevance“): In der **Grundlagenfor-**

schung ist mit inhaltlicher Relevanz vor allem die wissenschaftliche Relevanz im jeweiligen Forschungsfeld gemeint. Sie bemisst sich als wissenschaftsinternes Kriterium daran, inwiefern der wissenschaftliche Kenntnisstand erweitert bzw. eine Forschungslücke geschlossen wird. Vordringlich ist die empirisch gestützte Bildung, Prüfung und Überarbeitung von wissenschaftlichen **Theorien,** weshalb statt von „wissenschaftlicher" auch von „theoretischer" Relevanz gesprochen wird. Geringe theoretische Relevanz haben z. B. Studien, die zu wenig fokussiert sind, um zur Theoriebildung beizutragen oder die bereits Altbekanntes wiederholen, wobei die Beurteilung oft stark von der Betrachtungsperspektive abhängt. Darüber hinaus können Forschungsthemen in der sozial- und humanwissenschaftlichen Grundlagenforschung auch nach wissenschaftsexternen inhaltlichen Relevanzkriterien bewertet werden, etwa wenn man der wissenschaftlichen Forschung zu aktuellen Problemen, neuen Phänomenen, gesellschaftlich umstrittenen Fragestellungen etc. Vorrang einräumt – dies ist oft mit Erwartungen an die praktische Relevanz der Forschung verbunden.

- **Praktische Relevanz** („practical relevance", „practical usefulness"): In der **Anwendungsforschung** (z. B. Evaluationsforschung, Marktforschung, Auftragsforschung) ist mit inhaltlicher Relevanz vor allem die praktische Relevanz bzw. Nützlichkeit angesprochen. Damit eine wissenschaftliche Studie für Praktiker und Entscheidungsträger nützlich ist, müssen aus den Befunden der Studie möglichst konkrete Handlungsempfehlungen ableitbar sein bzw. bereits von den Forschenden abgeleitet werden. Praktischer Nutzen kann z. B. darin liegen, dass die Befunde helfen, bestimmte Praxisprobleme zu lösen, etablierte Maßnahmen zu optimieren, neue Techniken zu entwickeln und zu vermarkten und/oder gesellschaftliche Verhältnisse zu verbessern. Im Bereich der Evaluationsforschung werden Studien ausdrücklich anhand von Nützlichkeitsstandards beurteilt (▶ Abschn. 18.1.9).

2. **Methodische Strenge** („rigor", „methodological rigor"): Die Wissenschaftlichkeit verlangt eine Orientierung an wissenschaftlichen Methodologien und Methoden, die zur Bearbeitung des Forschungsproblems geeignet sind. Eine **gute** wissenschaftliche Studie zeichnet sich dabei durch einen hohen Grad an methodischer Strenge aus, d. h. es werden anspruchsvolle bzw. aussagekräftige Methodologien und Methoden gewählt und die einzelnen methodologischen und methodischen Aspekte der Studie werden sehr stringent, regelgetreu und unverzerrt umgesetzt. Bei der methodischen Strenge handelt es sich um ein **rein wissenschaftsimmanentes Kriterium**, das in zahlreiche Unterkriterien ausdifferenziert ist, wobei diese im quantitativen, im qualitativen und im Mixed-Methods-Paradigma unterschiedlich ausfallen. Unteraspekte der methodischen Strenge sind stark voneinander abhängig: Fehler und Ungenauigkeiten bei der Datenerhebung oder Datenbereinigung schränken die Datenqualität ein und führen dazu, dass auch die Ergebnisse der anschließenden Datenanalyse verzerrt bzw. in ihrer Aussagekraft reduziert sind. In der Grundlagenforschung wird auf methodische Strenge sehr großen Wert gelegt (▶ Abschn. 7.2). In der Anwendungsforschung werden hier Abstriche eher zugestanden, sofern Mindeststandards der Wissenschaftlichkeit erfüllt sind und praktische Relevanz gegeben ist.
3. **Ethische Strenge** („ethical rigor", „ethical acceptability", „ethical justifiability", „ethical defendability"): Die Wissenschaftlichkeit verlangt eine Orientierung an Prinzipien der Wissenschafts- und Forschungsethik. Eine **gute** wissenschaftliche Studie weist dabei einen hohen Grad an ethischer Strenge auf, d. h. sie erfüllt Standards der Wissenschafts- und Forschungsethik (▶ Kap. 4) besonders umfassend. Neben Studien, in denen potenziell beeinträchtigende oder gefährliche Verfahren eingesetzt werden (z. B. Gabe unterschiedlicher Medikamente; Befragung zu traumatischen Erlebnissen) sind z. B. auch Evaluationsstudien forschungsethisch besonders brisant, etwa wenn von ihren Ergebnissen existenzielle Entscheidungen für die Betroffenen abhängen (z. B. Evaluation von Organisationen oder von Personal). Für die Evaluationsforschung wurden eigene forschungsethische Richtlinien ausformuliert, sog. Fairness-Standards (▶ Abschn. 18.1.9). Verletzungen der Wissenschafts- und Forschungsethik können neben dem Schaden für die betroffenen Untersuchungsteilnehmenden auch rechtliche Konsequenzen sowie Reputationsverlust für die Forschenden bzw. die Forschungseinrichtungen nach sich ziehen und nicht zuletzt das Vertrauen der Öffentlichkeit in empirische Sozialforschung erschüttern. Häufig werden ethische Fragen nur negativ hinsichtlich Verstößen gegen ethische Prinzipien betrachtet. Dass sie hier als positive Dimension auftauchen und eine graduelle Abstufung zwischen ethisch strengerem und weniger strengem Vorgehen vorgeschlagen wird, soll angesichts der wachsenden Beachtung von Wissenschafts- und Forschungsethik das Augenmerk darauf lenken, dass zur Bearbeitung ethischer Probleme im Forschungsprozess oft mehrere Lösungsmöglichkeiten bestehen (z. B. Art und

Weise der Aufklärung der Untersuchungsteilnehmenden über ihre Rechte) und dass es lohnt, diese genauer zu reflektieren und methodisch weiterzuentwickeln.

4. **Präsentationsqualität**: Die Wissenschaftlichkeit verlangt es, dass eine Studie so umfassend dokumentiert wird, dass die zentralen Arbeitsschritte nachvollziehbar sind. Eine **gute** wissenschaftliche Studie zeichnet sich darüber hinaus durch hohe Präsentationsqualität aus, das betrifft die Vollständigkeit, Gewichtung und Strukturierung der Darstellung des Forschungsprojektes, die Lesbarkeit des sprachlichen Stils oder die Anschaulichkeit der verwendeten Grafiken etc. (► Kap. 13). Eine hohe Präsentationsqualität soll dafür sorgen, dass die Studie von den Zielgruppen verstanden wird und ihre Erkenntnisse somit theoretisch und praktisch nutzbar werden. Insbesondere durch die vermehrten Bemühungen um Forschungssynthese, also die Zusammenfassung des gesamten Forschungsstandes zu jeweils einem bestimmten Forschungsproblem, steigen die Anforderungen an die Präsentationsqualität empirischer Studien: Sollen in einer Metaanalyse (► Kap. 16) die statistischen Befunde verschiedener Studien aggregiert werden, muss sichergestellt sein, dass alle Studien alle relevanten statistischen Kennwerte in exakt vergleichbarer Weise berichten. Fachgesellschaften wie die American Psychological Association geben deswegen detaillierte Richtlinien dazu heraus, in welcher Weise empirische Studien in wissenschaftlichen Publikationen darzustellen sind (z. B. APA Publication Manual: APA, 2009; Reporting Standards for Research in Psychology: APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). Auch Gütekriterien der qualitativen Forschung verlangen detaillierte Angaben zu Methodenentscheidungen und Methodenmerkmalen in allen Phasen des Forschungsprozesses.

Von jeder wissenschaftlichen Studien ist eine **Selbstreflexion ihrer wissenschaftlichen Qualität** zu fordern: Es sollte – meist im Methoden- sowie im Diskussionsteil der Arbeit – angegeben werden, ob und inwiefern die genannten Kriterien wissenschaftlicher Qualität erfüllt werden und an welchen Punkten Einschränkungen vor allem der methodischen Strenge zu verzeichnen sind. Ein offensiver Umgang mit den Limitationen der eigenen Studie entspricht den Regeln guter wissenschaftlicher Praxis, während unkritische Verallgemeinerungen und Überinterpretationen der eigenen Befunde als allgemeingültige Erklärungen oder unumstößliche Gewissheiten die wissenschaftliche Seriosität bezweifeln lassen.

Unabhängig davon, wie die Autorinnen und Autoren selbst die wissenschaftliche Qualität ihrer Studie und deren Grenzen einordnen, kann und soll sich die Leserschaft selbst ein Bild machen und eine **Qualitätsbewertung** vornehmen. Eine umfassende Kenntnis relevanter Gütekriterien kann z. B. dazu führen, dass man Limitationen bemerkt, die von den Forschenden selbst gar nicht genannt wurden.

Während Wissenschaftlichkeit als kategoriales Merkmal fungiert (wissenschaftliche Studie oder nicht?), wird die wissenschaftliche Qualität als **kontinuierliches Merkmal** aufgefasst (mehr oder minder gute oder schlechte wissenschaftliche Studie?). Dabei ist die wissenschaftliche Qualität wie dargestellt ein **mehrdimensionales Konzept**, d. h. eine Studie kann innerhalb des Bereiches methodischer Strenge in bestimmter Hinsicht eine hohe wissenschaftliche Qualität aufweisen (z. B. sehr hohe Datenqualität durch aufwändige Stichprobenziehung und sorgfältige Datenerhebung und Datenbereinigung), gleichzeitig in anderer Hinsicht aber Schwächen haben (z. B. geringe Analysequalität durch die Wahl teilweise ungeeigneter statistischer Verfahren oder Fehlinterpretationen statistischer Ergebnisse). Auch eine in ihrer Gesamtqualität limitierte Studie kann unter Umständen noch substanziell zum wissenschaftlichen Erkenntnisgewinn beitragen, z. B. wenn die Daten im Rahmen einer Sekundäranalyse (► Abschn. 7.4) erneut und korrekt ausgewertet und interpretiert werden. Umgekehrt kann jedoch die exzellente Erfüllung eines einzelnen Kriteriums (z. B. sehr gute Präsentationsqualität) anderweitige Mängel oft nicht beheben (z. B. Ziehung einer stark verzerrten Stichprobe), so dass in der Gesamtbilanz die Qualität der wissenschaftlichen Erkenntnis als gering einzustufen ist.

Wie **einzelne Stärken und Schwächen** einer Studie exakt zu bemessen sind und in welcher Gewichtung sie in die Gesamtbewertung der wissenschaftlichen Qualität eingehen, ist nicht genau standardisiert und auch nicht standardisierbar, sondern muss vor dem Hintergrund der **Besonderheiten der Studie** innerhalb des jeweiligen Forschungsfeldes beurteilt werden: So mag man Abstriche bei der methodischen Strenge eher hinnehmen und der Arbeit eine gewisse Qualität zuschreiben, wenn die Studie über herausragende inhaltliche Relevanz in der Weise verfügt, dass es zu dem betreffenden Forschungsgegenstand bislang kaum Daten gibt. Dagegen wird man dieselben methodischen Limitationen zum Anlass nehmen, eine Studie als weitgehend unbrauchbar einzuordnen, wenn bereits viele Studien in dem Forschungsgebiet vorliegen, die eine höhere Qualität aufweisen.

Letztlich sind Beurteilungen der wissenschaftlichen Qualität auch wesentlich von den **Perspektiven der Urteilenden** abhängig, etwa ihren eigenen inhaltlichen und methodischen Ausrichtungen. Aus diesem Grund werden Beurteilungen der wissenschaftlichen Qualität in vielen Fällen auf der Basis von **mindestens zwei Expertenur-**

teilen getroffen (z. B. mehrere Gutachten für Doktorarbeiten, Forschungsanträge oder eingereichte Zeitschriftenaufsätze). Wenn im Rahmen systematischer Forschungssynthese die Qualität von Studien zu bewerten ist, um zu entscheiden, ob und ggf. mit welchem Gewicht sie in ein systematisches Review oder eine Metaanalyse eingehen sollen, empfiehlt z. B. die Cochrane-Collaboration (▶ Abschn. 16.1.2) insbesondere bei qualitativen Studien, die auf sehr heterogenen Methodologien basieren können, das Hinzuziehen von mehreren Fachexperten, die sowohl in der jeweiligen Methodologie als auch dem inhaltlichen Forschungsfeld ausgewiesen sind (Hannes, 2011, S. 6).

3.2 Gütekriterien für quantitative Forschung

Bei der Durchführung einer eigenen wissenschaftlichen Studie ist es empfehlenswert, von Anfang an und über alle Phasen des Forschungsprozesses hinweg die einzelnen Arbeitsschritte an den vier Kriterien wissenschaftlicher Qualität zu orientieren, die in der quantitativen Forschung noch in spezifischer Weise ausdifferenziert sind (◘ Tab. 3.3 und ▶ Abschn. 3.2.2). Das betrifft vor allem das Kriterium der methodischen Strenge, mit dem herausgehobenen Unterkriterium der Validität (▶ Abschn. 3.2.2).

3.2.1 Quantitative Gütekriterien im Überblick

Für die neun Phasen des empirischen Forschungsprozesses in der quantitativen Sozialforschung (▶ Abschn. 1.3.1) stellt ◘ Tab. 3.4 überblicksartig dar, welche der vier Kriterien wissenschaftlicher Qualität inwiefern und mit welchen Unteraspekten jeweils zu beachten sind. In sieben der neun Phasen steht dabei die methodische Strenge als Qualitätskriterium im Mittelpunkt.

3.2.2 Vier Typen der Validität in der Campbell-Tradition

Unter den vier Kriterien der Wissenschaftlichkeit nimmt die **methodische Strenge** einen zentralen Platz ein. In der quantitativen Forschung wird die methodische Strenge vor allem mit dem Konzept der Validität beschrieben. Validität meint den **Grad der Gültigkeit wissenschaftlicher Aussagen**, die aus einer quantitativ-empirischen Studie abgeleitet werden:

» We use the term *validity* to refer to the appropriate truth of an inference. When we say something is valid we make a judgement about the extent to which relevant evidence supports that inference as being true or correct. […] Validity is a property of inferences. It is *not* a property of designs or methods, for the same design may contribute to more or less valid inferences under different circumstances. (Shadish et al., 2002, S. 34).

Die **tatsächliche** Gültigkeit einer wissenschaftlichen Aussage ist durch die verwendete Methodik **nicht** eindeutig vorbestimmt, da auch bei einer mit großer Strenge durchgeführten Datenerhebung und Datenanalyse am Ende das Ergebnis von den Forschenden immer noch falsch interpretiert werden kann. Auch wenn sich die Validität als Kriterium auf Aussagen bezieht, wird sie in verkürzter Sprechweise doch auch den verwendeten Methodologien und Methoden zugeschrieben. Etwa wenn man sagt, dass ein Experiment einer vorexperimentellen Studie hinsichtlich interner Validität oder ein etabliertes psychometrisches Testverfahren einer Ad-hoc-Skala hinsichtlich Konstruktvalidität überlegen ist.

In der Tradition des US-amerikanischen Psychologen und Methodikers **Donald T. Campbell** (1916–1996) werden die bereits angesprochenen vier Typen der Validität differenziert (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish et al., 2002): **Konstruktvalidität**, **interne Validität**, **externe Validität** und **statistische Validität**.

Die Validitätstypologie in der Campbell-Tradition ist auf experimentelle und quasi-experimentelle Studien zugeschnitten, deren Ziel es ist, Ursache-Wirkungs-Relationen durch Variation der Untersuchungsbedingungen zu

◘ **Tabelle 3.3** Die vier Kriterien der wissenschaftlichen Qualität und die wichtigsten Gütekriterien für quantitative Forschung

Kriterien der wissenschaftlichen Qualität	Gütekriterien für quantitative Forschung
1. Inhaltliche Relevanz	Wissenschaftliche Relevanz
	Praktische Relevanz
2. Methodische Strenge	**Konstruktvalidität**
	Interne Validität
	Externe Validität
	Statistische Validität
3. Ethische Strenge	Forschungsethik
	Wissenschaftsethik
4. Präsentationsqualität	Standards der Berichterstattung

Tabelle 3.4 Die neun Phasen des empirisch-quantitativen Forschungsprozesses und die jeweils zu berücksichtigenden Gütekriterien

Phasen des quantitativ-empirischen Forschungsprozesses	Kriterien wissenschaftlicher Qualität	Konkretisierungen und Kommentare
1. Forschungsthema (► Kap. 5)	Inhaltliche Relevanz	**Wissenschaftliche und praktische Relevanz**: Bei der Themenwahl ist die inhaltliche Relevanz, d. h. in der Grundlagenforschung vor allem die wissenschaftliche bzw. theoretische Relevanz und in der Anwendungsforschung die praktische Relevanz vordringlich. Aber auch methodische und ethische Strenge müssen bedacht werden. Denn es lohnt sich nicht, überambitionierte Forschungsideen zu verfolgen, die mit den vorhandenen zeitlichen, finanziellen und personellen Ressourcen letztlich nicht mit ausreichender methodischer Strenge realisierbar sind oder an ethische Grenzen stoßen. Eine besser eingegrenzte Studie läuft weniger Gefahr, am Ende aufgrund gravierender Mängel bei der methodischen Strenge als wissenschaftlich schwaches oder gar pseudowissenschaftliches Projekt zu enden (für genauere Hinweise zur Auswahl von Forschungsthemen ► Abschn. 5.3). Die Wahl und Zuspitzung des zu untersuchenden Forschungsproblems erfolgt auf der Basis der Erarbeitung des Forschungsstandes und theoretischen Hintergrundes (► nachfolgenden Punkt 2), zudem ist gerade bei Anwendungsforschung sowie Evaluationsforschung auch die direkte Rücksprache mit Auftraggebern, Betroffenen und weiteren Anspruchsgruppen sinnvoll, um Fragestellungen zu entwickeln, die hohe praktische Relevanz aufweisen.
2. Forschungsstand und theoretischer Hintergrund (► Kap. 6)	Methodische Strenge	**Qualität der Literaturrecherche**: Sie sollte umfassend und systematisch über wissenschaftliche Datenbanken anhand einer Auswahl einschlägiger Suchbegriffe erfolgen und dabei deutsch- und englischsprachige Quellen einbeziehen (► Abschn. 6.1).
		Qualität der Erarbeitung des Forschungsstandes: Er sollte im Hinblick auf das Forschungsproblem erschöpfend und sinnvoll strukturiert sein sowie auch kritisch bewertet werden (► Abschn. 6.2).
		Qualität der Theoriebildung: Sie sollte an einschlägige Theoriemodelle des Forschungsfeldes anknüpfen (► Abschn. 6.3) und Qualitätskriterien wissenschaftlicher Theorien genügen (► Abschn. 2.2.10).
		Schlüssigkeit der aus Forschungsstand und Theorierahmen abgeleiteten Forschungsfragen bzw. Forschungshypothesen: Alle Forschungshypothesen und/oder Forschungsfragen müssen inhaltlich mit Quellen nachvollziehbar begründet sein (► Abschn. 6.4). Zudem lassen sich Forschungsfragen in der Anwendungsforschung aber auch anhand der konkreten Anforderungen der Auftraggeber oder Betroffenen einer Maßnahme pragmatisch begründen (z. B. kann in der Evaluationsforschung die Festlegung der Forschungsfragen gemeinsam mit den wichtigsten Stakeholdern erfolgen; ► Abschn. 18.1.4).
3. Untersuchungsdesign (► Kap. 7)	Methodische Strenge	**Interne Validität:** Eine Studie – bzw. genauer: die aus ihr gewonnenen Erkenntnisse – gelten als intern valide, wenn die untersuchten Variablenzusammenhänge mit hoher Sicherheit als kausale Ursache-Wirkungs-Relationen zu interpretieren sind. Dies ist vorrangig eine Frage des Untersuchungsdesigns: – **Experimentelle Studien**, in denen die Untersuchungsbedingungen aktiv variiert (experimentelle Variation), die Untersuchungspersonen den Bedingungen (mindestens eine Experimental- und eine Kontrollgruppe) zufällig zugewiesen (Randomisierung) und die Variablenveränderungen gemessen werden (Vorher-Nachher-Messung), ermöglichen Kausalinterpretationen mit der höchsten internen Validität. Randomisierte Kontrollgruppenstudien gelten deswegen als „Goldstandard" wissenschaftlicher Designs, weil sie einen klaren Kausalitätsnachweis liefern können (► Abschn. 7.6.1). – **Quasi-experimentelle Studien**, bei denen zwar eine experimentelle Variation und oft auch eine Vorher-Nachher-Messung, aber keine Randomisierung realisiert wird, können ebenfalls Kausalinterpretationen mit recht hoher interner Validität ermöglichen, wenn die durch fehlende Randomisierung auftretenden Störeinflüsse anderweitig umfassend kontrolliert werden (► Abschn. 7.6.2).

Tabelle 3.4 (Fortsetzung)

Phasen des quantitativ-empirischen Forschungsprozesses	Kriterien wissenschaftlicher Qualität	Konkretisierungen und Kommentare
3. Untersuchungsdesign (▶ Kap. 7)	Methodische Strenge	– **Vorexperimentelle Studien**, in denen keine experimentelle Variation, keine Randomisierung und oft auch keine Vorher-Nachher-Messung durchgeführt wird, erlauben kaum kausale Rückschlüsse (▶ Abschn. 7.6.3). Vor- bzw. **nicht-experimentelle Studien** können nur Zusammenhänge zwischen Variablen aufzeigen, ohne dass empirisch eindeutig klärbar ist, wie diese ursächlich zustande kommen. Dementsprechend ist die interne Validität kausaler Interpretationen gering bzw. nicht vorhanden. Eine Ausnahme bilden vorexperimentelle Studien mit Messwiederholungen (▶ Abschn. 7.8.2).
		Externe Validität: Eine Studie – bzw. genauer: die aus ihr gewonnenen Erkenntnisse – gelten in dem Maße als extern valide, wie sich die Ergebnisse auf andere Orte, Zeiten, Operationalisierungen der abhängigen und unabhängigen Variablen oder auf andere Personen als die konkret untersuchten verallgemeinern lassen. – **Feldstudien** wird z. B. im Hinblick auf die Generalisierbarkeit auf unterschiedliche Orte eine höhere externe Validität zugeschrieben als Laborstudien (▶ Abschn. 7.7). – **Längsschnittstudien** mit mehreren Messzeitpunkten besitzen meist eine höhere externe Validität im Hinblick auf die Generalisierbarkeit auf unterschiedliche Zeiten als Querschnittstudien mit nur einem Messzeitpunkt (▶ Abschn. 7.8). – Studien, in denen die zentralen Konzepte parallel in **unterschiedlicher Weise operationalisiert** werden, haben höhere externe Validität im Hinblick auf die Generalisierbarkeit auf unterschiedliche Messverfahren als Studien, die sich auf ein einziges Messinstrument stützen. – Im Vergleich zu einer einzelnen empirischen Untersuchung sichert der Studientyp der **Metaanalyse** eine höhere externe Validität der Ergebnisinterpretation, denn hier werden die Befunde mehrerer Studien zum selben Thema statistisch aggregiert und es wird auch betrachtet, unter welchen Bedingungen sich ähnliche oder unterschiedliche Effekte zeigen (▶ Kap. 16). – Die externe Validität hängt maßgeblich vom Untersuchungsdesign, aber auch von der Stichprobenziehung ab, denn **repräsentative Stichproben** steigern die externe Validität (▶ unten Punkt 5 Stichprobenziehung).
4. Operationalisierung (▶ Kap. 8)	Methodische Strenge	**Konstruktvalidität**: Die im Zusammenhang mit dem Forschungsproblem interessierenden theoretischen Konstrukte bzw. Konzepte müssen anhand des Forschungsstandes und des theoretischen Hintergrundes möglichst exakt definiert sein (**Konzeptspezifikation;** ▶ Abschn. 8.2). Auf dieser Basis muss bei der Operationalisierung angegeben werden, über welche beobachtbaren Merkmale **(Indikatoren)** das jeweilige theoretische Konzept mit welchen standardisierten **Messinstrumenten** erfasst werden soll (z. B. Einzelitem, psychometrische Skala, Index; ▶ Kap. 8). Entscheidend für die Qualität der Studie bzw. der aus ihr abgeleiteten Erkenntnisse ist die Konstruktvalidität der verwendeten Messinstrumente, also die Frage, wie gut das jeweilige **Messinstrument** tatsächlich genau das Zielkonstrukt inhaltlich erfasst bzw. genauer: wie gut sich die mit dem Messinstrument gewonnenen Daten und aus ihnen abgeleiteten Schlussfolgerungen theoretisch und empirisch rechtfertigen lassen (▶ Abschn. 10.4.1 „Hauptgütekriterien psychologischer Tests“). Ein Messinstrument kann nur dann inhaltlich gültige bzw. valide Messwerte liefern, wenn es auch hohe **Reliabilität** (Messgenauigkeit) und **Objektivität** aufweist. Mit standardisierten Messinstrumenten werden vor allem die abhängigen Variablen, teilweise (bei nicht-experimentellen Studien) auch die unabhängigen Variablen erhoben. Bei experimentellen und quasi-experimentellen Studien werden die **Untersuchungsbedingungen** absichtlich variiert als Ausprägungen von unabhängigen Variablen (z. B. verschiedene Therapie- oder Lernmethoden). Auch hier muss die Operationalisierung valide sein, d. h. die verschiedenen Bedingungen dürfen sich nur in genau den interessierenden Aspekten unterscheiden und müssen angemessen dosiert sein (▶ Abschn. 8.3.3).

Tabelle 3.4 (Fortsetzung)

Phasen des quantitativ-empirischen Forschungsprozesses	Kriterien wissenschaftlicher Qualität	Konkretisierungen und Kommentare
5. Stichprobenziehung (► Kap. 9)	Methodische Strenge	**Repräsentativität**: Eine Stichprobe gilt in der quantitativen Forschung als repräsentativ, wenn sie die Population, aus der sie stammt und über die Aussagen getroffen werden sollen, möglichst unverzerrt „im Kleinen" abbildet (► Abschn. 9.1.3). Verschiedene Stichprobenarten unterscheiden sich im Grad ihrer Repräsentativität, wobei probabilistische Stichproben den nicht-probabilistischen überlegen sind (► Abschn. 9.3). Die Repräsentativität einer Stichprobe ist ein Unteraspekt der **externen Validität** (► oben Punkt 3 Forschungsdesign).
6. Datenerhebung (► Kap. 10)	Methodische Strenge	**Datenqualität I**: Um eine hohe Datenqualität sicherzustellen, reicht eine entsprechende Operationalisierung durch objektive, reliable und vor allem valide Messinstrumente nicht aus (► oben Punkt 4 Operationalisierung), sondern es muss auch sichergestellt werden, dass die Datenerhebung im Rahmen der jeweiligen Erhebungsmethode (Beobachtung, Interview, Fragebogen, psychologischer Test, physiologische Messung, Dokumentenanalyse; ► Kap. 10) sorgfältig und sachgerecht durchgeführt wird. Das betrifft in der quantitativen Forschung z. B. die Herstellung vergleichbarer und anonymer Untersuchungsbedingungen, den Einsatz geschulter Interviewender oder Versuchsleitender, die Verblindung bei Experimenten etc. Mangelnde Datenqualität kann sich in reduzierter **Konstruktvalidität** (► oben Punkt 4 Operationalisierung) sowie in weiteren Problemen (z. B. viele fehlende Fälle und fehlende Werte im Datensatz durch Versuchsabbruch oder Antwortverweigerung) niederschlagen.
		Dokumentation: In der Phase der Datenerhebung ist die Dokumentation besonders wichtig, insbesondere die vollständige Sicherung der Rohdaten und ihrer Entstehungsbedingungen (z. B. exakte Beobachtungsorte und -zeiten; Angaben zu Anzahl, Merkmalen und Schulung der Versuchsleitenden etc.).
	Ethische Strenge	**Forschungsethik**: Die Forschungsethik verlangt es, dass Untersuchungspersonen durch die Teilnahme an einer wissenschaftlichen Studie nicht beeinträchtigt oder gar geschädigt werden, dass sie über die Modalitäten der Datenerhebung informiert werden und ihr Einverständnis erklären (► Abschn. 4.1).
7. Datenaufbereitung (► Kap. 11)	Methodische Strenge	**Datenqualität II**: Eine Studie kann sich nur dann auf hohe Datenqualität berufen, wenn das sorgfältig erhobene Rohdatenmaterial (► oben Punkt 6 Datenerhebung) einer systematischen und dokumentierten **Datenbereinigung** unterzogen wurde (z. B. Ausschluss unplausibler Werte und Fälle, sachgerechte Behandlung fehlender Werte; ► Abschn. 11.4.3). Einbußen der Datenqualität, die durch Fehler bei der Datenerhebung (z. B. mangelnde Interviewerschulung) oder Fehler bei der Operationalisierung (z. B. Nutzung eines Messinstrumentes mit zu geringer Messgenauigkeit) zustande gekommen sind, lassen sich im Zuge der Datenaufbereitung jedoch nicht mehr kompensieren. Mangelnde Datenqualität kann sich in reduzierter **Konstruktvalidität** (► oben Punkt 4 Operationalisierung) niederschlagen.
8. Datenanalyse (► Kap. 12)	Methodische Strenge	**Statistische Validität:** Eine quantitativ-empirische Studie kann ein hohes Maß an statistischer Validität beanspruchen, wenn die Ergebnisse der Studie (v. a. Aussagen über Effekte in der Population und ihre Größe) auf einer korrekten deskriptiv- und inferenzstatistischen Analyse basieren. Die statistische Validität ist z. B. eingeschränkt, wenn die inferenzstatistische Datenanalyse nicht hypothesenbasiert stattfindet, sondern theorielos nach Signifikanzen gesucht wird („Signifikanzenfischen"; ► Abschn. 12.5.2) und wenn ungeeignete statistische Verfahren verwendet werden (Frage der **Indikation** sowie der **Voraussetzungen** statistischer Signifikanztests für unterschiedliche Zusammenhangs-, Unterschieds- und Veränderungshypothesen, ► Abschn. 12.5). Weitere Bedrohungen der statistischen Validität sind mangelnde **Teststärke** von Signifikanztests z. B. durch zu geringen Stichprobenumfang in Relation zur interessierenden Effektgröße oder die Wahl eines suboptimalen **Effektgrößenmaßes** (► Kap. 14).

Tabelle 3.4 (Fortsetzung)

Phasen des quantitativ-empirischen Forschungsprozesses	Kriterien wissenschaftlicher Qualität	Konkretisierungen und Kommentare
9. Ergebnispräsentation (► Kap. 13)	Inhaltliche Relevanz	**Wissenschaftliche und praktische Relevanz**: Die Ergebnisse der Datenanalyse müssen hinsichtlich ihrer wissenschaftlichen und praktischen Relevanz interpretiert werden. Eine Studie ist umso aussagekräftiger, je genauer die Hinweise zur Theorieentwicklung und für Anschlussstudien und/oder die Handlungsempfehlungen für die Praxis ausfallen, die stringent aus den Befunden der Studie abgeleitet werden.
	Ethische Strenge	**Forschungsethik**: Im Zuge der Ergebnisdarstellung sind Untersuchungspersonen vor allem vor ungewollter Identifizierbarkeit zu schützen. Bei Evaluationsprojekten (► Kap. 18) sollten die Befunde aus Gründen der Fairness vor einer Publikation den Betroffenen bekannt gemacht bzw. mit ihnen diskutiert werden.
		Wissenschaftsethik: Die Publikation der Befunde darf keine Überinterpretationen enthalten, vielmehr müssen die Grenzen der Aussagekraft ausdrücklich benannt werden. Vorarbeiten müssen korrekt zitiert und alle Mitautorinnen und Mitautoren sowie Sponsoren der Studie genannt werden (► Abschn. 4.2).
	Präsentationsqualität	**Standards der Berichterstattung**: Hohe Präsentationsqualität ist gegeben, wenn die Studie in ihrem Ablauf und ihren Ergebnissen umfassend und nachvollziehbar sowie gut lesbar dargestellt wird, wobei die konkreten fach- und zeitschriftenspezifischen Standards der Berichterstattung einzuhalten sind (z. B. APA, 2009; ► Kap. 13). Da in Buchkapiteln oder Zeitschriftenartikeln oft der Platz begrenzt ist, sind notwendige Hintergrundinformationen, die in der Publikation selbst nicht unterzubringen sind (z. B. Instruktionen für die Versuchspersonen oder Fragebögen sowie Codebücher im Wortlaut), ergänzend z. B. über Websites verfügbar zu machen.

prüfen. Sie ist jedoch auch auf vor- bzw. nicht-experimentelle Studien anwendbar, in denen die vorgefundenen Bedingungen untersucht werden. Hohe Validität ist nicht nur in der **Grundlagenforschung** wichtig, sondern ebenso in der **Anwendungsforschung**: Soll z. B. die Wirksamkeit von Kriminalitätspräventionsprogrammen, Medikamenten oder computergestützten Lernsystemen empirisch im Zuge von Evaluationsstudien geprüft werden, so wird man auch von diesen Studien methodische Strenge und vor allem hohe Validität fordern, wobei bis heute maßgeblich auf die Validitätskonzepte in der Campbell-Tradition Bezug genommen wird (zur Bedeutung der Validität in der Evaluationsforschung s. z. B. Farrington, 2003). Für eine praxisnahe Darstellung von Validitätsbedrohungen in der quantitativen Forschung, die ein anderes Strukturierungsprinzip zugrunde legt und 1. propositionale, 2. ethische, 3. strukturelle/methodologische, 4. funktionale und 5. technische Restriktionen der Validität differenziert s. Keller & Casadevall-Keller (2010).

Konstruktvalidität – Konstruktvalidität („construct validity"): Repräsentieren die verwendeten Messinstrumente (abhängige Variablen) sowie die hergestellten Untersuchungsbedingungen (unabhängige Variablen) inhaltlich in hohem Maße jeweils genau die interessierenden theoretischen Konstrukte? Kann von den empirisch-statistischen Ergebnissen auf die interessierenden theoretischen Konstrukte zurückgeschlossen werden? Die Konstruktvalidität ist vor allem von der Qualität der Theoriearbeit (Konzeptspezifikation) sowie der Operationalisierung abhängig.

Interne Validität – Interne Validität („internal validity"): Wie zweifelsfrei kann für die interessierenden Effekte tatsächlich der vermutete Kausaleinfluss der unabhängigen Variable/n auf die abhängige/n Variable/n belegt werden? Die interne Validität ist vor allem von der Qualität des Untersuchungsdesigns (experimentelle, quasi-experimentelle oder vorexperimentelle Studie) und dessen Umsetzung abhängig.

Externe Validität – Externe Validität („external validity"): Inwiefern können die Ergebnisse der Studie – vor allem im Sinne der Wirkungen einer Maßnahme – auf andere Orte, Zeiten, Wirkvariablen, Treatmentbedingungen oder Personen als die konkret untersuchten verallgemeinert werden? Die externe Validität ist vor allem von der Qualität des Untersuchungsdesigns und der Stichprobenziehung (in der Umfrageforschung idealerweise probabilistische bzw. repräsentative Stichprobe; in der Experimentalforschung dagegen meist nicht-probabilistische Stichproben) abhängig.

Statistische Validität – Statistische Validität („statistical inference validity"): Wurden die deskriptiv- und inferenzstatistischen Analysen korrekt durchgeführt, so dass für die interessierenden Variablenzusammenhänge bzw. Effekte mit hoher Sicherheit nachgewiesen wurde, ob sie a) statistisch überzufällig (signifikant) sind oder nicht und b) ob sie eine theoretisch und/oder praktisch relevante Effektgröße aufweisen. Die statistische Validität ist vor allem von der Qualität der statistischen Datenanalyse abhängig. Aber auch untersuchungsplanerische Aspekte (z. B. Stichprobenumfang, Messgenauigkeit der gewählten bzw. entwickelten Messinstrumente) sind relevant.

Zielsetzung guter wissenschaftlicher Forschung im quantitativen Paradigma sollte es sein, möglichst hohe Validität sicherzustellen bzw. Bedrohungen der vier Arten der Validität durch sorgfältige Planung und Umsetzung der

Studie in allen Phasen des Forschungsprozesses auszuschließen oder zumindest zu reduzieren. Eine Studie, bei der die vier Arten der Validität stark ausgeprägt sind, kann große methodische Strenge für sich beanspruchen und gilt als wissenschaftlich besonders saubere und anspruchsvolle Arbeit. Einschränkungen der Validität, von denen fast jede empirische Studie aufgrund forschungspraktischer Gegebenheiten mehr oder minder stark betroffen ist, müssen offengelegt und bei der Ergebnisinterpretation berücksichtigt werden.

Im Folgenden werden die vier Typen der Validität genauer dargestellt, wobei vor allem auf die jeweiligen Bedrohungen der Validität, die es im Zuge guter Studien zu minimieren gilt, eingegangen wird. Die Reihenfolge der Darstellung orientiert sich an der **Chronologie des Forschungsprozesses**, denn die Sicherstellung der Konstruktvalidität betrifft vor allem die Phase der Theoriebildung und Operationalisierung, die Sicherstellung der internen Validität die Phase der Festlegung des Untersuchungsdesigns und die Sicherstellung der externen Validität ebenfalls die Phase des Untersuchungsdesigns sowie die Phase der Stichprobenziehung. Über das Ausmaß der statistischen Validität entscheidet vor allem das Vorgehen in der Phase der Datenanalyse.

Forschungslogisch wird allerdings die **Konstruktvalidität als Teilaspekt der externen Validität** verstanden (Campbell, 1957, S. 38; Shadish et al., 2002): Denn externe Validität im Sinne einer Verallgemeinerbarkeit der empirischen Befunde impliziert zunächst, dass von den Messwerten auf die theoretischen Konstrukte generalisiert werden kann (Konstruktvalidität), bevor darüber hinaus der Effekt noch auf andere Kontexte übertragen wird. Ebenso wird forschungslogisch die **statistische Validität als Teilaspekt der internen Validität** betrachtet. Denn interne Validität im Sinne der Gültigkeit einer Kausalinterpretation ist nur möglich, wenn zunächst überhaupt ein überzufälliger Variablenzusammenhang mit einer relevanten Effektgröße zuverlässig nachgewiesen wurde (statistische Validität).

Die Validitätstypologie in der Campbell-Tradition ist seit Dekaden in der quantitativen Sozialforschung etabliert (zu wissenschaftstheoretischen Grundlagen und Gütekriterien der psychologischen Experimentalforschung s. z. B. Westermann, 2000), wird jedoch auch kritisch betrachtet. Kritikpunkte betreffen z. B. die teilweise komplizierte Terminologie sowie die Abgrenzung der vier Validitätstypen (ist es wirklich sinnvoll, die Konstruktvalidität von der externen Validität abzuheben?). Für die Evaluationsforschung werden anstelle der vier Validitätstypen der Campbell-Tradition z. B. drei Gütekriterien vorgeschlagen: Validität (analog interne Validität), Präzision (Genauigkeit der Effektgrößenschätzung) und Generalisierbarkeit (Verallgemeinerbarkeit der Effekte in Anlehnung an externe Validität; Reichardt, 2011). Wir führen die vier Validitätstypen sowie die zugehörigen Bedrohungen in der Campbell-Tradition vollständig auf, da sie – unabhängig von Aspekten ihrer Benennung, Zuordnung und Abgrenzung – insgesamt ein nützliches und detailliertes Beurteilungsraster bilden, um die methodische Strenge von quantitativen Studien zu beurteilen.

Konstruktvalidität

Repräsentieren die in (quasi-)experimentellen Studien verwendeten Messinstrumente (abhängige Variablen) sowie die hergestellten Untersuchungsbedingungen (unabhängige Variablen) inhaltlich in hohem Maße jeweils genau die interessierenden theoretischen Konstrukte? Können also anhand der Daten der Studie gültige Aussagen über die interessierenden theoretischen Konstrukte getroffen werden? Oder ist das letztlich nicht möglich, weil die Daten die theoretischen Konzepte ungenügend, ausschnitthaft oder verzerrt widerspiegeln? Das Kriterium der Konstruktvalidität bezieht sich auf die **Gültigkeit der Interpretation empirischer Daten als Indikatoren für theoretische Konstrukte.** In verkürzter Sprechweise wird die Konstruktvalidität oft direkt den Messinstrumenten zugeschrieben, obwohl sie sich streng genommen auf die Interpretation der Messwerte bezieht (▶ Abschn. 10.4.1).

Die Konstruktvalidität ist ein grundlegendes Gütekriterium für jede quantitativ-empirische Studie, denn wenn die Messwerte letztlich nicht theoretisch sinnvoll interpretierbar sind, ist die gesamte Studie nicht aussagekräftig. Jene Grundsatzkritik am quantitativen Paradigma der empirischen Sozialforschung (▶ Abschn. 2.2), die den Sinngehalt numerischer Messwerte zur Abbildung der sozialen Wirklichkeit pauschal infrage stellt, adressiert im Grunde das Gütekriterium der Konstruktvalidität. Im Rahmen der Wissenschaftstheorie des Kritischen Rationalismus ist mit der Konstruktvalidität das **Korrespondenzproblem** angesprochen (▶ Abschn. 2.2.3).

Innerhalb des quantitativen Paradigmas besteht Einigkeit darüber, dass Messwerte nur dann sinnvoll und nützlich für wissenschaftlichen Erkenntnisgewinn in der Sozialforschung sind, wenn ihre inhaltliche Aussagekraft sichergestellt und geprüft ist. Dies ist ein methodisch und theoretisch anspruchsvoller und in der Forschungspraxis aufwändiger Prozess, der größtenteils in der Phase der Aufarbeitung von Forschungsstand und theoretischem Hintergrund, Planung des Untersuchungsdesigns und vor allem der Operationalisierung zu leisten ist – also lange vor der eigentlichen Datenerhebung. Erfahrungsgemäß besteht ein typischer Anfängerfehler darin, zu schnell mit der Datenerhebung zu starten und die Vorarbeiten nicht gründlich genug durchzuführen. In der Konsequenz entstehen dadurch nicht selten Datensätze mit Variablen, die

mangels Konstruktvalidität kaum sinnvoll interpretierbar sind.

Die theoretische Basis der Konstruktvalidität ist die **Konzeptspezifikation** (► Abschn. 8.2). Methodisch müssen auf dieser Basis die **Operationalisierungen** der abhängigen Variablen sowie der unabhängigen Variablen mit entsprechenden Gütekriterien umgesetzt werden. Dazu sollte – sofern inhaltlich möglich – auf Instrumente und Untersuchungsbedingungen zurückgegriffen werden, die bereits in der Forschungsliteratur etabliert und in ihrer Güte bestätigt sind. Wird mit neuen Operationalisierungen gearbeitet, so muss deren Validität geprüft werden (etwa mit testtheoretisch fundierten Methoden der Konstrukt- oder auch Kriteriums- und Inhaltsvalidierung; ► Abschn. 10.4.1). Theoretisch sinnvoll interpretierbare Daten liefert eine Studie nur, wenn auf der Basis der ausgearbeiteten Operationalisierungen dann auch die Datenerhebung und die Datenaufbereitung sorgfältig durchgeführt werden, so dass **hohe Datenqualität** resultiert. Letztlich sind sogar die Modalitäten des Untersuchungsdesigns ausschlaggebend für die Konstruktvalidität: So kann z. B. eine nominell unbehandelte Kontrollgruppe inhaltlich nicht als unbehandelte Gruppe interpretiert werden, wenn die Mitglieder der Kontrollgruppe Kenntnis über die Behandlung der Experimentalgruppe haben und somit ihr Verhalten ändern, etwa weil sie sich benachteiligt fühlen oder in Wettstreit zur Experimentalgruppe treten (► Vierzehn Bedrohungen der Konstruktvalidität).

Das Gütekriterium der Konstruktvalidität in der Campbell-Tradition geht weit über die **testtheoretische Konstruktvalidität eines Messinstrumentes** (z. B. eines Fragebogens oder eines psychologischen Tests) hinaus. Denn es ist auch die Operationalisierung unabhängiger Variablen angesprochen, was bei experimentellen und quasiexperimentellen Studien die **Gestaltung der Untersuchungsbedingungen** betrifft (z. B. Entscheidungen über Stimulusmaterial, Dauer der Versuchsdurchgänge, Verhalten der Versuchsleitenden, Verschwiegenheitsregelungen, damit Kontrollgruppenmitglieder nichts über die Experimentalbedingungen erfahren etc.; ► Vierzehn Bedrohungen der Konstruktvalidität).

Interne Validität

Ein zentrales Anliegen sozialwissenschaftlicher Forschung im quantitativen Paradigma ist die Erklärung sozialer Sachverhalte, d. h. die Prüfung von Ursache-Wirkungs-Relationen. Kausalerklärungen stehen im Zentrum wissenschaftlichen Erkenntnisgewinns und bilden den Kern wissenschaftlicher Theorien in der quantitativen Sozialforschung auf der Basis der Wissenschaftstheorie des Kritischen Rationalismus (► Abschn. 2.2). Zudem sind Kausalitätsprüfungen praxisrelevant: Denn nur wenn wir zuverlässig wissen, wodurch bestimmte negative Effekte hervorgerufen werden, können wir sie durch entsprechende Gegenmaßnahmen verhindern. Und nur wenn wir wissen, wie wünschenswerte Effekte zustande kommen, können wir sie zielgerichtet fördern. Das Kriterium der internen Validität bezieht sich auf die **Gültigkeit der Kausalerklärungen, die aus einer Studie abgeleitet werden.** Hohe interne Validität liegt vor, wenn die Ausprägungen bzw. Veränderungen in den abhängigen Variablen eindeutig auf den Einfluss der unabhängigen Variablen zurückzuführen sind.

Für den **wissenschaftlichen Kausalitätsnachweis** muss 1.) die Ursache der Wirkung zeitlich vorausgehen, 2.) die Ursache mit der Wirkung statistisch systematisch und mit relevanter Effektgröße zusammenhängen (statistische Validität) und 3.) die konkrete Kausalerklärung der Untersuchungshypothese allen anderen möglichen Alternativerklärungen eindeutig überlegen sein. Die Sicherung der internen Validität (sofern statistische Validität gegeben ist) dreht sich vor allem darum, mögliche **Alternativerklärungen** zu den Ergebnissen, die der eigentlichen Untersuchungshypothese widersprechen, von vornherein systematisch auszuschließen. Dies geschieht in der Experimentalforschung, indem möglichst alle relevanten Einflussfaktoren kontrolliert oder eliminiert werden und gezielt nur die interessenden Ursachenfaktoren variiert werden. Die aus der Studie abgeleitete Kausalaussage bezieht sich dann auf eben die untersuchten Ursachenfaktoren – alle anderen Einflussfaktoren bleiben zunächst unberücksichtigt. Dieses der Logik der Experimentalforschung zugrunde liegende Analysemodell wird auch als **Ceteris-Paribus-Prinzip** (lat. „alles andere bleibt gleich" bzw. „unter sonst gleichen Bedingungen") bezeichnet (Westermann, 2000). Beispiel: Um den Einfluss von körperlichem Ausdauertraining auf Depressivität zu untersuchen, werden zwei vergleichbare Gruppen von Patienten mit Depression unter gleichen Bedingungen über einen Zeitraum von acht Wochen hinweg untersucht, wobei der einzige Unterschied darin besteht, dass eine Gruppe am Ausdauertraining teilnimmt (Untersuchungsgruppe) und die andere nicht (Kontrollgruppe). Zeigt sich nun in der Untersuchungsgruppe eine Besserung der depressiven Beschwerden, so lässt sich mit hoher interner Validität schlussfolgern, dass – ceteris paribus bzw. unter sonst gleichen Bedingungen – ein Ausdauertraining zur Besserung depressiver Symptome wirksam ist. Damit wird ausdrücklich nicht gesagt, dass nur oder vor allem ein Ausdauertraining zur Depressionsminderung wirksam ist. Unter anderen Bedingungen – z. B. Kombination des Ausdauertrainings mit anderen Verfahren – könnten bessere oder schlechtere Effekte erzielt werden.

Wissenschaftstheoretisch wird teilweise hinterfragt, ob es überhaupt möglich ist, soziale Sachverhalte kausal zu erklären. Schließlich agieren Menschen sehr indivi-

Vierzehn Bedrohungen der Konstruktvalidität

In der Campbell-Tradition werden heute vierzehn Bedrohungen der Konstruktvalidität im Hinblick auf die abhängigen sowie unabhängigen Variablen (**Treatment**) unterschieden (Shadish et al., 2002, S. 73) Die hier aufgeführten Validitätsbedrohungen 11–14 wurden früher (d. h. bei Cook & Campbell, 1976) noch als Bedrohungen der internen Validität eingeordnet.

1. **Inadäquate Konzeptspezifikation** („inadequate explication of constructs"): Die interessierenden theoretischen Konstrukte wurden hinsichtlich ihrer Bedeutungsdimensionen und Unteraspekte ungenügend ausgearbeitet (► Abschn. 8.2), dementsprechend kann auch nicht erschöpfend und schlüssig angegeben werden, durch welche Indikatoren die Konstrukte zu erfassen sind bzw. welche Konstrukte die erhobenen Indikatoren repräsentieren.
   Beispiel: Es soll die Hypothese geprüft werden, dass körperliches Ausdauertraining die Depressivität reduziert. Dazu muss theoretisch genau erarbeitet werden, was unter „körperlichem Ausdauertraining" und unter „Depressivität" zu verstehen ist. Andernfalls kann nicht entschieden werden, ob z. B. wöchentliche Wanderausflüge als Ausdauertraining gelten können oder nicht.
2. **Konfundierungseffekte** („construct confounding"): Bestimmte Operationalisierungen erfassen mehrere Konstrukte gleichzeitig, so dass keine eindeutige Aussage über das Zielkonstrukt möglich ist. So ist die Konstruktvalidität eines Treatments bedroht, wenn es mit untersuchungsfremden Treatments gekoppelt einhergeht.
   Beispiel: Die Wirkung des Ausdauertrainings auf die Depressivität wird im Rahmen einer stationären Rehabilitationsmaßnahme geprüft, die auch eine Ernährungsumstellung beinhaltet. Somit ist dann im Falle eines Rückgangs der Depressivität nicht klar, ob es sich um die Wirkung des Ausdauertrainings und/oder der Ernährungsumstellung und/oder der Kombination von Ausdauertraining und Ernährungsumstellung handelt.
3. **Mono-Operationalisierungs-Bias** („mono-operation bias"): Eine valide Erfassung theoretischer Konstrukte kann noch besser sichergestellt werden, wenn anstelle eines einzelnen validen Messinstrumentes parallel mehrere valide Instrumente zum Einsatz kommen (z. B. Depressivität wird mit zwei unterschiedlichen Depressionsskalen erhoben). Denn jede einzelne Operationalisierung theoretischer Konstrukte durch empirische Indikatoren birgt gewisse Verzerrungen. Der Mono-Operationalisierungs-Bias limitiert nicht nur die Konstruktvalidität auf Seiten der Wirkungen bzw. abhängigen Variablen, sondern auch auf Seiten der Ursachenfaktoren bzw. unabhängigen Variablen als **Mono-Treatment-Bias**: Die zentralen unabhängigen Variablen werden jeweils nur durch eine Variante des Treatments operationalisiert. Valider wird die Studie, wenn mehrere Treatment-Varianten geprüft werden.
   Beispiel: Das Ausdauertraining als unabhängige Variable wird vergleichend über Laufen, über Schwimmen und über Radfahren realisiert, statt einer Treatmentgruppe werden also drei Treatmentgruppen untersucht.
4. **Mono-Methoden-Bias** („mono method bias"): Auch wenn mehrere Operationalisierungen zur Messung eines Konstruktes verwendet werden, kann dieses immer noch verzerrt repräsentiert sein, wenn die Operationalisierungen alle mit derselben Datenerhebungsmethode (z. B. nur Befragungsmethoden) erfolgen. Die Konstruktvalidität wird gesteigert, wenn mehrere Operationalisierungen auf der Basis verschiedener Datenerhebungsmethoden Anwendung finden (zum Multitrait-Multimethod-Ansatz ► Abschn. 10.4.4).
   Beispiel: Die Depressivität wird nicht über zwei Depressionsskalen – die beide auf Selbstauskünften basieren – erfasst, sondern durch eine Depressionsskala (Befragungsmethode) sowie durch Expertenurteile der behandelnden Psychologen (Beobachtungsmethode).
5. **Konfundierung von Treatment und Treatmentausprägungen** („confounding constructs and levels of constructs"): Die Konstruktvalidität des Treatments ist bedroht, wenn es nur in einer bestimmten Ausprägung untersucht wird.
   Beispiel: Es wird 2-mal pro Woche jeweils 30 Minuten Ausdauertraining durchgeführt. Erweist sich dieses als wirkungslos, ist damit eine Aussage über genau diese Trainingsdosierung getroffen, nicht über alle anderen möglichen höheren oder niedrigeren Trainingsintensitäten.
6. **Interaktion zwischen Treatment und Dimensionalität des Messinstrumentes** („treatment sensitive factorial structure"): Durch die Behandlung verändert sich die faktorielle Struktur des Messinstrumentes.
   Beispiel: Ein zunächst bei der untersuchten Gruppe eindimensionales Messinstrument für politische Einstellungen wird im Zuge des Treatments (z. B. politische Informationsveranstaltung) mehrdimensional, d. h. die Antworten der Befragten sind in zwei oder drei Dimensionen aufzusplitten. Wird diese Veränderung nicht bemerkt und das Instrument jeweils eindimensional ausgewertet, wird das gemessene Konstrukt unzureichend erfasst und der Treatmenteffekt der Einstellungsdifferenzierung unterschätzt oder übersehen.
7. **Reaktive Veränderung von Selbstauskünften** („reactive self-report changes"): Selbstauskunftsmaße können im Zuge einer Studie ihre Validität einbüßen und nicht mehr das Zielkonstrukt messen, wenn die Untersuchungspersonen taktisch falsche Angaben machen (z. B. ihre Symptome oder Probleme übertreiben), etwa um in eine bestimmte Behandlungsgruppe zu kommen. Designs, in denen die Teilnehmenden wissen, dass bestimmte Angaben bestimmte Vor- oder Nachteile für sie haben könnten (z. B. Einteilung in verschiedene Behandlungsgruppen), sind deswegen zur Sicherung der Konstruktvalidität zu vermeiden. ►

duell und nicht deterministisch nach „Naturgesetzen". Dazu ist festzuhalten, dass der Erklärungsanspruch in der quantitativen Sozialforschung weder deterministisch (sondern statistisch) noch universell (sondern lokal) erhoben wird: Auf der Basis empirischer Studien können keine Gesetzmäßigkeiten aufgestellt werden, die für jeden einzelnen Menschen auf der ganzen Welt zu jeder Zeit gleichermaßen gelten. Vielmehr erheben Kausalerklärungen mit hoher interner Validität einen von vorne herein begrenzten Erklärungsanspruch, weshalb die interne Validität präziser auch als **„local molar causal validity"** bezeichnet wird (Shadish et al., 2002): Es soll

8. **Reaktivität der experimentellen Situation** („reactivity to the experimental situation“): Die interessierenden Zielkonstrukte werden durch die erhobenen Messwerte nicht valide abgebildet, wenn die Untersuchungspersonen ihr Verhalten ändern, weil sie untersucht werden (unabhängig davon, ob das Treatment wirkt). Man spricht auch vom **Hawthorne-Effekt**, der anhand von Experimenten in der Hawthorne-Fabrik in Chicago entdeckt wurde: Eigentlich sollte die Wirkung der Beleuchtung auf die Arbeitsleistung untersucht werden und tatsächlich stieg die Arbeitsleistung bei verbesserter Beleuchtung in der Experimentalgruppe an. Doch sie stieg auch in der Kontrollgruppe und blieb auch bei der Rückkehr zur vorherigen Beleuchtung bestehen (Roethlisberger, Dickson & Wright, 1939). Offenbar führte vor allem die Tatsache, dass eine Studie durchgeführt und der Situation der Fabrikarbeiterinnen Aufmerksamkeit geschenkt wurde, dazu, dass sie ihre Arbeitsleistung steigerten. Eine andere Form der Reaktivität der Untersuchungssituation ist **Bewertungsangst** („evaluation apprehension“): Untersuchungspersonen können Angst vor einer negativen Beurteilung haben, was sich auf ihr Verhalten auswirkt (z. B. könnten durch Bewertungsangst ihre Leistungen schlechter werden). Eine Gegenmaßnahme sind strikt anonyme Datenerhebungen. Schließlich kann die Untersuchungssituation auch dazu veranlassen, **Vermutungen über die Forschungshypothese** anzustellen: Untersuchungspersonen können z. B. anhand der Merkmale des Untersuchungssettings die Richtung der Forschungshypothese raten und sich daraufhin hypothesenkonform verhalten. Damit ist dann die Konstruktvalidität der abhängigen Variablen gefährdet (z. B. wenn nach dem Ausdauertraining geringere Depressivität angegeben wird, weil die Probanden denken, dass genau dieser Effekt nachgewiesen werden soll). Eine Gegenmaßnahme ist die Modifikation der Untersuchungssituation, z. B. könnten neben der Depressivität noch ganz andere Merkmale vor und nach dem Ausdauertraining erfasst werden, um die Versuchspersonen von der eigentlichen Hypothese abzulenken. Derartige Methoden stoßen aber an forschungsökonomische und ggf. auch ethische Grenzen.
9. **Versuchsleiter-Erwartungseffekt** („experimenter expectancies“): Untersuchungspersonen können die Erwartungen der Versuchsleitenden erahnen und sich diesen mehr oder minder bewusst anpassen (auch dies kann im Beispiel zu einer absichtlich verzerrten Angabe der Depressivität führen, wenn die Teilnehmenden der untersuchten Reha-Maßnahme ihre Depressivität herunterspielen, um den Forschenden einen Gefallen zu tun). Eine Gegenmaßnahme ist das Arbeiten mit „blinden“ Versuchsleitern, die selbst die Hypothese nicht kennen bzw. nicht wissen, ob sie die Experimental- oder Kontrollgruppe vor sich haben). Der Versuchsleiter-Erwartungseffekt wird nach ihrem Entdecker auch als **Rosenthal-Effekt** bezeichnet und ist durch Primärstudien sowie Metaanalysen belegt (► Abschn. 16.1.1). Zudem existieren Versuchsleitereffekte, die nicht von den Erwartungen, sondern von anderen Merkmalen der Versuchsleitenden abhängig sind (z. B. können sich Untersuchungspersonen anders verhalten, je nachdem, ob sie mit – z. B. hinsichtlich soziodemografischer Variablen – sehr ähnlichen oder sehr unähnlichen Versuchsleitern konfrontiert sind, denen sie in unterschiedlichem Maße Vertrauen schenken).
10. **Novitäts- und Störungseffekte** („novelty and disruption effects“): Untersuchungspersonen reagieren anfangs meist besonders positiv auf interessante neuartige Veränderungen in ihrem Umfeld (Novitätseffekt) sowie besonders negativ auf Störungen ihrer gewohnheitsmäßigen Abläufe (Störungseffekt). Diese Effekte sind vorübergehender Natur, was bei der Interpretation der Messwerte der abhängigen Variablen zu beachten ist. So kann eine vermeintlich positive Treatmentwirkung, die aus den erhöhten Messwerten der abhängigen Variable herausgelesen wird, in Wirklichkeit ein vorübergehender Novitätseffekt sein.
11. **Kompensatorischer Ausgleich in der Kontrollgruppe** („compensatory equalization“): Dieses Problem – sowie die drei folgenden Punkte – treten vor allem bei experimentellen oder quasi-experimentellen Feldstudien auf, da dort eher ein Kontakt zwischen Teilnehmenden der Kontroll- und Experimentalgruppe stattfindet als bei Laborstudien. Untersuchungspersonen der Kontrollgruppe, die nicht in den „Genuss“ der Experimentalbedingung kommen, werden im Zuge der Studie von den Projektbeteiligten mehr oder minder bewusst zum Ausgleich besonders freundlich behandelt oder erhalten gewisse Vorteile. Damit sind sie faktisch nicht mehr Mitglieder einer unbehandelten Kontrollgruppe.
12. **Kompensatorischer Wettstreit der Kontrollgruppe mit der Experimentalgruppe** („compensatory rivalry“): Untersuchungspersonen der Kontrollgruppe können durch Kenntnis ihrer Gruppenzugehörigkeit in ihrem Ehrgeiz angestachelt werden, genauso gute oder bessere Ergebnisse zu zeigen als die Experimentalgruppe. Ihre Messwerte spiegeln somit nicht die Ausprägung des Konstruktes korrekt wieder, sondern sind verzerrt.
13. **Empörte Demoralisierung der Kontrollgruppe** („resentful demoralization“): Untersuchungspersonen der Kontrollgruppe erfahren, dass die Treatmentgruppe eine vorteilhaftere Behandlung erhält und zeigen deshalb durch Neid, Ablehnung oder Empörung beeinträchtigte Reaktionen, was wiederum verzerrte Messwerte erzeugt.
14. **Treatmentdiffusion in die Kontrollgruppe** („treatment diffusion“): Die Kontrollgruppe erhält Kenntnis darüber, was in der Experimentalgruppe geschieht, und versucht die Reaktionen in der Experimentalgruppe zu antizipieren und zu imitieren. Auch hier ist die Konstruktvalidität gefährdet, weil die Kontrollgruppenmitglieder sich eben faktisch nicht in einer unbehandelten Kontrollgruppe befinden, sondern in einer künstlichen Vergleichssituation mit der Experimentalgruppe.

geprüft werden, ob ein bestimmtes Bündel von Einflussfaktoren – ein einzelner Ursachenfaktor kann nie völlig isoliert untersucht werden, sondern immer im Gesamtpaket einer Maßnahme (deswegen: „molar“ und nicht „molekular“) – bestimmte Unterschiede in den gemessenen abhängigen Variablen verursacht („causal“) und zwar in dem jeweils konkreten Setting, Zeitrahmen und bei den untersuchten Personengruppen („local“). In welchem Maße sich diese Kausalerklärung auf andere als die konkret untersuchten Personen, Settings, Zeiten, Messinstrumen-

te, Einflussfaktoren etc. verallgemeinern lässt, ist eine Frage der **externen Validität**. Zur Kennzeichnung verschiedener Untersuchungspläne – etwa im Kontext der Diskussion ihrer Güte und Aussagekraft – hat sich in der Campbell-Tradition auch eine spezielle Notation etabliert (► Versuchsplan-Formalisierung). Auf unterschiedliche Varianten von Untersuchungsdesigns geht ► Kap. 7 ausführlich ein.

Wie belastbar wissenschaftliche Kausalaussagen sind, hängt maßgeblich vom Untersuchungsdesign ab (► Neun Bedrohungen der internen Validität): **Echte Experimente** bzw. randomisierte Kontrollgruppenstudien gelten als **Goldstandard** in der Grundlagenforschung, aber auch in der Anwendungsforschung (z. B. Wirksamkeitsstudien zu Therapien oder Medikamenten). **Quasi-Experimente** sind mangels Randomisierung in ihrer internen Validität eingeschränkt, hier ist es schwieriger, Experimental- und Kontrollbedingungen mit exakt vergleichbaren Personengruppen zu untersuchen. Die interne Validität von Schlussfolgerungen aus Quasi-Experimenten steigt in dem Maße, in dem bereits vor der Behandlung bestehende Unterschiede zwischen den Personen in der Experimental- und der Kontrollgruppe identifiziert und aus der Analyse eliminiert werden können (z. B. durch Konstanthaltung oder Erhebung von Störvariablen). **Vor-Experimente** ohne experimentelle Variation, ohne randomisierte Gruppenzuteilung und oft auch ohne Messwiederholung (z. B. es wird ohne Vorher-Messung nur im Nachhinein die Outcome-Variable erhoben), weisen keine interne Validität auf: Gültige Kausalschlüsse sind hier angesichts der Fülle möglicher Alternativerklärungen nicht möglich. Die einzige Ausnahme bilden **nicht-experimentelle Designs mit Messwiederholungen**, aus denen sich durch Variablenveränderungen über die Zeit kausale Rückschlüsse ziehen lassen (z. B. Cross-Lagged-Panel-Design; ► Abschn. 12.5.2).

Externe Validität

Eine empirische Studie weist hohe methodische Strenge im Sinne der externen Validität auf, wenn Schlussfolgerungen aus der Studie (vor allem der nachgewiesene Kausaleffekt) auf andere Personen, Varianten der unabhängigen Variablen, Messungen der abhängigen Variablen, Settings und Zeiten verallgemeinerbar bzw. **generalisierbar** sind. Eine erste Voraussetzung für externe Validität ist die Generalisierbarkeit von den konkret gemessenen Variablen auf die interessierenden theoretischen Konstrukte, die auf einer höheren Abstraktionsebene angesiedelt sind: Die Konstruktvalidität ist deswegen eine wichtige Voraussetzung der externen Validität.

Forschungslogisch ist die Verallgemeinerbarkeit der Befunde nicht nur durch Grenzen der Konstruktvalidität, sondern darüber hinaus dadurch eingeschränkt, dass der Kausaleffekt mit den entsprechenden Bedingungen der konkreten Studie in einer Wechselwirkung stehen kann, so dass sich der Effekt unter anderen als den untersuchten Bedingungen in anderer Weise zeigen würde: Er könnte in verstärkter oder abgeschwächter **Intensität** auftreten

Versuchsplan-Formalisierung

Das Design einer Studie kann mit einer Versuchsplan-Formalisierung („experimental design notation“) grafisch veranschaulicht werden. Diese wurde von Cook und Campbell (1979; s. auch Shadish, Cook & Campbell, 2002) eingeführt. Die Notation ist mittlerweile weit verbreitet und enthält u. a. folgende Elemente:

X = Treatmentbedingung/unabhängige Variable,

O = Beobachtung/Messung/abhängige Variable und

R = Randomisierung.

Verschiedene Treatmentkombinationen werden in parallelen Zeilen dargestellt. Werden diese Zeilen durch eine gestrichelte Linie getrennt, hat keine Randomisierung stattgefunden, wurden die Untersuchungspersonen jedoch randomisiert zugeteilt, werden die Zeilen der einzelnen Treatmentbedingungen **nicht** mit einer gestrichelten Linie voneinander getrennt. Die zeitliche Abfolge der einzelnen Schritte wird innerhalb einer Zeile von links nach rechts dargestellt. Alles, was innerhalb einer Spalte dargestellt wird, passiert zum gleichen Zeitpunkt. Diese Formalisierung wird nun anhand von ausgewählten Beispielen dargestellt.

Beispiel 1: One-Shot Case Study

X O

Treatment mit nachfolgender Messung der abhängigen Variable/n ohne Kontrollgruppe.

Beispiel 2: Pretest-Posttest-Design

O1 X O2

Vorher-Messung, Treatment und Nachher-Messung der abhängigen Variable/n, keine Kontrollgruppe.

Beispiel 3: Pretest-Posttest-Kontrollgruppendesign in einem Quasi-Experiment

O1 X O2

O1 O2

Vorher- und Nachher-Messung, eine Experimentalgruppe, eine unbehandelte Kontrollgruppe, keine randomisierte Zuweisung.

Beispiel 4: Pretest-Posttest-Kontrollgruppendesign in einem Experiment

R O1 X O2

R O1 O2

Vorher- und Nachher-Messung, eine Experimentalgruppe, eine unbehandelte Kontrollgruppe, Randomisierung

Neun Bedrohungen der internen Validität

In der Campbell-Tradition werden heute neun Bedrohungen der internen Validität unterschieden (Cook & Campbell, 1979; Shadish et al., 2002, S. 55). Diese Bedrohungen der internen Validität führen dazu, dass der statistisch nachgewiesene Zusammenhang zwischen Variablen (dies setzt statistische Validität voraus) nicht eindeutig als Ursache-Wirkungs-Zusammenhang interpretiert werden kann.

1. **Unklare zeitliche Abfolge** („ambigous temporal precedence"): Ein kausaler Ursachenfaktor muss zeitlich immer der Wirkung vorgelagert sein. In Experimenten und Quasi-Experimenten, in denen das Treatment vorgegeben und nachher die Wirkung gemessen wird, ist die untersuchte Reihenfolge klar. Bei nicht-experimentellen korrelativen Studien, in denen potenzielle Ursachenfaktoren nicht beeinflusst, sondern in ihren Ausprägungen vorgefunden und nicht zeitlich vor, sondern gleichzeitig mit den abhängigen Variablen gemessen werden, kann es bei der kausalen Interpretation zu Fehlern kommen. Die Kausalrichtung könnte verkehrt herum behauptet oder wechselseitige kausale Beeinflussung könnte übersehen werden.
   Beispiel: Es zeigt sich ein positiver Zusammenhang zwischen Selbstwertgefühl und Schulleistung. Wie ist dieser kausal zu interpretieren? Führt hohes Selbstwertgefühl zu guter Schulleistung oder fördert die Schulleistung das Selbstwertgefühl? Sofern zur Steigerung der internen Validität kein experimentelles oder zumindest quasi-experimentelles Design sowie keine Messwiederholungs- bzw. Langzeitstudie realisierbar ist, sollte bei nicht-experimentellen Studien jede Kausalaussage kritisch hinsichtlich der behaupten Kausalrichtung hinterfragt werden.
2. **Selektionseffekte** („selection"): In den verschiedenen Untersuchungsgruppen müssen sich vergleichbare Untersuchungspersonen befinden. Unterscheiden sich die Untersuchungspersonen in den einzelnen Gruppen bereits vor der Behandlung, so können diese bereits bestehenden Unterschiede später fälschlich als kausale Treatmentwirkung interpretiert werden. Um vergleichbare Untersuchungsgruppen ohne Selektionseffekte zu erzeugen, wird bei echten Experimenten mit Randomisierung (zufällige Zuordnung der Probanden zu den Untersuchungsgruppen) gearbeitet (dies setzt aber Mindestgruppengrößen voraus). Bei quasi-experimentellen Studien oder Experimenten mit sehr kleinen Gruppen arbeitet man vor allem mit Matching, Parallelisierung oder Konstanthalten personengebundener Störvariablen (► Abschn. 7.6.2).
3. **Externe zeitliche Einflüsse** („history"): Geht bei einer Studie das untersuchte Treatment mit anderen, untersuchungsunabhängigen äußeren Veränderungen in den situativen oder Umweltvariablen der Untersuchungspersonen einher, so können letztere fälschlich als kausaler Treatmenteffekt interpretiert werden. Um den interessierenden Ursachenfaktor von externen zeitlichen Einflüssen abzugrenzen, ist bei einem Design mit Vorher-Nachher-Messung neben der Untersuchungsgruppe auch eine Kontrollgruppe einzuplanen. Untersuchungs- und Kontrollgruppe unterliegen denselben externen zeitlichen Einflüssen, aber nur in der Untersuchungsgruppe wirkt das Treatment, so dass bei Überlegenheit der Treatmentgruppe gegenüber der Kontrollgruppe die Alternativerklärung alleiniger externer zeitlicher Einflüsse auszuschließen ist.
4. **Reifungsprozesse** („maturation"): Da sich Untersuchungspersonen genau wie alle Menschen im Verlauf der Zeit von ganz allein in ihren Zuständen und Merkmalen verändern (z. B. älter, reifer, kränker, gesünder, hungriger, entspannter, weniger aufmerksam etc. werden), können diese „Reifungsprozesse" fälschlich als kausale Treatmenteffekte interpretiert werden. Um den interessierenden Ursachenfaktor von personenbezogenen Reifungsprozessen abzugrenzen, ist bei einem Design mit Vorher-Nachher-Messung neben der Untersuchungsgruppe auch eine Kontrollgruppe vorzusehen.
5. **Statistische Regressionseffekte** („regression"): Werden für eine Studie Personen mit extrem hohen oder extrem niedrigen Merkmalsausprägungen ausgesucht, so ist bei wiederholter Messung automatisch eine Tendenz zur Mitte zur erwarten, d. h. die Extremausprägungen (z. B. extreme Schmerzen, extreme Einstellungen) schwächen sich ab. Dieser statistische Regressionseffekt kann fälschlich als kausale Treatmentwirkung interpretiert werden. Als Gegenmaßnahme sollte möglichst nicht mit Extremgruppen, sondern mit Untersuchungspersonen üblicher Merkmalsausprägungen gearbeitet werden.
6. **Experimentelle Mortalität** („mortality", „subject attrition"): Im Verlaufe einer Studie fallen immer wieder Untersuchungspersonen aus, etwa weil sie bestimmte Antworten verweigern, die Teilnahme an der Studie aus Protest abbrechen, umziehen, erkranken oder versterben (dies gilt besonders bei Längsschnittstudien; ► Abschn. 7.8). Wenn ein Vorher-Nachher-Vergleich nicht auf derselben Gruppe basiert, sondern aus der Nachher-Gruppe selektiv bestimmte Personen herausgefallen sind, kann dieser „Mortalitätseffekt" fälschlich als kausaler Treatmenteffekt interpretiert werden. So kann die Wirksamkeit eines Trainings stark überschätzt werden, wenn bei der Nachher-Messung nur noch die besonders motivierten und leistungsstarken Teilnehmenden präsent sind, weil die anderen die Maßnahme inzwischen abgebrochen haben. An welchen Stellen des Untersuchungsablaufs aus welchen Gründen welche Untersuchungspersonen aus einer Studie herausgefallen sind, ist genau zu dokumentieren und bei der Interpretation zu berücksichtigen.
7. **Testübung** („testing"): Wird im Zuge einer Studie mehrfach dasselbe Messinstrument eingesetzt (z. B. Vorher-Nachher-Messung), kann die Erfahrung mit dem Instrument das Antwortverhalten der Untersuchungspersonen verändern. Dieser Testübungseffekt (auch **Pretest-Effekt**) kann fälschlich als kausale Treatmentwirkung interpretiert werden. Um Testübung auszuschließen kann z. B. mit unterschiedlichen Messinstrumenten gearbeitet (z. B. Intelligenztests mit verschiedenen Aufgaben, um Lerneffekte auszuschließen) oder z. B. ein Solomon-Vierguppen-Plan realisiert werden (► Abschn. 12.5.4).
8. **Mangelnde instrumentelle Reliabilität** („instrumentation"): Im Verlaufe der Zeit kann sich die Qualität eines Messinstrumentes verändern, so dass Messwertunterschiede entstehen, die fälschlich als kausale Treatmentwirkung interpretiert werden können. Bei einer Datenerhebung mittels Beobachtung kann es z. B. durch Konzentrationsschwankungen oder Voreingenommenheiten der Beobachter zu Veränderungen in den Messwerten kommen. Als Gegenmaßnahme sind standardisierte Messinstrumente mit hoher Reliabilität zu nutzen und z. B. nur geschulte Beobachter für jeweils kurze Beobachtungsspannen einzusetzen (z. B. um Ermüdung zu verhindern; ► Abschn. 10.1.3).
9. **Kombination der genannten Bedrohungen der internen Validität** („additive and interactive effects of threats to internal validity"): Die acht vorgenannten Bedrohungen der internen Validität können auch in Kombination wirksam sein und dabei additiv oder in anderer Weise zusammenwirken.

Fünf Bedrohungen der externen Validität

In der Campbell-Tradition werden heute **fünf Bedrohungen der externen Validität** unterschieden (Cook & Campbell, 1976; Shadish et al., 2002, S. 87).

1. **Wechselwirkung des Kausaleffektes mit den Untersuchungspersonen** („interaction of the causal relationship with units"): Ein Effekt, der in einer Studie mit einer bestimmten Personengruppe nachgewiesen wurde, lässt sich bei spezifischem Zusammenwirken zwischen Personenmerkmalen und Kausaleffekt nicht auf andere Personengruppen übertragen. Beispiel: Ein Medikament, dessen Wirksamkeit an jungen Männern getestet wurde, kann bei älteren Menschen sowie bei Frauen anders wirken. Deswegen wird die externe Validität gesteigert, wenn der Effekt an Stichproben aus unterschiedlichen Populationen nachgewiesen wurde.
2. **Wechselwirkung des Kausaleffektes mit den Treatmentbedingungen** („interaction of the causal relationship over treatment variations"): Ein Effekt, der bei einer bestimmten Treatmentbedingung (Ausprägung der unabhängigen Variablen) gefunden wird, kann sich bei einer anderen Variante desselben Treatments, bei Kombination mit anderen Treatments oder durch begrenzte Implementierung unterschiedlich ausprägen.
   Beispiel: Effekte der Verkleinerung von Schulklassen sollen untersucht werden. Dabei mögen sich sehr unterschiedliche Effekte zeigen, je nach dem zu welchem Zeitpunkt im Schuljahr die Klassen aufgeteilt werden, in welchen Räumen und von welchen Lehrkräften die neu gebildeten kleineren Klassen unterrichtet werden oder wie die Maßnahme unter Lehrkräften und Eltern diskutiert wird. Die externe Validität lässt sich durch Berücksichtigung unterschiedlicher Settings steigern.
3. **Wechselwirkung des Kausaleffektes mit den abhängigen Variablen** („interaction of the causal relationship with outcomes"): Ein Effekt, der sich bei einer bestimmten abhängigen Variable in bestimmter Weise zeigt, kann bei einer anderen abhängigen Variablen anders zutage treten.
   Beispiel: Der Behandlungserfolg einer Krebstherapie zeigt sich unterschiedlich, je nachdem, ob und wie die subjektive Lebensqualität, die Überlebensrate oder die metastasenfreie Zeit als Wirkung erfasst wird. Die externe Validität lässt sich durch Berücksichtigung mehrerer Operationalisierungen der abhängigen Variablen steigern.
4. **Wechselwirkungen des Kausaleffektes mit den Settings** („interactions of the causal relationships with settings"): Ein Effekt, der in einem bestimmten Setting gefunden wurde, mag sich nicht auf andere Settings übertragen lassen.
   Beispiel: Ein Anti-Drogenprogramm, das in ländlichen Regionen erfolgreich ist, mag in Großstädten nicht entsprechend wirken. Die externe Validität lässt sich durch Datenerhebungen in unterschiedlichen Settings steigern.
5. **Kontextabhängige Mediatoreffekte** („context-dependent mediation"): Eine Mediatorvariable, die den interessierenden Kausaleffekt mit hervorbringt, mag in einem Setting wirken, jedoch nicht in einem anderen.
   Beispiel: Welches Studienfach gewählt wird, hängt von den individuellen Interessen ab, wird aber durch Geschlechtsrollenerwartungen vermittelt, wobei diese als Mediatoren in unterschiedlicher Weise wirken, je nach kulturellem Hintergrund oder historischem Zeitpunkt der Studie.

(Frage der Effektgröße), er könnte aber auch seine **Qualität** ändern (z. B. Veränderung der Kausalrichtung, Hinzutreten neuer Mediatorvariablen). Die Bedrohungen der externen Validität beziehen sich dementsprechend auf die Existenz und Stärke entsprechender **Interaktionseffekte zwischen dem Kausaleffekt und Aspekten des Untersuchungsdesigns** (► Fünf Bedrohungen der externen Validität). Dabei sind hier nicht nur statistische Interaktionseffekte im engeren Sinne gemeint (► Abschn. 12.5.3), sondern inhaltliche Wechselwirkungen.

Eine Möglichkeit, die externe Validität einer Studie zu erhöhen, besteht darin, möglichst umfangreiche und vielfältige Informationen zu gewinnen, d. h. das Untersuchungsdesign zu erweitern. Also z. B. anstelle eines einfachen Vergleichs zwischen **einer** Behandlungs- und **einer** Kontrollgruppe mehrere Varianten der Behandlung zu untersuchen sowie große und möglichst repräsentative Stichproben zu erheben oder verschiedene Zielgruppen in die Studie einzubeziehen. Entsprechende **Erweiterungen des Designs** stoßen jedoch in der Praxis oft an forschungsökonomische Grenzen, da sie zusätzlichen Aufwand bedeuten.

Jede erfolgreiche **Replikation** (► Abschn. 7.3) einer Studie bekräftigt die externe Validität des Kausaleffektes. Dementsprechend wird kritisch hinterfragt, ob wir in den einzelnen Wissenschaftsdisziplinen oft genug Replikationsstudien durchführen. In der Psychologie berichten offenbar nur rund 1 % der Zeitschriftenartikel über Replikationsstudien. Die Replikationsrate in anderen Sozial- und Humanwissenschaften scheint ebenfalls niedrig zu sein (Makel, Plucker, & Hegarty, 2012). Bei allen Hypothesen bzw. Effekten, die in einem Fach große Beachtung finden (z. B. oft zitiert werden), sind Replikationsstudien durch andere Forschergruppen als die Ursprungsautoren dringend zu fordern, um das Ergebnis abzusichern (z. B. gegenüber Wissenschaftsfälschungen; ► Abschn. 4.2.2) und seine Generalisierbarkeit zu prüfen (z. B. existiert der Effekt in vergleichbarer, stärkerer oder schwächerer Ausprägung auch in anderen Kulturkreisen?). Wirklich gut einschätzbar ist die Generalisierbarkeit von Effekten erst durch systematische Forschungssynthese etwa in Form von **Metaanalysen**, die zahlreiche Studien zur selben Forschungsfrage statistisch zusammenfassend analysieren (► Kap. 16). Durch eine Moderatorvariablen-Analyse

Neun Bedrohungen der statistischen Validität

In der Campbell-Tradition werden heute **neun Bedrohungen der statistischen Validität** unterschieden (Cook & Campbell, 1976; Cook & Campbell, 1979; Shadish et al., 2002, S. 45). Diese Bedrohungen führen dazu, dass die auf der Basis einer empirischen Studie gezogenen Schlussfolgerungen über den statistischen Zusammenhang zwischen Variablen bzw. die Effektgröße in der Population nicht korrekt sind. Manche der Bedrohungen der statistischen Validität fördern einen Alpha-Fehler bzw. **falsch-positive Entscheidungen** gegenüber Forschungshypothesen (z. B. das sog. „Signifikanzfischen"), andere einen Beta-Fehler bzw. **falsch-negative Entscheidungen** gegenüber Forschungshypothesen (z. B. zu geringe Teststärke, mangelnde Reliabilität von Messinstrumenten, beschränkter Wertebereich von Variablen etc.). Weitere Bedrohungen der statistischen Validität können sowohl Alpha- als auch Beta-Fehler beeinflussen (z. B. Verletzung der Voraussetzungen der verwendeten Signifikanztests) und wieder andere führen zu falschen Effektgrößenschätzungen.

1. **Zu geringe Teststärke** („low statistical power"): Statistische Signifikanztests liefern keine Grundlage für die wissenschaftliche Hypothesenprüfung, wenn ihre Teststärke nicht ausreichend groß ist. Idealerweise wird eine Studie so geplant (z. B. durch optimalen Stichprobenumfang), dass sie über ausreichende Teststärke in Relation zur interessierenden Effektgröße verfügt (► Abschn. 14.4). Nicht immer wird dies jedoch bedacht und teilweise lassen sich optimale Stichprobenumfänge aus forschungsökonomischen Gründen auch nicht realisieren. Im Zuge der inferenzstatistischen Datenanalyse ist neben dem jeweiligen Signifikanztest immer auch eine Post-hoc-Teststärkeanalyse durchzuführen, insbesondere bei nichtsignifikantem Ergebnis (► Abschn. 14.1), um sicherzustellen, dass Tests mit mangelnder Teststärke nicht zur Grundlage von Fehlinterpretationen im Sinne falsch-negativer Befunde werden. Viele sozialwissenschaftliche Studien leiden unter geringer statistischer Validität mangels Teststärke.
2. **Verletzung der Voraussetzungen der verwendeten Signifikanztests** („violated assumptions of statistical tests"): Unterschiedliche statistische Signifikanztests sind an jeweils bestimmte Voraussetzungen gebunden (z. B. Normalverteilung der abhängigen Variablen oder Varianzhomogenität zwischen Gruppen), um valide Ergebnisse zu liefern und Effekte weder zu über- noch zu unterschätzen. Dass diese Voraussetzungen erfüllt sind, ist durch entsprechende Voraussetzungstests sicherzustellen (im Rahmen der Darstellung der statistischen Signifikanztests ► Abschn. 12.5 wird auch auf deren Voraussetzungen eingegangen). Bei Verletzung der Voraussetzungen gängiger statistischer Datenanalysemethoden sind alternative Testverfahren (z. B. sog. nonparametrische Verfahren) zu wählen.
3. **Signifikanzfischen und Alpha-Fehler-Kumulation** („fishing and the error rate problem"): Wenn die statistische Analyse nicht darauf hinausläuft, gezielt einige wenige theoretisch abgeleitete Forschungshypothesen zu prüfen, sondern wenn alle möglichen Variablenzusammenhänge durchgetestet werden in dem Bestreben, signifikante Effekte zu finden („Signifikanzfischen", ► Abschn. 12.5.2), wird die Logik deduktiv-hypothesenprüfender quantitativer Forschung verletzt und manipulativ gearbeitet. In Studien, in denen multiple Tests durchgeführt werden (allerdings begründet und nicht im Rahmen puren Signifikanzfischens), muss man sich gegen die Alpha-Fehler-Kumulation absichern und z. B. mit einem strengeren Signifikanzniveau arbeiten, um Einbußen der statistischen Validität zu vermeiden.
4. **Mangelnde Reliabilität von Messinstrumenten** („unreliability of measures"): Messinstrumente mit geringer Reliabilität liefern definitionsgemäß Messwerte, die durch Messfehler verzerrt sind und somit die statistisch untersuchten Effekte nur ungenau abbilden. Es empfiehlt sich, etablierte Messinstrumente mit hoher Reliabilität zu nutzen und bei selbst entwickelten Instrumenten die Messgenauigkeit zu prüfen.
5. **Beschränkter Wertebereich von Variablen** („restriction of range"): Wird bei der Operationalisierung von Variablen der Wertebereich stark eingeschränkt, reduziert sich die Varianz und somit auch die mögliche Kovarianz mit anderen Variablen. Statistische Zusammenhänge werden somit unterschätzt.
   Beispiel: Der Wertebereich ist beschränkt, wenn die Intensität des Zigarettenrauchens kategorial erfasst wird (Nichtraucher, Gelegenheitsraucher, regelmäßiger Raucher) im Unterschied zur Erfassung der genauen Zahl der gerauchten Zigaretten (► Abschn. 8.4).
6. **Mangelnde Reliabilität der Treatment-Implementierung** („unreliability of treatment implementation"): Der statistisch nachweisbare Effekt eines Treatments wird unterschätzt, wenn dieses unvollständig oder inkorrekt umgesetzt wird.
   Beispiel: Die Wirkung einer Trainingsmethode kann nicht statistisch valide geprüft werden, wenn das Training nicht im üblichen Umfang, sondern nur in verkürzter Form durchgeführt wurde.
7. **Störeinflüsse im experimentellen Setting** („extraneous variance in the experimental setting"): Der statistisch nachweisbare Effekt eines Treatments wird unterschätzt, wenn die Untersuchungsbedingungen durch Störeinflüsse verzerrt sind (z. B. unterschiedliche räumliche Bedingungen oder Verhaltensweisen der Versuchsleitenden).
8. **Heterogenität der Untersuchungspersonen bzw. -einheiten innerhalb der Untersuchungsgruppen** („heterogeneity of units"): Um den Effekt einer experimentellen oder quasi-experimentellen Variation zu prüfen, müssen die Untersuchungseinheiten innerhalb der Gruppen vergleichbar und möglichst homogen sein. Bei zu großer Heterogenität innerhalb der Gruppen können Unterschiede zwischen den Gruppen statistisch nicht eindeutig identifiziert werden.
9. **Falsche Effektgrößenbestimmung** („inaccuarate effect size estimation"): Effekte sind nicht nur hinsichtlich ihrer statistischen Signifikanz, sondern auch hinsichtlich ihrer Effektgröße zu betrachten, wofür unterschiedliche nicht-standardisierte sowie standardisierte Effektgrößenmaße zur Verfügung stehen (► Abschn. 14.2). Die statistische Validität ist gefährdet, wenn ein falsches Effektgrößenmaß berechnet wird. So ist z. B. bei dichotomen abhängigen Variablen die Odds-Ratio den Korrelationskoeffizienten vorzuziehen (► Abschn. 14.2.1).

betrachtet die Metaanalyse auch die Bedingungen, unter denen der interessierende Effekt in unterschiedlicher Stärke auftritt.

Häufig wird die Differenzierung zwischen **Labor- und Feldstudien** mit dem Gütekriterium der externen Validität verknüpft. Dahinter steht die Überlegung, dass ein Effekt, der sich im Labor unter hochkontrollierten Bedingungen zeigt, vermutlich unter den Bedingungen des Alltags in anderer Weise zutage tritt. Eine Feldstudie, die unter „natürlichen Bedingungen" abläuft, würde dementsprechend eher auf den Alltag verallgemeinerbare Befunde erbringen als eine Laborstudie unter „künstlichen Bedingungen".

Beispiel: In einer experimentellen Laborstudie zur Wirkung von gewalthaltigen Computerspielen wird die Aggressivität nach dem Spielen dadurch gemessen, dass die Spieler andere Personen durch das Applizieren von Geräuschen bestrafen sollen, wobei die jeweils eingestellte Lautstärke der Geräusche als Maß für den Grad der Aggressivität dient (Anderson & Dill, 2000). Ob diese Form der Operationalisierung wirklich in alltagsrelevanter Weise Aggressivität erfasst, ist zunächst eine Frage der Konstruktvalidität.

Die Laborbedingungen, unter denen über eine sehr kurze Zeitspanne (z. B. 10 Minuten) ein vorgegebenes Game gespielt, dieses dann abgebrochen wird und von den Probanden eine andere Aufgabe zu bearbeiten ist, unterscheiden sich deutlich vom Spielen im Alltag in häuslicher Umgebung, wo selbst gewählte Spiele über längere Zeitperioden hinweg gespielt werden. Man könnte somit im Sinne externer Validität argumentieren, dass die möglichen aggressionssteigernden Wirkungen gewalthaltiger Spiele am besten in der natürlichen Umgebung zu erfassen seien, schließlich könnte die Laborsituation als solche die Spielenden frustrieren und ihre Aggressivität künstlich steigern. Hier würde also das Untersuchungssetting (Labor) mit dem eigentlich interessierenden Kausaleffekt (gewalthaltige Spiele steigern Aggressivität) eine Wechselwirkung eingehen, die die Verallgemeinerbarkeit der Befunde der Studie auf Nicht-Laborsituationen reduziert.

Untersuchungsbedingungen, die denen des natürlichen Alltags möglichst nahe kommen (sog. **ökologische Validität**) sind anzustreben. Sie stoßen dort an Grenzen, wo der Versuch, die externe Validität zu steigern, durch die Vielzahl der Einflussfaktoren im Feld die interne Validität gefährdet: Eine Feldstudie, bei der heimisches Spielen selbstgewählter Games für selbstgewählte Zeitspannen mit unterschiedlichen Nebenbeschäftigungen (z. B. essen, trinken, fernsehen) untersucht wird, wäre angesichts der Vielzahl der Einflussfaktoren, die mit der unabhängigen Variablen verknüpft sind, in ihrer internen Validität stark beschränkt. Darüber hinaus ist bei Feldstudien (ähnlich wie bei Laborstudien) eine Verallgemeinerbarkeit auf andere als genau die untersuchten Feldbedingungen fraglich: Was für heimisches Computerspielen gilt, mag dann wiederum nicht auf andere Spielkontexte (z. B. auf einer LAN-Party oder bei Freunden) generalisierbar sein.

Statistische Validität

Eine quantitative empirische Studie weist hohe statistische Validität auf, wenn die statistischen Datenanalysen korrekt durchgeführt wurden, so dass für die interessierenden Variablenzusammenhänge bzw. Effekte mit hoher Sicherheit nachgewiesen wurde, ob sie statistisch überzufällig (d. h. statistisch signifikant) sind oder nicht. Invalide sind Analysen, in denen fälschlich die Existenz eines Effekts angenommen wird (Fehler erster Art = Alpha-Fehler, d. h. falsche Entscheidung für die Forschungshypothese) sowie Analysen, in denen fälschlich die Existenz eines Effektes zurückgewiesen wird (Fehler zweiter Art = Beta-Fehler, d. h. falsche Entscheidung für die Nullhypothese; ► Abschn. 12.5.1). Zudem erfordert eine statistisch valide Analyse, dass mit hoher Sicherheit angegeben werden kann, ob im Falle eines signifikanten Effektes eine theoretisch und/oder praktisch bedeutsame Effektgröße (► Abschn. 14.2) in der Population vorliegt. Die statistische Validität ist vor allem von der Qualität der statistischen Datenanalyse (► Kap. 12) abhängig (► Neun Bedrohungen der statistischen Validität). Statistische Validität ist eine wesentliche Voraussetzung der internen Validität, denn nur wenn ein Populationseffekt relevanter Größe statistisch nachgewiesen ist, kann er im zweiten Schritt im Sinne interner Validität sinnvoll als Kausaleffekt interpretiert werden.

3.3 Gütekriterien für qualitative Forschung

Während in der quantitativen Forschung die zentralen Gütekriterien für methodische Strenge sowie für die anderen Qualitätskriterien der Wissenschaft (inhaltliche Relevanz, ethische Strenge und Präsentationsqualität) relativ konsensfähig und in vielen Bereichen detailliert ausformuliert sind, wird die Debatte um geeignete Qualitätskriterien in der qualitativen Forschung sehr viel kontroverser geführt. Wir gehen kurz auf diese Kontroverse ein, bevor wir zwei Kriterienkataloge exemplarisch vorstellen.

3.3.1 Kontroversen über Gütekriterien qualitativer Forschung

Mit der wachsenden Anerkennung des qualitativen Ansatzes in der empirischen Sozialforschung wuchs in den

letzten Dekaden die Nachfrage nach verbindlichen Qualitätskriterien, anhand derer gute von weniger guten qualitativen Studien abgegrenzt werden können. Denn von entsprechenden Bewertungen hängt es ab, ob eine qualitative Studie in einer renommierten Fachzeitschrift publiziert werden kann oder nicht, ob für eine geplante qualitative Grundlagenstudie Fördergelder eingeworben werden können oder ob Auftraggeber und Entscheidungsträger von der Tragfähigkeit der Ergebnisse einer qualitativen Evaluationsstudie überzeugt werden können (zur Bedeutung von Gütekriterien qualitativer Sozialforschung im Kontext der Einwerbung von Drittmitteln s. z. B. Reichertz, 2000). Es herrscht heute weitgehende Übereinstimmung dahingehend, dass es sinnvoll und notwendig ist, die wissenschaftliche Qualität qualitativer Forschungsprozesse und Forschungsergebnisse einer Bewertung zu unterziehen. Damit diese intersubjektiv nachvollziehbar ist, muss man sich auf die Grundlagen einer solchen Bewertung einigen.

Ein Ansatz zur Definition von Gütekriterien für die qualitative Forschung besteht darin, sich **an den Gütekriterien quantitativer Forschung zu orientieren** bzw. diese auf den qualitativen Ansatz zu übertragen. Dabei werden hinsichtlich methodischer Strenge meist „Objektivität, Reliabilität und Validität", „Reliabilität und Validität" oder „Objektivität, Reliabilität, interne und externe Validität" als Kernkriterien quantitativer Forschung zugrunde gelegt. Dieser Ansatz birgt zwei Probleme. Zum einen werden bei diesen Übertragungsversuchen nicht selten die Systematik sowie die Definitionen quantitativer Gütekriterien unscharf verwendet: Oft findet eine Vermischung der Testgütekriterien Objektivität, Reliabilität und Validität (die sich auf Messinstrumente beziehen) mit den Validitätskriterien wissenschaftlicher Aussagen statt (v. a. interne und externe Validität), die primär von Merkmalen des Untersuchungsdesigns abhängen. Zum anderen ist zu hinterfragen, ob es sinnvoll ist, Kriterien für Studien, die einem ganz anderen wissenschaftstheoretischen Paradigma folgen, in das qualitative Paradigma zu importieren (zu verschiedenen Ansätzen von Gütekriterien und ihrer historischen Entwicklung s. Breuer & Reichertz, 2001; Emden & Sandelowski, 1998; Flick, 2008).

Größere Akzeptanz als die Übernahme quantitativer Gütekriterien hat ein zweiter Ansatz, der darauf abzielt, aus der Logik der qualitativen Forschung heraus **eigene Gütekriterien zu entwickeln** und Techniken ihrer Sicherstellung anzugeben. Dies hat zu einem Boom an qualitativen Kriterienkatalogen geführt, so dass hier eine gewisse Unübersichtlichkeit vorherrscht: Offenbar existieren in der Fachliteratur **mehr als einhundert verschiedene Kriterienkataloge** (Noyes, Popay, Pearson, Hannes, & Booth, 2008, S. 580). Dementsprechend wurden und werden immer wieder Anläufe unternommen, die wichtigsten vorliegenden Kriterienkataloge für gute qualitative Forschung zu ordnen und zu einem integrativen Modell zusammenzufassen (Malterud, 2001; Meyrick, 2006; Parker, 2004; Reid & Gough, 2000; Tracy, 2010; Whittemore, Chase, & Mandle, 2001). Auch für qualitative Evaluationsstudien liegt ein integratives Kriterienraster vor (Spencer, 2003). Ein konsensfähiger einheitlicher Kanon von Kernkriterien ist indessen bislang nicht entstanden.

Auch ist bei den vorliegenden Kriterien-Checklisten teilweise unklar, wie praxistauglich sie sind: Lassen sich vorliegende qualitative Publikationen z. B. anhand der jeweiligen Guidelines und Checklisten eindeutig und intersubjektiv nachvollziehbar hinsichtlich ihrer wissenschaftlichen Güte einordnen? Ein umfassender Kriterienkatalog stammt z. B. von Elliott, Fischer und Rennie (1999), der sieben Kriterien der Wissenschaftlichkeit vorschlägt, die auf quantitative wie qualitative Studien zutreffen sollen, sowie sieben spezifische Qualitätskriterien für gute qualitative Forschung. Dieser Kriterienkatalog wurde anhand einer Stichprobe deutschsprachiger Publikationen in der Psychologie auf seine Handhabbarkeit untersucht, woraus Hinweise zur Bedeutung und Gewichtung der einzelnen Kriterien ableitbar sind (Ilg & Boothe, 2010).

Angesichts der Tatsache, dass qualitative Forschung hinsichtlich ihrer methodologischen und methodischen Vorgehensweisen und wissenschaftstheoretischen Grundlagen sehr ausdifferenziert ist (▶ Abschn. 2.3), ziehen manche Autoren in Zweifel, ob überhaupt einheitliche Kriterienkataloge für das gesamte Feld der qualitativen Sozialforschung formulierbar sind. Zudem wird befürchtet, dass die durch das wissenschaftstheoretische Paradigma vorgegebene Offenheit und Flexibilität qualitativer Forschung zu sehr beschnitten wird, wenn allgemeinverbindliche Qualitätskriterien vorgegeben werden, an die sich dann alle Forschenden halten sollen. Aus diesen Bedenken wird entweder abgeleitet, dass vorliegende Kriterienkataloge nicht mechanisch anzulegen und schematisch abzuarbeiten, sondern individuell abgestimmt auf die jeweilige Studie zu nutzen sind. Oder es wird grundlegend das Arbeiten mit Kriterienkatalogen infrage gestellt und nach Alternativen für die Bewertungspraxis gesucht, etwa im Sinne von Maßnahmen des Qualitätsmanagements im Forschungsprozess (Flick, 2008).

Wir vermuten, dass sich auch in der qualitativen Forschung mittel- bis langfristig verbindliche Qualitätskriterien etablieren und durchsetzen werden. Denn obschon alternative Vorgehensweisen wissenschaftsintern erfolgreich sein können, werden vor allem wissenschaftsexterne Zielgruppen verstärkt Bedarf nach klaren Kriterien anmelden, etwa im Bereich der Anwendungsforschung wie z. B. der qualitativen Evaluationsforschung, wenn es darum geht, aus Studienergebnissen praktische Konsequenzen abzuleiten. Auch die wachsende Bedeutung

der Forschungssynthese, bei der die Befunde qualitativer Studien sowie qualitativer und quantitativer Studien zur selben Forschungsfrage zusammengefasst werden sollen (► Kap. 16), erfordert die nachvollziehbare Einstufung aller Studien gemäß ihrer Qualität – vor allem ihrer methodischen Strenge (Noyes et al., 2008).

Neben methodologischen Diskussionen rund um die **normative Festlegung von Gütekriterien** in der qualitativen Sozialforschung mehren sich Studien, die **empirisch untersuchen** und beschreiben, ob und wie bislang in der Fachliteratur die Qualität qualitativer Studien gekennzeichnet wird. Anhand einer Analyse von 60 Zeitschriftenartikeln qualitativer Studien aus Soziologie und Erziehungswissenschaft zeigte sich z. B., dass die Qualität der publizierten Beiträge trotz erfolgreich durchlaufendem Peer-Review noch sehr heterogen ausfällt und insbesondere die Datenanalyse nicht immer nachvollziehbar beschrieben wird (Grunenberg, 2007). Offenbar wird sie auch bis heute nur selten mithilfe qualitativer Datenanalyse-Software (► Abschn. 12.1.3) durchgeführt, was die Dokumentation und Transparenz des Vorgehens möglicherweise fördern würde.

Wir empfehlen, bei der Durchführung eigener qualitativer Studien – ebenso wie bei quantitativen Studien – von Anfang an bewusst die Zielsetzung zu verfolgen, eine besonders aussagekräftige Studie zu realisieren und die verschiedenen Qualitätsdimensionen in den einzelnen Phasen des Forschungsprozesses zu beachten. Auf welche Kriterien bzw. welchen Kriterienkatalog man sich dabei aus welchen Gründen im Einzelnen stützt, sollte angesichts der Vielfalt der Ansätze und Kriterien ausdrücklich begründet werden.

Nicht zuletzt sei erwähnt, dass innerhalb der qualitativen Forschung auch Strömungen existieren, die – z. B. vor dem Hintergrund eines postmodernen oder eines bestimmten konstruktivistischen Wissenschaftsverständnisses – Qualitätskriterien für qualitative Forschung generell ablehnen, weil sie infrage stellen, dass es überhaupt einen übergeordneten Standpunkt gäbe, von dem aus entsprechende Kriterien zu legitimieren seien. Diese Position wird hier nicht vertieft. Sie stellt auch innerhalb des qualitativen Ansatzes eine Minderheitenposition dar (Steinke, 1999). Denn Qualitätskriterien lassen sich durchaus auch mit einem – in der qualitativen Forschung verbreiteten – (sozial-)konstruktivistischen Wissenschaftsverständnis vereinbaren (zur Wissenschaftstheorie ► Abschn. 2.3), allerdings gehört dazu dann die **Reflexion der Konstruiertheit der jeweiligen Qualitätskriterien**: Wer favorisiert welche Kriterien aus welchen Gründen, und welche Vor- und Nachteile entstehen dadurch für welche Forschenden bzw. Forschungsbereiche? Ein gänzlicher Verzicht auf die Qualitätsbewertung qualitativer Studien anhand nachvollziehbarer Kriterien oder Prinzipien wäre im Sinne der Wissenschaftlichkeit nicht vertretbar, denn er würde schließlich Beliebigkeit nach sich ziehen und auch eine Immunisierung gegen Kritik am jeweiligen Forschungsprozess darstellen.

3.3.2 Vier Kriterien der Glaubwürdigkeit von Lincoln und Guba (1985)

Der in der internationalen Fachliteratur mit Abstand am häufigsten zitierte Kriterienkatalog geht auf die US-amerikanische Erziehungswissenschaftlerin Yvonna Lincoln und ihren Kollegen Egon Guba zurück (Lincoln & Guba, 1985). Lincoln ist Mitautorin und Mitherausgeberin zahlreicher einschlägiger qualitativer Methodenhandbücher (Denzin & Lincoln, 1994, 2000, 2005, 2012; Denzin, Lincoln, & Smith, 2008; Lincoln & Denzin, 2003). Guba und Lincoln haben zudem umfassend zur qualitativen Evaluationsforschung publiziert und den Ansatz der konstruktivistischen Evaluation bzw. „Evaluation der vierten Generation“ entwickelt, bei dem die aktive Partizipation der Betroffenen eine wichtige Rolle spielt (zu partizipativen Evaluationsmodellen ► Abschn. 18.2.4).

Gemäß Lincoln und Guba (1985) muss gute qualitative Forschung das Ober-Kriterium der **Glaubwürdigkeit** („trustworthiness“) erfüllen, also das Publikum überzeugen, dass die Ergebnisse der jeweiligen qualitativen Studie wirklich aussagekräftig sind und etwas über die untersuchte soziale Wirklichkeit aussagen. Es werden vier Kriterien der Glaubwürdigkeit vorgeschlagen (1. **Vertrauenswürdigkeit**, 2. **Übertragbarkeit**, 3. **Zuverlässigkeit** und 4. **Bestätigbarkeit**), für die jeweils konkrete Techniken der Qualitätssicherung sowie die logischen Entsprechungen im quantitativen Paradigma genannt werden (■ Tab. 3.5). Das wichtigste Gütekriterium im Glaubwürdigkeitsmodell ist die Vertrauenswürdigkeit, die das qualitative Pendant zur internen Validität darstellt.

Ein Grund für den Erfolg des Kriterienkatalogs von Lincoln und Guba (1985) liegt vermutlich darin, dass durch die Nennung der jeweiligen Pendants im quantitativen Paradigma Anschlussfähigkeit an etablierte Diskussionen zur wissenschaftlichen Qualität hergestellt wurde: In ihrer Darstellung nehmen Lincoln und Guba (1985) einerseits ausdrücklich auf die Validitätstypologie der Campbell-Tradition sowie auf die testtheoretischen Gütekriterien der Objektivität, Reliabilität und Validität (► Abschn. 10.4.1) Bezug und grenzen sich andererseits durch die Begründung und Festlegung eigener Kriterien davon ab. Zudem gibt das Modell relativ genaue Praxishinweise zur Sicherung der vier Gütekriterien. Allerdings werden keine klaren Standards formuliert, anhand derer einzuordnen wäre, ob und unter welchen Bedingungen ei-

Tabelle 3.5 Vier Gütekriterien für qualitative Forschung von Lincoln & Guba (1985. Copyright © 1985 by SAGE Publications. Reprinted by Permission of SAGE Publications) inklusive Techniken zu ihrer Sicherung sowie den von Schou, Høstrup, Lyngsø, Larsen, & Poulsen (2012. Mit freundlicher Genehmigung von John Wiley and Sons) entwickelten Bewertungskriterien für Zeitschriftenartikel. (Eigene Ergänzungen)

Allgemeine Qualitätsaspekte wissenschaftlicher Forschung (Lincoln & Guba, 1985, S. 290)	Qualitätskriterien in der qualitativen Sozialforschung: Vier Kriterien der *Glaubwürdigkeit* (Trustworthiness)	Techniken zur Sicherung der vier Kriterien der Glaubwürdigkeit im qualitativen Forschungsprozess (Lincoln & Guba, 1985, S. 301)	Checkliste zur Einhaltung der vier Kriterien der Glaubwürdigkeit in Zeitschriftenartikeln (Schou et al., 2012, S. 2090)	Entsprechende Qualitätskriterien in der quantitativen Sozialforschung
„**Wahrheitsgehalt**“ („Truth Value“): How can one establish confidence in the „truth“ of the findings of a particular inquiry for the subjects (respondents) with which and the context in which the inquiry was carried out?	**1. Vertrauenswürdigkeit** (Credibility): Die Ergebnisse und Interpretationen der Studie auf der Basis der Daten sind vertrauenswürdig.	– **Umfassende Datenerhebung** durch lange Zeiten im Feld (Prolonged Engagement, Persistent Observation) und Triangulation im Sinne der Kombination von verschiedenen Daten, Methoden und Forschenden (Triangulation) – **Nachbesprechung der Studie mit außenstehenden Fachkollegen** (Peer Debriefing) – **Überdenken der Vorannahmen anhand von Gegenbeispielen in den Daten** (Negative Case Analysis) – **Überprüfung der Interpretationen anhand der Rohdaten** (Referential Adequacy) – **Kommunikative Validierung durch Rücksprache mit den untersuchten Personen** (Member Checking)	1. The purpose is described clearly. 2. The method is described. 3. Arguments for choice of method have been made. 4. The methods suits the purpose. 5. There is a description of how data were registered. 6. Triangulation has been applied. 7. The research process is described.	**Interne Validität**
Anwendbarkeit (Applicability): How can one determine the extent to which the findings of a particular inquiry have applicability in other contexts or with other subjects (respondents)?	**2. Übertragbarkeit** (Transferability): Die Ergebnisse bzw. Schlussfolgerungen aus der Studie sind auf andere Kontexte übertragbar.	**Dichte Beschreibung** der untersuchten Personen und Kontextbedingungen, damit eine mögliche Übertragbarkeit der Ergebnisse auf andere Personen und Kontexte abschätzbar ist (Thick Description).	1. Selection of informants or sources is described. 2. There is a description of the informants. 3. It is argued why these informants are selected. 4. The context (place and connection of research) is described. 5. The relationship between the researcher(s) and the context (in which the research takes place) and the informants is described.	**Externe Validität**
Konsistenz (Consistency): How can one determine whether the findings of an inquiry would be repeated if the inquiry were replicated with the same (or similar) subjects (respondents) in the same (or similar) context?	**3. Zuverlässigkeit** (Dependability): Der Forschungsprozess wurde nachvollziehbar gestaltet und durchlaufen.	– Forschungs-Audit, bei dem das Forschungsteam die Vorgehensweise gegenüber Experten detailliert erläutert und begründet (**Inquiry Audit**). – Die Zuverlässigkeit kann auch durch **Triangulation** geprüft werden, indem bestimmte Forschungsschritte von verschiedenen Mitgliedern des Teams parallel durchgeführt werden (Stepwise Replication) oder indem im Sinne von Triangulation die mit einer Methode gewonnenen Daten durch eine andere Methode geprüft werden (Overlap Methods).	1. A logical connection between data and themes is described. 2. The process of analysis is described. 3. There is a clear description of the findings. 4. The findings are credible. 5. Any quotation is reasonable/supporting the interpretation. 6. There is agreement between the findings of the study and the conclusions.	**Reliabilität**

Tabelle 3.5 (Fortsetzung)

Allgemeine Qualitätsaspekte wissenschaftlicher Forschung (Lincoln & Guba, 1985, S. 290)	Qualitätskriterien in der qualitativen Sozialforschung: Vier Kriterien der *Glaubwürdigkeit* (Trustworthiness)	Techniken zur Sicherung der vier Kriterien der Glaubwürdigkeit im qualitativen Forschungsprozess (Lincoln & Guba, 1985, S. 301)	Checkliste zur Einhaltung der vier Kriterien der Glaubwürdigkeit in Zeitschriftenartikeln (Schou et al., 2012, S. 2090)	Entsprechende Qualitätskriterien in der quantitativen Sozialforschung
Neutralität (Neutrality): How can one establish the degree to which the findings of an inquiry are determined by the subjects (respondents) and conditions of the inquiry and not by the biases, motivations, interests, or perspectives of the inquirer.	**4. Bestätigbarkeit** (Confirmability): Die Studienergebnisse sind nicht durch Vorurteile, Interessen, Perspektiven der einzelnen Forschenden vorbestimmt.	Forschungs-Audit, bei dem das Forschungsteam die vorliegenden Daten und Dokumentationen des Forschungsprozesses detailliert vorlegt und erläutert (**Confirmability Audit**). Dabei sind Ergebnisse im Rahmen der Triangulation sowie Dokumente der Selbstreflexion der Forschenden (z. B. Forschungstagebücher: Reflexive Journals) besonders nützlich. Bestätigbarkeit und Zuverlässigkeit lassen sich gemeinsam in einem einzigen Forschungs-Audit prüfen.	1. The researcher has described his background and perceptions or preunderstanding. 2. There are references to theories (clear who has inspired the analysis). 3. Description of whether themes were identified from data or formulated in advance. 4. It is described who conducted the study. 5. It is described how the researcher participated in the process. 6. The researcher has described whether his/her position is important in relation to the findings.	**Objektivität**
			Zusätzliche formale Anforderungen der Wissenschaftlichkeit: 1. Background of the study is described through existing literature. 2. It appears why the study is relevant. 3. It is described how demands to informed consent have been met (Helsinki Declaration) 4. It is described if there are relevant approvals. 5. The researcher has described whether the study can affect the informants. 6. The researcher has described what will be done if the study affects the participants.	**Inhaltliche Relevanz** **Ethische Strenge**

ne bestimmte Maßnahme zur Sicherung eines Gütekriteriums (z. B. kommunikative Validierung zur Sicherung der Vertrauenswürdigkeit; dichte Beschreibung zur Sicherung der Übertragbarkeit) in ausreichendem, gutem oder gar exzellentem Maße umgesetzt wurde. Doch entsprechende Standards sind auch für die Gütekriterien quantitativer Forschung kaum präzise und allgemeingültig anzugeben, sondern erfordern Experteneinschätzungen und argumentative Einordnungen.

Für die Orientierung an dem Glaubwürdigkeitsmodell sprechen seine **weltweite Anerkennung** (u. a. auch durch die Cochrane-Collaboration, die Forschungsergebnisse in der Medizin systematisch zusammenfasst; ▶ Abschn. 16.1.2; Hannes, 2011) sowie auch die **Verfügbarkeit konkreter Indikatoren und Checklisten**, die es erleichtern sollen, eigene oder publizierte Studien hinsichtlich der vier Qualitätsdimensionen einzuordnen (z. B. Schou et al., 2012; Tab. 3.5). Zu beachten ist, dass die vorge-

schlagenen Techniken zur Sicherstellung der qualitativen Gütekriterien teilweise **sehr anspruchsvoll** sind. Die geforderten Forschungs-Audits werden z. B. in der Praxis der qualitativen Forschung kaum umgesetzt. Generell sind alle Techniken der Qualitätssicherung kontextsensibel einzusetzen. So kann kommunikative Validierung („member checking") an Grenzen stoßen, wenn die Forschung Interpretationen liefert, die die Interviewpartner nicht wahrhaben wollen (Lincoln & Guba, 1985, S. 325) oder wenn die Konfrontation mit den Befunden für die Teilnehmenden belastend wäre.

Beispiel: Eine qualitative Beobachtungsstudie arbeitet heraus, dass und wie in einem bestimmten Kontext Menschen mit Handicaps von ihren Ehepartnern entmündigt und wie Kinder behandelt werden. Sollte man die Betroffenen (welche?) um eine ausdrückliche Bestätigung dieses Befundes bitten, der möglicherweise kränkend oder selbstwertbedrohend wirkt? Hier sind ethische Aspekte zu bedenken (für weitere Praxiserfahrungen mit verschiedenen Techniken zur Sicherung der qualitativen Gütekriterien s. Krefting, 1991).

3.3.3 Sieben Kernkriterien zur Bewertung qualitativer Forschung nach Steinke (1999)

Auf der Basis einer gründlichen Betrachtung der Gütekriterien quantitativer Forschung und ihrer möglichen Übertragbarkeit auf qualitative Forschung sowie auf der Grundlage nationaler und internationaler Vorschläge zu eigenständigen Gütekriterien qualitativer Forschung hat die Psychologin Ines Steinke einen Katalog von **sieben Kernkriterien zur Bewertung qualitativer Forschung** zusammengestellt (Steinke, 1999, 2009). Die Kriterien lassen sich im Sinne einer Checkliste nutzen, um eigene Untersuchungen mit hoher Qualität zu planen, umzusetzen und darzustellen sowie vorliegende Publikationen zu beurteilen. Dabei sollen die Kriterien nicht mechanisch abgearbeitet, sondern in Abstimmung auf die jeweiligen Untersuchungsbedingungen ausgewählt, konkretisiert, modifiziert oder erweitert werden. Für alle sieben Kriterien werden auch Methoden zu ihrer Sicherung angegeben (Tab. 3.6).

Während sich die vier Kriterien der Glaubwürdigkeit von Lincoln und Guba (1985) ganz auf die methodische Strenge als eines von vier allgemeinen Kriterien der wissenschaftlichen Qualität (▶ Abschn. 3.1.2) konzentrieren, wird mit den sieben Kernkriterien nach Steinke (1999) neben der **methodischen Strenge** (zu der Indikation, empirische Verankerung, Limitation, reflektierte Subjektivität und Kohärenz gehören) auch die Frage der **Relevanz** sowie der **Repräsentationsqualität** angesprochen (intersubjektive Nachvollziehbarkeit; Tab. 3.7). Fragen der **ethischen Strenge** werden nicht elaboriert; sie sind beim Kriterium der reflektierten Subjektivität indirekt angesprochen (etwa wenn es um die Vertrauensbeziehung zwischen Forschenden und Beforschten geht).

Die **Triangulation**, d. h. die Kombination unterschiedlicher Theorien, Methoden, Daten und/oder Forschender zur wechselseitigen Absicherung und Ergänzung des qualitativen Forschungsprozesses (Flick, 2004) ist im Kriterienkatalog nicht explizit enthalten, da sie gemäß Steinke (1999, S. 247, 251) primär der **Generierung** von (möglichst) allgemeingültigen Theorien dient. Zur Validierung bzw. Prüfung der Grenzen der entwickelten Theorie dienen die dem Kriterium der Limitation zugeordneten Techniken (Tab. 3.6). Andere Ansätze fassen jedoch die Triangulation ausdrücklich als ein Gütekriterium qualitativer Forschung auf (z. B. Mayring, 2002) bzw. betrachten Triangulation als eine wichtige Technik zur Sicherstellung von Gütekriterien (z. B. der Zuverlässigkeit im Kriterienkatalog von Lincoln & Guba, 1985).

Ein genuin qualitatives Kriterium zur Bewertung wissenschaftlicher Studien ist das Kriterium der **reflektierten Subjektivität**, weil gemäß qualitativem Wissenschaftsverständnis (▶ Abschn. 2.3) die subjektiven Sicht- und Verhaltensweisen der Forschenden nicht durch Standardisierung des Vorgehens neutralisiert, sondern für den Erkenntnisprozess genutzt werden sollen. Die subjektiven Sichtweisen der Forschenden sollen als solche anerkannt und in den Dienst der Forschung gestellt werden: Gerade aufgrund ihrer subjektiven Vorerfahrungen, Werthaltungen oder persönlichen Eigenschaften können bestimmte Forschende Zugang zu bestimmten Praxisfeldern erhalten und/oder ein besonderes Vertrauensverhältnis zu bestimmten Informantinnen und Informanten aufbauen, was die Art und Qualität der erhobenen Daten sowie deren Interpretation stark beeinflusst. Merkmale der subjektiven Sichtweisen der Forschenden sowie ihr Einfluss auf den Forschungsprozess müssen im Zuge jeder seriösen qualitativen Forschung umfassend dokumentiert und kritisch reflektiert werden (zur Qualität qualitativer Interviews und zur Selbstreflexion der Forschenden bei der Durchführung und Auswertung qualitativer Interviews s. Hellferich, 2005).

Betrachtet man die sieben Kernkriterien für gute qualitative Forschung, so fällt auf, dass viele von ihnen übergreifende Kriterien wissenschaftlicher Qualität darstellen, die ebenso an quantitative Forschungsprozesse anzulegen sind (Tab. 3.7). Dies ist insofern nicht verwunderlich, als a) qualitative wie quantitative Forschung übergeordneten Kriterien der Wissenschaftlichkeit verpflichtet sind (▶ Abschn. 3.1.2) und b) bei der Erarbeitung der Kernkriterien von Steinke (1999, 2009) auf etablierte Gütekriterien quantitativer Forschung aufgebaut wurde,

Tabelle 3.6 Checkliste der sieben Kernkriterien zur Beurteilung der Qualität qualitativer Studien. (Mod. nach Steinke, 1999, S. 252ff. Mit freundlicher Genehmigung von Beltz Juventa)

Kernkriterien	Unterkriterien	Umsetzung
1. Intersubjektive Nachvollziehbarkeit Wie gut können Außenstehende den gesamten qualitativen Forschungsprozess der Studie anhand der Studiendokumentation im Detail nachvollziehen und somit auch bewerten?	1.1 Transparenz 1.2 Explizitheit	– Umfassende Dokumentation: a) des Vorverständnisses über den Gegenstand b) der Erhebungsmethoden und Erhebungskontexte c) der Transkriptionsregeln d) der Daten e) der Auswertungsmethoden f) der Informationsquellen g) der Entscheidungen und Probleme im Forschungsprozess h) der angelegten wissenschaftlichen Gütekriterien i) der Reflexion der eigenen Subjektivität – Interpretationen in Gruppen – Anwendung bzw. Entwicklung kodifizierter Verfahren
2. Indikation Wie gut sind die einzelnen methodischen Entscheidungen im qualitativen Forschungsprozess der Studie hinsichtlich ihrer Angemessenheit für das Forschungsproblem bzw. den Forschungsgegenstand begründet? (Das Kriterium der Indikation geht über die Gegenstandsangemessenheit hinaus, indem die Angemessenheit des gesamten Forschungsprozesses in den Blick genommen wird.)	2.1 Indikation qualitativen Vorgehens angesichts der Fragestellung	Allgemeine Begründung dazu, warum ein qualitativer Ansatz für das Forschungsproblem gut geeignet bzw. notwendig ist
	2.2 Indikation der Methodenwahl	Konkrete Begründungen dazu, inwiefern die gewählten qualitativen Methoden und die Art ihrer Anwendung gegenstandsangemessen sind. Positive Antworten auf folgende Fragen sind zu liefern: a) Wurde den Äußerungen und Bedeutungen der Untersuchten ausreichend Spielraum eingeräumt? b) Waren die Forschenden längere Zeit im Feld anwesend? c) Besteht ein Arbeitsbündnis zwischen Forschenden und Informanten? d) Wurden die Methoden gegenstandsangemessen ausgewählt? e) Wurden die Methoden gegenstandsangemessen entwickelt? f) Ermöglichen die verwendeten Verfahren Irritationen des Vorwissens?
	2.3 Indikation der Transkriptionsregeln	Begründung dazu, warum die gewählten Transkriptionsregeln für das Forschungsproblem angemessen sind
	2.4 Indikation der Samplingstrategie	Begründung dazu, warum die gewählte Samplingstrategie für das Forschungsproblem angemessen ist
	2.5 Indikation methodischer Einzelentscheidungen	Begründung von methodischen Einzelentscheidungen im Kontext der gesamten Untersuchung. Positive Antworten auf folgende Fragen sind zu geben: a) Sind die verschiedenen genutzten qualitativen Datenerhebungs- und Datenanalysemethoden koindiziert, d. h. passen sie sinnvoll zusammen? b) Ist das geplante Untersuchungsdesign unter forschungspragmatischer Perspektive (Zeit-, Kosten-, Personalressourcen, Aufwand für die Beforschten) realisierbar bzw. stehen bei einer bereits realisierten Studie Aufwand und Nutzen in vertretbarem Verhältnis?
	2.6 Indikation von Bewertungskriterien	Erläuterung dazu, warum die in den Methoden-Begründungen herangezogenen Qualitätskriterien qualitativer Forschung der Fragestellung, den Methoden und dem Gegenstand angemessen sind

sofern sich eine entsprechende Übertragbarkeit auf den qualitativen Ansatz forschungslogisch rechtfertigen ließ. So ist von der Theoriebildung in der qualitativen wie in der quantitativen Forschung Konsistenz gefordert. Ebenso müssen Methodenwahlen im qualitativen wie im quantitativen Forschungsprozess in Bezug zum jeweiligen Forschungsproblem ausdrücklich hinsichtlich ihrer Eignung begründet werden (sog. Indikationsfrage).

Der in Tab. 3.7 dargestellte Überblick der bislang vorgestellten quantitativen und qualitativen Kriterienkataloge zeigt, dass sich deren Gütekriterien schwerpunktmäßig auf methodische Strenge beziehen. Das trifft auch

Tabelle 3.6 (Fortsetzung)

Kernkriterien	Unterkriterien	Umsetzung
3. Empirische Verankerung Wie gut sind die gebildeten und/oder geprüften Hypothesen und Theorien auf der Basis der empirischen Daten begründet?	3.1 Empirische Verankerung für Theoriebildung	Einsatz von kodifizierten Verfahren zur empirisch verankerten Theoriebildung
	3.2 Empirische Verankerung für Theorieprüfung	– Einsatz von kodifizierten Verfahren zur empirisch verankerten Theorieprüfung: a) Gibt es hinreichende Textbelege für die generierte Theorie? b) Wie wurde mit Widersprüchen, abweichenden bzw. negativen Fällen, Ereignissen etc. umgegangen? – Analytische Induktion – Prognosen: a) innerhalb des Textes b) über künftiges Verhalten – Kommunikative Validierung
4. Limitation Wie genau wird angegeben, auf welche weiteren Bedingungen (Kontexte, Fälle, Untersuchungsgruppen, Phänomene etc.) sich die Forschungsergebnisse verallgemeinern lassen bzw. inwiefern die Verallgemeinerbarkeit limitiert ist?	4.1 Beschreibung der Kontexte 4.2 Identifikation von für das Untersuchungsphänomen relevanten Bedingungen und Kontexten	– Technik des Fortdenkens – Erfahrungsprobe – Fallkontrastierung – Suche und Analyse von abweichenden, negativen und extremen Fällen
5. Reflektierte Subjektivität Wie umfassend und überzeugend reflektieren die Forschenden ihre eigenen subjektiven Positionen und Rollen (z. B. Forschungsinteresse, sozialer und biografischer Hintergrund) im Verhältnis zum untersuchten Phänomen und zu den untersuchten Personen(gruppen)?	5. Reflektierte Subjektivität wird nicht in Unterkriterien differenziert; sie bezieht sich auf die Selbstreflexion der Forschenden über den gesamten Forschungsprozess hinweg.	Erläuterung dazu, wie die Forschenden ihre subjektiven Perspektiven und Rollen im Forschungsprozess reflektieren. Dabei sind verschiedene Aspekte und Unteraspekte anzusprechen: – Wird der Forschungsprozess durch Selbstbeobachtungen begleitet? – Werden persönliche Voraussetzungen für die Erforschung des Untersuchungsgegenstandes reflektiert? a) Sind das methodische Vorgehen und der Untersuchungsgegenstand der Person der Forschenden angemessen? b) Werden die eigenen Voraussetzungen als Angehörige einer bestimmten Berufsgruppe reflektiert? c) Wird die kulturelle Herkunft der Forschenden reflektiert? – Liegt gleichschwebende Aufmerksamkeit als Einstellung zum Untersuchungsthema und zur Rolle des Forschers vor? – Wird eine reflektierte Beziehung zum Untersuchten durch Oszillation zwischen Annäherung und Distanz zum Untersuchten hergestellt?
6. Kohärenz Wie stimmig und widerspruchsfrei ist die Theorie bzw. sind die Interpretationen auf der Basis der Daten?	6. Kohärenz wird nicht in Unterkriterien differenziert; sie bezieht sich auf die durch die Studie hervorgebrachten Interpretationen und Theorien.	– Erläuterung dazu, inwieweit die generierte Theorie kohärent ist – Erläuterung dazu, wie mit Widersprüchen in den Daten und bei den Interpretationen umgegangen wurde
7. Relevanz Wie groß ist der Beitrag einer grundlagenwissenschaftlichen qualitativen Studie für den wissenschaftlichen Erkenntnisfortschritt im Sinne von Gegenstandsbeschreibung und Theoriebildung (theoretische Relevanz)? Bei angewandten qualitativen Studien wie z. B. qualitativer Evaluationsforschung ist stattdessen die praktische Relevanz hervorzuheben.	7.1 Relevanz der Fragestellung	Erläuterung dazu, inwiefern die gewählte Fragestellung bzw. der untersuchte Gegenstand relevant ist
	7.2 Relevanz der entwickelten Theorie: Welchen Beitrag leistet die entwickelte Theorie?	Erläuterung zum Wert der Studienergebnisse, insbesondere der neu gebildeten Hypothesen und Theorien: – Werden durch die Theorie neue Deutungen zur Verfügung gestellt? – Beinhaltet die Theorie Erklärungen für das interessierende Phänomen? – Inwiefern regt die Theorie zur Lösung von Problemen an? – Sind die Ergebnisse verallgemeinerbar? – Ist die Darstellung der Theorie überschaubar?

Tabelle 3.7 Kriterien wissenschaftlicher Qualität im quantitativen und qualitativen Paradigma im Überblick

Standards der Wissenschaftlichkeit	Kriterien der wissenschaftlichen Qualität	Gütekriterien quantitativer Forschung: Vier Arten der Validität (Shadish et al., 2002)	Gütekriterien qualitativer Forschung: Vier Kriterien der Glaubwürdigkeit (Lincoln & Guba, 1985)	Sieben Kernkriterien zur Bewertung qualitativer Forschung (Steinke, 1999)
1. Wissenschaftliches Forschungsproblem	Inhaltliche Relevanz	–	–	Relevanz
2. Wissenschaftlicher Forschungsprozess	Methodische Strenge	(Objektivität)	Bestätigbarkeit	Reflektierte Subjektivität
		(Reliabilität)	Zuverlässigkeit	–
		Konstruktvalidität	–	–
		Interne Validität	Vertrauenswürdigkeit	Indikation Empirische Verankerung Kohärenz
		Externe Validität	Übertragbarkeit	Limitation
		Statistische Validität	–	–
3. Wissenschafts- und Forschungsethik	Ethische Strenge	–	–	–
4. Dokumentation des Forschungsprojekts	Präsentationsqualität	–	–	Intersubjektive Nachvollziehbarkeit

auf viele der anderen qualitativen Kriterienkataloge zu. So formuliert der Psychologe **Philipp Mayring sechs Gütekriterien** (Mayring, 2002, S. 144ff.):

1. Verfahrensdokumentation,
2. Interpretationsabsicherung mit Argumenten,
3. Regelgeleitetheit,
4. Nähe zum Gegenstand,
5. kommunikative Validierung und
6. Triangulation.

Diese sechs Gütekriterien widmen sich primär der methodischen Strenge und hier spezieller der empirischen Verankerung qualitativer Forschung. Die Kernkriterien guter qualitativer Forschung nach Steinke (1999) weisen darüber hinaus auf die Bedeutung der reflektierten Subjektivität im qualitativen Forschungsprozess hin, verlangen die ausdrückliche Begründung aller Entscheidungen im Forschungsprozess (Indikation), fordern bedeutsame Fragestellungen und Ergebnisse (Relevanz) sowie eine ausdrückliche Behandlung der Generalisierbarkeit der Ergebnisse (Limitation). Gerade wegen seiner Breite kann das Kriterienraster von Steinke (1999, 2009) eine nützliche Arbeitsgrundlage darstellen.

Abschließend sei noch erwähnt, dass es in der qualitativen Forschung neben übergeordneten Qualitätskriterien meist für jede einzelne Methodologie und Methode (z. B. ethnografische Feldbeobachtung, Grounded-Theory-Methodologie, Metaphern-Analyse) noch verfahrensspezifische Kriterien gibt, die vor allem auf die konkrete Sicherstellung der methodischen Strenge beim jeweiligen Verfahren abheben und somit für die Planung eigener Studien hilfreich sein können. Entsprechende Hinweise finden sich jeweils in der Methoden-Fachliteratur zu den einzelnen Verfahren (► Abschn. 12.1).

3.4 Gütekriterien für Mixed-Methods-Forschung

Da in der Mixed-Methods-Forschung qualitative und quantitative Forschungsprozesse und Methoden miteinander verknüpft werden (► Abschn. 2.4 und ► Abschn. 7.1), kommen sowohl die **Gütekriterien der quantitativen Sozialforschung** als auch die **Gütekriterien der qualitativen Sozialforschung** zur Anwendung. Denn es ist sicherzustellen, dass die quantitativen und qualitativen Bestandteile der Mixed-Methods-Studie jeweils für sich genommen hohe wissenschaftliche Qualität aufweisen.

Darüber hinaus müssen **spezifische Gütekriterien der Mixed-Methods-Forschung** angelegt werden. Sie beziehen sich auf zwei Arbeitsschritte, die nur in Mixed-Methods-Studien vorkommen:

- Die **Mixed-Methods-Designqualität** („mixed methods design quality“) adressiert die Art und Weise der Verknüpfung qualitativer und quantitativer Teilstudien

oder Phasen des Forschungsprozesses zu einem spezifischen Mixed-Methods-Design. Diese Verknüpfung erfolgt im Zuge der Planung sowie der Umsetzung des Mixed-Methods-Designs und muss mit Blick auf das Forschungsproblem möglichst aussagekräftig sein. Es muss begründet werden, warum für das konkrete Forschungsproblem überhaupt eine Mixed-Methods-Studie indiziert ist und warum welcher Typus von Mixed-Methods-Design ausgewählt wurde (für unterschiedliche Mixed-Methods-Designtypen ► Abschn. 7.1). Für das jeweilige Mixed-Methods-Design muss dann eine stringente Umsetzung (z. B. bei der wechselseitigen Abstimmung quantitativer und qualitativer Datenerhebung und Stichprobenziehung) nachgewiesen werden (genauer zur Mixed-Methods-Designqualität s. Teddlie & Tashakkori, 2009, S. 301).

- Die **Mixed-Methods-Interpretations-/Meta-Interpretationsqualität** („mixed methods interpretive rigor“) bezieht sich auf Auswertungs- und Interpretationsphase und die Art und Weise, wie qualitative und quantitative Teilergebnisse herausgearbeitet, in geeigneter Weise aufeinander bezogen und zu einer schlüssigen Gesamtinterpretation verdichtet werden. Für die Güte einer Mixed-Methods-Studie sind dabei gerade die Konsistenz zwischen quantitativen und qualitativen Teilergebnissen sowie der Umgang mit möglichen Inkonsistenzen im Rahmen der Gesamtinterpretation von Bedeutung. Ebenso muss die Meta-Interpretation die Zielsetzungen einlösen, die ursprünglich von den Forschenden mit der Mixed-Methods-Studie verfolgt wurden (genauer zur Mixed-Methods-Interpretationsqualität s. Teddlie & Tashakkori, 2009, S. 301).

Hohe Mixed-Methods-Design- sowie Mixed-Methods-Interpretationsqualität sorgen gemeinsam für eine hohe **Inferenzqualität** („inference quality“), dem Mixed-Methods-Pendant der internen Validität im quantitativen Paradigma bzw. der Vertrauenswürdigkeit im qualitativen Paradigma.

Ist hohe Inferenzqualität sichergestellt (also von einem tragfähigen Gesamtergebnis der Mixed-Methods-Studie auszugehen), stellt sich noch die Frage nach der externen Validität bzw. Übertragbarkeit auf andere Settings/Kontexte, Personen/Populationen, Zeiten oder Erhebungsverfahren. Dieses Kriterium wird im Mixed-Methods-Paradigma auch als **Inferenzübertragbarkeit** („inference transferability“) bezeichnet (Teddlie & Tashakkori, 2009, S. 311).

Während mit Inferenzqualität und Inferenzübertragbarkeit zwei zentrale Gütekriterien der **methodischen Strenge** angesprochen sind, existieren auch breiter angelegte Kriterienlisten für Mixed-Methods-Forschung, die **Gütekriterien für alle Phasen des Forschungsprozesses** formulieren (O'Cathain, 2010; O'Cathain, Murphy, & Nicholl, 2008):

1. **Planungsqualität**: z. B. Ist die Mixed-Methods-Studie durchführbar geplant?
2. **Designqualität**: z. B. Ist die Wahl eines Mixed-Methods-Designs mit Bezug zum Forschungsproblem inhaltlich und methodisch nachvollziehbar begründet Mixed-Methods-Designqualität)?
3. **Datenqualität**: z. B. Werden aussagekräftige qualitative und quantitative Stichproben und Datenerhebungsverfahren gewählt?
4. **Interpretationsqualität**: z. B. Werden qualitative und quantitative Daten bei der Auswertung und Interpretation sinnvoll aufeinander bezogen Mixed-Methods-Interpretationsqualität)?
5. **Inferenzübertragbarkeit**: z. B. Lassen sich die Ergebnisse der Studie schlüssig auf andere Kontexte übertragen?
6. **Präsentationsqualität**: z. B. Werden die Details der Mixed-Methods-Studie in einer Publikation detailliert und nachvollziehbar dargestellt?
7. **Synthetisierbarkeit**: z. B. Sind die Studienergebnisse so umfassend dargestellt, dass die Mixed-Methods-Studie in eine Forschungssynthese (z. B. Metaanalyse) aufgenommen werden könnte?
8. **Nützlichkeit**: z. B. Werden die Ergebnisse der Mixed-Methods-Studie in der Praxis genutzt (z. B. in bestimmten Berufsfeldern oder als Basis politischer Entscheidungen)?

Aufschlussreich sind auch die Anmerkungen von Creswell und Clark (2011, S. 268) zu der Frage, wie sie als Herausgeber des *Journal of Mixed Methods Research* eingereichte Artikel hinsichtlich ihrer Qualität einer ersten Bewertung unterzogen. Sie unterschieden klassisch zwischen den Aspekten der inhaltlichen Relevanz („relevance“) und der methodischen Strenge („rigor“). Um Hinweise auf die methodische Strenge einer Mixed-Methods-Studie zu finden, orientierten sie sich vor allem am Methodenteil: Welches Mixed-Methods-Design (für unterschiedliche Designvarianten in der Mixed-Methods-Forschung ► Abschn. 7.1) wurde mit welcher Begründung und welcher Stringenz umgesetzt? Dann betrachteten sie den Ergebnis- und Diskussionsteil hinsichtlich der Schlüssigkeit der Interpretationen und insbesondere der wechselseitigen Bezüge zwischen quantitativen und qualitativen Teilbefunden. Indikatoren, die darauf hindeuten, dass es sich um eine Studie handelt, die quantitative und qualitative Einzelbefunde nicht nebeneinander stellt, sondern im Sinne des Mixed-Methods-Paradigmas eine durch die Forschungsfrage geleitete Integration bietet, waren z. B. Tabellen und Grafiken, die qualitative und quantitative Befunde gemeinsam darstellten sowie die Verwendung

korrekter Terminologie aus dem Mixed-Methods-Paradigma (etwa korrekte Fachbegriffe für Mixed-Methods-spezifische Designtypen oder Qualitätskriterien).

Lernquiz

1. Grenzen Sie verschiedene Formen der Pseudowissenschaft voneinander ab.
2. Nennen Sie vier Standards der Wissenschaftlichkeit und die zugehörigen Kriterien wissenschaftlicher Qualität!
3. Welche Aussagen stimmen?
   a) Eine zu geringe Stichprobengröße gefährdet die statistische Validität einer Studie.
   b) Durch eine repräsentative Stichprobenauswahl wird die interne Validität einer Studie gesteigert.
   c) Forschungsethische Qualitätskriterien sind vor allem bei der Datenerhebung und Ergebnispublikation zu beachten.
   d) In den meisten Phasen des empirisch-quantitativen Forschungsprozesses ist die methodische Strenge das zentrale Qualitätskriterium.
   e) Im quantitativen und qualitativen Paradigma werden hinsichtlich methodischer Strenge dieselben Gütekriterien angelegt.
4. Welche Formen der Validität werden in der quantitativen Forschung in der Campbell-Tradition unterschieden und auf welche Phasen des Forschungsprozesses bzw. auf welche Aspekte einer Studie beziehen sie sich hauptsächlich?
5. Welche Typen der Validität einer Studie werden durch geringe Reliabilität von Messinstrumenten gefährdet?
6. Welche Aussagen sind falsch?
   a) Randomisierte Kontrollgruppenstudien garantieren maximale externe Validität.
   b) Testübung bedroht vor allem die interne Validität.
   c) Zu kleine Stichproben bedrohen die statistische Validität.
   d) Die Konzeptspezifikation ist die theoretische Basis der Konstruktvalidität.
   e) Interne Validität ist die Voraussetzung für externe Validität.
   f) Statistische Validität ist die Voraussetzung für interne Validität.
7. Welche Gütekriterien müssen qualitative Studien gemäß Lincoln und Guba (1985) erfüllen und was sind die jeweiligen Pendants der Kriterien im quantitativen Ansatz?
8. Was ist gemäß den Kernkriterien zur Beurteilung qualitativer Forschung nach Steinke (1999) unter den Kriterien „Indikation“ und „Limitation“ zu verstehen?
9. Was stimmt?
   a) Aufgrund des wissenschaftstheoretischen Hintergrundes qualitativer Studien werden intersubjektiv nachvollziehbare Gütekriterien generell abgelehnt.
   b) Es existieren nur sehr wenige Kriterienkataloge zur Beurteilung qualitativer Studien.
   c) Nur wenn jedes einzelne Kriterium aus einem Kriterienkatalog qualitativer Gütekriterien erfüllt ist, handelt es sich um eine hochwertige qualitative Studie.
   d) Aufgrund des Prinzips der Offenheit spielt methodische Strenge als Qualitätskriterium in der qualitativen Forschung eine untergeordnete Rolle.
   e) Das Kriterium der intersubjektiven Nachvollziehbarkeit gilt nur im qualitativen Forschungsansatz, nicht im quantitativen.
   f) Das Kriterium der reflektierten Subjektivität ist spezifisch für qualitative Forschung.
10. Was versteht man im Kontext der Mixed-Methods-Forschung unter Inferenzqualität und in welche Unterkriterien teilt sie sich auf?

Übungsaufgaben

1. Beschaffen Sie sich folgenden Artikel:
   Li, Q. (2006). Cyberbullying in schools. A research of gender differences. *School Psychology International, 27*(2), 157–170.
   http://clubs.ucalgary.ca/~qinli/publication/cyberGender2005SPIJ.pdf
   Beurteilen Sie diese quantitative Studie hinsichtlich Konstruktvalidität, interner Validität, externer Validität und statistischer Validität anhand der in diesem Kapitel vorgestellten Listen von Gefährdungen der Validität gemäß der Campbell-Tradition.
2. Wählen Sie sich aus der Open-Access-Zeitschrift *Forum Qualitative Sozialforschung* (*FQS;* http://www.qualitative-research.net/) einen Beitrag zu einer qualitativen Studie aus und bewerten Sie dessen Qualität anhand der Checkliste von Letts et al. (2007). Reflektieren Sie neben der Qualität des Beitrags auch die Qualität der Checkliste – welche Kernkriterien zur Beurteilung qualitativer Forschung nach Steinke (1999) werden abgedeckt?
3. Laden Sie sich folgenden Artikel herunter:
   Humphries, N., Brugha, R. & McGee, H. (2009). Sending money home: a mixed-methods study of

remittances by migrant nurses in Ireland. *Human Resources for Health*, 7: 66. doi:10.1186/1478-4491-7-66
http://www.human-resources"=health.com/content/7/1/66
Beurteilen Sie diesen Artikel zu einer Mixed-Methods-Studie hinsichtlich seiner methodischen Strenge bzgl. des Designs und dessen Begründung sowie die Integration qualitativer und quantitativer Ergebnisse anhand der Hinweise von Creswell und Clark (2011) zur Begutachtung von Einreichungen beim *Journal of Mixed Methods Research.*

Weiterführende Literatur

Gütekriterien für quantitative Forschung:

- Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.
- Girden, E. R. & Kabacoff, R. (2010). *Evaluating research articles from start to finish.* Thousand Oaks: Sage.
- Keller, D. K. & Casadevall-Keller, M. L. (2010). *The Tao of research: A path to validity.* Thousand Oaks: Sage.
- Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.
- Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik: Ein Lehrbuch zur Psychologischen Methodenlehre.* Göttingen: Hogrefe.

Gütekriterien für qualitative Forschung:

- Flick, U. (2008). *Managing the quality of qualitative research.* London: Sage.
- Seale, C. (1999). *The quality of qualitative research.* London: Sage.
- Steinke, I (1999). *Kriterien qualitativer Forschung.* Ansätze zur Bewertung qualitativ-empirischer Sozialforschung. München: Juventa.

Gütekriterien für Mixed-Methods-Forschung:

- O'Cathain, A. (2010). Assessing the quality of mixed methods research: Toward a comprehensive framework. In A. Tashakkori & C. Teddlie (Eds.), *The Sage handbook of mixed methods in social and behavioral research* (2nd ed., pp. 531–555). Thousand Oaks: Sage.
- O'Cathain, A., Murphy, E., & Nicholl, J. (2008). The quality of mixed methods studies in health services research. *Journal of Health Services Research Policy, 13*(2), 92–98.
- Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating quantitative and qualitative approaches in the social and behavioral sciences.* Thousand Oaks: Sage.

Literatur

Ambach, W. (2012). *Experimentelle Psychophysiologie in Grenzgebieten.* Würzburg: Ergon.

Anderson, C. A. & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology, 78*(4), 772–790.

APA (American Psychological Association). (2009). Publication manual of the American Psychological Association (6th ed.). Washington: American Psychological Association.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology. Why do we need them? What might they be? Washington: American Psychological Association.

Barker, C. & Pistrang, N. (2005). Quality criteria under methodological pluralism: Implications for conducting and evaluating research. *American Journal of Community Psychology, 35(*3), 201–212.

Bortz, J., Lienert, G. A. & Boehnke, K. (2008). Verteilungsfreie Methoden in der Biostatistik (3. Aufl.). Berlin: Springer.

Boykoff, M. T. & Boykoff, J. M. (2004). Balance as bias: Global warming and the US prestige press. *Global Environmental Change, 14*(2), 125–136.

Breuer, F. & Reichertz, J. (2001). Standards of social research. FQS Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 2(3), Art. 24. Retrieved November 29, 2013, from http://www.qualitative-research.net/index.php/fqs/article/view/919/2008.

Bryman, A., Becker, S., & Sempik, J. (2008). Quality criteria for quantitative, qualitative and mixed methods research: A view from social policy. *International Journal of Social Research Methodology, 11*(4), 261–276.

Campbell, D. T. (1957). factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–311.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Carpenter, J. C. (2012). *First sight: ESP and parapsychology in everyday life.* Lanham: Rowman & Littlefield.

Cook, T. D. & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), *Handbook of industrial and organizational research* (pp. 223–326). Chicago: Rand McNally.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cooper, H. (2010). *Reporting research in psychology: How to meet journal article reporting standards.* Washington: American Psychological Association.

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks: Sage.

Denzin, N. K. & Lincoln, Y. S. (Eds.). (1994). *The SAGE handbook of qualitative research* (1st ed.). Thousand Oaks: Sage.

Denzin, N. K. & Lincoln, Y. S. (Eds.). (2000). *The SAGE handbook of qualitative research* (2nd ed.). Thousand Oaks: Sage.

Denzin, N. K. & Lincoln, Y. S. (Eds.). (2005). *The SAGE handbook of qualitative research* (3rd ed.). Thousand Oaks: Sage.

Denzin, N. K. & Lincoln, Y. S. (Eds.). (2012). *The SAGE handbook of qualitative research* (4th ed.). Thousand Oaks: Sage.

Denzin, N. K., Lincoln, Y. S., & Smith, L. T. (2008). Handbook of Critical and Indigenous Methodologies. Thousand Oaks: Sage.

Elliott, R., Fischer, C. T., & Rennie, D. L. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology, 38*(3), 215–229.

Emden, C. & Sandelowski, M. (1998). The good, the bad and the relative. Part One: Conceptions of goodness in qualitative research. *International Journal of Nursing Practice, 4*(4), 206–212.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science 587*(1), 49–68.

Flick, U. (2004). *Triangulation. Eine Einführung.* Wiesbaden: VS Verlag.

Flick, U. (2008). *Managing the quality of qualitative research.* Thousand Oaks: Sage.

Geiger, W. & Kotte, W. (2007). Handbuch Qualität: Grundlagen und Elemente des Qualitätsmanagements: Systeme – Perspektiven (5.Aufl.). Wiesbaden: Vieweg & Teubner.

Girden, E. R. & Kabacoff, R. (2010). *Evaluating research articles from start to finish.* Thousand Oaks: Sage.

Grunenberg, H. (2007). Empirische Befunde zur Qualität qualitativer Sozialforschung. Resultate einer Analyse von Zeitschriftenartikeln. In U. Kuckartz, H. Grunenberg & T. Dresing (Hrsg.), Qualitative Datenanalyse: computergestützt (S. 210–226): VS Verlag.

Haldeman, D. C. (1999). The pseudo-science of sexual orientation conversion therapy. *Angles, 4*(1), 1–4.

Haldeman, D. C. (2002). Gay rights, patient rights: The implications of sexual orientation conversion therapy. *Professional Psychology: Research and Practice, 33*(3), 260–264.

Hannes, K. (2011). Chapter 4: Critical appraisal of qualitative research. In: Noyes J, Booth A, Hannes K, Harden A, Harris J, Lewin S, Lockwood C (Eds.), Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions. Version 1 (updated August 2011). Cochrane Collaboration Qualitative Methods Group, 2011. Retrieved November 29, 2013, from URL http://cqrmg.cochrane.org/supplemental-handbook-guidance.

Hellferich, C. (2005). Die Qualität qualitativer Daten. Manual für die Durchführung qualitativer Interviews (2. Aufl.). Wiesbaden: VS Verlag.

Ickinger, J. (2006). Methodisches Vorgehen bei UFO-Falluntersuchungen. *Zeitschrift für Anomalistik, 6,* 116–162.

Ilg, S. & Boothe, B. (2010). Qualitative Forschung im psychologischen Feld: Was ist eine gute Publikation? FQS Forum Qualitative Sozialforschung/Forum: *Qualitative Social Research, 11*(2), Art. 25. Retrieved November 29, 2013, from http://www.qualitative-research.net/index.php/fqs/article/view/1371/2975.

Keller, D. K. & Casadevall-Keller, M. L. (2010). The Tao of research: A path to validity. Thousand Oaks: Sage.

Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy, 45*(3), 214–222.

Lamnek, S. (2010). Qualitative Sozialforschung: Lehrbuch. Mit Online-Materialien (5.Aufl.). Weinheim: Beltz.

Letts, L., Wilkins, S., Law, M., Stewart, D., Bosch, J., & Westmorland, M. (2007). Critical Review Form – Qualitative Studies (Version 2.0). Retrieved November 4th, 2013 from http://www.srs-mcmaster.ca/Portals/20/pdf/ebp/qualreview_version2.0.pdf.

Lilienfeld, S. O., Lynn, S. J., & Lohr, J. M. (Eds.). (2003). *Science and pseudoscience in clinical psychology: Concluding thoughts and constructive remedies.* New York: Guilford.

Lilienfeld, S. O., Ruscio, J., & Lynn, S. J. (Eds.). (2008). *Navigating the mindfield: A user's guide to distinguishing science from pseudoscience in mental health.* Amherst: Prometheus.

Lincoln, Y. S. & Denzin, N. K. (Eds.). (2003). *Turning Points in Qualitative Research: Rifts, Ruptures and Revolutions in Interpretive Inquiry.* Walnut Creek: Altamira.

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic Inquiry.* Newbury Park: Sage.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542.

Malterud, K. (2001). Qualitative research: Standards, challenges, and guidelines. *The Lancet, 358*(9280), 483–488.

Mayring, P. (2002). *Einführung in die qualitative Sozialforschung: Eine Anleitung zu qualitativem Denken.* Weinheim: Beltz.

Meyer, C. & Reiter, S. (2004). Impfgegner und Impfskeptiker. Geschichte, Hintergründe, Thesen, Umgang. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz, 47*(12), 1182–1188.

Meyrick, J. (2006). What is good qualitative research? A first step towards a comprehensive approach to judging rigour/quality. *Journal of Health Psychology, 11*(5), 799–808.

Noyes, J., Popay, J., Pearson, A., Hannes, K., & Booth, A. (2008). Chapter 20: Qualitative research and cochrane reviews. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions*. Chichester: Wiley. Retrieved December 6, 2013, from http://handbook.cochrane.org/.

O'Cathain, A. (2010). Assessing the quality of mixed methods research: Toward a comprehensive framework. In A. Tashakkori & C. Teddlie (Eds.), *The Sage handbook of mixed methods in social and behavioral researc*h (2nd ed., pp. 531–555). Thousand Oaks: Sage.

O'Cathain, A., Murphy, E., & Nicholl, J. (2008). The quality of mixed methods studies in health services research. *Journal of Health Services Research Policy, 13*(2), 92–98.

Olatunji, B. O., Parker, L. M., & Lohr, J. M. (2005). Pseudoscience in contemporary psychology: Professional issues and implications. *The Scientific Review of Mental Health Practice, 4*(2), 19–36.

Parker, I. (2004). Criteria for qualitative research in psychology. *Qualitative Research in Psychology, 1*(2), 95–106.

Rahmstorf, S. (2007). Alles nur Klimahysterie? *Universitas, 9,* 895–913.

Reichardt, C. S. (2011). Criticisms of and an alternative to the shadish, cook, and campbell validity typology. New directions for evaluation. Special issue: Advancing Validity in Outcome Evaluation. *Theory and Practice, 2011*(130), 43–53.

Reichertz, J. (2000). Zur Gültigkeit von Qualitativer Sozialforschung. FQS Forum Qualitative Sozialforschung/Forum: *Qualitative Social Research, 1*(2), Art. 32. Abgerufen 6. Dezember, 2013, unter http://www.qualitative-research.net/index.php/fqs/article/view/1101/2427.

Reid, A. & Gough, S. (2000). Guidelines for reporting and evaluating qualitative research: What are the alternatives? *Environmental Education Research, 6*(1), 59–91.

Roethlisberger, F. J., Dickson, W. J., & Wright, H. A. (1939). Management and the worker: An account of a research program conducted by the Western Electric Company, Hawthorne Works, Chicago. Havard: Harvard University Press.

Sales, B. D. & Folkman, S. (Eds.). (2000). Ethics in research with human participants. Washington: American Psychological Association.

Schou, L., Høstrup, H., Lyngsø, E. E., Larsen, S., & Poulsen, I. (2012). Validation of a new assessment tool for qualitative research articles. *Journal of Advanced Nursing, 68*(9), 2086–2094.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Spencer, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence.* Retrieved December 6, 2013, from http://www.civilservice.gov.uk/wp-content/uploads/2011/09/a_quality_framework_tcm6-38740.pdf.

Steinke, I. (1999). *Kriterien qualitativer Forschung.* München: Juventa.

Steinke, I. (2009). Die Güte qualitativer Marktforschung. In R. Buber & H. Holzmüller (Hrsg.), Qualitative Marktforschung. Konzepte – Methoden – Analysen (2.Aufl.,S. 261–289). Wiesbaden: Gabler.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136*(4), 471–485.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating quantitative and qualitative approaches in the social and behavioral sciences.* Thousand Oaks: Sage.

Tracy, S. J. (2010). Qualitative quality: Eight „Big-Tent" criteria for excellent qualitative research. *Qualitative Inquiry 16*(10), 837–851.

Truzzi, M. (1996). Pseudoscience. In G. Stein (Ed.), *The Encyclopedia of the Paranormal* (pp. 560–575). Amherst: Prometheus.

Wendt, A. & Duvall, R. (2008). Sovereignty and the UFO. *Political Theory, 36*(4), 607–633.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik: Ein Lehrbuch zur Psychologischen Methodenlehre.* Göttingen: Hogrefe.

Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research, 11*(4), 522–537.

Wunder, E. (2001). Die Wahrnehmung der Struktur der deutschsprachigen UFO-Szene: Eine multidimensionale Skalierung von Expertenurteilen. *Zeitschrift für Anomalistik, 1*(1), 75–101.

Forschungs- und Wissenschaftsethik

Nicola Döring, Jürgen Bortz

4.1 Forschungsethik – 123

4.2 Wissenschaftsethik – 132

Literatur – 138

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_4, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Die zentralen ethischen Richtlinien zum Umgang mit Untersuchungspersonen in der human- und sozialwissenschaftlichen Forschung kennen.
- Die wichtigsten Regeln guter wissenschaftlicher Praxis erläutern können.
- Eigene Forschungsaktivitäten an Prinzipien der Forschungs- und Wissenschaftsethik ausrichten können.
- Vorliegende Studien hinsichtlich möglicher ethischer Probleme bewerten können.
- Wissen, wie man forschungs- bzw. wissenschaftsethische Fragen selbst zum Gegenstand empirischer Forschung machen kann.

Das Streben nach wissenschaftlicher Erkenntnis besitzt in unserer Gesellschaft einen sehr hohen Wert. Die **Wissenschaftsfreiheit**, konkretisiert als Freiheit von Forschung und Lehre, hat in Deutschland den Stellenwert eines Grundrechtes gemäß **Artikel 5 (Absatz 3) des Grundgesetzes**. Auch Österreich und die Schweiz haben die Wissenschaftsfreiheit in ihren Verfassungen verankert. Analog zu Pressefreiheit und Kunstfreiheit soll die Wissenschaftsfreiheit in einer demokratischen Gesellschaft die in diesen Bereichen Tätigen vor Fremdbestimmung, Zensur oder gar einer Instrumentalisierung durch Machthaber schützen, wie sie in Diktaturen nicht unüblich (gewesen) ist.

Wissenschaftliche Erkenntnisinteressen müssen jedoch trotz ihres grundsätzlich hohen Stellenwertes immer mit anderen gesellschaftlichen Werten und Normen sowie Gesetzen abgewogen werden. Von einzelnen Wissenschaftlerinnen und Wissenschaftlern sowie von Forschungsinstitutionen wird verlangt, dass sie verantwortungsvoll und nach ethischen Prinzipien handeln, dies betrifft vor allem zwei Bereiche, die in diesem Kapitel separat behandelt werden:

1. **Forschungsethik:** Untersuchungsteilnehmende sollen in ihrer Würde und ihrem Wohlergehen geschützt werden. Umgekehrt: Weder Menschen noch Tiere sollen für wissenschaftliche Zwecke im Zuge empirischer Studien in unethischer Weise missbraucht oder geschädigt werden.
2. **Wissenschaftsethik:** Wissenschaftlicher Erkenntnisgewinn soll den Regeln guter wissenschaftlicher Praxis folgen und überprüfbar sein. Umgekehrt: Forschungsergebnisse sollen nicht unkritisch behauptet, erfunden oder gestohlen werden.

In grundlegenden ethischen Fragen – etwa der Wahrung der Menschen- und Persönlichkeitsrechte von Untersuchungsteilnehmenden oder der Verurteilung von Wissenschaftsfälschungen und Plagiaten – herrscht heute international und interdisziplinär Einigkeit. Ethische Detailfragen – z. B. zur Legitimität verdeckter Feldbeobachtung in unterschiedlichen Online-Foren oder zur fairen Festlegung der Autorenreihenfolge bei wissenschaftlichen Publikationen – werden jedoch nicht selten kontrovers diskutiert und müssen kontextspezifisch entschieden werden.

Generell ist in den letzten Jahren und Jahrzehnten die **Bedeutung der Forschungs- und Wissenschaftsethik** stetig gewachsen. Dies hat unterschiedliche Gründe:

- **Technisierung:** Die rasante technische Entwicklung eröffnet ständig neue Forschungsmöglichkeiten, die mit kaum überschaubaren Chancen und Risiken für Mensch und Umwelt behaftet sind und deren ethische Beurteilung ausgesprochen komplex ist. Man denke etwa an Embryonen-, Stammzell-, Gen-, Nano- oder Weltraumforschung. So beschränkt z. B. das ethisch begründete deutsche Embryonenschutzgesetz die Wissenschaftsfreiheit in der Embryonenforschung.
- **Globalisierung:** Forschungsprojekte sind heute oftmals international ausgerichtet. Dementsprechend müssen ethische Richtlinien kulturübergreifend diskutiert werden. Besonders relevant ist die Sicherung von Forschungsethik in bislang geringer entwickelten Ländern, weil Armut z. B. das Risiko steigert, dass die Rechte von Untersuchungsteilnehmenden unzureichend gewahrt werden (z. B. ethisch problematische Durchführung von Medikamententests in Afrika, die Krankheiten betreffen, die vor allem in den reichen Industrienationen auftreten).
- **Emanzipation:** Im Zuge gesellschaftlicher Emanzipationsprozesse wird Wissenschaft zunehmend kritisch hinterfragt im Hinblick auf ungerechtfertigte Autoritäts- und Machtansprüche. So wehren sich z. B. gesellschaftliche Minderheiten verstärkt gegen die Definitionsmacht der Wissenschaft, wenn sie sich nicht angemessen repräsentiert fühlen. Insbesondere wird stärker problematisiert, inwiefern die soziale Position der Forschenden – oft weiße männliche Angehörige der bürgerlichen Mittelschicht – zu systematischen Verzerrungen bei der Themenwahl und Ergebnisinterpretation führt. Der Ethikkodex der Australian Association for Research in Education (AARE; http://www.aare.edu.au/ethics/ethcfull.htm) geht z. B. ausdrücklich auf die Problematik ein, dass Forschung zu benachteiligten sozialen Gruppen, die mit guten Absichten erfolgt, dennoch Vorurteile und Stereotype verstärken kann und empfiehlt deswegen, offizielle Vertreter betroffener Gruppen beratend in die Planung entsprechender Studien einzubeziehen. In den Qualitätsstandards der Evaluationsforschung

(► Abschn. 18.1.9) ist festgelegt, dass die Interessen und Sichtweisen der Beteiligten und Betroffenen einer evaluierten Maßnahme von den Forschenden aufzugreifen sind.
- **Wissenschaftsskandale:** Nicht zuletzt sorgen Wissenschaftsskandale rund um Fälschungen und Plagiate immer wieder dafür, dass bestehende Maßnahmen zur Verhinderung unethischen Verhaltens von Forschenden überprüft und optimiert werden.

Die wachsende Bedeutung von Forschungs- und Wissenschaftsethik spiegelt sich darin wider,
- dass auf internationaler wie nationaler sowie institutioneller Ebene immer mehr Expertengremien an immer neuen ethischen Richtlinien und Kodizes arbeiten,
- dass immer mehr Forschungsprojekte vor ihrer Durchführung formal von Ethikkommissionen überprüft werden,
- dass in Forschungsberichten immer ausführlicher auch ethische Aspekte diskutiert werden,
- dass in der akademischen Methodenausbildung die Forschungs- und Wissenschaftsethik immer größeren Raum einnimmt und nicht zuletzt,
- dass der Umgang mit forschungs- und wissenschaftsethischen Fragen seinerseits zu einem boomenden empirischen Forschungsgebiet geworden ist.

Relevante Forschungsfragen lauten z. B.: Wie bekannt sind geltende ethische Richtlinien bei unterschiedlichen Gruppen von Forschenden? Wie werden die Interessen und Rechte der Untersuchungsteilnehmenden in unterschiedlichen Ländern gewahrt? Wie verbreitet sind Plagiate in Studium und Wissenschaft? Einschlägige Fachzeitschriften zur Forschung über Forschungs- und Wissenschaftsethik sind z. B.:
- *Research Ethics* (Sage),
- *Journal of Empirical Research on Human Research Ethics* (University of California Press) oder
- *IRB: Ethics & Human Research* (The Hastings Center).

4.1 Forschungsethik

Die Abgrenzung zwischen einer Forschungsethik, die sich auf den Umgang mit Untersuchungsteilnehmenden bei der Datenerhebung und Datenanalyse konzentriert, und einer Wissenschaftsethik, die sich auf die Interpretation und Veröffentlichung wissenschaftlicher Erkenntnisse bezieht, ist nicht ganz trennscharf. Oft werden auch beide Bereiche unter der Bezeichnung „Forschungsethik" zusammengefasst. In der Praxis besteht der Unterschied jedoch darin, dass der ethisch verantwortungsvolle Umgang mit Untersuchungsteilnehmenden in Forschungsberichten ausdrücklich thematisiert und im Vorfeld oft durch Ethikkommissionen bestätigt werden muss. Demgegenüber wird in Studien nicht explizit diskutiert, ob und wie Regeln guter wissenschaftlicher Praxis – z. B. Selbstkritik, Förderung des wissenschaftlichen Nachwuchses oder faire Festlegung der Autorenreihenfolge – umgesetzt wurden. Forschungs- und Wissenschaftsethik betreffen das wissenschaftliche Qualitätskriterium der ethischen Strenge (► Abschn. 3.1.2).

Forschungsethik – Die Forschungsethik („research ethics", „ethics of research") umfasst alle ethischen Richtlinien, an denen sich Forschende bei ihrer Forschungstätigkeit – und zwar hier speziell bei der Datenerhebung und Datenanalyse im Rahmen empirischer Studien – orientieren sollen. Im Mittelpunkt stehen der verantwortungsvolle Umgang mit menschlichen und tierischen Untersuchungsteilnehmenden und ihr Schutz vor unnötigen oder unverhältnismäßigen Beeinträchtigungen durch den Forschungsprozess. Dass und wie der Forschungsethik in einer konkreten Studie Genüge getan wird, sollte im Vorfeld genau überlegt und im Forschungsbericht erläutert werden. Immer häufiger wird in verschiedenen human- und sozialwissenschaftlichen Disziplinen auch verlangt, dass geplante Studien mit menschlichen oder tierischen Probanden vor ihrer Durchführung durch eine Ethikkommission geprüft und genehmigt werden.

Im Folgenden werden der ethische Umgang mit Untersuchungspersonen (► Abschn. 4.1.1), die forschungsethischen Richtlinien einzelner Fachdisziplinen (► Abschn. 4.1.2) sowie Maßnahmen zur Einhaltung der Forschungsethik (► Abschn. 4.1.3) behandelt.

4.1.1 Ethischer Umgang mit Untersuchungspersonen

Ein ethisch verantwortungsvoller Umgang mit Untersuchungspersonen sowie mit gesammelten Dokumenten als Untersuchungsmaterial umfasst – gemäß nahezu allen Ethikrichtlinien der einzelnen sozial- und humanwissenschaftlichen Fachdisziplinen (► Abschn. 4.1.2) – vor allem drei Prinzipien (Sales & Folkman, 2000):
1. Freiwilligkeit und informierte Einwilligung,
2. Schutz vor Beeinträchtigung und Schädigung und
3. Anonymisierung und Vertraulichkeit der Daten.

Dahinter steht jeweils der **Schutz der Menschenwürde, der Menschenrechte, der Persönlichkeitsrechte sowie der Integrität der Untersuchungspersonen**. Regelungen für Tierversuche (z. B. rechtliche Bestimmungen zur artgerechten Haltung von Versuchstieren in Laboren) sind ein gesondertes Thema, das hier nicht vertieft wird (Akins, Panicker, & Cunningham, 2004).

Freiwilligkeit und informierte Einwilligung

Typischerweise sollten sich Untersuchungsteilnehmende freiwillig an wissenschaftlichen Studien beteiligen, nachdem sie über Zielsetzung und Ablauf der jeweiligen Studie gründlich und in verständlicher Weise aufgeklärt wurden und auf dieser Basis ihr ausdrückliches Einverständnis erklärt haben (informierte Einwilligung; auf Aufklärung basierende Einwilligung: „**informed consent**"). Ausnahmen von dieser Regel stellen Studien dar, die für die Untersuchungspersonen keinerlei Risiken bergen **(„no risk research")** oder allenfalls minimale, dem Alltag entsprechende Belastungen mit sich bringen **(„minimal risk research")**. In den beiden letztgenannten Fällen kann auf umfangreiche Aufklärung und informierte Einwilligung verzichtet werden.

Bei Studien, die mehr als minimale Risiken oder Belastungen bergen, ist die informierte Einwilligung dagegen notwendig **(„greater than minimal risk research")**. Die **Einverständniserklärung** wird in der Regel mündlich oder schriftlich abgegeben. Bei schriftlicher Form lässt man im Vorfeld der Untersuchung ein entsprechendes Formular („consent form") unterschreiben. Die Art der Aufklärung der Untersuchungsteilnehmenden sowie die Einwilligungen der Untersuchten sind von den Forschenden zu **dokumentieren**. Dabei wird zuweilen die mündliche Form der Einwilligung bevorzugt, weil schriftliche Einwilligungen wiederum sensible personalisierte Dokumente sind.

Laut **Ethikrichtlinien der Deutschen Gesellschaft für Psychologie** (DGPs) sollten beim Einholen der informierten Einwilligung die potenziellen Untersuchungsteilnehmerinnen und -teilnehmer über folgende Sachverhalte **aufgeklärt** werden (DGPs, 2004, Punkt 3d; http://www.dgps.de/dgps/aufgaben/ethikrl2004.pdf):

1. den Zweck der Forschung;
2. die erwartete Dauer der Untersuchung und das Vorgehen;
3. ihr Recht darauf, die Teilnahme abzulehnen oder sie zu beenden, auch wenn die Untersuchung schon begonnen hat;
4. absehbare Konsequenzen der Nicht-Teilnahme oder der vorzeitigen Beendigung der Teilnahme;
5. absehbare Faktoren, von denen man vernünftigerweise erwarten kann, dass sie die Teilnahmebereitschaft beeinflussen, wie z. B. potenzielle Risiken, Unbehagen oder mögliche anderweitige negative Auswirkungen, die über alltägliche Befindlichkeitsschwankungen hinausgehen;
6. den voraussichtlichen Erkenntnisgewinn durch die Forschungsarbeit;
7. die Gewährleistung von Vertraulichkeit und Anonymität sowie ggf. deren Grenzen;
8. eine mögliche Aufwandsentschädigung und/oder Bonus für die Teilnahme und
9. an wen sie sich mit Fragen zum Forschungsvorhaben und zu ihren Rechten als Forschungsteilnehmende wenden können.

Folgende Sondersituationen sind bei der Aufklärung der Untersuchungspersonen zu beachten:

- Studien, bei denen **innovative Maßnahmen** oder Untersuchungsmethoden eingesetzt werden: Hier müssen die Untersuchungspersonen besonders sorgfältig über Chancen und Risiken der Erprobung neuartiger Behandlungsmethoden aufgeklärt werden. Bei Untersuchungsverläufen, die für die Probanden im Vorfeld schwer einschätzbar sind, spielt zudem der Hinweis auf die Möglichkeit, die Teilnahme bei Bedarf jederzeit abbrechen zu können, aus ethischer Sicht eine wichtige Rolle.
- Studien, bei denen **Ton- und/oder Bildmaterial** erzeugt wird: Hier müssen z. B. Modalitäten des Umgangs mit Audio-/Videoaufzeichnungen der Untersuchungspersonen genau festgelegt werden. Aus forschungsethischer Sicht (Schutz der Privatsphäre) ist die Archivierung dieses identifizierbaren Rohdatenmaterials problematisch, aus Sicht der Wissenschaftsethik (Schutz vor Wissenschaftsfälschungen; ► Abschn. 4.2.2) jedoch notwendig. Hier gilt es, jeweils kontextspezifisch sinnvolle Lösungen zu finden.
- Studien, bei denen zwischen Forschenden und potenziellen Untersuchungsteilnehmenden ein **Abhängigkeitsverhältnis** besteht (z. B. Untersuchungen an den eigenen Patienten, Mitarbeitern, Studierenden): Hier muss im Sinne der Freiwilligkeit sichergestellt und zugesichert werden, dass bei Nicht-Teilnahme keine nachteiligen Konsequenzen entstehen oder zu befürchten sind. Eine Ausnahme stellt die Ableistung von Versuchspersonenstunden dar, die für Studierende z. B. in der Psychologie verpflichtend ist, wobei eine Wahlmöglichkeit zwischen verschiedenen Studien gegeben sein sollte.
- Studien, bei denen die **Einwilligungsfähigkeit** („consent capability") potenzieller Untersuchungspersonen eingeschränkt oder nicht vorhanden ist, etwa aufgrund des Lebensalters (z. B. Säuglingsforschung), geringer Bildung, verminderter Intelligenzleistung oder psychiatrischer Erkrankungen (z. B. Schizophrenie, Demenz). In diesen Fällen muss die Einwilligungsfähigkeit eingeschätzt werden, zuweilen kann alternativ auf die Einwilligung von rechtlichen Vertretern zurückgegriffen werden (z. B. bei minderjährigen Untersuchungspersonen von Erziehungsberechtigen; bei dementen Personen von Angehörigen etc.). Eine Um-

frage unter Mitgliedern von Ethikkommissionen zeigte z. B., dass die Experten bei depressiven Patienten eine verminderte Einwilligungsfähigkeit vermuteten (Tait, Chibnall, Iltis, Wall, & Deshields, 2011).

Im Zuge der Aufklärung über die Studie, die meist am Anfang **(Briefing)** und/oder am Ende **(Debriefing)** einer Untersuchung stattfindet, sollte den Teilnehmenden die Möglichkeit für Rückfragen gegeben werden (z. B. mündlich im Labor; per E-Mail bei Online-Untersuchungen). Auch wird empfohlen, den Untersuchungsteilnehmenden einen **Aufklärungs- bzw. Informationsbogen** zur Studie mitzugeben, der Kontaktdaten der Forschenden, Ziele des Forschungsprojekts, Hinweise zur Anonymität und zum Datenschutz zu entnehmen sind. Im Falle von belastenden Themen (z. B. Interviewstudie zu traumatischen Erfahrungen) sollte der Aufklärungsbogen auch auf Unterstützungsangebote (z. B. Selbsthilfegruppen, Beratungsstellen, Hotlines etc.) verweisen.

Die Freiwilligkeit der Teilnahme ist forschungsethisch geboten, schränkt jedoch die Generalisierbarkeit von Befunden ein. Denn über Personen, die sich nicht freiwillig an wissenschaftlichen Studien beteiligen, erfahren wir nichts **(„participation bias“)**. Dementsprechend sind Studien zur Motivation freiwilliger Untersuchungsteilnehmender sowie zur Wirksamkeit unterschiedlicher – ethisch vertretbarer – Methoden der Steigerung der Teilnahmebereitschaft sehr nützlich.

Freiwilligkeit, vorherige Aufklärung und ausdrückliche Einwilligung sind nicht unter allen Umständen möglich oder nötig. Folgende **Ausnahmen und Besonderheiten** sind zu berücksichtigen: 1. Legitimer Verzicht auf informierte Einwilligung, 2. legitime Täuschung der Untersuchungsteilnehmenden und 3. angemessene Anreize.

1. **Legitimer Verzicht auf informierte Einwilligung bei risikolosen Studien:**
   So genannte **non-reaktive Methoden** der Datenerhebung, bei denen kein direkter Kontakt zwischen Forschenden und Untersuchten stattfindet, sind in vielen Fällen ohne ausdrückliche informierte Einwilligung der Betroffenen zulässig. Dies betrifft z. B. die **Verhaltensbeobachtung an öffentlichen Orten** (z. B. Auszählung der Häufigkeit von Handynutzung am Steuer durch Verkehrsbeobachtung auf der Straße; ► Abschn. 10.1.3) oder das Sammeln vorfindbarer **öffentlicher Dokumente bzw. von Archivmaterial** (z. B. Auswertung von Zeitungsartikeln; ► Abschn. 10.6). In Zeiten des Internet haben sich die Möglichkeiten non-reaktiver Online-Datenerhebung per Beobachtung und Dokumentenanalyse exponentiell vervielfacht, gleichzeitig müssen jedoch die ethischen Implikationen kontextspezifisch betrachtet und diskutiert werden: Unter welchen Bedingungen dürfen Forschende verdeckt in Online-Communitys beobachten? Unter welchen Bedingungen dürfen öffentliche Online-Dokumente wie z. B. Diskussionsbeiträge in Online-Foren oder Profile auf Social-Networking-Sites ohne Wissen und Einverständnis ihrer jeweiligen Autorinnen und Autoren zum Gegenstand wissenschaftlicher Studien gemacht werden? Die **Ethik der Online-Forschung** („ethics of online research“) ist deswegen zu einem wachsenden neuen Forschungsfeld geworden (Buchanan & Williams, 2010; McKee & Porter, 2009; Nosek, Banaji, & Greenwald, 2002). Auch für die Berufspraxis hat die Online-Kommunikation zahlreiche ethische Implikationen, etwa hinsichtlich der verantwortungsvollen Durchführung von Online-Beratung oder der Frage, ob Psychotherapeuten ihre Patienten „googeln“ sollten (Kaslow, Patterson, & Gottlieb, 2011). Auch bei **reaktiven Datenerhebungsverfahren** wie Befragungsmethoden kann ein Vorgehen ohne ausdrückliche informierte Einwilligung legitim sein, etwa wenn ein **anonymer Fragebogen** verteilt wird.
   Gerechtfertigt ist ein Verzicht auf ausdrückliche informierte Einwilligung bei den genannten Datenerhebungsmethoden, wenn für die Untersuchungsteilnehmenden durch die Datenerhebung keine besonderen Belastungen entstehen und wenn die Dokumentation, Analyse und Publikation der Daten für die Betroffenen keine negativen Konsequenzen hat. In Kontexten, in denen routinemäßige anonyme Befragungen durchgeführt werden (z. B. Lehrveranstaltungsevaluation an Hochschulen, Mitarbeiterbefragungen in Unternehmen) ist informierte Einwilligung ebenfalls in der Regel obsolet.
2. **Legitime Täuschung der Untersuchungsteilnehmenden über das Thema der Studie zwecks unverzerrter Ergebnisse**:
   In gesonderten Ausnahmefällen ist es notwendig und ethisch legitim, Untersuchungsteilnehmende über die eigentliche Zielsetzung der Studie im Unklaren zu lassen (**passive Täuschung**, „passive deception“) oder ihnen sogar aktiv falsche Informationen zu geben (**aktive Täuschung**, „active deception“), um unverzerrte Ergebnisse zu erhalten. Will man z. B. natürliches Flirtverhalten im Labor beobachten, ist es nicht sinnvoll, den Probanden vorab zu sagen, man wolle gleich ihr Flirtverhalten auswerten und werde ihnen dazu einen Flirtpartner zuordnen. Dies würde vermutlich zu sehr untypischen Verhaltensmustern führen. Hier wird stattdessen eine Coverstory genutzt, dergemäß man den Versuchspersonen z. B. mitteilt (aktive Täuschung), sie nähmen an einer Studie zum Leseverhalten teil und sie bittet, im Warteraum Platz zu nehmen, bevor der Versuch startet. Im angeblichen

Bekannte sozialwissenschaftliche Studien mit forschungsethischen Problemen
Beispiele für Täuschungen der Untersuchungspersonen

Stanley Milgram (1963, 1974): Milgram Experiment on Obedience to Authority Figures

Der Sozialpsychologe Stanley Milgram führte ab 1961 an der Yale Universität eine Serie von Experimenten zum Gehorsam gegenüber Autoritäten durch. Milgrams Forschungsziel war es zu überprüfen, ob „ganz normale" Bürger tatsächlich aus Gehorsam gegenüber Autoritätspersonen in extremer Weise unethisch handeln. Motiviert war die Forschungsfrage durch die Suche nach psychologischen Erklärungen für die breite Mittäterschaft im Nationalsozialismus.
Im Rahmen eines angeblichen Lernexperimentes wurde den Versuchspersonen vermeintlich die Rolle des Lehrers übertragen, der einen Schüler in einem anderen Raum für Fehler mit Elektroschocks wachsender Stärke bestrafen sollte (15 Volt bis 450 Volt). Den Versuchspersonen wurde mitgeteilt, angebliches Ziel der Studie sei es, den Effekt von Bestrafungen auf das Lernverhalten zu untersuchen. Mit steigender Stärke der Stromstöße reagierte der angebliche Schüler im Nachbarraum für die Versuchsperson wahrnehmbar mit immer heftigeren Reaktionen (Stöhnen, Schreien, an die Wand schlagen) bis bei sehr starken Stromstößen schließlich Stille eintrat. Wenn die Versuchspersonen angesichts der steigenden Stromstärken und heftigeren Reaktionen des Schülers mit Unsicherheit und Zögern reagierten, wurden sie vom Versuchsleiter in standardisierter Weise aufgefordert, den Versuch fortzusetzen. Das erstaunliche und erschütternde Hauptergebnis bestand darin, dass freiwillige Versuchspersonen in einem wissenschaftlichen Experiment, das sie ohne Negativfolgen jederzeit hätten abbrechen können, mehrheitlich bereit waren, einem anderen Menschen durch Stromstöße starke Schmerzen zuzufügen und ihn schließlich zu töten. Das inzwischen vielfach replizierte Milgram-Experiment gehört zu den klassischen Studien der Sozialpsychologie und ist auch in das Alltagswissen eingegangen. Ein Forschungsüberblick zeigte, dass in verschiedenen Studien jeweils rund 60 % der männlichen ebenso wie der weiblichen Probanden tödliche Stromstöße verabreichten (Blass, 1999; diverse Videos von Replikationen sind auf der Videoplattform YouTube zu finden).
Ethisch kritisiert wurde, dass im Rahmen des Täuschungsexperimentes die Probanden starkem emotionalem Stress ausgesetzt waren. Dies betrifft nicht nur die Phase der Versuchsdurchführung, die einen Gewissenskonflikt erzeugte, sondern auch die rückblickende Verarbeitung und Reflexion des eigenen unmoralischen Verhaltens. Zudem wurde kritisiert, dass ohne Einwilligung der Probanden Videoaufzeichnungen der Versuchsdurchgänge als Lehrmaterial veröffentlicht wurden, wobei die Versuchspersonen erkennbar sind und inhaltlich mit Verbrechern des Naziregimes in Verbindung gebracht werden.

Laud Humphreys (1970): Tearoom Trade

Im Rahmen seiner Doktorarbeit an der Universität Washington untersuchte der Soziologe Laud Humphreys die Situation von Männern, die in öffentlichen Toiletten („tearoom") Sex mit anderen Männern suchten, und dabei angesichts der damaligen Kriminalisierung von Homosexualität in den USA von polizeilicher Verfolgung und Verhaftung bedroht waren. Humphreys' Forschungsziel war es, die soziale Situation einer stigmatisierten gesellschaftlichen Gruppe zu beschreiben.
In der Rolle eines Homosexuellen führte Humphreys verdeckte teilnehmende Feldbeobachtungen auf öffentlichen Toiletten durch (► Abschn. 10.1.2), identifizierte rund 50 der beteiligten Männer über ihre Autokennzeichen und interviewte diese ein Jahr später in der Rolle eines Marktforschers persönlich in ihrem häuslichen Umfeld. Hauptergebnis seiner Studie war, dass die von ihm untersuchten Männer nicht dem Klischee einer gesellschaftlichen „Randgruppe" entsprachen, sondern mehrheitlich verheiratet, gut gebildet, beruflich erfolgreich und sozial anerkannt waren. Die Publikation seiner Doktorarbeit wurde mit dem „C. Wright Mills Award of the Society for the Study of Social Problems" ausgezeichnet. Humphreys gilt als Pionier der Homosexualitätsforschung, sein eigenes Outing erfolgte 1974 auf dem amerikanischen Soziologenkongress.
Als ethisch problematisch wurde kritisiert, dass Humphreys die unfreiwilligen Untersuchungsteilnehmer zweimal aktiv täuschte, indem er eine falsche Rolle spielte, dass er durch die Nachforschung von Adressen über Autokennzeichen die Privatsphäre verletzte und dass er in seiner Publikation durch die Beschreibung der Lebensumstände der Untersuchungspersonen deren Identifizierung riskierte. Im soziologischen Institut der Universität Washington kam es zu intensiven Auseinandersetzungen um die Arbeit. Der Entzug des Doktortitels wurde gefordert, aber letztlich nicht umgesetzt. Der öffentlich stark wahrgenommene Vorfall führte letztlich zur Auflösung des Instituts (Allen, 1997; Galliher, Brekhus, & Keys, 2004).

Warteraum findet dann jedoch das eigentliche Experiment statt, indem das Interaktions- und Flirtverhalten mit anderen scheinbar zufällig dort wartenden Personen – in Wirklichkeit Mitwisser des Versuchsleiters – beobachtet wird (für ein entsprechendes Studienbeispiel ► Abschn. 10.1.1).
Eine Täuschung der Probanden über die Zielsetzung der Studie widerspricht dem Prinzip der bewussten Einwilligung und könnte zudem zu einem Negativ-Image der Forschung in der Öffentlichkeit beitragen. Täuschungsstudien sind deswegen nur in Ausnahmefällen und mit guter Begründung ihres besonderen wissenschaftlichen Wertes und ihrer methodischen Alternativlosigkeit einerseits sowie des Ausschlusses von Nachteilen für die Probanden andererseits durchzuführen (für bekannte, aber ethisch aufgrund des Einsatzes von Täuschung kritisierte Studien ► Bekannte sozialwissenschaftliche Studien mit forschungsethischen Problemen). Bei Täuschungsstudien ist die Aufklärung aller Untersuchungspersonen am Ende des Versuchs nachzuholen (**„debriefing"**). Es muss ihnen dabei die Gelegenheit gegeben werden, rückwirkend ihr Einverständnis zu entziehen und die Verwendung ihrer Daten zu untersagen.

3. **Angemessene Anreize**:
   Um Untersuchungsteilnehmende zu rekrutieren, werden oft Aufwandsentschädigungen und Anreize („incentives“) geboten, etwa Gutscheine, Gewinnverlosungen oder kleine Geldbeträge. Unverhältnismäßig große Anreize sind aus ethischer Sicht zu vermeiden, wenn diese dazu führen könnten, dass bestimmte Personen(gruppen) nur aus finanziellen Interessen oder aus einer finanziellen Notlage heraus an der Studie teilnehmen und somit keine echte Freiwilligkeit vorliegt. Empirische Studien zur Incentivierung befassen sich oft damit, welche Anreize die Teilnahmebereitschaft in welcher Weise beeinflussen bzw. steigern. Ethisch relevant ist aber auch die Frage, ob und wie Untersuchungspersonen in unterschiedlichen Studien durch Entlohnungen überhaupt in angemessener Weise für ihren tatsächlichen Aufwand (Zeit, Anreisekosten etc.) entschädigt werden (Ripley, Macrina, Markowitz, & Gennings, 2010). Aus Sicht der Forschenden besteht die Herausforderung darin, Finanzierungsquellen für die Anreize zu erschließen.

Schutz vor Beeinträchtigung und Schädigung

Generell sollten Untersuchungsteilnehmende durch den Forschungsprozess keine besonderen physischen oder psychischen Beeinträchtigungen oder gar irreversible Schädigungen erleiden. Als Vergleichsmaßstab dienen die **üblichen Befindlichkeitsschwankungen im Alltag**. Demgemäß wäre es also ethisch zumutbar, Studienteilnehmende zu problematischen Themen (z. B. Krankheitsgeschichte, Arbeitslosigkeit, Beziehungsprobleme) zu interviewen, auch wenn sie dadurch kurzfristig in eine negative Stimmungslage versetzt werden. Doch Vergleichbares geschieht eben auch durch Alltagsgespräche, in denen immer wieder unangenehme Themen aufkommen, weshalb entsprechende Studien dann als Minimal-Risk-Forschung einzuordnen sind.

Anders sieht es aus, wenn bestimmte Personengruppen mit potenziell sehr belastenden Befragungsthemen konfrontiert werden. So wäre es ethisch fragwürdig, Patienten mit Traumata unvorbereitet mit Untersuchungsmaterial zu konfrontieren, das eine Re-Traumatisierung auslösen könnte. Es sollte also z. B. nicht leichtfertig ein Fragebogen zu Missbrauchserfahrungen an Betroffene verteilt werden. Hier wären aus forschungsethischer Sicht entsprechende Vorkehrungen zu treffen, um nur geeignete und stabile Personen zur Untersuchung zuzulassen, das Untersuchungsmaterial zu entschärfen, informierte Einwilligung der Untersuchten einzuholen und im Falle auftretender akuter Belastungen geeignete Unterstützungsangebote (z. B. telefonische Hotline) bereitzustellen (zu ethischen Aspekten der Erforschung von Suizidalität s. z. B. Lakeman & FitzGerald, 2009). Bei Interventions- und Evaluationsstudien sowie Experimenten ist mit besonders starken Beeinträchtigungen zu rechnen, die nicht immer vorher absehbar sind (► Unvorhergesehen starke Beeinträchtigung der Probanden).

In Studien, in denen bestimmte Belastungen und Risiken nicht auszuschließen sind (z. B. Medikamententests in der **Medizin**), müssen zumindest geeignete Maßnahmen zu ihrer Kontrolle ergriffen und die Untersuchungspersonen detailliert und in verständlicher Form aufgeklärt werden. Während in der medizinischen Forschung durch invasive Methoden (z. B. Medikamenteneinnahme, Blutabnahme) **körperliche Risiken** ein ethisches Thema sind, spielt dies in der sozialwissenschaftlichen Forschung kaum eine Rolle. Doch auch im Rahmen sozialwissenschaftlicher Experimente kann es zu Belastungen kommen, die aus ethischen Gründen bestimmte Studien verbieten (z. B. können Wirkungen gewalthaltiger Medienangebote auf Kinder und Jugendliche nur sehr bedingt experimentell geprüft werden, da man hier eine mögliche Schädigung der minderjährigen Untersuchungspersonen befürchtet).

Zudem sind in medizinischen sowie sozialwissenschaftlichen Interventions- und Evaluationsstudien ethische Fragen der **Benachteiligung oder Bevorzugung** relevant. Dies betrifft z. B. die Zuteilung von Probanden zu unterschiedlichen Untersuchungsgruppen sowie zur Kontrollbedingung. Bei einem echten Experiment erfolgt diese Zuteilung per Zufallsverfahren (randomisiert), so dass keine systematische Benachteiligung einzelner Personen stattfindet. Bei Interventions- und Evaluationsstudien, deren Effekte über kurzfristige Wirkungen

Unvorhergesehen starke Beeinträchtigung der Probanden

Philip Zimbardo (1973): Stanford Prison Experiment
Der Sozialpsychologe Philip Zimbardo untersuchte im Rahmen eines Feldexperimentes das Verhalten von Aufsehern und Gefangenen in einem simulierten Gefängnis, das im Keller des psychologischen Instituts der Stanford Universität eingerichtet wurde (http://www.prisonexp.org; Haney, Banks, & Zimbardo, 1973; ► Abschn. 7.7). Als Probanden fungierten 24 männliche Studenten ohne psychische Auffälligkeiten, die zufällig entweder der Gruppe der Aufseher oder der Gefangenen zugeteilt wurden. Obwohl das auf zwei Wochen angelegte Experiment zuvor durch eine Ethikkommission offiziell genehmigt worden war, eskalierte die Situation so stark (ein Teil der Aufseher zeigte aggressive Verhaltensweisen; vier der Gefangenen erlitten einen emotionalen Zusammenbruch, ein weiterer entwickelte psychosomatische Beschwerden), dass der Versuch nach sechs Tagen vorzeitig abgebrochen werden musste. Im Nachgang organisierte Zimbardo Gesprächsrunden mit den Probanden, um die Erlebnisse aufzuarbeiten.

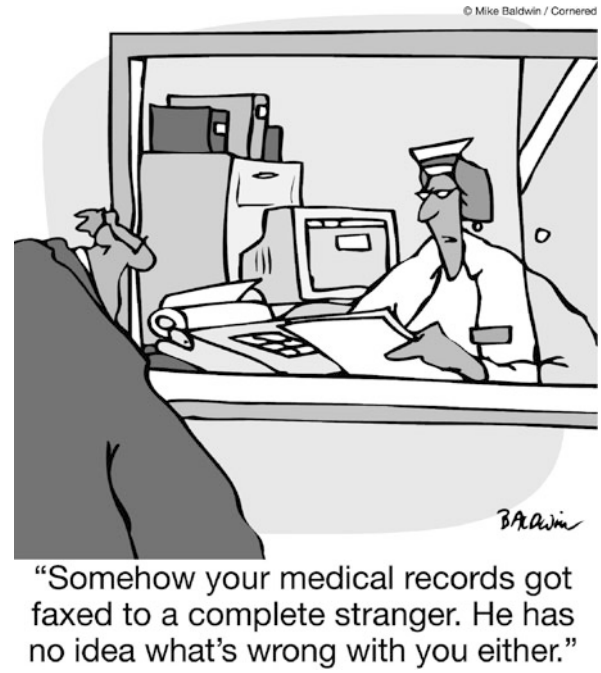
im Labor hinausgehen (z. B. Vergleich unterschiedlicher Lehr- oder Therapiemethoden in Feldstudien), kann die Teilnahme an der Kontrollbedingung eine Benachteiligung gegenüber den Personen in der Untersuchungsbedingung darstellen, die möglicherweise von einer innovativen Maßnahme (z. B. neues Therapie- oder Lernprogramm) profitieren. Hier wird zuweilen den Probanden der Kontrollgruppe im Nachgang ebenfalls eine entsprechende Behandlung angeboten (z. B. Teilnehmer auf der Warteliste erhalten nach Abschluss der Studie dieselbe psychotherapeutische Intervention wie die Experimentalgruppe).

Schädigungen können nicht nur während, sondern auch nach der Datenerhebung im Zuge der **Datenanalyse und Publikation der Daten** entstehen. Diese sollen vor allem durch eine gründliche Anonymisierung im Zuge der Datenaufbereitung (▶ Kap. 11) sowie durch vertrauliche Behandlung sensibler Daten verhindert werden.

Anonymisierung und Vertraulichkeit

Um Privatsphäre und Persönlichkeitsrechte der Untersuchungsteilnehmenden zu wahren und sie vor negativen Effekten zu schützen (z. B. soziale Stigmatisierung, berufliche Nachteile, Rufschädigung etc.), werden Forschungsdaten üblicherweise **anonym erhoben** (z. B. anonyme Fragebogenerhebung). Dies gilt insbesondere für Befragungs- oder Testdaten, die sich auf heikle Aspekte beziehen (z. B. deviantes oder kriminelles Verhalten, Erkrankungen, Leistungsdefizite; ◘ Abb. 4.1).

Teilweise ist eine komplett anonyme Datenerhebung jedoch nicht möglich. Werden etwa im Rahmen von Face-to-Face-Interviews oder Gruppendiskussionen Audio- und Videoaufzeichnungen erstellt, Verhaltensbeobachtungen per Foto und/oder Video dokumentiert oder personalisierte Online-Dokumente gesammelt (z. B. persönliche Homepages, Online-Profile), so lassen sich die Untersuchungspersonen anhand des Rohdatenmaterials identifizieren. In diesen Fällen ist dafür Sorge zu tragen, dass das **personalisierte Rohdatenmaterial** vertraulich behandelt wird und **nur autorisierten Personen zugänglich** ist (Vertraulichkeit, „confidentiality"). Das bedeutet z. B., dass entsprechende Videoaufzeichnungen in abschließbaren Schränken bzw. auf passwortgeschützten Computern aufbewahrt werden und nur ausgewählten, in Forschungsethik geschulten Mitgliedern des Forschungsteams zugänglich sind, die z. B. mit der Transkription des Materials betraut sind. Im Zuge der Datenaufbereitung (▶ Kap. 11) ist dann für eine **Anonymisierung der aufbereiteten Datensätze** zu sorgen (z. B. Entfernung von Namen und identifizierbaren Merkmalen aus Interview-Transkripten). Die sorgfältige Anonymisierung der aufbereiteten Datensätze ist heute auch deswegen von großer Bedeutung, weil Datensätze immer häufiger in **Online-Archiven** öffentlich zugänglich für Sekundäranalysen bereitgestellt werden. Rechtliche **Datenschutzvorschriften** sind hier zu beachten und bei Bedarf die Datenschutzbeauftragten der Forschungseinrichtung sowie der untersuchten Institutionen zu konsultieren.



◘ **Abb. 4.1** Personalisiertes Rohdatenmaterial muss vertraulich behandelt werden und darf nur autorisierten Mitgliedern des Forschungsteams zugänglich sein. (© Mike Baldwin/Search ID: mban1505, Rights Available from CartoonStock.com)

4.1.2 Ethische Richtlinien der einzelnen Wissenschaftsdisziplinen

Die **wissenschaftlichen Fachgesellschaften** aller human- und sozialwissenschaftlichen Disziplinen haben mittlerweile eigene Ethik-Richtlinien verabschiedet, die auf die Besonderheiten des jeweiligen Faches eingehen. Teilweise werden Ethik-Richtlinien auch gemeinsam mit den **Berufsverbänden** formuliert, so dass sie nicht nur Hinweise dazu enthalten, wie sich z. B. Ärzte und Psychologen in der Rolle von Forschenden gegenüber Untersuchungsteilnehmenden verhalten sollen, sondern auch darauf eingehen, wie Ärzte und Psychologen in der Rolle von Behandlern ihren Patienten gegenüberzutreten haben. In der Praxis – z. B. bei **klinischen Studien**, in denen in Behandlung befindliche Patienten als Untersuchungsteilnehmende rekrutiert werden – ist ethische Verantwortung hinsichtlich beider Rollen, Forscherin und Ärztin oder Psychotherapeutin, gleichzeitig Genüge zu tun.

Alle einschlägigen Ethik-Richtlinien sind im Internet über die **Websites der Fachgesellschaften** öffentlich zugänglich. Für einige ausgewählte human- und sozialwissenschaftliche Fächer sollen zentrale Ethik-Kodizes im Folgenden kurz vorgestellt werden. Dabei wird der – aus fachlichen wie historischen Gründen – besonders fortgeschrittenen forschungsethischen Diskussion in Medizin und Psychologie mehr Raum gegeben. Auch die privatwirtschaftliche Markt- und Meinungsforschung orientiert sich an Ethik-Kodizes ihrer Branchenverbände (z. B. Market Research Society: http://www.mrs.org.uk/standards/codeconduct.htm; Rat der Deutschen Markt- und Sozialforschung e.V.: http://www.rat-marktforschung.de).

Forschungsethik in der Medizin

Der **Weltärztebund** (World Medical Association, WMA) vertritt Ärztinnen und Ärzte aller Fachrichtungen weltweit und setzt sich seit seiner Gründung 1947 dafür ein, unethisches ärztliches Handeln, wie es besonders eklatant in Zeiten des Nationalsozialismus bei Humanexperimenten in Konzentrationslagern zutage trat, zukünftig zu verhindern (▶ Der Nürnberger Code). Grundlage sollen dabei stets die 1948 von den Vereinten Nationen bestätigten **Menschenrechte** sein.

Ethische Richtlinien speziell für die **biomedizinische Forschung** verabschiedete der Weltärztebund 1964 mit der **Deklaration von Helsinki** (Declaration of Helsinki, DoH), die u. a. festlegt: „In der medizinischen Forschung am Menschen muss das Wohlergehen der einzelnen Versuchsperson Vorrang vor allen anderen Interessen haben." (Weltärztebund, 2008, S. 2). Die Deklaration von Helsinki wird durch internationale Abstimmungsprozesse fortlaufend weiterentwickelt und in Spezialfragen durch zahlreiche Stellungnahmen ergänzt. Um die **Ethikbildung unter Medizinern** zu fördern, hat der Weltärztebund im Jahr 2005 ein umfassendes, mit Fallbeispielen didaktisch aufbereitetes **„Handbuch ärztlicher Ethik"** herausgegeben, das in mehreren Sprachen online kostenlos zur Verfügung steht (http://www.wma.net/en/30publications/30ethicsmanual/).

Forschungsethik in der Psychologie

Die internationale Führungsrolle im Fach Psychologie nimmt mit der **APA** (American Psychological Association) bislang ein nordamerikanischer Fachverband ein, der sowohl als wissenschaftliche Fachgesellschaft als auch als Berufsverband fungiert (▶ Weltweite Ethik-Deklaration für die Psychologie). Ebenso wie die Deklaration von Helsinki basiert auch der **Ethik-Kodex der APA** (APA Ethics Code) zentral auf den Menschenrechten und wurde in seiner ersten Fassung Mitte der 1960er-Jahre publiziert. Seitdem unterliegt der Kodex fortlaufenden Überarbeitungen. Die aktuelle Fassung der **Ethical Prin-**

Der Nürnberger Code

Die grausamsten Auswüchse unethischen Verhaltens von Wissenschaftlern sind aus der Zeit des Nationalsozialismus bekannt (Mitscherlich & Mielke, 1949): Zwischen 1939 und 1945 wurden im Rahmen von „Forschungsprojekten" mit Hunderten von unfreiwilligen Teilnehmenden **medizinische Humanexperimente in Konzentrationslagern** durchgeführt. Unter dem Anschein der Wissenschaftlichkeit setzten approbierte Ärzte die „Probanden" wissenschaftlich unsinnigen Prozeduren aus, die im Grunde Foltermethoden entsprachen und oft zu schwersten Verletzungen oder zum Tod (sog. „finale Experimente") führten. So wurde z. B. „untersucht", wie oft Knochen gebrochen werden können, bis sie nicht mehr heilen oder wie lange ein Mensch unbekleidet bei Minusgraden im Freien aushält, bevor der Erfrierungstod einsetzt (United States Holocaust Memorial Museum: Medical Experiments: http://www.ushmm.org).

Im Rahmen des **Nürnberger Ärzteprozesses**, der von 1946 bis 1947 vor dem US-Militärgericht stattfand, wurden die „Humanexperimente" in Konzentrationslagern (KZ) als verbrecherisch eingestuft und einige der verantwortlichen KZ-Ärzte angeklagt und verurteilt. Sieben Todesstrafen wurden vollstreckt und neun langjährige Haftstrafen verhängt, die jedoch alle in eine vorzeitige Entlassung bereits Anfang der 1950er-Jahre mündeten.

Im Zuge der Aufarbeitung der nationalsozialistischen Menschenexperimente formulierte das **US-Militärgericht** im Jahr 1947 den **Nürnberger Code** (Nuremberg Code), der zehn Richtlinien zum ethischen Umgang mit Untersuchungsteilnehmenden vorschreibt. Er ist bis heute für medizinische und psychologische Experimente richtungsweisend. Zentral verankert im Nürnberger Code sind

- die informierte Einwilligung der Versuchspersonen samt Möglichkeit, die Teilnahme im laufenden Versuch jederzeit abzubrechen, die Verhinderung von unnötigen körperlichen und seelischen Leiden und Verletzungen im Zuge der Untersuchung,
- die Beschränkung auf wissenschaftlich sinnvolle Untersuchungen,
- die einen gesellschaftlich wertvollen Erkenntnisgewinn liefern,
- eine Kosten-Nutzen-Abwägung in dem Sinne, dass Studien, deren Risiken größer sind als ihr erwarteter Nutzen, unterbleiben sollen.

Medizinhistorische Studien zeigen, dass nicht nur im Nationalsozialismus, sondern auch in anderen Kulturen und Epochen unfreiwillige Versuche an Erwachsenen und Kindern durchgeführt wurden, die große Risiken bis hin zum Tod bargen (zu medizinischen Menschenversuchen in der Weimarer Zeit s. Reuland, 2004; zu zwischen 1949 und 1969 an der Zivilbevölkerung durchgeführten Tests nuklearer, chemischer und biologischer Waffen durch die US-Regierung s. Moreno, 2000). Bezeichnungen wie „Humanexperiment" oder „Menschenversuch" sowie „Versuchsperson" haben durch ethisch verwerfliche, unmenschliche Untersuchungen einen sehr negativen Klang bekommen. Sich durch eine Umbenennung – z. B. aus „Versuchsperson" wird „Versuchspartner" – sprachlich von dieser Vergangenheit zu distanzieren, scheint weniger hilfreich als eine konsequente Orientierung aller wissenschaftlichen Untersuchungen an aktuellen disziplinspezifischen forschungsethischen Richtlinien.

Weltweite Ethik-Deklaration für die Psychologie

Die starke Fokussierung im Fach Psychologie auf den nordamerikanischen Forschungsraum und die APA wird den Anforderungen einer zunehmend globalisierten Forschungswelt nicht gerecht. Dementsprechend hat die International Union of Psychological Science **(IUPsyS)** als globaler Dachverband es sich zur Aufgabe gemacht, psychologische Ethik-Kodizes unterschiedlicher Länder zu vergleichen und eine weltweit gültige Ethik-Deklaration für die Psychologie zu verabschieden. Seit 2008 liegt die **Universal Declaration of Ethical Principles for Psychologists** vor (http://www.am.org/iupsys/resources/ethics). Da sie jedoch in Forschungsfragen nicht sehr detailliert ist, dürften der APA-Ethics Code bzw. die DGPs-Richtlinien weiterhin die zentrale Referenz für psychologisch Forschende im deutschsprachigen Raum darstellen.

ciples of Psychologists and Code of Conduct (APA, 2010; http://www.apa.org/ethics/code/) widmet einen von zehn Abschnitten ausdrücklich der psychologischen Forschungsethik.

Die Deutsche Gesellschaft für Psychologie **(DGPs)** hat in ihren zusammen mit dem Berufsverband Deutscher Psychologinnen und Psychologen **(BDP)** herausgegebenen **Ethischen Richtlinien der DGPs und des BDP** mittlerweile die APA-Richtlinien zur Forschungsethik nahezu identisch übernommen (http://www.dgps.de/dgps/kommissionen/ethik/). Die DGPs versteht sich als „internationale Gesellschaft im deutschsprachigen Raum" und ist somit auch für die psychologisch Forschenden in Österreich und in der Schweiz eine Anlaufstelle. Die Österreichische Gesellschaft für Psychologie (ÖGP) hat bislang keine eigenen Ethikrichtlinien verabschiedet (http://www.oegp.net), während die Schweizerische Gesellschaft für Psychologie **(SGP)** auf der Basis der APA- und DGPs-Richtlinien eigene **Ethische Richtlinien für Psychologinnen und Psychologen der SGP** formuliert hat (http://www.ssp-sgp.ch/ethik.htm).

Forschungsethik in der Soziologie

Ethik-Kodizes für die Soziologie wurden z. B. von der International Sociological Association (**ISA**; http://www.isa-sociology.org/about/isa_code_of_ethics.htm) sowie von der Deutschen Gesellschaft für Soziologie (**DGS**) zusammen mit dem Berufsverband Deutscher Soziologen (**BDS**) (http://www.soziologie.de/index.php?id=19) herausgegeben. Hierbei wird u. a. auf die „Rechte der Untersuchten" eingegangen, vor allem auf die Wahrung der Persönlichkeitsrechte von Untersuchungsteilnehmenden.

Forschungsethik in der Erziehungswissenschaft

Ethikrichtlinien für die Erziehungswissenschaft, die neben wissenschaftsethischen auch dezidiert forschungsethische Fragen des Umganges mit Untersuchungspersonen behandeln, liegen u. a. von der Australian Association for Research in Education (**AARE**; http://www.aare.edu.au/ethics/ethcfull.htm), von der American Educational Research Association (**AERA**; http://www.aera.net/AboutAERA/Default.aspx?menu_id=90&id=717) sowie von der Deutschen Gesellschaft für Erziehungswissenschaft (**DGfE**; http://www.dgfe.de/service/satzung.html) vor.

Forschungsethik in der Kommunikationswissenschaft

Die International Communication Association **(ICA)** hat sich bewusst entschlossen, keinen eigenen Ethik-Kodex zu verabschieden und verweist stattdessen auf die Richtlinien zur Forschungsethik benachbarter Fachdisziplinen (http://www.icahdq.org/aboutica/ethics.asp). Die Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft **(DGPuK)** bietet eine kurzgefasste Ethik-Erklärung (http://www.dgpuk.de/index.cfm?id=3377) und überwacht forschungs- und wissenschaftsethisches Fehlverhalten durch einen eigenen Ethik-Ausschuss. Tatsächlich sind forschungsethische Fragen in der Kommunikationswissenschaft meist weniger brisant als in der Medizin oder Psychologie, in deren Studien teilweise weitreichende Eingriffe in psychische und/oder physische Vorgänge bei Untersuchungsteilnehmenden vorgenommen werden. Eine Ausnahme stellen kommunikationswissenschaftliche Experimente im Rahmen der Medienwirkungsforschung dar: Einerseits ist eine Erforschung möglicher Negativwirkungen von Medien besonders relevant, andererseits ist es in experimentellen Studien ethisch problematisch, Probanden (insbesondere Minderjährige) mit potenziell schädlichen Medieninhalten zu konfrontieren.

4.1.3 Einhaltung ethischer Richtlinien

Wie wird bei einer konkreten Studie die Einhaltung ethischer Richtlinien im Umgang mit Untersuchungsteilnehmenden und Untersuchungsmaterial sichergestellt? Zum einen wird heute zunehmend verlangt, die forschungsethischen Aspekte einer Studie in deren Forschungsbericht bzw. Publikation dezidiert zu diskutieren. Zum anderen wird immer häufiger gefordert, Studien vorab durch eine

Ethikkommission formal genehmigen zu lassen. Sollten trotz dieser Maßnahmen ethische Richtlinien verletzt worden sein, so gilt es den Schaden zu beheben sowie dessen Ursachen zu ergründen und für die Zukunft auszuschließen. Dazu zählen dann auch Sanktionsmaßnahmen gegenüber den verantwortlichen Forscherinnen und Forschern.

Begründung des Vorgehens im Forschungsbericht

Ethische Überlegungen sind bereits bei der Planung einer Studie anzustellen. Auf der Basis welcher Ethik-Kodizes dabei welche Entscheidungen getroffen und welche Maßnahmen ergriffen wurden, ist im Untersuchungsbericht zu dokumentieren und zu begründen. Dabei ist auf entsprechende Quellen zur Forschungsethik sowie auf vergleichbare Studien des jeweiligen Forschungsfeldes zu verweisen. Diese Transparenz erlaubt es, eine ethisch problematische Studie zumindest im Zuge des Peer-Review-Verfahrens von der Publikation auszuschließen. Eine bereits publizierte Studie kann im wissenschaftlichen Diskurs hinsichtlich möglicher ethischer Schwächen beleuchtet und kommentiert werden. Auf diese Weise entwickelt sich dann auch das Ethik-Verständnis des betreffenden Forschungsfeldes weiter.

Genehmigung durch Ethik-Kommissionen

Bei Studien mit menschlichen oder tierischen Untersuchungsteilnehmenden, bei denen ein mehr als minimales Risiko von Beeinträchtigungen besteht – etwa Medikamententests in der Medizin – ist die Begutachtung durch eine Ethikkommission der Normalfall (Arbeitskreis medizinischer Ethikkommissionen: http://www.ak-med-ethik-komm.de/). Auch in den Sozialwissenschaften wird vermehrt auf die Freigabe von Studien durch Ethikkommissionen Wert gelegt **(„ethics commitee approval“)**. Manche Fachzeitschriften verlangen laut ihren Autorenrichtlinien die Vorlage entsprechender Nachweise als Voraussetzung der Publikation einer Studie.

Sofern in der jeweiligen Fachdisziplin bzw. an der Forschungseinrichtung die formale Genehmigung durch eine Ethik-Kommission gefordert wird, so ist diese über die jeweilige Forschungseinrichtung abzuwickeln. Hochschulen verfügen z. B. in der Regel über entsprechende Kommissionen. In den USA ist jede staatlich geförderte Hochschule und Forschungseinrichtung verpflichtet, alle Forschungsarbeiten durch eine Ethikkommission (IRB, „**institutional review board**“) genehmigen zu lassen. Dies dient auch der Absicherung gegenüber rechtlichen Folgen im Falle tatsächlicher oder vermeintlicher Schädigung von Probanden. Es ist zu beachten, dass das Begutachtungs- und Genehmigungsverfahren einige Wochen in Anspruch nehmen kann, zuweilen Rückfragen eingehen oder Überarbeitungen des Untersuchungsplans notwendig sind. Bei internationalen Studien sind die jeweils lokalen Ethik-Anforderungen einzubeziehen.

Die Begutachtung durch eine Ethik-Kommission erfolgt auf der Basis eines standardisierten Fragebogens bzw. Formulars, in dem detailliert über die zu beurteilende geplante Studie berichtet werden muss. Auch bei Studien, bei denen keine formale Genehmigung notwendig ist, kann es hilfreich sein, ein **Ethikformular** durchzugehen, um sicherzustellen, dass man alle entscheidenden ethischen Punkte bedacht hat und das eigene Vorgehen ethisch begründen kann.

Der verstärkte Einsatz von Ethik-Kommissionen ist jedoch auch kritisch zu hinterfragen. Er hat eine weitere **Bürokratisierung** der Forschung mit sich gebracht und zur Etablierung einer regelrechten „Ethik-Industrie“ geführt. Zudem ist auch zu beobachten, dass konservative Ethik-Kommissionen unter dem Vorwand „ethischer Bedenken“ Studien zu bestimmten Themen oder mit bestimmten Methoden erschweren können (Langdrigde & Hagger-Johnson, 2009, S. 510f.). Dies betrifft z. B. Befragungsstudien zu sexuellen Themen oder Untersuchungen im Internet, die oft pauschal als besonders riskant eingestuft und verstärkt abgelehnt werden (Orton-Johnson, 2010; Sherry & Amidon, 2010).

Umgang mit Verletzungen der Forschungsethik

Sollte es trotz entsprechender Vorkehrungen zu Verletzungen ethischer Richtlinien und/oder unvorhergesehenen Beeinträchtigungen von Untersuchungsteilnehmenden kommen, so steht die **Schadensbegrenzung und -behebung** an erster Stelle. Bei körperlichen und/oder seelischen Beeinträchtigungen im Zuge von Studien sollten z. B. Nachuntersuchungen oder zusätzliche kostenlose Beratungs- und Behandlungsangebote unterbreitet werden.

Im schlimmsten Fall haben Verletzungen forschungsethischer Richtlinien so gravierende Negativfolgen für die betroffenen Untersuchungsteilnehmenden (z. B. gesundheitliche Beeinträchtigung durch experimentelle Intervention; Arbeitsplatzverlust aufgrund Publikation identifizierbarer Interviewäußerungen), dass dies auch **zivil- oder strafrechtliche Konsequenzen** hat.

Wissenschaftsinterne Sanktionen können z. B. auf den temporären oder dauerhaften Ausschluss aus Fachgesellschaften oder von Forschungsförderungen hinauslaufen. Allein das Bekanntwerden entsprechenden Fehlverhaltens in der Fachcommunity (z. B. durch öffentliche Rüge der Fachgesellschaft einer Wissenschaftsdisziplin oder der Deutschen Forschungsgemeinschaft) stellt bereits eine gravierende Sanktion dar, weil damit der Ruf und die Karriere der Betroffenen beschädigt werden.

Nicht zuletzt sind die Ursachen des Problems zu ergründen und durch entsprechende **Präventionsmaßnahmen** möglichst für die Zukunft auszuschließen. Beispielsweise könnte eine Forschungseinrichtung einen entsprechenden Vorfall zum Anlass nehmen, um Forschungsethik stärker in der Ausbildung zu verankern, die Genehmigung aller Studien durch eine neue institutionseigene Ethik-Kommission zu verlangen oder die Arbeitsweise einer vorhandenen Ethik-Kommission durch Anpassung an internationale Standards zu optimieren.

4.2 Wissenschaftsethik

Forschungsethische Richtlinien sind nur relevant bei Studien, die Untersuchungsteilnehmende direkt involvieren, also lediglich bei empirischen Primärstudien. Bei der Sekundäranalyse bereits vorliegender Datensätze sowie bei Metaanalysen früherer Studien oder sonstigen reinen Theoriearbeiten kommen Forschende schließlich gar nicht mehr mit Untersuchungsteilnehmenden in Kontakt (zu verschiedenen Untersuchungsdesigns und Studientypen ▶ Kap. 7). Wissenschaftsethische Regeln sind dagegen – ganz unabhängig vom Studientyp – bei jeder einzelnen wissenschaftlichen Arbeit von Bedeutung, das gilt auch für studentische Arbeiten im Zuge des Studiums.

Wissenschaftsethik – Die Wissenschaftsethik umfasst alle ethischen **Regeln guter wissenschaftlicher Praxis** („codes of scholarly conduct", „codes of academic conduct"), an denen sich Forschende bei ihrer wissenschaftlichen Tätigkeit jenseits des Umgangs mit Untersuchungspersonen (dieser ist Gegenstand der Forschungsethik) orientieren sollen, um **wissenschaftliches Fehlverhalten** („scientific misconduct") zu verhindern. Im Mittelpunkt steht der verantwortungsvolle Umgang mit wissenschaftlichen Ergebnissen, ihrer Generierung, Interpretation, Bewertung und Veröffentlichung.

Im Folgenden werden zunächst die Regeln guter wissenschaftlicher Praxis vorgestellt (▶ Abschn. 4.2.1) und anschließend Wissenschaftsfälschungen (▶ Abschn. 4.2.2) sowie Plagiate (▶ Abschn. 4.2.3) als besonders gravierendes wissenschaftliches Fehlverhalten angesprochen. Auch ethische Aspekte der Autorschaft wissenschaftlicher Publikationen (▶ Abschn. 4.2.4) werden thematisiert.

4.2.1 Regeln guter wissenschaftlicher Praxis

Die Verletzung der Regeln guter wissenschaftlicher Praxis wird als **wissenschaftliches Fehlverhalten** („scientific misconduct") bezeichnet und wissenschaftsintern sanktioniert (z. B. Entzug des Doktortitels bei nachgewiesenem Plagiarismus; Ausschluss aus wissenschaftlichen Fachgesellschaften bei nachgewiesener Wissenschaftsfälschung etc.). Darüber hinaus kann wissenschaftliches Fehlverhalten auch **rechtliche Konsequenzen** haben (z. B. Arbeitsrecht, Disziplinarrecht, Urheberrecht). Die Deutsche Forschungsgemeinschaft gibt Empfehlungen dazu ab, wie Forschungseinrichtungen die Umsetzung guter wissenschaftlicher Praxis fördern können (DFG, 1998), etwa durch entsprechende Sensibilisierung und Qualifizierung von Studierenden und wissenschaftlichem Nachwuchs sowie die Einrichtung von Selbstkontrollorganen (Vertrauenspersonen, Ombudsman der DFG: http://www.ombudsman-fuer-die-wissenschaft.de etc.).

Zu den Regeln guter wissenschaftlicher Praxis gehören vor allem folgende Prinzipien wissenschaftlicher Integrität und Redlichkeit (DFG, 1998; HRK, 1998), die sich teilweise in den wissenschaftlichen Qualitätskriterien der methodischen und ethischen Strenge sowie der Präsentationsqualität wiederfinden (▶ Abschn. 3.1.2):

1. Es soll nach den „Regeln der Wissenschaft" **(„lege artis")** gearbeitet werden; dies betrifft vor allem die Anwendung üblicher wissenschaftlicher Methoden. Idealerweise geschieht dies auf der Basis des jeweils aktuellen Forschungsstandes, d. h. gemäß dem „**state of the art**". Dieses Kriterium guter wissenschaftlicher Praxis verlangt vor allem nach der Einhaltung der wissenschaftlichen Gütekriterien (▶ Kap. 3). Unsaubere Forschung auf der Basis fragwürdiger Theorien oder veralteter Methoden ist insofern unethisch, als damit der Öffentlichkeit ungültige und somit nutzlose Forschungsergebnisse präsentiert werden. Dies kann zu falschen Entscheidungen in der Praxis führen und stellt außerdem eine Vergeudung von Forschungsmitteln dar. Die gravierendste Verletzung dieser Regel ist die **Wissenschaftsfälschung**.
2. Wissenschaftliche Ergebnisse – insbesondere auch eigene – sind im Sinne der Wissenschaftsethik stets kritisch zu hinterfragen. Die **(selbst)kritische Reflexion** von Forschungsergebnissen soll – neben der Orientierung am aktuellen Forschungsstand – der Qualitätssicherung dienen und die Öffentlichkeit vor fehl- oder überinterpretierten Befunden schützen. Eine kollektive kritische Reflexion wissenschaftlicher Ergebnisse bietet der **Peer-Review-Prozess**, der wissenschaftlichen Konferenzvorträgen und Zeitschriftenartikeln vorgeschaltet ist. Vor diesem Hintergrund ist es z. B. wissenschaftsethisch zu kritisieren, wenn Forschende ihre Befunde in unkritischer Weise ohne Kenntlichmachung der Limitationen ihrer Studie in der Öffentlichkeit verbreiten, womöglich sogar ohne vorherige kritische Überprüfung durch einen Peer-Review-Prozess. Auch der Peer-Review-Prozess muss ethischen Kriterien genügen und auf einer Selbstverpflichtung zur sachgerechten und neutralen Begutachtung basieren. Reviewende sollten dementsprechend eine Begutachtung ablehnen, wenn sie gerade zu wenig Zeit

haben, um den Beitrag gründlich zu prüfen, wenn der Beitrag aus einem ihnen fremden Forschungsfeld stammt oder wenn Interessenskonflikte vorliegen. Zudem müssen Reviewende verantwortungsvoll mit den ihnen zugänglich gemachten noch unveröffentlichten Manuskripten umgehen und dürfen aus ihnen keinen persönlichen Nutzen ziehen.

3. **Primärdaten sollen gesichert** und mindestens zehn Jahre lang aufbewahrt werden (typischerweise an der Forschungseinrichtung, an der die Studie durchgeführt wurde), um Nachvollziehbarkeit zu gewährleisten und z. B. Wissenschaftsfälschungen aufdecken zu können. Diese Sicherung muss unter Anwendung des Datenschutzes erfolgen. Der Datenschutz darf indessen keine Ausrede dafür sein, Primärdaten zu beseitigen und damit die Überprüfbarkeit einer Studie zu verhindern.
4. Die wissenschaftlichen **Beiträge von Kooperationspartnern, Konkurrenten und Vorgängern** zur eigenen Arbeit sind in Forschungsberichten und Publikationen wahrheitsgemäß als solche kenntlich zu machen. Das betrifft korrekte Zitation anstelle von **Plagiarismus**, aber auch eine faire Festlegung der **Autorschaft** von Publikationen. Die Mitautorschaft einer angesehenen Kollegin oder eines angesehenen Kollegen darf nicht ohne deren ausdrückliches Einverständnis beansprucht werden. Umgekehrt dürfen Personen, die maßgeblich zur Publikation beigetragen haben, auch nicht ungenannt bleiben.
5. Die **Sabotage der Forschungstätigkeit** von Konkurrenten etwa durch Beschädigung, Zerstörung oder Manipulation von Versuchsanordnungen, Geräten, Unterlagen, Hardware oder Software widerspricht den Regeln guter wissenschaftlicher Praxis.

Wissenschaftliches Fehlverhalten müssen sich nicht nur diejenigen anlasten lassen, die selbst Regeln guter wissenschaftlicher Praxis missachten oder sich aktiv am Fehlverhalten anderer beteiligen. Auch Mitwisserschaft, die Beteiligung an plagiats- oder fälschungsbehafteten Publikationen oder die Vernachlässigung der Aufsichtspflicht stellen eine Mitverantwortung für wissenschaftliches Fehlverhalten dar.

4.2.2 Wissenschaftsfälschungen

Wissenschaftsfälschungen stellen ein gravierendes wissenschaftliches Fehlverhalten dar (Abb. 4.2; Goodstein, 2010). Als Motive der Fälschenden werden u. a. akademischer Erfolgsdruck hinsichtlich Publikationen und Drittmitteleinwerbungen, aber auch Geltungsdrang und

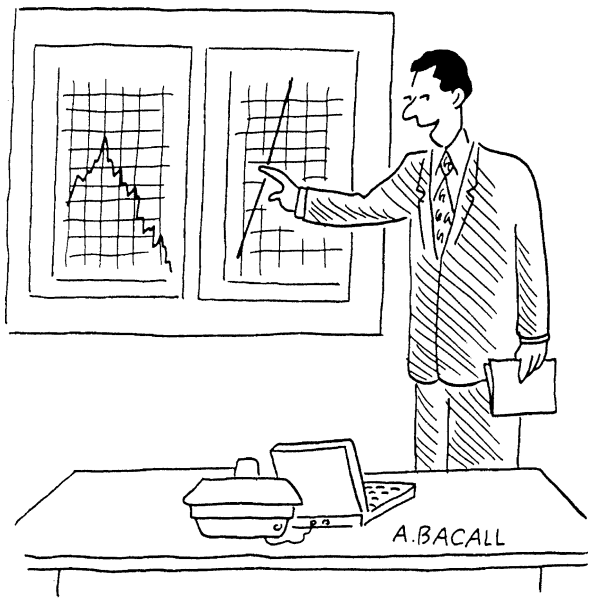


Abb. 4.2 Eine in Richtung auf das Wunschergebnis manipulierte Darstellung von Forschungsbefunden wird als Wissenschaftsfälschung gewertet. (© Aaron Bacall/Search ID: aba0105, Rights Available from CartoonStock.com)

das Streben nach Ruhm und Anerkennung sowie Bequemlichkeit vermutet.

Wissenschaftsfälschung – Eine **Wissenschaftsfälschung** („scientific fraud") liegt vor, wenn a) empirische Forschungsergebnisse bewusst in Richtung eines Wunschergebnisses manipuliert werden (z. B. gezielte Eliminierung hypothesenkonträrer Fälle oder geschönte Darstellung von Befunden; „data falsification", „data massaging") oder wenn b) gar keine empirischen Daten erhoben, sondern die Ergebnisse erfunden wurden (z. B. simulierte Datensätze; „data fabrication").

Verantwortlich gemacht für einen – oft längere Zeit unbemerkt gebliebenen – Wissenschaftsbetrug werden jedoch nicht nur die Forschenden selbst, sondern unter bestimmten Umständen auch ihr Umfeld (z. B. Kollegen, Mitarbeiter, denen Unregelmäßigkeiten aufgefallen sein müssten), die betreffende Forschungseinrichtung (die durch entsprechende Ausbildung und Kontrolle für die Einhaltung guter wissenschaftlicher Praxis sorgen muss) sowie die Peer Reviewer und Zeitschriften-Herausgeber (die zumindest sehr auffällige Manipulationen und Fälschungen in den von ihnen begutachteten Manuskripten bemerken sollten). Laut einer Metaanalyse von 21 Umfragen unter Wissenschaftlerinnen und Wissenschaftlern, berichteten 2 %

der Befragten, bereits mindestens einmal selbst Daten manipuliert oder gefälscht zu haben, 14 % hatten Wissenschaftsfälschungen bei Kollegen mitbekommen (Fanelli, 2009). Zu beachten ist, dass unethisches Verhalten nicht nur Extremfälle wie die Komplettfälschung von ganzen Datensätzen von Studien betrifft, sondern bereits bei kleineren Manipulationen beginnt, etwa wenn bei einer quantitativen Fragebogenstudie einzelne „unpassende" Bögen eliminiert, einige Bögen mit „Wunschantworten" hinzugefügt oder bei der Datenbereinigung (► Abschn. 11.4.3) selektiv im Sinne eines Wunschergebnisses vorgegangen wird. Das „Herauskitzeln" von Wunschergebnissen kann bei einer Veränderung der Datensätze und/oder bei der verzerrten Ergebnisdarstellung für die Publikation ansetzen (z. B. ist es unethisch und manipulativ, von einer Studie nur die hypothesenkonformen Befunde in die Publikation einzubeziehen und die konträren Befunde zu verschweigen).

Wenn Wissenschaftsfälschungen aufgedeckt werden, dann zuweilen durch Hinweise aus dem Umfeld der Forschenden sowie vor allem durch mangelnde Replizierbarkeit oder offenkundige Unstimmigkeiten in Veröffentlichungen. In Deutschland sind spektakuläre Wissenschaftsfälschungen in Medizin und Naturwissenschaften aus den 1980er, 1990er und 2000er Jahren bekannt geworden (► Beispiele für spektakuläre Wissenschaftsfälschungen in Deutschland). Der Fälschungsfall aus der Medizin hat dabei die Deutsche Forschungsgemeinschaft im Jahr 2000 zur Veröffentlichung ihrer Regeln guter wissenschaftlicher Praxis veranlasst.

Doch nicht nur Medizin und Naturwissenschaften, auch die Sozialwissenschaften und insbesondere die **Psychologie** sind von Wissenschaftsfälschungen betroffen. Der **niederländische Sozialpsychologe Diederik Stapel** musste sich 2011 aus dem Universitätsleben zurückziehen und gab auch seinen Doktortitel zurück, nachdem durch Mitglieder seines Teams bekannt gemacht und von einer Untersuchungskommission bestätigt worden war, dass er über Jahre hinweg in großem Stil Daten erfunden hatte. Mehr als fünfzig seiner Publikationen in renommierten Fachzeitschriften wurden offiziell zurückgezogen (Dokumentation auf **Retraction Watch**: http://www.retractionwatch.com/category/by-author/diederik-stapel/). Dieser Fall erregte weltweit Aufmerksamkeit. Zeitgleich häuften sich weitere anekdotische und empirische Hinweise darauf, dass ein unsorgfältiges bis manipulatives Vorgehen bei der Datenerhebung und/oder Datenanalyse in der Psychologie erschreckend verbreitet ist (Pashler & Wagenmakers, 2012).

Daraufhin setzte eine intensive Debatte über die **Glaubwürdigkeitskrise der psychologischen Forschung** ein, die u. a. in einem Sonderheft der Zeitschrift „Perspectives on Psychological Science" dokumentiert ist (Pashler & Wagenmakers, 2012). Es geht darum, die tatsächlichen und vermeintlichen Missstände rund um Manipulationen und Fälschungen in der Psychologie aufzudecken und zu erklären, sowie sinnvolle Gegenmaßnahmen zu entwickeln, umzusetzen und zu evaluieren. Offenbar besteht das Problem eben nicht nur darin, dass einzelne Forschende aus persönlichen Gründen Fehlverhalten an den Tag legen, sondern dass bestimmte **Mechanismen im Wissenschaftssystem**, die eigent-

Beispiele für spektakuläre Wissenschaftsfälschungen in Deutschland

Medizin

In Deutschland gilt der Fall des **Krebsforscherpaares Friedhelm Herrmann und Marion Brach** als einer der gravierendsten Fälle von Wissenschaftsfälschung (DFG, 2000): Eine Task Force identifizierte für die Zeitspanne von 1988–1996 insgesamt 29 Publikationen mit gefälschten und 65 Publikationen mit manipulierten Daten, darüber hinaus 121 Publikationen, bei denen der Verdacht auf Fälschung nicht erhärtet, aber auch nicht ausgeräumt werden konnte aufgrund mangelnder Dokumentation bzw. Archivierung der Rohdaten. Die Deutsche Forschungsgemeinschaft (DFG) nahm den Fall Herrmann/Brach zum Anlass, im Jahr 1998 eindeutige **Empfehlungen zur Sicherung guter wissenschaftlicher Praxis für DFG-Forschungsprojekte** zu formulieren und fortan strenger zu überprüfen.

Physik

Die Physik wurde im Jahr 2002 international erschüttert durch die Fälschungen des deutschen Physikers **Jan Hendrik Schön**. Aufgefallen war, dass er in verschiedenen hochrangigen Publikationen identische Messwertreihen präsentiert hatte, die sich im Nachhinein als computergenerierte Simulationsdaten herausstellten. Die Konsequenzen dieses wissenschaftlichen Fehlverhaltens waren gravierend: Der Forscher verlor seine Stelle an einer Forschungseinrichtung, zwei wissenschaftliche Preise wurden ihm aberkannt und mehr als 25 seiner Publikationen von den jeweiligen Fachzeitschriften zurückgezogen. Der Fall löste eine intensive Debatte über die Qualität des **Peer-Review-Prozesses** aus, weil fragwürdige Paper in so hochrangigen Zeitschriften wie *Science* und *Nature* erschienen und den Fachgutachtenden nicht negativ aufgefallen waren.

Anthropologie

Der deutsche Anthropologe **Reiner Rudolf Robert Protsch** hatte mit gefälschten Datierungen den Eindruck erweckt, über spektakuläre Schädelfunde aus der Altsteinzeit zu verfügen. Der im Jahr 2005 vorgelegte Abschlussbericht der Kommission zum Umgang mit wissenschaftlichem Fehlverhalten der Universität Frankfurt wirft dem Wissenschaftler Manipulationen und Fälschungen wissenschaftlicher Daten **über eine Zeitspanne von rund dreißig Jahren hinweg** vor. Dieser Fall lenkte das Augenmerk auf die **Mitwissenden im Umfeld**, etwa Mitgliedern im Forschungsteam, denen Unregelmäßigkeiten aufgefallen sein müssten, die aber vermutlich aus Angst vor persönlichen Folgen geschwiegen hatten.

lich wünschenswerte gute wissenschaftliche Praxis nicht genügend fördern oder gar verhindern. Als Faktoren diskutiert werden unter anderem der stark wachsende Publikations- und Vermarktungsdruck sowie auch das mangelnde Ansehen von Replikationsstudien, die als unoriginell gelten, aber ein wichtiges Mittel sind, um verzerrte oder verfälschte Befunde zu identifizieren (zu Replikationsstudien ▶ Abschn. 7.3). Nicht zuletzt wurde unzureichende Präsentationsqualität der Befunde (▶ Abschn. 3.1.2) im Sinne unvollständiger Darstellung des methodischen Vorgehens moniert. In kurzer Zeit starteten daraufhin im Jahr 2012 innerhalb der Psychologie eine Reihe von Initiativen, um die Wissenschaftspraxis in der Weise zu verändern, dass **Wissenschaftsfälschungen** und sonstige **falsch-positive Befunde** (d. h. hypothesenkonforme Befunde, die durch methodische Ungenauigkeit, Manipulation oder puren Zufall zustande kommen) effektiver verhindert werden und vorläufig gesicherte Befunde besser erkennbar sind.

- Das **Reproducability-Project in der Psychologie** hat sich zum Ziel gesetzt, mit einem weltweiten ehrenamtlichen Team von Forschenden ausgewählte Studien aus drei renommierten psychologischen Fachzeitschriften systematisch nach einem festgelegten Protokoll zu replizieren, um auf dieser Basis fundierte Aussagen über die Verlässlichkeit psychologischer Befunde zu treffen (Carpenter, 2012; Open Science Collaboration, 2012; http://www.openscienceframework.org). Damit stellt sich die Wissenschaftsdisziplin selbst auf den Prüfstand.
- Auf der Online-Plattform **PsychFileDrawer.org** können Ergebnisse gelungener sowie misslungener Replikationsstudien aus der Experimentalpsychologie in Kurzform dargelegt werden, die von Forschenden gemäß „Schubladenproblem“ bzw. Publication Bias (▶ Abschn. 16.1.3) meist nicht in Fachartikeln veröffentlich werden (können), etwa weil die Ergebnisse vermeintlich zu unoriginell oder uninteressant für Fachzeitschriften sind.
- Auf der Online-Plattform **PsychDisclosure.org** werden ergänzend zu publizierten Fachartikeln jeweils zusätzliche Details des methodischen Vorgehens offengelegt. Etwa soll dargestellt werden, welche Daten einer Studie aus welchen Gründen nicht in die Publikation eingingen. Damit soll u.a. aufgedeckt und verhindert werden, dass Forschende sich aus einem sehr umfangreichen Datensatz willkürlich nur die besten Befunde heraussuchen und die anderen ignorieren („Rosinenpicken“, „Signifikanzenfischen“; ▶ Abschn. 12.5.2). Die Initiative zieht angesichts breiter Resonanz eine positive erste Bilanz und empfiehlt, in Fachzeitschriften generell transparentere Methodendarstellungen zu verlangen (LeBel et al., 2013).
- Die neu gegründete APA-Fachzeitschrift **Archives of Scientific Psychology** folgt nicht nur dem Open-Access-Konzept (d. h. alle Artikel stehen kostenlos im Internet zur Verfügung), sondern auch dem **Open-Method- und Open-Data-Konzept**: Mit jedem eingereichten Artikel sind alle Details des methodischen Vorgehens über eine standardisierte Eingabemaske exakt anzugeben, darüber hinaus sind sämtliche Datenerhebungsinstrumente und Datensätze abzugeben und werden im Falle einer positiven Begutachtung zusammen mit dem Beitrag publiziert. Auf diese Weise haben nicht nur die Reviewenden einen vollständigen Überblick, auch die Leserschaft kann die präsentierten Befunde jederzeit kritisch prüfen, so dass Manipulationen und Fälschungen kaum noch möglich sein sollten.

4.2.3 Plagiarismus in der Wissenschaft

Beim Fremdplagiat eignet man sich fremdes geistiges Eigentum an und gibt dieses als eigene wissenschaftliche Leistung aus. Das Spektrum reicht vom Komplettplagiat (z. B. wenn Studierende eine Hausarbeit aus dem Internet herunterladen und als eigene Arbeit einreichen) über die Aneignung fremder Textpassagen bis hin zur Verwendung einzelner Ideen, Sätze, Tabellen oder Abbildungen ohne ausreichende Zitierung der Ursprungsquelle. Eine sorgfältige Verwaltung der eigenen Literaturquellen (z. B. mit einem Literaturverwaltungsprogramm; ▶ Abschn. 6.1.3) sowie eine akribische Einhaltung von Zitationsregeln ist beim wissenschaftlichen Arbeiten unerlässlich, um versehentliche Plagiate zu vermeiden. Je mehr Stellen ohne korrekte Quellenangabe in einer Arbeit auftreten, umso unglaubwürdiger wird die Erklärung eines versehentlichen Zitationsfehlers und umso eher verdichtet sich der Verdacht eines bewussten Plagiats.

Plagiarismus – Plagiarismus („plagiarism“) liegt vor, wenn in wissenschaftlichen Werken Ideen oder Textteile anderer Autorinnen und Autoren übernommen werden, ohne dies entsprechend durch Zitation der Originalquellen kenntlich zu machen. Neben dem Fremdplagiat existiert auch das Selbstplagiat, wenn ohne Kennzeichnung auf eigene frühere Texte zurückgegriffen wird.

Das Hauptmotiv für Plagiarismus dürfte die Arbeitsersparnis sein. Bei **Qualifikations- und Abschlussarbeiten** kann nachgewiesener Plagiarismus zum Nichtbestehen der Prüfung, zur Exmatrikulation sowie nachträglich zur Aberkennung des Titels führen. In einer Studie, in der Studierende im Fach Soziologie befragt sowie soziologische Hausarbeiten geprüft wurden, stellte sich heraus, dass rund 20 % der Studierenden schon plagiiert hatten (Sattler, 2007, 2008). Es häufen sich zudem Fälle, in denen

Plagiate in Doktorarbeiten von Personen des öffentlichen Lebens (z. B. von Politikern) nachgewiesen werden. Plagiarismus in der Wissenschaft liegt auch vor, wenn Personen sich ihre Abschlussarbeiten von **Ghostwritern** erstellen lassen und als eigene Arbeiten ausgeben.

Nicht zuletzt existieren neben Fremdplagiaten auch **Selbstplagiate**, bei denen man sich ohne ausreichende Quellenangaben aus eigenen Werken bedient. Ein Selbstplagiat liegt z. B. vor, wenn in einer studentischen Masterarbeit große Teile der eigenen Bachelorarbeit wortwörtlich und ohne Zitation wiederverwendet werden. Auch in wissenschaftlichen Publikationen kommt es zu Eigenplagiaten, etwa wenn Forschende unter Publikationsdruck versuchen, ihre Publikationszahl zu erhöhen, indem sie – statt genuin neue Daten und Ideen zu präsentieren – aktuelle Publikationen ohne entsprechende Kennzeichnung aus alten Publikationen zusammensetzen. Die Grenzziehung zwischen der legitimen Weiterentwicklung und Aktualisierung eigener Arbeiten sowie der Verbreitung einer Theorie oder Studie in unterschiedlichen disziplinären bzw. sprachlichen Fachcommunitys einerseits und ethisch problematischen Selbstplagiaten andererseits kann dabei im Einzelfall schwierig sein. Die Inhaltsanalyse einer Stichprobe von $n = 62\,213$ Abstracts aus der Literaturdatenbank Medline ergab, dass 0.04 % der Publikationen von unterschiedlichen Autoren sowie 1.35 % der Publikationen von denselben Autoren nahezu identische Abstracts aufwiesen, was ein Hinweis auf Fremd- bzw. Selbstplagiate sein könnte (Errami, Hicks, & Fisher, 2008).

Es wird empfohlen, in der akademischen Ausbildung zu verdeutlichen, dass Plagiate keine Bagatelle sind, sondern eine ernst zu nehmende Verletzung der Regeln guter wissenschaftlicher Praxis darstellen und somit dem Selbstanspruch aller Forschenden widersprechen sollten. Für die Suche nach Plagiaten in wissenschaftlichen Texten liegen entsprechende Software-Tools vor. Diese **Plagiatserkennungs-Systeme** weisen jedoch unterschiedliche Schwächen auf und stellen somit nur eine grobe Hilfe dar (eine Übersicht bietet das Informationsportal der Berliner Medieninformatik-Professorin Debora Weber-Wulf: http://plagiat.htw-berlin.de/). So sind in erster Linie wortwörtliche Übernahmen **(Copy-Paste-Plagiat)** aus gut zugänglichen, digitalen Quellen nachweisbar. Die Übernahme von Textstellen aus abseitigen, schwer zugänglichen bzw. nicht-digitalen Quellen, die sinngemäße Übernahme von Ideen **(Ideenplagiat)** oder die Übernahme von Textstellen bzw. Ideen aus fremdsprachlichen Quellen **(Übersetzungsplagiat)** sind dagegen in automatisierter Form kaum erkennbar. Entsprechende Plagiate können jedoch auffallen, wenn in einem Text plötzlich inhaltliche oder formale Stilbrüche, ungewöhnliche Formulierungen oder selten verwendete Begriffe auftauchen. Es ist empfehlenswert, jeden selbstproduzierten wissenschaftlichen Text, der nach den Regeln guter wissenschaftlicher Praxis erstellt wurde, vor Abgabe bzw. Einreichung zur Sicherheit routinemäßig einem automatisierten Plagiats-Check zu unterziehen und das Ergebnis zu archivieren. Forschungsinstitute bzw. Fakultäten, Hochschulrechenzentren und Bibliotheken verfügen heute meist über Plagiatserkennungs-Software.

4.2.4 Wissenschaftliche Autorschaft

Es gehört zu den Regeln guter wissenschaftlicher Praxis, bei wissenschaftlichen Publikationen alle an der jeweiligen Studie sowie dem veröffentlichen Paper maßgeblich beteiligten Personen als Autorinnen und Autoren auszuweisen. Dies ist umso komplizierter, je größer und heterogener die Gruppe der an einem – z. B. interdisziplinären und/oder internationalen – Forschungsprojekt beteiligten Wissenschaftlerinnen und Wissenschaftler ist und je länger ein Publikationsprojekt dauert. Es wird empfohlen, Absprachen zur Autorschaft frühzeitig zu treffen und den sich ggf. im Laufe der Zeit ändernden Bedingungen anzupassen (z. B. unerwartete Mehrarbeiten durch Revisionsauflagen im Peer-Review-Prozess, personeller Wechsel im Forschungsteam etc.).

Wissenschaftliche Autorschaft – Wissenschaftliche Autorschaft („academic authorship“) meint die namentliche Nennung aller Forschenden, die maßgeblich an der Durchführung und Veröffentlichung einer wissenschaftlichen Studie beteiligt waren. Da human- und sozialwissenschaftliche Studien meist in Teamarbeit durchgeführt werden, muss gemäß den Regeln guter wissenschaftlicher Praxis in sachlich angemessener und fairer Weise festgelegt werden, welche Art von Mitarbeit eine Person zur Mit-Autorschaft qualifiziert und wie bei kollektiver Autorschaft die Autorenreihenfolge festzulegen ist.

Ein Streitpunkt bei kollektiver Autorschaft kann dabei sein, wer überhaupt in die **Autorenliste** („authorship credit“) aufgenommen wird:

- **Geringfügige Beiträge ohne Berechtigung zur Autorschaft**: Mitglieder des Forschungsteams, die nur kleine Teilarbeiten im Rahmen der Studie geleistet, aber nicht an der Publikation direkt mitgewirkt haben (z. B. studentische Hilfskräfte, die unter Anleitung die Datenerhebung – z. B. Durchführung von Interviews –, die Datenaufbereitung – z. B. Transkription – und/oder die Datenauswertung – z. B. Kodierung – unterstützt haben), sind gemäß APA-Richtlinien keine Mitautoren (APA, 2009). Sie können aber bei Bedarf in einer Fußnote, in der ihnen für ihre Unterstützung gedankt wird, namentlich genannt werden. Entscheidend für die Mitautorschaft ist der substanzielle wissenschaftliche Beitrag. Besondere Sorgfalt in

der Abwägung ist geboten, wenn es um Publikationen von Wissenschaftlerinnen und Wissenschaftlern geht, in die Ergebnisse der von ihnen betreuten studentischen Arbeiten einfließen. Je nach Umfang der Betreuungsleistung einerseits und eigenständigen Arbeit der Studierenden andererseits sowie je nach Arbeitsanteil beim Schreiben des Manuskriptes kann das Spektrum reichen von studentischer Erstautorschaft, Mitautorschaft bis zu keiner Autorschaft (Fine & Kurdek, 1993). Studierende als Mitautoren aufzunehmen, die dafür gar nicht die notwendige wissenschaftliche Leistung erbracht haben, ist ebenso unethisch wie Studierende, die entsprechend substanzielle Beiträge einbringen, nicht zu nennen.

- **Ehrenautorschaft** („honorary authorship"): Forschungsgruppenleitende sind nicht automatisch qua beruflicher Position Mitautoren jeder einzelnen Publikation aus ihrer Abteilung. Eine solche Ehrenautorschaft widerspricht den Regeln guter wissenschaftlicher Praxis. Denn alle Mitautoren einer Publikation müssen aktiv an deren Erstellung mitgewirkt haben und auch inhaltlich für das Ergebnis Verantwortung tragen können (z. B. wenn sich im Extremfall der Beitrag im Nachhinein als plagiiert oder gefälscht erweist). Eine Auswertung medizinischer Fachzeitschriften zeigte, dass der Anteil der Publikationen mit Ehrenautoren zwischen 60 % und 4 % pro Zeitschrift schwankte (Bates, Anić, Marušić, & Marušić, 2004).
- **Ghostwriter:** Die tatsächliche Autorschaft eines Beitrags nicht offenzulegen, stellt ebenfalls eine Verletzung der Regeln guter wissenschaftlicher Praxis dar. Entsprechende Fälle treten z. B. auf, wenn Interessensgruppen oder Unternehmen ungenannt bleiben und die von ihnen lancierten Beiträge unter dem Namen von Wissenschaftlern publizieren lassen, um den Anschein der Neutralität zu wahren.

Wurden alle substanziell an der Publikation beteiligten Personen identifiziert, die zur Mitautorschaft legitimiert sind, so muss unter diesen noch die **Autorenreihenfolge** („authorship order") festgelegt werden. Üblicherweise stellt die Reihenfolge gleichzeitig eine Rangfolge hinsichtlich der Bedeutung der Autoren für das betreffende Paper dar. Besonders umkämpft ist die Platzierung als Erstautor, da bei mehr als zwei Autoren bei Zitaten mit „et al." (et alii/et aliae = und andere Mitautoren bzw. Mitautorinnen) abgekürzt wird und somit nur der Name der Erstautorin bzw. des Erstautors auftaucht und bekannt wird. Die Abwägung des Gewichtes der Einzelbeiträge der Mitautoren kann dabei schwierig sein. Wer hat den wichtigeren Beitrag geleistet: Die Person, die die Idee zu der Studie hatte und den Theorieteil zum Paper beigesteuert hat? Die Person, die für die Datenerhebung und Datenauswertung verantwortlich zeichnet? Oder die Person, die den größten Teil des Manuskripts geschrieben hat? Hier obliegt es dem Forschungsteam, zu fairen Lösungen zu kommen und am besten bereits bei der Untersuchungsplanung entsprechende Vereinbarungen zu treffen. Im Zweifelsfall kann auch per Los entschieden werden. Will man signalisieren, dass alle Mitautoren vergleichbare Zuarbeiten geleistet haben, wird zuweilen die **alphabetische Reihenfolge** gewählt.

Nicht zuletzt sind bei wissenschaftsethischen Fragen rund um die Autorschaft auch die Konventionen der unterschiedlichen Fachdisziplinen, Institutionen, Arbeitsgruppen und Fachzeitschriften zu beachten. Während in den Geistes- und Kulturwissenschaften die alleinige Autorschaft dominiert, sind in der Medizin Autorenzahlen im zweistelligen Bereich nicht ungewöhnlich. Ein einzelner medizinischer Zeitschriftenartikel brachte es sogar auf 972 Mitautoren (The GUSTO Investigators, 1993). Die Sozialwissenschaften nehmen hier eine Mittelposition ein. Soll die Leistung einzelner Forschender über die Anzahl ihrer Publikationen quantifiziert werden, so gehen Publikationen in Mitautorschaft nur anteilig gemäß der jeweiligen **Autorenzahl des Artikels** ein.

In jedem Fall kommt der **Erstautorin bzw. dem Erstautor** („primary author") eine besondere Verantwortung zu, weil sie oder er meist auch als Kontaktperson zur Zeitschrift fungiert („corresponding author") und in dieser Rolle die Einhaltung bestimmter Regeln guter wissenschaftlicher Publikationspraxis sicherstellen muss (z. B. dass alle Mitautoren genannt werden, dass alle Mitautoren die Endfassung des Manuskripts autorisiert haben, dass das Manuskript nicht zeitgleich an anderer Stelle eingereicht wurde etc.).

Lernquiz

1. Wie lassen sich Forschungsethik und Wissenschaftsethik voneinander abgrenzen?
2. Welche Formen der Täuschung von Untersuchungspersonen im Rahmen von empirischen Studien werden differenziert und unter welchen Umständen sind sie ethisch legitim?
3. Was versteht man unter informierter Einwilligung und wann ist sie notwendig?
4. Welche Beiträge zu einer wissenschaftlichen Publikation qualifizieren gemäß guter wissenschaftlicher Praxis in der Regel für die Mitautorschaft?
   a. Die extrem zeitaufwändige Transkription aller Interviews,
   b. die umfassende Analyse der Daten samt Ergebnisinterpretation,

   c. die Akquise der Drittmittel für das betreffende Forschungsprojekt und/oder
   d. die Idee zu Studienthema und Untersuchungsdesign sowie die Überarbeitung des Manuskripts.
5. Welche Art von Plagiaten lässt sich durch Software automatisch entdecken?
6. Welche zwei Varianten der Wissenschaftsfälschung werden unterschieden?
7. Im Rahmen einer empirischen Studie werden Fokusgruppen-Diskussionen durchgeführt und auf Video aufgezeichnet. Welches Risiko birgt diese Studie für die Teilnehmenden und warum:
   a. Kein Risiko,
   b. minimales Risiko oder
   c. größer als minimales Risiko.

   Was folgt aus der Risikobewertung für den ethischen Umgang mit den Fokusgruppen-Teilnehmenden?
8. Welche Ethikrichtlinien sind für die psychologische Forschung a) im deutschsprachigen Raum und b) international besonders relevant?
9. Nennen Sie drei Datenerhebungsmethoden, die als risikolose Forschung einzustufen sind und somit in der Regel keine formale informierte Einwilligung erfordern.
10. Im Rahmen eines Experimentes soll der Einfluss von Stimmungen auf die Gedächtnisleistung untersucht werden. Die Probanden werden zufällig in drei Gruppen eingeteilt, die sich unter dem Vorwand der Überbrückung von Wartezeit jeweils entweder einen lustigen, neutralen oder traurigen Filmclip anschauen, um positive, neutrale oder negative Stimmung zu erzeugen. Wie ist mit dieser Situation aus forschungsethischer Sicht umzugehen?
    a. Die Probanden müssen vorher ihr Einverständnis zu der Stimmungsbeeinflussung erklären.
    b. Da Stimmungsänderungen im Alltag normal sind, handelt es sich um risikolose Forschung, bei der gar keine besonderen ethischen Vorkehrungen zu treffen sind.
    c. Das experimentelle Setting beinhaltet aktive Täuschung, so dass die Probanden am Ende des Versuchs über die wahre Zielsetzung der Studie aufzuklären sind.
    d. Den Probanden der neutralen und negativen Stimmungsbedingung sollte am Ende des Versuchs ebenfalls der lustige Filmclip gezeigt werden, damit sie in positiver Stimmung entlassen werden.

Übungsaufgaben

1. Lesen Sie sich die ethischen Richtlinien der führenden Fachgesellschaft Ihrer Wissenschaftsdisziplin gründlich durch!
2. Eruieren Sie, an welche Stelle innerhalb Ihrer Hochschule oder Forschungseinrichtung Sie sich mit forschungs- und wissenschaftsethischen Fragen wenden können!
3. Suchen Sie im Internet eine Einverständniserklärung zu einer psychologischen Studie (Suchbegriffe: „+Einverständniserklärung+Psychologie“ oder „+'consent form'+psychology“) und prüfen Sie diese auf Allgemeinverständlichkeit und Vollständigkeit!

Weiterführende Literatur

- Buchanan, T, Williams, J. E. (2010). Ethical issues in psychological research on the internet. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 255–271). Washington, DC: American Psychological Association.
- DGPs (Deutsche Gesellschaft für Psychologie). (2004). Revision der auf die Forschung bezogenen ethischen Richtlinien. Abgerufen 17. Juli, 2014, unter http://www.dgps.de/fileadmin/documents/ethikrl2004.pdf
- Goodstein, David. (2010). *On fact and fraud: Cautionary tales from the front lines of science*. New Jersey: Princeton University Press.
- Informationsportal des APA Research Ethics Office. (2013). Retrieved July 17, 2014, from http://www.apa.org/ethics/.
- Nagy, T. F. (2010). *Essential ethics for psychologists: A primer for understanding and mastering core issues*. Washington, DC: American Psychological Association.

Literatur

Akins, C. K., Panicker, S., & Cunningham, C. L. (Eds.). (2004). *Laboratory animals in research and teaching: Ethics, care, and methods*. Washington, DC: American Psychological Association.

Allen, C. (1997). Spies like us, when sociologists deceive their subjects. *Lingua Franca, 7*(9), 30–39.

APA (American Psychological Association). (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: APA.

APA (2010). Ethical principles of psychologists and code of conduct. 2010 Amendments. Retrieved November 20, 2013, from http://www.apa.org/ethics/code/

Bates, T., Anić, A., Marušić, M., & Marušić, A. (2004). Authorship criteria and disclosure of contributions. Comparison of 3 general medical journals with different author contribution forms. *JAMA, 292*(1), 86–88.

Blass, T. (1999). „The Milgram paradigm after 35 years: Some things we now know about obedience to authority". *Journal of Applied Social Psychology, 29*(5), 955–978.

Buchanan, T. & Williams, J. E. (2010). Ethical issues in psychological research on the internet. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (S. 255-271). Washington, DC, US: American Psychological Association.

Carpenter, S. (2012). Psychology's bold initiative. *Science, 335*, 1558–1560.

DFG (Deutsche Forschungsgemeinschaft). (1998). Sicherung guter wissenschaftlicher Praxis. Denkschrift. Abgerufen unter http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

DFG (2000). Task Force legt Abschlussbericht vor. Abgerufen 20. November, 2013, unter http://www.dfg.de/service/presse/pressemitteilungen/2000/pressemitteilung_nr_26/index.html

DGPs (Deutsche Gesellschaft für Psychologie). (2004). Revision der auf die Forschung bezogenen ethischen Richtlinien. Abgerufen 20. November, 2013, unter http://www.dgps.de/fileadmin/documents/ethikrl2004.pdf

Errami, M., Hicks, J. M., & Fisher, W. (2008). „Déjà vu – a study of duplicate citations in Medline". *Bioinformatics, 24*(2), 243–249.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE, 4*(5).

Fine, M. A. & Kurdek, L. A. (1993). Reflections on determining authorship credit and authorship order on faculty-student collaborations. *American Psychologist, 48*(11), 1141–1147.

Galliher, J. F., Brekhus, W., & Keys, D. P. (2004). *Laud Humphreys: prophet of homosexuality and sociology*: Madison: University of Wisconsin Press.

Goodstein, D. (2010). *On fact and fraud: Cautionary tales from the front lines of science*. New Jersey: Princeton University Press.

Haney, C., Banks, W. C., & Zimbardo, P. G. (1973). Study of prisoners and guards in a simulated prison. *Naval Research Reviews, 9*, 1–17.

HRK (1998). Zum Umgang mit wissenschaftlichem Fehlverhalten in den Hochschulen. Abgerufen November 20, 2013, unter Forschungsethik http://www.hrk.de/de/beschluesse/109_422.php.

Humphreys, L. (1970). *Tea-room trade*. Chicago: Aldine.

Kaslow, F. W., Patterson, T., & Gottlieb, M. (2011). Ethical dilemmas in psychologists accessing internet data: Is it justified? *Professional Psychology: Research and Practice, 42*(2), 105–112.

Lakeman, R. & FitzGerald, M. (2009). The ethics of suicide research: The views of ethics committee members. *Crisis: The Journal of Crisis Intervention and Suicide Prevention, 30*(1), 13–19.

Langdrigde, D. & Hagger-Johnson, G. (2009). *Introduction to research methods and data analysis in psychology* (2nd ed.). München: Pearson.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., et al. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8*(4), 424–432.

McKee, H. A. & Porter, J. E. (2009). *The ethics of internet research: A rhetorical, case-based process.* (Digital formations). New York, NY: Lang.

Milgram, S. (1963). „Behavioral study of obedience". *Journal of Abnormal and Social Psychology, 67*(4), 371–378.

Milgram, S. (1974). *Obedience to authority; An Experimental View*. New York: Harpercollins.

Mitscherlich, A. & Mielke, F. (1949). *Wissenschaft ohne Menschlichkeit – medizinische und eugenische Irrwege unter Diktatur, Bürokratie und Krieg*. Heidelberg: Schneider.

Moreno, J. D. (2000). *Undue risk: Secret state experiments on humans*. London: Routledge.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues, 58*(1), 161–176.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660.

Orton-Johnson, K. (2010). Ethics in Online Research; Evaluating the ESRC Framework for Research Ethics Categorisation of Risk. *15*(4), 13. Retrieved November 20, 2013, from http://www.socresonline.org.uk/15/4/13.html

Pashler, H. & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science 7*(6), 528–530.

Reuland, J. (2004). *Menschenversuche in der Weimarer Republik*. Norderstedt: Books on Demand.

Ripley, E., Macrina, F., Markowitz, M., & Gennings, C. (2010). Who's doing the math? Are we really compensating research participants? *Journal of Empirical Research on Human Research Ethics, 5*(3), 57–65.

Sales, B. D. & Folkman, S. (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association.

Sattler, S. (2007). *Plagiate in Hausarbeiten. Erklärungsmodelle mit Hilfe der Rational Choice Theorie.* Mit einem Vorwort von Andreas Diekmann. Hamburg: Kovac.

Sattler, S. (2008). Unterschätztes Phänomen? Über den Umfang von und Umgang mit Plagiaten. *Forschung und Lehre*, 222–223.

Sherry, A. & Amidon, A. (2010). The ethics of sex research on the internet. In D. L. Streiner & S. Sidani (Eds.), *When research goes off the rails: Why it happens and what you can do about it* (pp. 27–33). New York, NY: Guilford.

Tait, R. C., Chibnall, J. T., Iltis, A., Wall, A., & Deshields, T. L. (2011). Assessment of consent capability in psychiatric and medical studies. *Journal of Empirical Research on Human Research Ethics, 6*(1), 39–50.

The GUSTO Investigators (1993). An international randomized trail comparing four thrombolytic strategies for acute myocardial infarction. *The New England Journal of Medicine*. Retrieved November 20, 2013, from http://www.nejm.org/doi/full/10.1056/NEJM199309023291001

Weltärztebund (WMA). (2008). Deklaration von Helsinki – Ethische Grundsätze für die medizinische Forschung am Menschen. Abgerufen 25. November, 2013, unter http://www.bundesaerztekammer.de/downloads/deklHelsinki2008.pdf.

Anwendung

Kapitel 5 **Forschungsthema – 143**
Nicola Döring, Jürgen Bortz

Kapitel 6 **Forschungsstand und theoretischer Hintergrund – 157**
Nicola Döring, Jürgen Bortz

Kapitel 7 **Untersuchungsdesign – 181**
Nicola Döring, Jürgen Bortz

Kapitel 8 **Operationalisierung – 221**
Nicola Döring, Jürgen Bortz

Kapitel 9 **Stichprobenziehung – 291**
Nicola Döring, Jürgen Bortz

Kapitel 10 **Datenerhebung – 321**
Nicola Döring, Jürgen Bortz

Kapitel 11 **Datenaufbereitung – 579**
Nicola Döring, Jürgen Bortz

Kapitel 12 **Datenanalyse – 597**
Nicola Döring, Jürgen Bortz

Kapitel 13 **Ergebnispräsentation – 785**
Nicola Döring, Jürgen Bortz

Forschungsthema

Nicola Döring, Jürgen Bortz

5.1 Vom Forschungsthema zum Forschungsproblem – 144

5.2 Forschungshypothesen und Forschungsfragen – 145

5.3 Sechs Kriterien für die Wahl des eigenen Forschungsthemas – 149

5.4 Planung des eigenen Forschungsprojektes – 153

Literatur – 155

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_5, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Forschungsthema und Forschungsproblem voneinander abgrenzen können.
- In der Lage sein, zu einem Forschungsproblem einen Satz von Forschungshypothesen und/oder Forschungsfragen zu formulieren.
- Abwägen können, nach welchen Kriterien ein Forschungsthema für eine Qualifikationsarbeit auszuwählen ist.
- Wissen, was bei der Erstellung eines Exposés und eines Zeitplans für ein Forschungsprojekt zu beachten ist.

Über welches Thema soll ich meine Bachelor-, Master- oder sonstige Abschlussarbeit schreiben? Diese Frage ist für viele Studierende schwer zu beantworten. Generell steht die geeignete Themenwahl am Anfang jedes empirischen Forschungsprojekts. Dabei ist es besonders wichtig, nach der Themenwahl das Forschungsproblem zuzuspitzen (▶ Abschn. 5.1). Zum Forschungsproblem wird bei jeder empirischen Studie ein Satz von detaillierteren Forschungshypothesen und/oder Forschungsfragen formuliert (▶ Abschn. 5.2). Anhand welcher Kriterien man die Eignung eines Forschungsthemas beurteilen kann (▶ Abschn. 5.3) und was bei der Zeit- und Ressourcenplanung eines Forschungsprojektes zu beachten ist (▶ Abschn. 5.4), wird im Folgenden näher erläutert.



Abb. 5.1 Manchmal ist ein radikaler Bruch mit herkömmlichen Denkmustern notwendig und manchmal das Anknüpfen an bestehendes Wissen überzeugender. (Aus Poskitt & Appleby, 1993. Mit freundlicher Genehmigung der Autoren)

5.1 Vom Forschungsthema zum Forschungsproblem

Studierenden, die noch auf **Themensuche** sind, empfehlen wir, für ihre Abschlussarbeiten auf Themenvorschläge zurückzugreifen, die an den Instituten ausgeschrieben sind. Es ist günstig, Themen zu wählen, bei denen man auf theoretische und methodische Vorkenntnisse aus dem Studium aufbauen kann. An Forschungstraditionen anzuknüpfen und z. B. eine bekannte Studie zu modifizieren oder zu replizieren, ist meist erfolgversprechender als ein übertriebener Originalitätsanspruch, der – insbesondere bei geringer Forschungserfahrung – oft zu fragwürdigen Resultaten führt (Abb. 5.1).

Das **Forschungsthema** beinhaltet in der Regel die Nennung eines bestimmten Untersuchungsgegenstandes bzw. Sachverhaltes aus der eigenen wissenschaftlichen Disziplin. So könnte sich eine Abschlussarbeit aus der Psychologie mit „Eifersucht in Freundschaften“, eine Arbeit aus der Kommunikationswissenschaft mit „Nachrichten im privaten und öffentlich-rechtlichen Fernsehen“ oder eine erziehungswissenschaftliche Qualifikationsarbeit mit „Online-Lernen an der Hochschule“ befassen. Viele Forschungsthemen sind so allgemein, dass sie sich aus verschiedenen fachlichen Perspektiven untersuchen lassen. So kann das Thema „Computerspiele“ z. B. aus psychologischem, soziologischem, kommunikationswissenschaftlichem, medienwissenschaftlichem oder erziehungswissenschaftlichem Blickwinkel untersucht werden.

Forschungsthema – Das Forschungsthema („research subject“) benennt einen Untersuchungsgegenstand.

Forschungsproblem – Das Forschungsproblem („research problem“) kennzeichnet, welche Erkenntnisse zu welchen Aspekten des Untersuchungsgegenstandes auf welcher theoretischen, empirischen und methodischen Basis gewonnen werden sollen. Das Forschungsproblem wird auch oft als **Fragestellung einer Studie** bezeichnet und in mehrere Forschungshypothesen und/oder Forschungsfragen ausdifferenziert.

Nach der Wahl des Themas ist es zwingend notwendig, das zu untersuchende **Forschungsproblem** zu formulieren bzw. – im Deutschen ist diese Bezeichnung geläufiger – die zentrale **Fragestellung** der Studie zuzuspitzen. Dies sollte üblicherweise theorie- und empiriegeleitet geschehen, d. h., auf der Basis vorliegender Theorien und empirischer Befunde. Doch auch erste Methodenentscheidungen tragen zur Transformation eines allgemeinen Forschungsthemas (hier im Beispiel: „Computerspiele“) in ein untersuchbares Forschungsproblem bei:

- **Inhaltliche Eingrenzung des Gegenstandes**: Welche Aspekte des Themas sollen untersucht werden (Eingrenzung von Zielgruppen, Zeiträumen, Effekten etc.)? Soll sich die Computerspiel-Studie auf Kinder oder auf Senioren beziehen? Auf Sport-, Strategie- oder Denkspiele? Geht es um einen interkulturellen Vergleich zwischen deutschen und japanischen Spielern? Oder um den didaktischen Vergleich zweier Englisch-Lernspiele?
- **Bezug zum empirischen Forschungsstand**: Die Überlegungen zur Eingrenzung des Gegenstandes sind auch empiriegeleitet anzustellen, d. h., mit Blick auf frühere Studien und vorliegende oder fehlende Befunde zu Einzelaspekten. Wenn bei der Aufarbeitung des bisherigen Forschungsstandes z. B. auffällt, das bestimmte Dimensionen des Gegenstandes bislang wenig bis gar nicht untersucht wurden, so kann diese empirische **Forschungslücke** („research gap" – man spricht auch vom **Forschungsdesiderat** („research desideratum") – zum Anlass der eigenen Studie genommen werden.
- **Wahl der Theorie(n)**: Welches sind die zentralen Theorien des eigenen Faches (und ggf. auch von Nachbardisziplinen), die auf das Thema anwendbar sind? Mit der Wahl der Theorie(n) ergeben sich Konsequenzen für die Betrachtung des Untersuchungsgegenstandes. So kann eine konkrete Theorie die Berücksichtigung oder auch die Vernachlässigung bestimmter Teilaspekte des Themas nahelegen. Die Nutzung von Computerspielen kann z. B. mit theoretischen Konzepten wie Motivation, Emotion, Sucht, Lernen, Aufmerksamkeit, Flow, Immersion, Präsenz, Sozialität wie Antisozialität etc. in Zusammenhang gebracht werden.
- **Wahl der Methode(n)**: Mit welchem qualitativen und/oder quantitativen Untersuchungsdesign und mit welchen Datenerhebungs- und Datenauswertungsmethoden soll das Thema untersucht werden? Generell sollte die Methodenwahl zum Forschungsproblem passen bzw. gegenstandsangemessen erfolgen. Aber auch forschungspraktische Aspekte sind zu berücksichtigen, etwa eigene Vorkenntnisse und vorhandene Ressourcen.

Ein typischer Anfängerfehler besteht darin, das Forschungsproblem bzw. die Untersuchungsfrage zu allgemein zu formulieren (z. B. „Welche psychologischen Wirkungen hat die Nutzung von Computerspielen bei Kindern?"), so dass sie letztlich nicht sinnvoll untersuchbar ist. Ein angemessen zugespitztes Forschungsproblem könnte stattdessen lauten „Welchen Einfluss haben Computerspiele kurzfristig auf die Konzentrationsleistung von Kindern im Grundschulalter?" Das Forschungsproblem wird üblicherweise im **Titel der Arbeit** genannt, wobei im **Untertitel** dann die methodische Herangehensweise konkretisiert werden kann, z. B. so: „Der kurzfristige Einfluss von Computerspielen auf die Konzentrationsleistung von Kindern im Grundschulalter: Ergebnisse einer Experimentalstudie mit einem Rennspiel".

5.2 Forschungshypothesen und Forschungsfragen

Über die notwendige Eingrenzung des Forschungsproblems sollte man sich von Anfang an Gedanken machen. Dabei können entsprechende Entscheidungen hinsichtlich der methodischen, theoretischen und inhaltlichen Ausrichtung der Arbeit meist erst im Zuge einer gründlichen Aufarbeitung des Forschungsstandes (▶ Abschn. 6.2) abschließend und detailliert getroffen werden. Erst dann ist es auch möglich, das Forschungsproblem bzw. die Fragestellung der Studie in einzelne Forschungshypothesen und/oder Forschungsfragen aufzugliedern, die im Rahmen der empirischen Studie dann auch einzelnen zu beantworten sind.

Typischerweise sind empirische Studien so angelegt, dass ihr jeweiliges Forschungsproblem bzw. ihre zentrale Fragestellung in etwa ein bis zehn verschiedene Forschungshypothesen und/oder Forschungsfragen aufgegliedert wird (hierbei handelt es sich um einen Erfahrungswert, nicht um eine feste Regel).

- **Forschungshypothesen** („research hypotheses"): Bei Forschungshypothesen handelt es sich um Aussagen, die aus Theorien abgeleitet wurden oder sich auf Ergebnisse früherer Studien stützen. Auf dieser Basis werden die Existenz sowie in der Regel auch die Richtung und Stärke von Effekten vorgegeben (Beispiel: „Der Einfluss des Computerspielens auf die Konzentrationsleistung ist bei jüngeren Kindern viel stärker als bei älteren Kindern."). Hypothesen dürfen nur dann formuliert werden, wenn sie sich eindeutig auf etablierte Theorien oder gesicherte empirische Befunde stützen und diese Quellen angegeben werden. Allein eine Vermutung „aus dem Bauch heraus" reicht als Basis für wissenschaftliche Hypothesen nicht aus. **Quantitative Studien** zielen meist darauf ab, theoretisch oder empirisch fundierte Hypothesen statistisch zu überprüfen. Die Forschungshypothesen einer Studie werden in der Regel mit „H" für Hypothese abgekürzt und durchnummeriert (H1, H2, H3 etc.).
- **Forschungsfragen** („research questions"): Fehlt es an ausreichend gesichertem Vorwissen zum Thema, so dass keine Forschungshypothesen aufgestellt werden können, formuliert man stattdessen Forschungsfragen. Bei Forschungsfragen handelt es sich oft um

W-Fragen (Beispiel: „Welche Computerspiele sind bei Vorschulkindern aus welchen Gründen besonders beliebt?“). Auf jeden Fall bleibt bei Forschungsfragen offen, in welche Richtung ein Effekt geht (Beispiel: „Spielen Kinder lieber alleine oder zu zweit am Computer oder ist beides gleichermaßen beliebt?“), während Hypothesen genau dies – mit triftiger Begründung – vorgeben würden (Beispiel: „Kinder spielen lieber zu zweit als alleine am Computer.“). **Qualitative Studien** zielen meist darauf ab, offene Forschungsfragen zu beantworten und auf dieser Basis neue Theorien zu bilden. Die Forschungsfragen einer Studie werden mit „F“ bzw. im Englischen mit „RQ“ abgekürzt und ebenfalls in der Regel durchnummeriert präsentiert (F1, F2, F3 etc. bzw. RQ1, RQ2, RQ3 etc.). In der qualitativen Forschung, in deren Logik Offenheit und Zirkularität des Forschungsprozesses wichtig sind, lassen sich Forschungsfragen im Laufe eines Forschungsprozesses bei Bedarf – in dokumentierter und begründeter Weise – noch verändern oder ergänzen (zu wissenschaftstheoretischen Grundlagen der qualitativen Forschung ► Abschn. 2.3).

Forschungshypothese – Die Forschungshypothese („research hypothesis“) wird aus etablierten Theorien und/oder gut gesicherten empirischen Befunden abgeleitet und postuliert die Existenz, Richtung und Stärke eines bestimmten Effekts. Anhand von Hypothesenprüfungen werden bestehende Theorien getestet und weiterentwickelt.

Forschungsfrage – Die Forschungsfrage („research question“) basiert auf dem bisherigen Forschungsstand und zielt v. a. auf Forschungslücken. Sie fordert dazu auf zu ermitteln, ob Regelhaftigkeiten vorliegen, welche Beschaffenheit sie haben, in welche Richtung sie weisen, wie stark sie sind, wie sie zustande kommen etc. Die Beantwortung von Forschungsfragen trägt zur Erkundung eines Sachverhaltes sowie zur Generierung neuer Theorien bei.

Werden Forschungshypothesen näher betrachtet, so sind v. a. drei Arten von Hypothesen nach der Art des postulierten Effekts zu unterscheiden, die jeweils mit unterschiedlichen statistischen Verfahren überprüft werden (► Kap. 12).

1. **Unterschiedshypothesen** („hypotheses of difference“) postulieren Unterschiede zwischen zwei oder mehr Gruppen. Die Gruppierungsvariable ist dabei die sog. **unabhängige Variable** (UV) („independent variable, IV“). Das Merkmal, auf dem sich der inhaltliche Gruppenunterschied zeigen soll, ist die **abhängige Variable** AV („dependent variable, DV“). Beispiel einer Unterschiedshypothese: Kinder, die ein eigenes Fernsehgerät im Kinderzimmer haben, schauen mehr fern als Kinder ohne eigenen Fernseher. Die UV ist hier der Gerätebesitz mit 2 Ausprägungen bzw. 2 Gruppen (eigenes TV-Gerät vs. kein eigenes TV-Gerät). Die AV ist die tägliche oder wöchentliche Fernsehnutzungszeit (in Stunden und Minuten). Zur statistischen Überprüfung von Unterschiedshypothesen sind z. B. der t-Test (2 Gruppen) und die Varianzanalyse (mehr als 2 Gruppen) indiziert (► Abschn. 12.5.3).
2. **Zusammenhangshypothesen** („correlation hypotheses“) postulieren Zusammenhänge zwischen zwei oder mehr Variablen. Statistische Zusammenhänge (Korrelationen) können keinesfalls automatisch als Beleg für Ursache-Wirkungs-Relationen (Kausalität) gelten. Dennoch stehen hinter Zusammenhangshypothesen oft Vorstellungen über Wirkungsmodelle. Variablen, die als Einflussfaktoren aufgefasst werden, bezeichnet man als **Prädiktoren** bzw. Prädiktorvariablen („predictor variables“). Die Variable, deren Ausprägung erklärt werden soll, heißt **Kriterium** bzw. Kriteriumsvariable („criterion variables“). Geht man etwa von der theoretischen Vorstellung aus, dass sich reale Sozialbeziehungen auf die Online-Welt übertragen, dass sich z. B. Jugendliche mit ihren Klassenkameraden im Internet austauschen, so könnte man folgende **bivariate Zusammenhangshypothese** (2 Variablen: 1 Prädiktor und 1 Kriterium) formulieren: Je mehr Freunde Jugendliche haben, umso mehr Zeit verbringen sie auf Social-Networking-Sites im Internet. Prädiktor: Freundeszahl, Kriterium: tägliche oder wöchentliche Nutzungszeit von Social-Networking-Plattformen in Stunden und Minuten. In diesem Fall wird ein gleichgerichteter bzw. **positiv gerichteter Zusammenhang** angenommen, d. h. je stärker die eine Variable bzw. der Prädiktor ausgeprägt ist (hier: Freundeszahl), umso stärker soll auch die andere Variable bzw. das Kriterium ausgeprägt sein (hier: Nutzungszeit von Social-Networking-Sites). Eine **negativ gerichtete Zusammenhangshypothese** würde dagegen einen – in diesem Fall inhaltlich nicht sinnvollen – gegenläufigen Zusammenhang postulieren: Je mehr Freunde Jugendliche haben, umso weniger Zeit verbringen sie auf Social-Networking-Sites. Neben bivariaten Zusammenhangshypothesen, die jeweils nur 2 Variablen einbeziehen, lassen sich auch **multiple Zusammenhangshypothesen** aufstellen, die Zusammenhänge zwischen zwei oder mehr Prädiktoren und einem Kriterium postulieren. Beispiel: Je mehr Freunde Jugendliche haben und je extravertierter sie sind, umso mehr Zeit verbringen sie auf Social-Networking-Sites im Internet. Im Rahmen von Zusammenhangshypothesen sind nicht nur Prädiktoren und Kriterium von Interesse, sondern oftmals auch sog. **Moderatorvariablen** („moderator variables“), die dafür verantwortlich sind, dass ein Zusam-

menhang zwischen 2 Variablen unter bestimmten Bedingungen stärker oder schwächer ausfällt. Beispiel: Der Zusammenhang zwischen Freundeszahl und der Intensität der Nutzung von Social-Networking-Sites ist enger, wenn der Freundeskreis geografisch stärker verstreut ist. Zur statistischen Überprüfung von Zusammenhangshypothesen sind z. B. die bivariate und die multiple Korrelations- und Regressionsanalyse (▶ Abschn. 12.5.2) sowie Strukturgleichungsmodelle (▶ Kap. 17) geeignet.

3. **Veränderungshypothesen** („hypotheses of change“) postulieren Veränderungen von Variablen bei denselben Individuen über die Zeit bzw. über zwei oder mehr Messzeitpunkte hinweg. Beispiel einer Veränderungshypothese: Die Aggressivität der Teilnehmenden eines Anti-Aggressionstrainings ist unmittelbar nach dem Training sowie 3 Monate später geringer ausgeprägt als vor dem Training. Veränderungshypothesen werden z. B. mit dem t-Test für abhängige Stichproben (2 Messzeitpunkte), der Varianzanalyse mit Messwiederholungen (mehr als zwei Messzeitpunkte) oder (bei sehr vielen Messzeitpunkten) über die Zeitreihenanalyse geprüft (▶ Abschn. 12.5.4).

Abgesehen davon, ob eine Hypothese inhaltlich auf Unterschiede, Zusammenhänge oder Veränderungen abzielt, ist noch zu beachten, ob die Hypothese die **Richtung des postulierten Effekts** vorgibt (**gerichtete Hypothese**, „one-sided hypothesis“) oder offenlässt (**ungerichtete Hypothese**, „two-sided hypothesis“). Wenn so viel theoretisches und empirisches Vorwissen vorliegt, dass überhaupt wohlbegründete Forschungshypothesen aufstellbar sind, wird man sich meist auch auf die Richtung des Effekts festlegen können. Denn ungerichtete Hypothesen, die z. B. aussagen würden, dass Kinder mit eigenem Fernseher entweder mehr oder weniger fernsehen als Kinder ohne eigenes TV-Gerät, dass Jugendliche mit wachsender Freundeszahl, entweder mehr oder weniger intensiv Social-Networking-Sites nutzen oder dass sich die Aggressivität von Teilnehmenden eines Anti-Aggressionstrainings hinterher entweder reduziert oder gesteigert hat, sind doch inhaltlich sehr vage. In der Forschungspraxis werden somit üblicherweise gerichtete Hypothesen aufgestellt und geprüft. Sie verlangen, dass das Stichprobenergebnis in die vorhergesagte Richtung weist und zudem der statistische Signifikanztest (▶ Abschn. 12.5.1) ein überzufälliges Ergebnis anzeigt.

Beim Formulieren der Forschungshypothesen stellt sich schließlich noch die Frage nach der Größe des Effekts, dessen Existenz und Richtung die Hypothese postuliert (**Effektgröße**, „effect size“): Wird vorhergesagt, dass Kinder mit eigenem TV-Gerät wenige Minuten, eine halbe Stunde oder mehrere Stunden mehr fernsehen als Kinder ohne eigenen Fernseher? Welche Größe die postulierten Effekte haben, ist sowohl grundlagenwissenschaftlich als auch anwendungswissenschaftlich von Bedeutung. Vernachlässigbar kleine Effekte weisen darauf hin, dass der Sachverhalt theoretisch noch nicht umfassend erklärt ist, und sie geben kaum Anlass, praktische Veränderungen vorzunehmen. Anders verhält es sich, wenn große Effekte vorhergesagt und dann auch empirisch bestätigt werden können. Die Bedeutung von Effektgrößen ist dabei jedoch nicht absolut zu sehen, sondern von der Art der Variablen abhängig. So können zuweilen auch sehr kleine Effekte eine hohe wissenschaftliche und praktische Bedeutsamkeit aufweisen (z. B. wenn es um Menschenleben geht). Ob man sich bei der Formulierung der Forschungshypothesen vorab auf eine präzise Effektgröße oder zumindest eine Größenordnung für den Effekt (kleiner, mittlerer oder großer Effekt) festlegen kann, hängt vom vorliegenden Forschungsstand ab.

Bereits bei der Aufstellung der Hypothesen auch **Voraussagen über die erwartete Effektgröße** zu machen, also **spezifische Forschungshypothesen** zu formulieren, ist notwendig, um den für eine aussagekräftige Studie mindestens notwendigen bzw. „**optimalen**“ **Stichprobenumfang** zu bestimmen. Denn kleine Effekte können empirisch überhaupt nur dann als statistisch signifikant nachgewiesen werden, wenn ein genügend großer Stichprobenumfang und somit ausreichende statistische **Teststärke** („power“) realisiert wird (▶ Kap. 14).

Forschungshypothesen – Forschungshypothesen („research hypotheses“) postulieren die Existenz, die Richtung und die Stärke unterschiedlicher Arten von Effekten. Zu jeder Forschungshypothese bzw. **Alternativhypothese** („research hypothesis“, „alternative hypothesis“: H_1) lässt sich eine komplementäre **Nullhypothese** („null hypothesis“: H_0) aufstellen, die den postulierten Effekt negiert. Das Hypothesenpaar bestehend aus H_1 und H_0 bildet die Grundlage des **Nullhypothesen-Signifikanztests** („null hypothesis significance testing“, NHST), einer sehr verbreiteten Methode der statistischen Hypothesenprüfung (▶ Kap. 12.5.1). Bei der Darstellung der Hypothesen in Forschungsberichten und Publikationen werden nur die Forschungshypothesen (nicht die Nullhypothesen) angegeben.

Nachfolgend sind verschiedene Typen von Forschungshypothesen nach Art, Richtung und Größe der postulierten Effekte definiert:

Art des Effekts – Bei Forschungshypothesen werden nach Art des postulierten Effekts drei inhaltliche Typen von Hypothesen differenziert: **Unterschiedshypothesen** („hypotheses of difference“), **Zusammenhangshypothesen** („correlation hypotheses“) und **Veränderungshypothesen** („hypotheses of change“).

Richtung des Effekts – Unterschieds-, Zusammenhangs- und Veränderungshypothesen können sowohl als **ungerichtete bzw. zweiseitige Hypothese** („non-directional hypothesis“, „two-sided hypothesis“) als auch – das ist in der Forschungspraxis der übliche Fall – als **gerichtete bzw. einseitige Hypothese** („directional hypothesis“, „one-sided hypo-

Tabelle 5.1 Darstellung von Forschungsthema, Forschungsproblem/Fragestellung der Untersuchung sowie dem zugehörigen Satz an Forschungshypothesen bzw. Forschungsfragen von drei Computerspiel-Studien. (Jeweils eigene Formulierungen)

Qualitative Studie	**Quandt, Grueninger, & Wimmer (2009)**
Forschungsthema	Computerspielende
Forschungsproblem	Forschungslücke in der Literatur: Jugendliche gelten als die typischen Computerspieler und werden oft untersucht, aber was kennzeichnet das Erleben und Verhalten **älterer Computerspielender**?
Vier Forschungsfragen	F1: Warum bzw. aus welchem Anlass haben die älteren Spieler ursprünglich mit dem Computerspielen begonnen? F2: Wie haben die älteren Spieler die für Computerspiele notwendigen Kenntnisse und Fähigkeiten erworben? F3: Wie integrieren die älteren Spieler das Computerspielen in ihren Alltag? F4: Wie und durch welche Einflussfaktoren (z. B. Familie, Beruf) verändert sich das Spielverhalten der älteren Spieler mit der Zeit?
Untersuchungsdesign und Datenerhebung	Explorative Interviewstudie, $n = 21$ Computerspielerinnen und -spieler im Alter zwischen 35 und 73 Jahren aus Deutschland
Quantitative Studie	**Bösche (2009)**
Forschungsthema	Computerspiele
Forschungsproblem	Zwei konkurrierende Thesen in der Literatur: Führen **gewalthaltige Computerspielinhalte** (weil sie negative Emotionen wecken) zu einer Verschlechterung der Spielleistung oder umgekehrt (weil sie positive Emotionen auslösen) zu einer **Verbesserung der Spielleistung**?
Eine Forschungshypothese auf der Basis der zweitgenannten Theorie	H1: Mit steigender Gewalthaltigkeit der Spielinhalte (nicht gewalthaltiges – moderat gewalthaltiges – sehr gewalthaltiges Spiel) steigt die Spielleistung (Schnelligkeit und Genauigkeit von Treffern). (Gerichtete Unterschiedshypothese mit 3 Gruppen, UV: Grad der Gewalthaltigkeit des Spiels mit 3 Ausprägungen, AV: Spielleistung)
Untersuchungsdesign und Datenerhebung	Laborexperiment mit $n = 50$ männlichen Studenten ohne Erfahrung mit gewalthaltigen Computerspielen aus Deutschland
Quantitative Studie	**Reinecke (2009)**
Forschungsthema	Computerspielen
Forschungsproblem	Forschungslücke in der Literatur: Die Nutzung von Computerspielen wurde bislang nicht systematisch daraufhin untersucht, inwiefern sie **Erholung vom Arbeitsstress** bietet. Das theoretisch gut elaborierte Konzept der „Erholung“ wird theorieprüfend auf Computerspiele angewendet.
Sechs Forschungshypothesen	H1: Je erholsamer das Computerspielen empfunden wird, umso häufiger wird für Erholungszwecke auf Computerspiele zurückgegriffen. (Bivariate positiv gerichtete Zusammenhangshypothese; Prädiktor: Erholsamkeitserleben des Computerspiels, Kriterium: Häufigkeit des Computerspielens zu Erholungszwecken) H2: Je mehr arbeitsbezogene Erschöpfung empfunden wird, umso häufiger wird für Erholungszwecke auf Computerspiele zurückgegriffen. (Bivariate positiv gerichtete Zusammenhangshypothese) H3: Je mehr Alltagsstress erlebt wird, umso häufiger wird für Erholungszwecke auf Computerspiele zurückgegriffen. (Bivariate positiv gerichtete Zusammenhangshypothese) H4: Personen mit emotionsbezogenen Bewältigungsstrategien greifen eher zu Erholungszwecken auf Computerspiele zurück als Personen mit problemlösungsorientierten Bewältigungsstrategien. (Diese Forschungshypothese postuliert keinen Gruppenunterschied, sondern enthält 2 Zusammenhangshypothesen. Denn jede Person weist sowohl emotions- als auch problemlösungsorientierte Bewältigungsstrategien zu einem gewissen Grad auf, so dass eine Gruppeneinteilung nicht möglich ist: Bivariate positiv gerichtete Zusammenhangshypothese zwischen der Ausprägung emotionsbezogener Bewältigungsstrategien [Prädiktor] und der Intensität des Computerspielens zu Erholungszwecken [Kriterium] sowie bivariate Hypothese, die keinen positiven Zusammenhang zwischen der Ausprägung problemlösungsorientierter Bewältigungsstrategien [Prädiktor] und der Intensität des Computerspielens zu Erholungszwecken [Kriterium] postuliert, also einen Nicht-Zusammenhang oder einen negativen Zusammenhang vorhersagt) H5: Der Zusammenhang zwischen arbeitsbezogener Erschöpfung und dem Spielen zu Erholungszwecken ist enger, wenn weniger soziale Unterstützung zur Verfügung steht. (Bivariate positiv gerichtete Zusammenhangshypothese mit sozialer Unterstützung als Moderatorvariable) H6: Der Zusammenhang zwischen Alltagsstress und dem Spielen zu Erholungszwecken ist enger, wenn weniger soziale Unterstützung zur Verfügung steht. (Bivariate positiv gerichtete Zusammenhangshypothese mit sozialer Unterstützung als Moderatorvariable)
Untersuchungsdesign und Datenerhebung	Online-Befragung von $n = 1\,614$ Internet-Usern (4 % Frauen; Durchschnittsalter 23 Jahre), rekrutiert über Game-bezogene Websites

thesis“) formuliert werden. Bei der gerichteten Hypothese wird jeweils die Richtung von Gruppenunterschieden, Merkmalszusammenhängen oder Veränderungen vorab postuliert. Bei der ungerichteten Hypothese bleibt die Richtung offen, lediglich die Existenz eines von Null abweichenden Effekts wird behauptet.

Größe des Effekts (Effektgröße) – („effect size“) Effekte, d. h. Unterschiede zwischen Gruppen, Zusammenhänge zwischen Variablen oder Veränderungen von Untersuchungsobjekten über die Zeit, können in ganz unterschiedlicher Größe auftreten. Eine Hypothese, bei der die erwartete Effektgröße (oder zumindest die Größenordnung: kleiner, mittlerer oder großer Effekt) angegeben ist, heißt **spezifische Hypothese** („specific hypothesis“). Wird dagegen die Effektgröße nicht spezifiziert, so handelt es sich um eine **unspezifische Hypothese** („non-specific hypothesis“).

Forschungsfragen – Bei Forschungsfragen („research question“) wird ergebnisoffen gefragt, ob ein Effekt existiert. Wenn ja, welche Richtung und Größe sowie welche qualitative Beschaffenheit hat er? Empirische Befunde zu Forschungsfragen dienen der Gegenstandsbeschreibung und Hypothesenbildung.

Das Verhältnis von Forschungsthema, Forschungsproblem sowie dem zugehörigen Satz von Forschungshypothesen bzw. Forschungsfragen wird in Tab. 5.1 noch einmal an je einem realen Studienbeispiel aus dem Themenfeld Computerspiele illustriert.

Mit der Konkretisierung des Forschungsproblems werden bereits erste Entscheidungen hinsichtlich des **Untersuchungsdesigns** getroffen (► Kap. 7). So kristallisiert sich meist schon heraus, ob ein qualitatives, quantitatives oder Mixed-Methods-Design gewählt wird, eine grundlagen- oder anwendungswissenschaftliche Studie durchgeführt werden soll oder welches primäre Erkenntnisinteresse eine empirische Studie verfolgt. Hinsichtlich des Erkenntnisinteresses unterscheidet man drei Typen von Studien:

1. **Explorative Studien** („explorative studies“): Gegenstandserkundende Studien untersuchen ein neues oder bislang nur ungenügend erforschtes Themenfeld und beantworten offene Forschungsfragen mit der Zielsetzung der Gegenstandsbeschreibung sowie der Hypothesen- und Theoriebildung. Sie arbeiten dabei mit quantitativen Methoden (z. B. Suche nach Mustern in einem großen Datensatz mit sehr vielen Variablen und Fällen; Myatt, 2007) und oft mit qualitativen Methoden (z. B. detaillierte Informationssammlung anhand relevanter Einzelfälle aus einer bewusst heterogen zusammengesetzten kleinen Stichprobe, Tab. 5.1 enthält als Beispiel für eine explorative Untersuchung eine Interviewstudie mit älteren Computer-Gamern).
2. **Deskriptive Studien** („descriptive studies“): Populationsbeschreibende Studien widmen sich der Verbreitung und Ausprägung einzelner Merkmale in größeren Grundgesamtheiten bzw. Populationen (z. B. Verbreitung von bestimmten Krankheiten oder politischen Einstellungen in der nationalen Bevölkerung). Nicht theoretische Erklärungen, sondern präzise Messungen stehen im Mittelpunkt. Populationsbeschreibende Studien arbeiten mit aufwändigen, möglichst repräsentativen Stichprobenverfahren und standardisierten Messinstrumenten (Best & Kahn 2005). Als Primärstudien sind sie für Abschlussarbeiten meist zu aufwändig. Allerdings können vorliegende bevölkerungsrepräsentative Großdatensätze im Rahmen einer Abschlussarbeit unter neuen Fragestellungen statistisch ausgewertet werden (sog. Sekundäranalyse, ► Abschn. 7.4).
3. **Explanative Studien** („explanatory studies“): Hypothesenprüfende Studien streben die empirische Prüfung von Forschungshypothesen und somit die Prüfung und Weiterentwicklung derjenigen Theorien an, aus denen die Hypothesen abgeleitet wurden. Explanative Untersuchungen arbeiten oft quantitativ und mit experimentellen oder quasi-experimentellen Designs (Tab. 5.1 enthält als Beispiel für eine explanative Studie ein Experiment zur Spielleistung bei mehr oder minder gewalthaltigen Computer-Games). Teilweise werden auch im Rahmen der qualitativen Forschung Hypothesen geprüft, so dient z. B. das qualitative Experiment (Kleining, 1991) der Prüfung von Hypothesen über Ursache-Wirkungs-Zusammenhänge (zu qualitativen Designs Mayring, 2010).

Neben Grundlagenforschung kommen auch Studien aus dem Bereich der angewandten Forschung bzw. **Auftragsforschung** für Qualifikationsarbeiten infrage, z. B. externe Evaluationen von Maßnahmen oder Produkten (zur Evaluationsforschung ► Kap. 18). So könnte ein Spielentwickler die Evaluation seines neuen Lernspiels im Hinblick auf Spielspaß und Wissenserwerb im Rahmen einer Qualifikationsarbeit untersuchen lassen.

5.3 Sechs Kriterien für die Wahl des eigenen Forschungsthemas

Wer ein Forschungsthema ausgewählt und ein Forschungsproblem eingegrenzt, vielleicht auch schon Ideen für den zugehörigen Satz an Forschungshypothesen bzw. Forschungsfragen entwickelt hat, sollte die eigene Themenwahl vor der endgültigen Entscheidung hinsichtlich folgender **sechs Kriterien** kritisch überdenken.

5.3.1 Persönliches Interesse am Thema

Empirische Studien sind meist zeitaufwändig und arbeitsintensiv. Es ist deshalb von Vorteil, wenn Studierende bzw. Forschende sich Themen zuwenden, deren Bearbeitung sie persönlich **interessiert** und motiviert. Dies hilft auch dabei, Durststrecken im Forschungsprozess zu über-

stehen und die wissenschaftliche Arbeit als sinnvoll zu empfinden.

Eine direkte **persönliche Betroffenheit** vom Forschungsthema (z. B. exzessive Computerspielerin untersucht Computerspielsucht, Veganer untersucht Veganismus) hat Vor- und Nachteile: Einerseits kann Forschung aus der **Insiderperspektive** besonders wertvolle neue Erkenntnisse hervorbringen, weil Fragen aufgeworfen und Daten erhoben werden, zu denen Außenstehende möglicherweise keinen Zugang haben (z. B. ethnografische Feldforschung innerhalb bestimmter Szenen, ▶ Abschn. 10.1.2). Andererseits kann persönliche Betroffenheit für die Forschenden auch als belastend erlebt werden und es ihnen erschweren, sich dem Gegenstand nüchtern, ergebnisoffen und wissenschaftlich fundiert zu nähern, weil bereits festgefügte Einstellungen bestehen. Will man ein Forschungsproblem bearbeiten, bei dem persönliche Betroffenheit besteht, so ist zur Vermeidung der genannten Nachteile eine sorgfältige **Selbstreflexion** notwendig, die in der qualitativen Forschung ohnehin als Teil des Forschungsprozesses gefordert wird (▶ Abschn. 2.3.5).

Schließlich ist bei der Berücksichtigung persönlicher Interessen im Rahmen der Themenwahl auch die weitere akademische und berufliche Entwicklung zu beachten: Im Sinne der **Karriereplanung** ist es sinnvoll, Themen zu wählen, mit denen man sich auch zukünftig verstärkt beschäftigen möchte und die durch ihre inhaltliche Ausrichtung und ihren methodischen Anspruch als Türöffner für Bewerbungen um Stipendien, Studien- oder Arbeitsplätze hilfreich sind. Der Titel der Arbeit, dem das Thema zu entnehmen ist, steht schließlich auch im Abschlusszeugnis und ist somit bei jeder Bewerbung erkennbar.

5.3.2 Theoretische und methodische Vorkenntnisse zum Thema

Für akademische Qualifikationsarbeiten sollten Themen gewählt werden, zu denen bereits theoretische und methodische Vorkenntnisse vorliegen, etwa durch den Besuch entsprechender Lehrveranstaltungen, Praktika oder das Verfassen von Hausarbeiten. Insbesondere die zur Bearbeitung des Forschungsproblems einzusetzenden Methoden der Datenerhebung und Datenanalyse (z. B. Interview, Fragebogen, qualitative oder quantitative Inhaltsanalyse) sollte man im Laufe des Studiums schon mindestens einmal **praktisch geübt** haben.

5.3.3 Wissenschaftliche Relevanz des Themas

Ein Forschungsthema sollte möglichst nicht nur von einem selbst interessant gefunden, sondern in der Scientific Community als bedeutsam erachtet werden. Dies schlägt sich in der Grundlagenforschung dann unter anderem darin nieder, dass andere Forschungsprojekte die eigene Studie zitieren und mit den gewonnenen Erkenntnissen weiterarbeiten. Oftmals wird von Studierenden fälschlich angenommen, dass ein Forschungsthema nur dann von großer wissenschaftlicher Bedeutung ist, wenn es bislang noch gar nicht untersucht wurde. Zwar stimmt es, dass Studien, die dazu dienen, zu historisch neuen oder **bislang vernachlässigten Phänomenen** erstmals wissenschaftliche Daten zu erheben, besonderen empirischen Wert besitzen. Allerdings ist auch eine Arbeit zu einem **gut erforschten Thema** im Sinne kumulativen wissenschaftlichen Erkenntnisgewinns wichtig. Sogar eine exakte Wiederholung einer bereits durchgeführten Untersuchung (**Replikationsstudie**) kann von großem wissenschaftlichem Nutzen sein. Denn nur, wenn eine Reihe von Studien zur selben Fragestellung zu ähnlichen Ergebnissen kommt, gilt ein Befund als gesichert. Im Diskussionsteil oder Ausblick wissenschaftlicher Fachpublikationen finden sich meist konkrete Hinweise dazu, welche weiteren oder vertiefenden Aspekte eines etablierten Themas besonders untersuchungswürdig sind.

Ob man lieber ein neues Phänomen wissenschaftlich erkundet und an der Theoriebildung arbeitet oder einer Forschungstradition folgt und theorieprüfend vorgeht, mag jeweils von der konkreten Forschungsfrage sowie von persönlichen Präferenzen abhängen – wissenschaftliche Relevanz haben beide Strategien.

5.3.4 Praktische Relevanz des Themas

Da sich die empirische Sozialforschung alltagsnahen Sachverhalten widmet, kann neben oder anstelle der wissenschaftlichen Bedeutsamkeit die praktische Relevanz für die Themenwahl ausschlaggebend sein. Praktisch bedeutsam sind Forschungsfragen, die sich ganz konkret auf die Gestaltung von Maßnahmen, das Fällen von Entscheidungen oder die Lösung von Praxisproblemen beziehen. Entsprechende Forschungsfragen spielen in der **Anwendungs- und Auftragsforschung** sowie der **Evaluationsforschung** eine große Rolle. Wird etwa der Frage nachgegangen, wie ein Online-Immobilienportal kundenfreundlicher gestaltet werden könnte, welche Zusatzwünsche die Besucherinnen und Besucher einer psychologischen Beratungsstelle haben oder wodurch sich die aktive Elternarbeit an einer Schule steigern ließe, so sind die für den konkreten Anwendungsfall gewonnenen Erkenntnisse von großer praktischer Bedeutung. Ihr wissenschaftlicher Erkenntnisnutzen im Sinne generalisierbarer Aussagen oder weiterer Theorieentwicklung ist jedoch aufgrund der praxisorientierten Zielsetzung

der Studie oft eingeschränkt. Damit eine praxisorientierte Studie dennoch wissenschaftlichen Charakter behält, ist der Rückgriff auf wissenschaftliche Theorien und Forschungsmethoden notwendig. Praxisorientierte Arbeiten entstehen in der Regel in Kooperation mit **Praxispartnern**, was den Forschenden die Befriedigung gibt, durch ihre Ergebnisse zu greifbaren praktischen Veränderungen beitragen zu können.

Forschungsarbeiten können nicht nur zur Lösung konkreter Praxisprobleme beitragen (z. B. Verbesserung des Spielspaßes eines bestimmten Mathematik-Lernspiels), sondern auch Beiträge zur **Lösung gesamtgesellschaftlicher Probleme** leisten (z. B. Identifikation und Prävention übermäßiger und inhaltlich schädlicher Mediennutzung bei Kindern). Themen aufzugreifen, welche die breite Öffentlichkeit bewegen, ist im Sinne der gesellschaftlichen Verantwortung empirischer Sozialforschung sinnvoll und wünschenswert. Dabei ist zu beachten, dass ein „Hype" um ein bestimmtes Thema auch rasch abflauen kann und dann der Eindruck eines „Modethemas" entsteht. Umgekehrt kann ein Thema, dem in der Scientific Community hohe Relevanz zugeschrieben wird (z. B. weil es in eine empirische Forschungslücke stößt und/oder zur Weiterentwicklung einer Theorie beiträgt), in der breiten Öffentlichkeit als völlig uninteressant und abgehoben gelten.

Abgesehen davon, dass es oft schwierig ist, bei der Beurteilung der Relevanz von Themen die öffentliche Meinung einerseits und die fachwissenschaftliche Betrachtung andererseits „unter einen Hut" zu bringen, kann der selbstgestellte Anspruch überfordern, mit einer eigenen Studie zur Verbesserung der Gesellschaft beizutragen. Aus empirischen Ergebnissen lassen sich in der Regel nicht unmittelbar politische Handlungsempfehlungen ableiten. Es besteht die Gefahr, dass empirische Befunde überinterpretiert werden. Zudem ist zu beachten, dass die hier analytisch getrennt aufgeführten Aspekte der wissenschaftlichen und praktischen (inklusive gesamtgesellschaftlichen) Relevanz eines Themas teilweise schwer voneinander abgrenzbar sind.

5.3.5 Empirische Untersuchbarkeit des Themas

Bei der Konzeption einer empirischen Studie geht es darum, wie möglichst großer Erkenntnisgewinn mit den gegebenen personellen, zeitlichen und finanziellen Mitteln und unter den sonstigen beschränkenden Rahmenbedingungen in der Forschungspraxis erreicht werden kann. Ein erkenntnistheoretisch ideales Wunschdesign lässt sich oft aus mindestens **sieben verschiedenen Gründen** nicht – oder nur beschränkt – umsetzen:

1. **Ethische Grenzen:** Die Forschungsethik (► Abschn. 4.1) verbietet alle Vorgehensweisen, durch die es für Untersuchungsteilnehmende zu unzumutbaren Beeinträchtigungen kommen kann. Besonders heikel sind deswegen z. B. Studien, bei denen die Untersuchungsteilnehmenden zunächst über die eigentlichen Studienziele getäuscht werden oder in denen sie mit negativen Reizen oder belastenden Situationen konfrontiert werden. Auch die verdeckte Beobachtung z. B. in Online-Communitys kann ethische Probleme aufwerfen im Sinne einer wahrgenommenen Verletzung der Privatsphäre. Ethisch besonders heikel sind Studien mit Kindern; hier wird die elterliche Erlaubnis benötigt. Erhebungen in Schulen verlangen meist eine offizielle Genehmigung durch das Schulamt. Diese wird nur gewährt, wenn die Studie einen für das Schulamt erkennbaren pädagogischen Nutzen hat. Entsprechende Einschränkungen aus ethischen Gründen sind bei der Zuspitzung des Forschungsproblems zu berücksichtigen und können z. B. eine Modifikation der Methode verlangen.
2. **Politisch brisante Themen:** Vor allem bei angewandten Forschungsarbeiten, deren Ergebnisse z. B. Entscheidungen in der Praxis beeinflussen können und sollen, haben viele Themen eine nicht zu unterschätzende politische Brisanz für die jeweiligen Anspruchsgruppen (Stakeholder). So wird eine Mitarbeiterbefragung zu Arbeitszufriedenheit und Betriebsklima unter anderem auch von Firmenleitung und Betriebsrat kritisch begleitet werden. Fragen, die aus etablierten Instrumenten (d. h. psychometrischen Skalen oder Tests, ► Abschn. 10.4.3) entnommen und für die Prüfung einer Theorie aus wissenschaftlichen Gründen sinnvoll und notwendig sind, können hier schnell der Kürzung zum Opfer fallen. So könnte eine Firmenleitung alle Fragen streichen lassen, die sich um Entlohnung drehen, weil sie hier hohe Unzufriedenheit befürchtet, die durch die Studie nicht geschürt werden soll. Alle wissenschaftlichen Studien mit Praxispartnern und insbesondere die Evaluationsforschung erfordern besonderes Geschick und Verantwortungsbewusstsein im Umgang mit den verschiedenen Anspruchsgruppen sowohl bei der Entwicklung der Untersuchungsinstrumente (z. B. Fragebögen), der Stichprobenziehung als auch der Interpretation und Veröffentlichung der Ergebnisse. Diese Prozesse sind erfahrungsgemäß sehr viel zeitaufwändiger und unkalkulierbarer als pure Grundlagenforschung, von deren Ergebnissen sich niemand unmittelbar betroffen oder gar bedroht und zur Einmischung berufen fühlt.
3. **Aufwand für die Untersuchungsteilnehmenden:** Viele experimentelle Computerspielstudien lassen die Probanden nur sehr kurze Zeit spielen, bevor dann

schon bestimmte Effekte im Labor gemessen werden. Kurze Spielsessions sind für manche Spielgattungen typisch (sog. Casual Games). Für andere Spielgattungen sind deutlich längere Spielsitzungen jedoch der Normalfall. Warum werden nicht mehr Experimentalstudien mit ausgedehnteren Spielzeiten durchgeführt? Obwohl diese sicher möglich und sinnvoll wären, kann der damit verbundene erhöhte Aufwand für die Probanden eine Grenze darstellen. Wie viele Versuchspersonen wären bereit, tatsächlich mehrere Stunden im Forschungslabor zu bleiben? Oft springen Untersuchungspersonen schon ab, wenn ein Online-Fragebogen mehr als zehn Minuten ihrer Zeit in Anspruch nimmt. Tagebuchstudien, bei denen über eine Woche hinweg regelmäßig Daten in Formulare einzutragen sind und die somit einen sehr guten Einblick in das Alltagsverhalten von Personen bieten (► Abschn. 10.3), leiden unter hohen Abbruchraten. Dennoch existieren Beispiele von Untersuchungen, bei denen die Teilnehmenden auch nach zwei oder drei Stunden die Interviews noch begeistert fortsetzen wollten. Welcher Aufwand von Studienteilnehmenden noch als akzeptabel empfunden wird, hängt stark von der jeweiligen Zielgruppe, vom Thema und von der Datenerhebungsmethode ab. Ob man bei der geplanten eigenen Studie die potenziellen Teilnehmenden hinsichtlich des Aufwands möglicherweise überfordert und dann durch mangelnde Teilnahmebereitschaft oder Abbrüche die Studie gefährdet wird, sollte im Vorfeld eruiert werden.

4. **Aufwand für die Forschenden:** Der Aufwand, den unterschiedliche Methoden der Datenerhebung für die Forschenden selbst mit sich bringen, wird nicht selten unterschätzt. So wären Computerspielstudien mit mehrstündigen Spielsessions nicht nur für die Versuchspersonen möglicherweise zu aufwändig, sondern mit hoher Wahrscheinlichkeit auch für die meisten Forschenden kaum zu bewältigen. Denn bei Laborexperimenten müssen nicht nur die Versuchsleiter während der kompletten Laufzeit der Versuche anwesend sein. Auch ist mit beträchtlichen Ausfallzeiten zu rechnen (wenn Versuchspersonen nicht erscheinen) sowie mit einem hohen Zeitaufwand, um die während des Versuchs erzeugten Audio- und Videodokumente zu sichten und auszuwerten. Damit der Aufwand bewältigbar bleibt, müssen also die Modalitäten der Datenerhebung, die Stichprobengröße sowie die notwendige Zeit für die Datenaufbereitung und Datenanalyse (z. B. Verschriftlichung von Interviews, Sichtung von Videomaterial etc.) im Vorfeld abgeschätzt und aufsummiert werden. Immer wieder kommt es vor, dass Studien weitgehend ergebnislos bleiben, weil wegen fehlerhafter Zeitplanung fast alle Ressourcen in die Datenerhebung geflossen sind und die letztlich entscheidende Datenauswertung und Ergebnisinterpretation dann gar nicht oder nur lückenhaft stattfindet. So manches spannende Forschungsproblem muss aufgegeben oder deutlich modifiziert werden, weil es mit den vorhandenen Ressourcen kaum untersuchbar ist. Im Rahmen der Methodenausbildung steht die wissenschaftliche Qualität von Studien und deren Optimierung im Vordergrund. Dennoch darf gleichzeitig nicht der Aspekt der **Forschungspragmatik bzw. Forschungsökonomie** („research economy“) außer Acht gelassen werden.
5. **Schwer erreichbare Zielgruppen bzw. schwer zugängliche Materialien:** Bei vielen Forschungsthemen sind Expertenbefragungen sehr aufschlussreich (zum Experteninterview ► Abschn. 10.2.2). Personen mit besonderer Fachexpertise sind jedoch oft schwer zu finden oder aufgrund ihrer Arbeitslast kaum erreichbar, Interviewtermine werden dementsprechend oft von vorne herein abgelehnt, fallen häufig aus oder werden mehrfach verschoben. Kleine oder sehr spezielle Zielgruppen untersuchen zu wollen (z. B. Heroinkonsumenten, Chefärztinnen), ist schwierig, es sei denn, man hat guten Zugang zum Feld oder es werden zur Rekrutierung spezielle Stichprobenverfahren erfolgreich eingesetzt (z. B. die Schneeball-Methode, ► Abschn. 9.3.1). Nicht nur Personen, sondern auch Dokumente, die für empirische Studien gebraucht werden, können schwer zugänglich sein (z. B. ältere Ausgaben von Publikumszeitschriften, Aufzeichnungen von TV-Programmen) und eine zeitaufwändige und kostspielige Phase der Materialbeschaffung (z. B. über Online-Auktionen) erfordern, deren Erfolg unsicher ist. Insbesondere bei Abschlussarbeiten mit fixem Abgabetermin und engem Zeitplan sollten derartige Probleme vorab bedacht und ggf. alternative Vorgehensweisen im Sinne eines „Plan B“ entwickelt werden, um den erfolgreichen Abschluss der Arbeit nicht zu gefährden.
6. **Abhängigkeit von Dritten bei der Datenerhebung:** Generell ist Vorsicht geboten, wenn die Möglichkeit zur Datenerhebung stark von der Kooperationsbereitschaft Dritter abhängt. Sei es, dass man auf eine Person angewiesen ist, die beim Programmieren des Online-Fragebogens hilft, auf eine Schlüsselperson, die Kontakte zu Interviewpartnern im Feld zu vermitteln versprochen hat oder auf die Firmenleitung, die das Zuliefern von Dokumenten zu einem bestimmten Termin zugesagt hatte. Erfahrungsgemäß kommt es hier immer wieder zu Verzögerungen. Im Extremfall werden entsprechende Zusagen gar nicht eingehalten. Hier sollte man ebenfalls alternative Vorgehensweisen ins Auge fassen.

7. **Verfügbarkeit technischer Hilfsmittel**: Wenn eine Studie geplant ist, bei der Videoaufzeichnungen erstellt, physiologische Messungen durchgeführt oder Beobachtungen im Labor durch eine Spiegelscheibe vorgenommen werden sollen, muss im Vorfeld sichergestellt werden, dass die entsprechenden Räumlichkeiten und technischen Hilfsmittel im Untersuchungszeitraum tatsächlich zur Verfügung stehen und funktionsfähig sind.

Entsprechende Vorüberlegungen und daraus folgende Entscheidungen zur Vorgehensweise können und sollen durchaus in den Forschungsbericht aufgenommen werden. Dabei ist ein ausdrücklicher Verweis auf forschungspragmatische bzw. forschungsökonomische Gründe zulässig, etwa indem argumentiert wird: „Die Berücksichtigung der Sichtweisen von Kindern zum Thema wäre sehr interessant gewesen, hätte allerdings eine schriftliche Zustimmung aller Eltern erfordert. Da dies zu aufwändig gewesen wäre, wurden in der vorliegenden Studie aus forschungsökonomischen Gründen ausschließlich Jugendliche ab 16 Jahre befragt."

5.3.6 Betreuungsangebot zum Thema

Typischerweise werden in den Sozialwissenschaften **Listen mit Themenangeboten** für Abschlussarbeiten an den jeweiligen Instituten bereitgestellt. Diese Themenvorschläge weisen meist enge Bezüge zu den laufenden Forschungsprojekten und Fachexpertisen der einzelnen Professuren auf. Es ist empfehlenswert, sich gemäß eigener Interessen und Vorkenntnisse ein passendes Thema aus diesen Listen auszusuchen. Teilweise besteht sogar die Möglichkeit, als wissenschaftliche Hilfskraft an größeren **Forschungsprojekten** mitzuwirken und in diesem Rahmen die eigene Abschlussarbeit zu schreiben. Dabei wird man üblicherweise ein Thema bearbeiten, das durch die Projektziele bereits vorgegeben ist. Der Vorteil dieses Vorgehens besteht darin, dass durch die Einbindung der eigenen Arbeit in ein größeres Forschungsvorhaben eine besonders intensive Betreuung erfolgt, man sich regelmäßig im Team austauscht und in der Regel professioneller forscht als im Alleingang. Auch ist hier am ehesten sichergestellt, dass die eigenen Forschungsergebnisse nicht „in der Schublade" landen, sondern in den Projektpublikationen berücksichtigt werden.

Wer ein **eigenes Thema** untersuchen möchte, das nicht im angebotenen Themenspektrum auftaucht, muss sich selbst rechtzeitig darum bemühen, eine entsprechend inhaltlich qualifizierte und motivierte Betreuung am Institut zu finden und gegebenenfalls **externe Betreuende und Gutachtende** hinzuziehen. Externe Betreuer kommen zudem zum Einsatz, wenn Arbeiten zusammen mit Praxispartnern geschrieben werden. Hier ist mit entsprechendem zeitlichem und organisatorischem Vorlauf zu rechnen (Abstimmungen und Absprachen mit allen Beteiligten sowie dem Prüfungsamt etc.).

Generell ist vorab sicherzustellen, dass interne und externe Betreuende zum Wunschthema während der Bearbeitungszeit in geeigneter Weise (z. B. persönlich, telefonisch, per E-Mail) erreichbar sind. Die Betreuung erfolgt nach dem **Holprinzip**, d. h. Studierende müssen sich selbst um entsprechende Beratungsgespräche bei ihren Betreuern bemühen und sollten diese gründlich (ggf. auch schriftlich) vorbereiten. Gefragt ist selbstständiges wissenschaftliches Arbeiten, so dass ständige Detailfragen oder Vorabkorrekturen nicht vorgesehen sind. Sinnvoll ist stattdessen eine genaue Absprache zu Beginn der Arbeit (meist anhand des schriftlichen Exposés) sowie bei wichtigen Zwischenergebnissen (z. B. Fertigstellung des Erhebungsinstruments). Zudem bieten sog. Kandidatenseminare bzw. Forschungskolloquien – die an vielen Instituten verpflichtend sind – die Gelegenheit, den eigenen Arbeitsstand der Bachelor-, Master- oder Doktorarbeit zu präsentieren und im Kollegenkreis zu diskutieren.

5.4 Planung des eigenen Forschungsprojektes

Generell lassen sich empirische Studien als **Projekte** auffassen: Es geht darum, innerhalb eines festgelegten Zeitraums mit begrenzten Ressourcen ein definiertes Ziel zu erreichen, das in der Regel aus mehreren klar abgrenzbaren Teilzielen besteht. Der Erfolg eines Forschungsprojekts leitet sich also nicht nur aus dem wissenschaftlichen Wert des Vorhabens ab, sondern auch aus einem professionellen **Projektmanagement**. Je umfassender das Forschungsprojekt, umso größer ist die Bedeutung des Projektmanagements. Verbund- und Großprojekte mit geografisch verstreuten Forschungsteams und mehrjährigen Projektlaufzeiten richten hierfür teilweise eigene Stellen ein. Bachelor- und Masterarbeiten, die in der Regel allein bearbeitet werden und eine Bearbeitungszeit von üblicherweise sechs Monaten haben, sind sehr viel übersichtlicher.

Vor Anmeldung der Bachelor-, Master- oder sonstigen Qualifikationsarbeit empfiehlt sich eine **Gesamtplanung** des eigenen Forschungsprojektes. Anhand der Sichtung der Themenangebote im Institut sowie einer ersten Literaturrecherche sollte man ein Forschungsthema wählen und das Forschungsproblem eingrenzen, also Entscheidungen hinsichtlich der zu betrachtenden inhaltlichen Aspekte des Gegenstandes, der genutzten Theoriemodelle sowie der präferierten Methoden treffen (▶ Abschn. 5.1

Tabelle 5.2 Grober Zeitplan für eine sechsmonatige empirische Bachelor- oder Masterarbeit

Zeitraum	Wichtige Aktivitäten	Ergebnis
Vorbereitung (ca. 4–6 Wochen)	Erste Literaturrecherche, Beratungsgespräche mit Betreuenden, Begründung des Forschungsthemas und Formulierung des Forschungsproblems, Erstellung des Exposés	Exposé
1. Monat	Vertiefte Literaturrecherche, Auswertung und Strukturierung der Quellen, Definition aller wichtigen Begriffe, strukturierte Darstellung des Forschungsstandes (theoretischer und empirischer Hintergrund der Studie), begründete Formulierung der einzelnen Forschungshypothesen und Forschungsfragen	Literaturverzeichnis (deutsch- und englischsprachige Quellen) sowie erste Manuskriptteile: Einleitung, Gegenstandsbeschreibung, Forschungsstand, Theorie, Satz von ca. ein bis zehn Forschungshypothesen bzw. Forschungsfragen
2. Monat	Entwicklung und Erprobung (Pretest) von Untersuchungsdesign und Untersuchungsinstrumenten	Fertiges und praktisch erprobtes Untersuchungsdesign (z. B. Versuchsablauf) sowie Untersuchungsinstrumente (z. B. Fragebogen, Interview-Leitfaden, inhaltsanalytisches Kategoriensystem)
3. Monat	Datenerhebung und Datenaufbereitung	Vollständiger qualitativer und/oder quantitativer Datensatz, in der Regel elektronisch (z. B. MaxQDA-, SPSS-File)
4. Monat	Datenauswertung: Kodierung und Interpretation von nichtnumerischen Daten bzw. statistische Analyse und Interpretation von numerischen Daten	Vollständige Ergebnisse zu allen Forschungsfragen bzw. Forschungshypothesen
5. Monat	Verschriftlichung des Methodenteils, des Ergebnisteils und des Diskussionsteils der Arbeit, Erstellung von Tabellen und Grafiken	Vollständiges Rohmanuskript
6. Monat	Arbeit mehrfach inhaltlich und sprachlich korrigieren, sorgfältig formatieren, drucken und binden lassen	Fertig korrigierte und gedruckte Arbeit

und ▶ Abschn. 5.2). Dabei ist die Eignung des Themas kritisch anhand der genannten sechs Kriterien der Themenwahl zu reflektieren (▶ Abschn. 5.3). Die Ergebnisse der Vorplanung sind schriftlich in einem **Exposé** festzuhalten, welches das Forschungsthema benennt, das Forschungsproblem formuliert und die Relevanz der Studie begründet. Das Exposé sollte zudem in groben Zügen den Forschungsstand wiedergeben (wichtige Theorien und einflussreiche Studien) und vielleicht schon erste konkrete Forschungshypothesen und/oder Forschungsfragen zum Forschungsproblem aufführen. Das geplante methodische Vorgehen (Untersuchungsdesign, Datenerhebungsmethode/n, Art und Umfang der Stichprobe, Datenauswertungsmethode/n) ist zu skizzieren. Insbesondere sollte dabei der Umfang der geplanten Studie – im Spannungsfeld zwischen möglichst großem Erkenntnisgewinn einerseits und praktischer Durchführbarkeit andererseits – abgesteckt werden. Dem vorläufigen Literaturverzeichnis sollte zu entnehmen sein, dass einschlägige und auch internationale Quellen herangezogen werden. Ein erster Entwurf des Inhaltsverzeichnisses der Arbeit ist nützlich. Nicht zuletzt ist dem Exposé ein wochengenauer Zeitplan beizulegen. Instruktionen zur Abfassung des Exposés und zur Anmeldung von Qualifikationsarbeiten sind den **Prüfungsordnungen der jeweiligen Studiengänge** zu entnehmen.

Einen groben Anhaltspunkt zur zeitlichen Gliederung bietet Tab. 5.2, wobei die Details in Abhängigkeit von den Eigenheiten der jeweiligen Studie variieren können. So wird beim Einsatz einer quantitativen Fragebogenmethode der Zeitaufwand für Datenerhebung und Datenaufbereitung sehr viel geringer ausfallen als beim Einsatz einer qualitativen Interviewmethode, weil sich hier die Dauer der einzelnen Interviews und deren Verschriftlichung jeweils aufsummiert (▶ Abschn. 10.2 zur Interviewmethode). Umgekehrt wäre dafür der Zeitaufwand für die Instrumentenentwicklung beim qualitativen Interviewleitfaden deutlich geringer anzusetzen als bei einem quantitativen psychometrischen Fragebogen, der auf einer gründlichen Suche nach vorliegenden Instrumenten basieren und mehrere Phasen von empirischen Pretests durchlaufen sollte (▶ Abschn. 10.3 zur Fragebogenmethode).

Die im Rahmen von Qualifikationsarbeiten anfallenden **finanziellen Aufwendungen** müssen meist von den Studierenden selbst getragen werden. Dazu zählen Kopier- und Druckkosten, aber auch Telefon- und Reisekosten (z. B. für Interviews) und zuweilen auch Ausgaben

für eine Incentivierung von Untersuchungsteilnehmenden (z. B. in Form von angebotenen Snacks, Verlosungen von Gutscheinen oder Aufwandsentschädigungen für die Anreise).

Lernquiz

1. Grenzen Sie Forschungsthema und Forschungsproblem voneinander ab!
2. Wann entscheidet man sich bei einer Studie für Forschungshypothesen und wann für Forschungsfragen?
3. Welche drei empirischen Studientypen lassen sich gemäß Erkenntnisinteresse unterscheiden?
4. Welche Vor- und Nachteile hat es, ein Thema wissenschaftlich zu untersuchen, zu dem man einen starken persönlichen Bezug hat?
5. Welche Inhalte sollte ein Exposé für eine wissenschaftliche Qualifikationsarbeit enthalten?
6. Was ist gemeint, wenn in einer Publikation zu lesen ist, dass eine bestimmte Methodenentscheidung aus „forschungsökonomischen“ oder „forschungspragmatischen“ Gründen getroffen wurde? Bitte erläutern Sie den Fachbegriff und geben Sie ein realistisches Beispiel!
7. Welche sieben Probleme können die empirische Untersuchbarkeit eines Forschungsthemas einschränken? Bitte die Probleme nennen und kurz erläutern!
8. Was kennzeichnet die praktische (inklusive gesamtgesellschaftliche) Relevanz eines Forschungsthemas und welche Grenzen ergeben sich bei der Bearbeitung?
9. „Einzelkinder sind sehr viel egoistischer als Geschwisterkinder.“ Charakterisieren Sie den Typ dieser Forschungshypothese mit einschlägigen Fachbegriffen und formulieren Sie die zum Hypothesenpaar gehörende Nullhypothese.
10. Welche Vor- und Nachteile hat es, ein bislang stark erforschtes oder sehr wenig erforschtes Thema zu bearbeiten?

Übungsaufgaben

1. Suchen Sie eine qualitative empirische Originalstudie heraus (z. B. aus dem qualitativen Open Access Journal FQS Forum Qualitative Sozialforschung) und geben Sie für diese a) Forschungsthema, b) Forschungsproblem sowie c) den Satz von Forschungsfragen an!
2. Suchen Sie eine quantitative empirische Originalstudie heraus (z. B. aus dem quantitativen Open Access Journal IJIS International Journal of Internet Science) und geben Sie für diese a) Forschungsthema, b) Forschungsproblem sowie c) den Satz von Forschungshypothesen an! Charakterisieren Sie zudem die Forschungshypothesen genauer im Hinblick auf den Hypothesentyp bzw. die Art, Richtung und Stärke der postulierten Effekte!
3. Ihre Hochschule beauftragt Sie mit einer Studie zum Thema Seniorenstudium (damit ist die Teilnahme von älteren Menschen an Lehrveranstaltungen von Hochschulen gemeint). Überführen Sie das Forschungsthema in ein empirisch untersuchbares Forschungsproblem. Erörtern Sie die wissenschaftliche und praktische Relevanz des Themas!

Weiterführende Literatur

- Karmasin, M. & Ribing, R. (2011). Die Gestaltung wissenschaftlicher Arbeiten: Ein Leitfaden für Seminararbeiten, Bachelor-, Master. und Magisterarbeiten sowie Dissertationen. Stuttgart: UTB.
- Töpfer, A. (2010). Erfolgreich Forschen: Ein Leitfaden für Bachelor-, Master-Studierende und Doktoranden. Berlin: Springer.
- White, P. (2008). Developing research questions: A guide for students, researchers and practitioners: A guide for social scientists. New York: Palgrave.

Literatur

Best, J. W. & Kahn, J. V. (2005). *Research in Education*. Boston: Allyn & Bacon.

Bösche, W. (2009). Violent content enhances video game performance. *Journal of Media Psychology: Theories, Methods, Applications, 21*(4), 145–150.

Kleining, G. (1991). Das qualitative Experiment. In U. Flick, E. von Kardoff, H. Keupp, L. von Rosenstiel & S. Wolff (Hrsg.), *Handbuch qualitative Sozialforschung: Grundlagen, Konzepte, Methoden und Anwendungen* (S. 263–266). München: Beltz – Psychologie Verlags Union.

Mayring, P. (2010). Design. In G. Mey & K. Mruck (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (S. 225–237). Wiesbaden: VS Verlag.

Myatt, G. J. (2007). *Making sense of data – A practical guide to exploratory data analysis and data mining*. Hoboken, NJ: Wiley & Sons.

Poskitt, K. & Appleby, S. (1993). Die 99 Lassedasse. Kiel: Achterbahn.

Quandt, T., Grueninger, H., & Wimmer, J. (2009). The gray haired gaming generation: Findings from an explorative interview study on older computer gamers. *Games and Culture, 4*(1), 27–46.

Reinecke, L. (2009). Games and recovery: The use of video and computer games to recuperate from stress and strain. *Journal of Media Psychology: Theories, Methods, Applications, 21*(3), 126–142.

Forschungsstand und theoretischer Hintergrund

Nicola Döring, Jürgen Bortz

6.1 Literaturrecherche – 158

6.2 Forschungsstand – 163

6.3 Theoretischer Hintergrund – 165

6.4 Ableitung von Forschungsfragen und Forschungshypothesen – 173

Literatur – 177

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_6, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Wissen, wie man eine wissenschaftliche Literaturrecherche durchführt.
- Wissen, wie die Ergebnisse einer Literaturrecherche zu verarbeiten sind.
- Den Forschungsstand zu einer geplanten Studie strukturiert darstellen können.
- Den theoretischen Hintergrund einer Studie entwickeln können.
- Die zu einem Forschungsproblem bzw. einer wissenschaftlichen Fragestellung gehörenden Forschungshypothesen und Forschungsfragen ableiten können.

Wissenschaftliche Studien basieren auf einer gründlichen Recherche nach der einschlägigen Fachliteratur (▶ Abschn. 6.1). Ziel ist es, den aktuellen Forschungsstand zu kennen und darzustellen (▶ Abschn. 6.2) sowie den theoretischen Hintergrund der Studie zu entwickeln (▶ Abschn. 6.3). Die Aufarbeitung von Forschungsstand und theoretischem Hintergrund ist notwendig, um den zum Forschungsproblem gehörenden Satz von Forschungsfragen und/oder Forschungshypothesen systematisch abzuleiten und schlüssig zu begründen (▶ Abschn. 6.4).

6.1 Literaturrecherche

Das Literaturverzeichnis einer wissenschaftlichen Arbeit soll keine Zusammenstellung von Zufallsfunden sein, sondern das Ergebnis einer zielgerichteten Suche. Bei einer systematischen Literaturrecherche kommt es darauf an, zunächst die Suchbegriffe festzulegen und mit ihrer Hilfe dann verschiedene wissenschaftliche Datenbanken systematisch zu durchsuchen. Die gefundenen Quellen müssen anschließend gesichtet und strukturiert sowie im Volltext beschafft werden. Die Literaturrecherche gehört zum grundlegenden wissenschaftlichen „Handwerkszeug". Sie wird an Hochschulen u. a. im Rahmen von Lehrveranstaltungen zu Techniken des wissenschaftlichen Arbeitens vermittelt. Auch bieten viele Hochschulbibliotheken entsprechende Kurse an.

6.1.1 Suchbegriffe

Hat man das Forschungsthema ausgewählt und begonnen, das Forschungsproblem einzugrenzen (▶ Abschn. 5.1), so sind für die Literaturrecherche die thematischen **Schlagwörter und Stichwörter** („key words") zusammenzustellen, mit denen nach einschlägigen wissenschaftlichen Quellen gesucht werden soll. Die Suchbegriffe lassen sich in **primäre Suchbegriffe** (Oberbegriffe, allgemeine und übergreifende Begriffe) sowie **sekundäre Suchbegriffe** (Unterbegriffe, spezifische und benachbarte Begriffe) einteilen. Wichtig ist jeweils auch eine englische Übersetzung aller Suchbegriffe, da eine Literaturrecherche neben deutschsprachigen – auch englischsprachige Quellen einbeziehen sollte, um den internationalen Forschungsstand zu erfassen.

Tabelle 6.1 Liste primärer und sekundärer Suchbegriffe zum Forschungsthema Computerspiele (Auszug)

Deutsch	Englisch
Primäre Suchbegriffe (Oberbegriffe)	
Computerspiel	computer game
Digitales Spiel	digital game
Videospiel	video game
Bildschirmspiel	TV game
EGame	electronic game
Konsolenspiel	console game
etc.	etc.
Sekundäre Suchbegriffe (Unterbegriffe)	
Rollenspiel	role playing game
World of Warcraft	World of Warcraft
Egoshooter	first person shooter
Counterstrike	Counterstrike
Lernspiel	educational game (edugame)
Online-Spiel	online game
Handyspiel	mobile game
Gelegenheitsspiel	casual game
Soziales Spiel	social game
etc.	etc.

Hat man z. B. „Computerspiele" als Forschungsthema gewählt, lässt sich vorab eine erste Liste primärer und sekundärer Suchbegriffe zusammenstellen (Tab. 6.1). Die Liste wird im Zuge der weiteren Recherchen verfeinert und ergänzt. Wichtig ist es in jedem Fall, die verwendeten Suchbegriffe zu dokumentieren, damit im Nachhinein das Vorgehen bei der Recherche rekonstruierbar ist. In die Liste können durchaus auch selten verwendete oder veraltet wirkende Begriffe einbezogen werden, um Quellenfunde aus früheren Dekaden oder technikfernen Disziplinen zu ermöglichen.

6.1.2 Datenbanken

Wissenschaftliche Arbeiten stützen sich fast ausschließlich auf wissenschaftliche Quellen und zwar in den Human- und Sozialwissenschaften primär auf Artikel in Fachzeitschriften sowie auf Fachbücher; sie gelten als akademische Primärquellen. Andere Quellen wie z. B. Presseartikel, Sachbücher, Lexika oder Websites sollten dagegen nur im begründeten Ausnahmefall zur Stützung einer wissenschaftlichen Arbeit herangezogen werden.

Es ist verbreitet und nützlich, im Rahmen wissenschaftlicher Arbeit auch in der Internet-Enzyklopädie **Wikipedia** nachzuschlagen (hier sind in der englischen Ausgabe die Beiträge zu sozialwissenschaftlichen Themen meist umfassender als in der deutschen Ausgabe). Ein Wikipedia-Artikel sollte dabei jedoch nur eine Anlaufstelle sein, um wissenschaftliche Primärliteratur zu finden. Wikipedia-Artikel selbst sind im wissenschaftlichen Kontext **nicht zitierfähig**, insbesondere nicht, um Begriffsdefinitionen zu entnehmen. Diese sollten immer der wissenschaftlichen Fachliteratur entstammen, denn Wikipedia-Artikel unterliegen nicht derselben systematischen Qualitätskontrolle wie die wissenschaftliche Fachliteratur. Ebenso sollten Begriffsdefinitionen in wissenschaftlichen Arbeiten nicht dem **Duden** oder einem allgemeinen **Konversationslexikon** bzw. einer Enzyklopädie für die breite Leserschaft entnommen werden, sondern ausschließlich wissenschaftlichen Quellen, etwa themenbezogenen Zeitschriftenartikeln oder **fachwissenschaftlichen Handbüchern, Lexika und Wörterbüchern** (für die Psychologie einschlägig sind z. B. „der Dorsch“: Lexikon der Psychologie und die Buchreihe „Enzyklopädie der Psychologie“).

Wissenschaftliche Fachzeitschriften sowie Fachbücher sind heute zum größten Teil mittels **wissenschaftlicher Literaturdatenbanken** elektronisch verfügbar. über die Hochschulbibliotheken wird den Hochschulmitgliedern inklusive der Studierenden ein kostenloser Zugriff auf diverse Datenbanken – oder zumindest Teilbereiche ihrer Inhalte (z. B. bestimmte Zeitschriften) – angeboten.

Bei der Auswahl der Datenbanken für die eigene Recherche muss man sich a) am lokalen Angebot der Hochschulbibliothek orientieren sowie b) die für die eigene Fachrichtung sowie das Forschungsthema besonders einschlägigen Angebote heraussuchen. Einige bekannte Datenbanken sind ◘ Tab. 6.2 zu entnehmen. Die Datenbanken enthalten neben den Referenzen auf die Quellen (**bibliografische Angaben** plus Kurzzusammenfassung bzw. Abstract) zunehmend auch Volltexte zum Download. Auch Artikel in Fachzeitschriften und Kapitel aus Fachbüchern, Konferenzbeiträge, Dissertationen, technische Berichte und Newsletter sind in den Datenbanken enthalten. Die Sammlungen reichen oft bis zum Jahr 1900 oder noch weiter zurück.

Neben den wissenschaftlichen Abstract-Datenbanken können noch andere Anlaufstellen für die Literaturrecherche nützlich sein, darunter vor allem die Websites von **Wissenschaftsverlagen** (z. B. Hogrefe/Huber: http://www.psycontent.com/psyjournals; Sage: http://online.sagepub.com; Springer: http://www.springerlink.com). Auch der **Online-Buchhandel** bietet Recherchehilfen, so kann man bei Amazon.com oder Amazon.de nach Büchern suchen, deren Inhaltsverzeichnisse sowie Probekapitel und teilweise auch Volltexte inspizieren und Leser-Rezensionen anschauen. Nicht zuletzt können auch Internet-Recherchen mit **Online-Suchmaschinen** insofern sehr ergiebig sein, als dass man u. a. ausgearbeitete themenbezogene Literaturlisten findet, die z. B. im Rahmen von Forschungsprojekten oder Lehrveranstaltungen erstellt wurden. Zum Auffinden von **Online-Bibliografien** werden die inhaltlichen Suchbegriffe mit dem Stichwort „Bibliografie“ bzw. „bibliography“ verknüpft.

Die von vielen Wissenschaftsorganisationen unterstützte **Open-Access-Initiative** setzt sich dafür ein, wissenschaftliche Publikationen im Volltext kostenlos online bereitzustellen (http://www.open-access.net). Diese Initiative steht in Konkurrenz zu den Wissenschaftsverlagen, die Publikationen vermarkten und im Internet teilweise bis zu 20 Euro und mehr pro heruntergeladenem Artikel verlangen. Im Rahmen der Open-Access-Aktivitäten werden die frei zugänglichen Volltexte auf speziellen Dokumentenservern bzw. **Repositorien** abgelegt und verwaltet. Man unterscheidet institutionelle Repositorien (z. B. Dokumentenserver einer Universität) sowie fachbezogene Repositorien (z. B. Dokumentenserver der Sozialwissenschaften: http://www.ssoar.info). Es existieren **Open-Access-Suchmaschinen**, die eine parallele Stichwortsuche über alle bekannten Open-Access-Repositorien zulassen (z. B. BASE: Bielefeld Academic Search Engine: http://www.base-search.net). Bei den Open-Access-Dokumenten handelt es sich um Buchkapitel und Zeitschriftenartikel, um Doktorarbeiten sowie teilweise auch um Forschungsberichte und Qualifikationsarbeiten. Der große Vorteil der Suche in Repositorien besteht definitionsgemäß darin, dass unmittelbar und kostenlos Volltexte zur Verfügung stehen. Nachteilig ist jedoch, dass Repositorien bislang nur einen sehr kleinen Ausschnitt relevanter Publikationen enthalten. Demgegenüber dokumentieren die wissenschaftlichen Literaturdatenbanken Quellen sehr viel umfassender, allerdings mit dem bereits genannten Nachteil, dass die Volltexte – je nach Zugangsberechtigung – oft nicht oder nicht kostenfrei abrufbar sind.

Tabelle 6.2 Wichtige wissenschaftliche Literaturdatenbanken (Auswahl)

Fachrichtung	Datenbank	Anzahl der Referenzen
Sozial- und Geisteswissenschaften, Naturwissenschaften	**Web of Science** (Bestandteil von ISI Web of Knowledge) Anbieter: Thomson Reuters http://www.isiknowledge.com	46 Mio.
Sozial- und Geisteswissenschaften, Medizin, Natur- und Ingenieurwissenschaften	**SciVerse Scopus** (Teil von SciVerse) Anbieter: Verlag Elsevier www.scopus.com	41 Mio.
Psychologie	**PsycINFO** Anbieter: American Psychological Association http://www.apa.org/psycinfo/	47 Mio.
Kommunikationswissenschaft	**CMMC** (Communication & Mass Media Complete) Anbieter: EBSCO Publishing http://search.ebscohost.com	300 000
Erziehungswissenschaft	**ERIC** (Education Resources Information Center) Anbieter: US Department of Education http://www.eric.ed.gov	1.3 Mio.
Medizin	**PubMed** Anbieter: US National Library of Medicine, National Institute of Health http://www.pubmed.gov	20 Mio.
Alle Disziplinen	**Google Scholar** Anbieter: Google http://scholar.google.com	keine Angabe

6.1.3 Datenbankabfrage und Schneeballsystem

Bei der Literatursuche wird die **systematische Abfrage** wissenschaftlicher Literaturdatenbanken mit einer Suche im Schneeballsystem kombiniert. Beim **Schneeballverfahren** sichtet man die Literaturverzeichnisse besonders einschlägiger aktueller Zeitschriftenartikel und übernimmt dort aufgeführte Quellen in die eigene Literatursammlung.

Bei der Datenbankabfrage greift man auf die Liste der Suchbegriffe zurück (▶ Abschn. 6.1.1), arbeitet diese systematisch ab und ergänzt oder kürzt sie. Viele Datenbanken bieten auch einen **Thesaurus** an, in dem inhaltsähnliche Wörter (Synonyme) und verwandte Suchbegriffe angezeigt werden. Grob lassen sich zwei Suchstrategien unterscheiden, die breite Recherche und die eingegrenzte Recherche:

- **Breite Recherche**: Recherchiert man zu einem Thema, zu dem bislang nur wenig wissenschaftliche Literatur zur Verfügung steht, kommt es vor allem auf eine Erweiterung der Suche an: Neben den primären müssen verstärkt auch sekundäre Suchbegriffe einbezogen werden. Durch **Trunkierung** können Begriffe auf einen Wortteil reduziert werden (z. B. „depressi*“), von dem dann alle Variationen gleichzeitig in die Suche einbezogen werden (z. B. depressiv, depressiven, Depression etc.). Auch Quellen, in denen die Suchbegriffe nicht zentral (z. B. im Titel oder im Abstract) vorkommen, sondern nur im Fließtext, sollten bei einer breiten Suche in die Abfrage einbezogen werden. Für eine breite Recherche wird man zudem mehrere Datenbanken nutzen und ggf. auch eine allgemeine Online-Suche mittels Internet-Suchmaschine durchführen. Bei der ergänzenden Suche nach dem Schneeballsystem werden alle Literaturverzeichnisse der gefundenen Quellen auf neue Literaturverweise abgesucht. Zudem werden die Websites der Autorinnen und Autoren gefundener Artikel auf der Suche nach weiteren Quellen (z. B. Forschungsberichte, Konferenzbeiträge) inspiziert.
- **Eingegrenzte Recherche**: Ist dagegen ein Thema Gegenstand der Recherche, zu dem bereits eine sehr große Zahl von wissenschaftlichen Publikationen existiert, muss die Suche von vorne herein sinnvoll eingegrenzt werden. Hierbei sind **Suchbegriffskombinationen** einschlägig, etwa UND-Verbindungen zwischen Suchbegriffen. Zudem sollte vorgegeben werden, dass sich die Suchbegriffe nicht im gesamten Text, sondern nur im Titel, in den Key-

Tabelle 6.3 Trefferzahlen einer exemplarischen Recherche zu Geschlechteraspekten bei Computerspielen in drei verschiedenen wissenschaftlichen Literaturdatenbanken. (Stand: März 2011)

Suchbegriffe	Web of Science (Topic)	PsycINFO (Key Words)	CMMC (Subject Terms)
computer AND games	2 113	2 260	461
computer AND games AND gender	185	116	13
computer AND games AND women	42	10	6
computer AND games AND girls	111	16	1
computer AND games AND men	40	9	1
computer AND games AND boys	110	28	1

CMMC Communication & Mass Media Complete

words oder im Abstract finden lassen dürfen. Weitere Beschränkungen, die über die Suchmasken der Datenbanken einzugeben sind, beziehen sich auf den Zeitraum (z. B. zunächst Beschränkung auf die letzten fünf Jahre) sowie auf die Publikationsart (z. B. nur Artikel aus Fachzeitschriften mit Peer-Review-Verfahren). Bei einem großen Literaturkorpus wird man das Schneeballverfahren auf einige wenige, besonders hochwertige und aktuelle Zeitschriftenbeiträge beschränken.

Angenommen, eine Studie zum Thema Computerspiele soll sich mit Geschlechteraspekten befassen. Die Trefferzahlen einer entsprechenden Datenbankrecherche sind Tab. 6.3 zu entnehmen: Während in der internationalen englischsprachigen Literatur sehr viele Quellen zum allgemeinen Stichwort „Computerspiele“ gefunden werden, grenzt die Kombination mit Geschlechterbegriffen die Menge der Treffer stark ein.

Als Ergebnis der Literaturrecherche entsteht eine Liste aller für das Forschungsthema relevant erscheinenden wissenschaftlichen Quellen. Oft wird man zunächst englischsprachige und dann noch deutschsprachige Quellen suchen. Es ist empfehlenswert, die Literaturliste mit einem **Literaturverwaltungsprogramm** zu erstellen. Hierfür stehen z. B. kostenpflichtige Programme wie Citavi, EndNote, Reference Manager und RefWorks zur Verfügung, die einen großen Funktionsumfang aufweisen. Des Weiteren gibt es kostenfreie Tools wie Mendeley oder Zotero, die der kostenpflichtigen Software hinsichtlich Funktionalität und Nutzerfreundlichkeit in nichts nachstehen. Schließlich lässt sich die recherchierte Literatur auch direkt in Textverarbeitungsprogrammen wie Word und LaTex verwalten. Dies ist jedoch wenig empfehlenswert, denn die Organisation der meist doch recht umfangreichen Literatur mittels speziell entwickelter Literaturverwaltungsprogramme ist insbesondere beim **Import von Rechercheergebnissen aus Datenbanken** komfortabler: Literaturverwaltungsprogramme haben den Vorteil, dass alle gefundenen Datenbankeinträge automatisch in die eigene elektronische Literaturliste übernommen (importiert) werden können, inklusive Abstracts. Zudem kann das Literaturverwaltungsprogramm beim Download von Referenzen aus verschiedenen Datenbanken die entstehende Liste automatisch um Dopplungen bereinigen. Die elektronische Literaturliste kann fortlaufend um Kommentare und Links erweitert und bequem durchsucht werden. Später hilft das Literaturprogramm auch bei der korrekten Erstellung des Literaturverzeichnisses für die Arbeit (zur Ergebnispräsentation ▶ Kap. 13).

6.1.4 Sichtung der Fundstellen

Es ist völlig normal, wenn als Ergebnis der Datenbankabfrage sowie der Suche im Schneeballverfahren eine Liste von mehreren Hundert englisch- und deutschsprachigen Quellen zum Forschungsthema vorliegt. Nachdem mithilfe des Literaturverwaltungsprogramms die Dubletten gesucht und gelöscht wurden, geht es darum, anhand der Titel und Abstracts unwichtige Quellen zu erkennen und aus der Liste zu streichen. Dazu gehören z. B. veraltete Artikel, kurze Editorials, thematisch oder fachlich zu weit entfernte Beiträge.

Neben der Eliminierung unbrauchbarer Quellen geht es bei der ersten Sichtung auch darum, die nützlichsten Quellen zu identifizieren (Tab. 6.4), denn diese haben

Tabelle 6.4 Ausgewählte besonders nützliche Quellen zu Geschlechteraspekten bei Computerspielen als Ergebnis einer exemplarischen Literaturrecherche. (Stand: März 2011)

Art der Quelle	Exemplarische Quellen mit bibliografischen Angaben
Review-Zeitschriftenartikel	Jenson, J. & Castell, S. (2010). Gender, Simulation, and Gaming: Research Review and Redirections. *Simulation & Gaming*, *41*(1), 51–71.
Review-Buchkapitel	Trepte, S. & Reinecke, L. (2010). Gender und Games – Medienpsychologische Gender-Forschung am Beispiel Video- und Computerspiele. In G. Steins (Hrsg.) *Handbuch Psychologie und Geschlechterforschung* (S. 229–248). Wiesbaden: VS Verlag.
Sammelbände	Cassell, J. & Henry Jenkins, H. (1998) (Eds.). *From Barbie to Mortal Kombat: Gender and Computer Games*. Cambridge: The MIT Press.
	Kafai, Y., Heeter, C.; Denner, J., & Sun, J.Y. (Eds.) (2008). *Beyond Barbie and Mortal Kombat: New Perspectives on Gender and Gaming*. Cambridge: The MIT Press.
Monografien	Graner-Ray, S. (2004). *Gender Inclusive Game Design. Expanding the Market.* Hingham: Charles River Media.
Aktuelle Forschungsbeiträge	Sonderheft „Women in Games“ des *International Journal of Gender, Science and Technology 2011 3*(1) siehe: http://genderandset.open.ac.uk

bei der Literaturbeschaffung und inhaltlichen Bearbeitung Priorität. Dazu zählen folgende Arten von Quellen:

- **Übersichtsartikel („Research Reviews“)**, die den Forschungsstand zum Thema zusammenfassen und in Form von Zeitschriftenartikeln oder Buchkapiteln erscheinen,
- **Metaanalysen**, die vorliegende statistische Befunde zu einem bestimmten Effekt aus dem Themenbereich statistisch aggregieren,
- **aktuelle Monografien und Sammelbände** mit großer Nähe zum Forschungsproblem,
- **aktuelle theoretische und/oder empirische Zeitschriftenbeiträge** mit großer Nähe zum Forschungsproblem, die teilweise gebündelt in Sonderheften zu finden sind, und nicht zuletzt
- **Pionierstudien bzw. „Klassiker“**, in denen ein Ansatz erstmals oder besonders richtungweisend publiziert wurde. Die Verarbeitung dieser grundlegenden Quellen kann vor Fehl- oder Uminterpretationen in der späteren Literatur schützen und zollt zudem den Wegbereiterinnen und Wegbereitern der jeweiligen Forschungstradition Respekt.

6.1.5 Literaturbeschaffung

Die Literaturrecherche bezieht sich zunächst auf eine Sammlung von bibliografischen Angaben und Abstracts. Für die laut Sichtung besonders einschlägigen Quellen müssen die Volltexte besorgt werden. Folgende **Beschaffungswege für Volltexte** sind einzuschlagen:

1. **Datenbanken**: Die in Tab. 6.2 genannten wissenschaftlichen Literaturdatenbanken bieten zum Teil Volltexte direkt zum Download an (der Zugang ist abhängig von den Lizenzverträgen der lokalen Hochschulbibliotheken).
2. **Repositorien**: Die Dokumentenserver der Open-Access-Initiative (http://www.open-access.net) stellen wissenschaftliche Volltexte kostenfrei bereit und lassen sich über eigene Suchmaschinen (z. B. http://www.base-search.net) parallel durchsuchen.
3. **Internet**: Steht der Volltext nicht über Datenbanken oder Repositorien zur Verfügung, ist eine Eingabe des Publikationstitels in eine Online-Suchmaschine empfehlenswert. Zahlreiche Volltexte sind im Internet verstreut zu finden, z. B. auf den persönlichen Homepages von Forschenden oder Forschungseinrichtungen.
4. **Lokaler Bibliotheksbestand:** Recherchierte Quellen können über den Offline- und/oder Online-Bestand (z. B. Print- und/oder Online-Zeitschrift) der lokalen Bibliothek zugänglich sein. Hierfür ist der elektronische Bibliothekskatalog zu konsultieren.
5. **Elektronische Zeitschriftenbibliothek (EZB)**: Viele elektronisch zugängliche Fachzeitschriften sind über die von der Universität Regensburg verwaltete Elektronische Zeitschriftenbibliothek EZB verlinkt (http://www.bibliothek.uni-regensburg.de/ezeit/ezb.phtml).
6. **Fernleihe**: Vor Ort nicht verfügbare Texte können über Fernleihe bestellt werden, was mehrere Wochen in Anspruch nehmen kann, da die Quellen aus anderen Bibliotheken verschickt werden.

7. **Dokumentenlieferdienste**: Für besonders schwer beschaffbare oder dringend benötigte Quellen bietet sich die kostenpflichtige Nutzung eines Dokumentenlieferdienstes an, der binnen kurzer Zeit Kopien liefert (z. B. Ingenta: http://www.ingentaconnect.com; Subito: http://www.subito-doc.de).
8. **Buchhandel**: Bücher lassen sich neu und gebraucht über den Online-Buchhandel beziehen, wobei über Online-Antiquariate auch nicht mehr lieferbare Bücher erstanden werden können (z. B. http://www.amazon.de; http://www.booklooker.de). Bezugsquellen für ein gesuchtes neues oder gebrauchtes Exemplar sind entsprechenden Bücher-Suchmaschinen zu entnehmen (z. B. http://www.findmybook.de).
9. **Autorinnen und Autoren**: Im Notfall kann man sich per E-Mail an die Autoren selbst wenden und eine Artikelkopie anfragen. Aufgrund der hohen Arbeitslast vieler Forschender haben hier studentische Anfragen jedoch eine deutlich geringere Bearbeitungswahrscheinlichkeit als Anfragen aus dem Kollegenkreis.

Immer wieder passiert es, dass sich einzelne Quellen als extrem schwer zugänglich erweisen. In diesem Fall sollte eventuell auf den entsprechenden Text verzichtet werden. Eine Verzögerung des gesamten Projekts, bloß weil eine Fernleihe noch nicht eingetroffen ist, dürfte in den meisten Fällen inhaltlich nicht gerechtfertigt sein.

6.2 Forschungsstand

Jede wissenschaftliche Studie sollte auf den aktuellen Forschungsstand („state of research“) eingehen, um das gewählte Forschungsproblem bzw. die Fragestellung der Untersuchung zu verorten (► Abschn. 5.2). Die Darstellung des Forschungsstandes kann mehr oder minder ausführlich geschehen – in wenigen Sätzen oder auf mehreren Seiten. Explorative und deskriptive Studien begründen ihren Satz an Forschungsfragen mit Verweisen auf frühere Studien sowie Forschungslücken. Demgegenüber werden Forschungshypothesen in explanativen Studien meist aus Theorien abgeleitet (► Abschn. 6.4.4 und ► Abschn. 6.4.5).

Bei der Darstellung des Forschungsstandes sollten folgende Fragen beantwortet werden:

- **Wie hat sich die Forschung zum Thema in den letzten Jahren bzw. Jahrzehnten entwickelt?** Seit wann existiert das Phänomen? Wie umfangreich wird es seitdem erforscht? Welche führenden Autorinnen und Autoren sowie Forschungsgruppen sind auszumachen? In welchen Ländern und an welchen Forschungseinrichtungen konzentrieren sich Forschungsaktivitäten zum betrachteten Thema? Welches sind die einschlägigen Fachzeitschriften und Konferenzen für das Themenfeld? (Dabei muss man sich immer bewusst sein, dass wir die internationale Forschungslandschaft meist nur in den Ausschnitten wahrnehmen, die in englischer Sprache zugänglich sind.)
- **Welche zentralen Forschungslinien und Befunde zum Thema sind auszumachen?** Zu welchen Aspekten des Untersuchungsgegenstandes liegen Studien vor? Auf welchen Theorien und Methoden basieren sie? Was sind gesicherte Erkenntnisse? Welche Fragen werden kontrovers diskutiert? Wo befinden sich Forschungslücken?
- **Welche Anwendungsfelder hat die Forschung zu diesem Themengebiet?** Welche praktischen Schlussfolgerungen werden aus den wissenschaftlichen Befunden gezogen? Inwiefern werden aus den Forschungsergebnissen bereits Praxismaßnahmen abgeleitet? Welche Anwendungsfragen sind wissenschaftlich noch ungeklärt?
- **Wie ist der Forschungsstand zum Thema insgesamt zu beurteilen?** Umfang, Qualität und Ausrichtung der bisherigen Forschungsbemühungen zum Untersuchungsgegenstand gilt es zusammenfassend zu bewerten.
- **Wie ordnet sich die eigene Studie in den Forschungsstand ein?** Abschließend ist die eigene Studie zu verorten, etwa dahingehend, inwiefern sie etablierte Forschungstraditionen weiterführt, theoretische oder methodische Innovation anstrebt oder dazu beitragen möchte, Forschungslücken zu schließen.

Wichtig ist es, den Forschungsstand klar strukturiert und verdichtet aufzubereiten – und schließlich auch kritisch zu bewerten. Eine lineare Darstellung und pure Nacherzählung ausgewählter Forschungsbeiträge in der Form: In Studie A wurde das untersucht, Studie B hat das herausgefunden, Autorin C hat das festgestellt, Autor D hat das gesagt etc. ist als Präsentation des Forschungsstandes unbrauchbar. Hier fehlt die Leistung einer sinnvollen Strukturierung und Bewertung.

Wir empfehlen deswegen dringend, bei der Literaturrecherche als erstes die zentralen Quellen und darunter vor allem aktuelle Review-Beiträge zu identifizieren und durchzuarbeiten, da diese bereits **Aufarbeitungen des Forschungsstandes** enthalten. Entsprechende Darstellungen sind vergleichend zu betrachten und ggf. zu ergänzen, um dann für die eigene Studie zu einer angemessenen Präsentation des Forschungsstandes zu kommen. Strukturieren lässt sich der bisherige Forschungsstand unter anderem chronologisch, nach Theorien, nach methodischen Ansätzen oder nach thematischen Aspekten.

Bei der Erarbeitung des Forschungsstandes wird man zunächst deduktiv vorgehen, d. h. Strukturierungskriterien

Tabelle 6.5 Grobstruktur der Darstellung des Forschungsstandes am Beispiel Geschlechteraspekte bei Computerspielen

Darstellung des Forschungsstandes: Wichtige Elemente und möglicher Aufbau	Beispiel: Forschungsstand zu Geschlechteraspekten bei Computerspielen
Allgemeiner historischer Rückblick auf die Entwicklung des Gegenstandes (hier: Computerspiele) sowie der zugehörigen Forschung. Dies kann ggf. unterlegt werden durch quantitative Ergebnisse der Literaturrecherche sowie Hinweise auf die Institutionalisierung des Forschungsgebietes.	Computerspiele wurden in den 1980er-Jahren mit der Verbreitung von Personalcomputern populär. Doch ein Forschungsboom setzte erst zur Jahrtausendwende ein: Die Literaturdatenbank PsycINFO weist für die Dekade von 1980–1989 insgesamt 177 wissenschaftliche Publikationen zu Computerspielen nach. Diese Zahl verdoppelte sich auf 350 Veröffentlichungen zwischen 1990–1999 und verachtfachte sich schließlich auf 1 368 in den 2000er-Jahren. Mit eigenen Fachzeitschriften, Konferenzreihen, Forschungsinstituten und Studiengängen haben sich Computerspiele als Mainstream-Thema in den Geistes-, Human- und Sozialwissenschaften etabliert. Sie gelten als wichtige Unterhaltungsmedien unserer Zeit und haben neben ihrer psychologischen und soziokulturellen insbesondere auch eine große wirtschaftliche Bedeutung.
Kennzeichnung und Strukturierung der **wichtigsten Forschungslinien und Befunde** mit Bezug zum gewählten Forschungsproblem (hier: Geschlechteraspekte bei Computerspielen). Dabei sollten zentrale Publikationen zitiert werden.	Computerspiele sind kein geschlechtsneutraler Gegenstand. Im Gegenteil: Verglichen mit anderen Unterhaltungsmedien (z. B. Musik, Kino) scheinen sich bei Computerspielen besonders deutliche Geschlechtereffekte zu zeigen und zwar in allen vier Phasen bzw. Bereichen des medialen Kommunikationsprozesses: **1. Entwicklung von Computerspielen:** Die Spielentwicklung stellt sich als Männerdomäne dar, es arbeiten international kaum Frauen als Game- Designerinnen (Consalvo, 2008; Prescott & Bogg, 2011). **2. Inhalte von Computerspielen:** Neben geschlechtsneutralen Spielgattungen (z. B. Denkspiele) existieren viele Spiele, die Geschlechterstereotype enthalten. Dies betrifft das Aussehen sowie die Rollenverteilung, etwa wenn männliche Charaktere als schwer bewaffnete Helden, weibliche dagegen als leicht bekleidete Randfiguren auftreten (Miller & Summers, 2007). **3. Nutzung von Computerspielen:** Jungen und Männer verbringen insgesamt deutlich mehr Zeit mit Computerspielen und greifen viel stärker als Spielerinnen auf gewalthaltige Spielgattungen (z. B. Ego-Shooter) zurück (Trepte & Reinecke, 2010). **4. Wirkung von Computerspielen:** Aufgrund ihrer intensiveren Nutzung profitieren Jungen und Männer einerseits stärker von positiven Wirkungen (z. B. Entwicklung von Medien- und Computerkompetenz; Sanford & Madill, 2007), sind aber gleichzeitig auch vermehrt von negativen Effekten betroffen (z. B. Aggressionssteigerung: Konijn, Nije Bijvank, & Bushman, 2007; Computerspielsucht: Peukert, Sieslack, Barth, & Batra, 2010; Verschlechterung von Schulleistungen: Weis & Cerankosky, 2010).
Darstellung der **Anwendungsfelder** und praktischen Konsequenzen der bisherigen Forschung	Die in der Forschung belegten Geschlechterunterschiede werden hinsichtlich des Nutzungsverhaltens von Mädchen und Jungen als Problem betrachtet. Praxismaßnahmen zum Abbau geschlechterdifferenter Nutzungsmuster verfolgen dabei primär zwei Zielsetzungen: **1. Intensivere Computerspielnutzung bei Mädchen:** Diegeringere Beteiligung von Mädchen und Frauen am Computerspielen schließt sie von entsprechenden Gratifikationen, Kompetenzen und Karrieremöglichkeiten in der Spielebranche aus, weshalb im Sinne der Gleichstellung eine Förderung weiblicher Beteiligung für wünschenswert gehalten wird. Mädchen und Frauen als Zielgruppen stärker anzusprechen, liegt gleichzeitig im ökonomischen Interesse der Computerspiel-Industrie. Zum einen werden medienpädagogische Projekte speziell für Mädchen durchgeführt, die Spielinteresse wecken und Kompetenz fördern sollen (Hayes, 2011). Zum anderen wird seit den 1990er-Jahren untersucht, ob und wie mädchengerechte Spiele zu entwickeln sind, die an geschlechtsspezifische Fähigkeiten und Vorlieben anknüpfen (Cassell & Henry Jenkins 1998; Graner-Ray, 2004). **2. Moderatere Computerspielnutzung bei Jungen:** Die starke Beteiligung von Jungen und Männern am Computerspielen erhöht tendenziell ihre Vulnerabilität für negative Effekte wie Sucht oder Aggressivität. Als Gegenmaßnahme wird eine stärkere rechtliche Regulierung von Spielen mit Gewaltinhalten sowie erhöhtem Suchtpotenzial diskutiert. Daneben werden pädagogische Interventionen für Jungen und deren Eltern empfohlen, die zu einem bewussten, kritischen und dosierten Computerspielgebrauch anleiten sollen (Kirsh, 2010; Sax, 2009).

aus bereits vorliegenden Darstellungen des Forschungsstandes entnehmen. Diese werden dann induktiv anhand des Durcharbeitens der eigenen Literaturliste ergänzt und verfeinert. Dabei werden anhand der Titel und Abstracts der gesammelten Referenzen die Beiträge zu inhaltsähnlichen Gruppen zusammengestellt. Hier ist zu erproben, ob man selbst besser mit **digitalen oder ausgedruckten Dokumenten** arbeiten kann. Um große Mengen an

Tabelle 6.5 (Fortsetzung)

Darstellung des Forschungsstandes: Wichtige Elemente und möglicher Aufbau	Beispiel: Forschungsstand zu Geschlechteraspekten bei Computerspielen
Kritische Beurteilung der bisherigen Forschung, vor allem mit Blick auf Theorien und Methoden. Aufzeigen aktueller Kontroversen und Forschungslücken	Die bisherige Forschung konzentriert sich stark auf Unterschiede zwischen männlichen und weiblichen Spielern, die zum Teil **biologisch** (z. B. schlechteres räumliches Vorstellungsvermögen bei Frauen), zum Teil **sozialisationsbedingt** (z. B. geringere Affinität zu Wettbewerb bei Frauen) erklärt werden (Trepte & Reinecke, 2010). Auf der Basis **konstruktivistischer Geschlechtertheorien i**st zu monieren, dass die Forschung, indem sie ein starres, binäres Geschlechtermodell zugrunde legt, ihrerseits Geschlechterstereotype reproduziert (Jenson & Castell, 2010). Insbesondere folgende Aspekte wurden bislang ungenügend untersucht: **1. Gemeinsamkeiten zwischen Spielerinnen und Spielern** (z. B. hinsichtlich der Qualitätsbewertung von Spielen; Lazarro, 2008). **2. Geschlechtsrollenkonträres Spielverhalten und Spielerleben** (z. B. Mädchen, die ausdrückliches Interesse an Gewaltinhalten bekunden, Jungen, die sich von hypersexualisierten männlichen Charakteren in ihrem körperbezogenen Selbstbild beeinträchtigt fühlen; Barlett & Harris, 2008; männliche Nichtspieler, weibliche Vielspielerinnen; Krause, 2010). **3. Intersektionalität** (das Zusammenwirken von Geschlecht mit anderen Personenvariablen wie Alter, Ethnizität, Bildungsstand, Persönlichkeitsmerkmalen etc. bei der Entwicklung des Spielverhaltens). **4. Umwelt- sowie Situationsfaktoren** (z. B. massenmediale Images von Spielerinnen und Spielern, elterliche Computerspielkontrolle bei Töchtern und Söhnen, Reaktionen von Peers auf Spielinteressen von Mädchen und Jungen etc.).
Einordnung des ausgewählten Forschungsproblems in den Forschungskontext	Die vorliegende Studie erforscht das Computerspiel „Die Sims 3“, eines der meistverkauften Spiele mit einem hohen Anteil an Nutzerinnen (Jansz, Avis, & Vosmeer, 2010). Konkret soll untersucht werden, was den Spielerinnen und Spielern an dem Spiel gefällt, ob und inwiefern sie welche Spielaktivitäten als weiblich/feminin oder männlich/maskulin wahrnehmen sowie ob und wie sie diese Spielaktivitäten selbst ausführen. Damit wird auf der Basis einer sozialkonstruktivistischen Gendertheorie der Frage nachgegangen, ob und wie durch den Umgang mit Computerspielen Geschlechteridentitäten ausgedrückt werden.

Quellen zu verarbeiten müssen Volltexte zunächst selektiv gelesen werden: Neben Titel und Abstract sind Einleitung und Diskussion besonders aufschlussreich, weil hier die Motivation der Studie erläutert sowie die wichtigsten Ergebnisse hervorgehoben und kritisch reflektiert werden.

Eine mögliche Grobstruktur für die Darstellung des Forschungsstandes zeigt Tab. 6.5 und illustriert diese an einem Beispiel. Das Augenmerk liegt hierbei auf der Strukturierung. In der Forschungspraxis stellt man die einzelnen Aspekte meist noch ausführlicher dar und belegt sie mit einer größeren Zahl an Quellen.

6.3 Theoretischer Hintergrund

Die eigene Studie sollte nicht nur in den bisherigen Forschungsstand eingeordnet werden, sondern auch theoretisch fundiert sein. Dies gilt insbesondere für **explanative (hypothesenprüfende) Untersuchungen**, bei denen aus der Theorie abgeleitete Forschungshypothesen empirisch getestet werden. Aber auch **explorative (gegenstandserkundende) Studien,** die offene Forschungsfragen empirisch beantworten, um den Gegenstand besser zu erfassen und Theoriebildung vorzubereiten, greifen oft auf etablierte theoretische Konzepte und Modelle zurück. In der qualitativen Forschung spricht man in diesem Zusammenhang unter anderem von „sensibilisierenden Konzepten“ (Blumer, 1973). So basieren beispielsweise viele Studien in der qualitativen bzw. ethnografischen Feldforschung auf der Theorie des symbolischen Interaktionismus, die sich zur Erklärung zahlreicher Alltagsphänomene eignet (▶ Abschn. 10.1.2). Schließlich sind auch **deskriptive (populationsbeschreibende) Studien** theoretisch unterlegt. Eine bevölkerungsrepräsentative Erhebung zur Verbreitung von „Depression“, „Politikverdrossenheit“ oder „Internetsucht“ setzt eine entsprechende theoretische Konzeptualisierung der jeweiligen Variablen voraus, um sie messbar zu machen (zur Operationalisierung ▶ Kap. 8).

In der Regel identifiziert man im Rahmen der Literaturrecherche eine oder sogar eine Reihe von etablierten wissenschaftlichen Theorien, die auf das gewählte Forschungsproblem anwendbar sind. Bei der **Erarbeitung des theoretischen Hintergrundes** der eigenen Studie sind folgende sechs Vorgehensweisen typisch, die wir im Folgenden genauer darstellen:

1. die Durchführung einer **reinen Theoriearbeit**, die verschiedene mehr oder minder etablierte Theorien vergleichend und bewertend einander gegenüberstellt und späteren empirischen Studien als Grundlage dient,
2. die Überprüfung und **Weiterentwicklung einer einzelnen Theorie** auf der Basis empirischer Daten,
3. die Formalisierung einer vorliegenden Theorie durch **Computersimulation**,
4. die Verknüpfung mehrerer Theorien zu einem **integrierten Theoriemodell**, das einer empirischen Prüfung unterzogen wird,
5. die **gegenstandsverankerte Theoriebildung** mithilfe qualitativer empirischer Methoden und
6. die **Entwicklung neuer Theorien** mit unterschiedlichen, mehr oder minder systematischen Strategien.

Die grundlegende Bedeutung von Theorien für den wissenschaftlichen Erkenntnisgewinn sowie Anforderungen an gute wissenschaftliche Theorien sind Gegenstand **wissenschaftstheoretischer Überlegungen** (▶ Kap. 2), die im Zuge der eigenen Theoriearbeit mitbedacht werden sollten. Im Sinne der **Wissenschaftlichkeit** als zentralem Qualitätskriterium empirischer Sozialforschung (▶ Abschn. 3.1) ist es notwendig, nicht nur den Forschungsstand und die vorliegenden Theorien zum gewählten Forschungsproblem umfassend zu recherchieren und zur Kenntnis zu nehmen, sondern die eigene Theoriearbeit ihrerseits zu reflektieren und deren Stärken und Schwächen zu diskutieren.

6.3.1 Durchführung einer reinen Theoriearbeit

Studien, in denen keine empirischen Daten erhoben werden, bezeichnet man als Literatur- bzw. Theoriearbeiten. Dazu zählen Review-Beiträge, die ausschließlich den Forschungsstand zu einem Thema resümieren und bewerten, sowie Metaanalysen, die die Ergebnisse quantitativer Studien zum gleichen Effekt statistisch zusammenfassen (▶ Kap. 16). Zudem werden Theoriearbeiten verfasst, die verschiedene Theorien vergleichend gegenüberstellen und/oder ein Theoriemodell weiterentwickeln oder neu entwerfen. Theorievergleichende Arbeiten als Sonderform des Forschungsüberblicks und insbesondere theorieentwickelnde Arbeiten erfordern umfassendes Überblickswissen und sind für Neulinge deswegen eher ungeeignet.

Bei der Einordnung von Theorien wird deren Reichweite differenziert; damit ist der Umfang des Gegenstandsbereiches gemeint, auf den die Theorie anwendbar ist: **Theorien geringer Reichweite** bzw. gegenstandsnahe Theorien rekonstruieren einzelne Sachverhalte in ihrem jeweiligen Kontext sehr differenziert. Gegenstandsnahe Theoriebildung spielt in der qualitativen Forschung eine wichtige Rolle (▶ Abschn. 6.3.5). In der quantitativen empirischen Sozialforschung werden meist **Theorien mittlerer Reichweite** geprüft („middle range theory"; Merton, 1949). Für Theoriearbeiten ist dagegen auch eine Betrachtung von **Großtheorien** bzw. **Metatheorien** von besonderer Bedeutung, d. h. von Theorien mit hohem Allgemeinheitsgrad, unter die sich dann jeweils diverse Theorien mittlerer und geringer Reichweite subsummieren lassen. Stichpunktartig sei auf einige der in den Sozialwissenschaften relevanten Großtheorien verwiesen:

- **Psychoanalyse**, **Behaviorismus** und **Kognitivismus** sind wichtige Großtheorien in der Psychologie mit unterschiedlichen Menschenbildern. Ihnen lassen sich entsprechende Entwicklungstheorien, Lerntheorien oder Therapiemodelle zuordnen.
- **Konstruktivismus** und **Instruktionismus** sind zentrale Großtheorien in der Erziehungswissenschaft, denen jeweils unterschiedliche Theorien des Lehrens und Lernens bzw. Unterrichts entsprechen.
- Menschliches Erleben und Verhalten führt die **Evolutionstheorie** auf biologische Ursachen zurück, während die **Sozialisationstheorie** kulturelle Faktoren in den Blick nimmt.
- Während die **Handlungstheorie** soziale Sachverhalte als Ergebnis der Handlungen einzelner Akteure analysiert, betrachtet die **Systemtheorie** sie als Resultat überindividueller Strukturen und Prozesse. Systemtheorien lassen sich auf der Makroebene anwenden, so erklärt z. B. Luhmann (1984) die Gesellschaft als ein komplexes System von Kommunikationen. Systemtheorien sind aber auch auf Phänomene auf der Mikroebene, wie z. B. Familien, anwendbar: Die systemische Familientherapie (z. B. Madanes & Haley, 1977) betrachtet Symptome von Familienmitgliedern (z. B. „Schulschwänzen" des Kindes) unter der Perspektive, welche Funktionen diese Symptome für das Gleichgewicht des Familiensystems haben könnten (z. B. Entschärfung der Eheprobleme der Eltern, die sich nun gemeinsam um das Problem des Kindes kümmern müssen). Eine handlungstheoretische Erklärung des Schulabsentismus würde dagegen die Motive des Kindes in den Blick nehmen, das z. B. aus Angst vor Mobbing die Schule meiden könnte.

Für eine theorievergleichende Arbeit können die Sichtweisen verschiedener Großtheorien auf einen ausgewählten Gegenstand systematisch kontrastiert und im Hinblick auf ihre Stärken und Schwächen diskutiert werden. Im Rahmen einer ausgewählten Großtheorie lassen sich auch

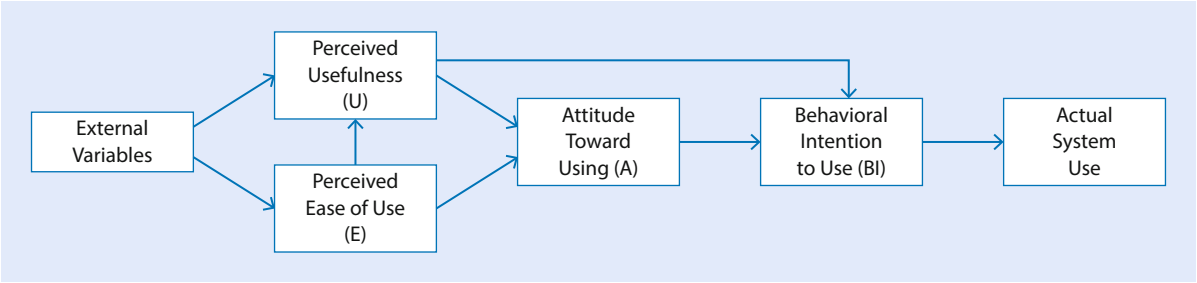


Abb. 6.1 Die erste Fassung des Technology Acceptance Model TAM. (Aus Davis, Bagozzi, & Warshaw 1989. Reprinted by permission, Davis, F. D., Bagozzi, R. P., & Warshaw P. R., User acceptance of computer technology: A comparison of two theoretical models. *Management Science,* volume 35, number 8, August 1989. Copyright 1989, the Institute for Operations Research and the Management Sciences [INFORMS], 7240 Parkway Drive, Suite 300, Hanover, MD 21076 USA)

neue theoretische Erklärungsansätze für einen bestimmten Sachverhalt entwickeln. Dabei können die historische Entwicklung von Großtheorien sowie die Kontroversen zwischen ihren Vertretern einbezogen werden.

Neben der Einordnung und Reflexion von Theorien geringerer und mittlerer Reichweite im Rahmen von etablierten Großtheorien, stellt die **grafische Darstellung** von Theoriemodellen ein wichtiges Hilfsmittel der Theoriearbeit dar. Grafische Darstellungen tragen dazu bei, die Struktur von Theorien sichtbar und vergleichbar zu machen sowie Inkonsistenzen und Lücken aufzudecken. Die Visualisierung kann sich dabei auf Teile von Theorien beziehen, etwa wenn ein zentrales theoretisches Konstrukt mit seinen Unterdimensionen in einem **konzeptuellen Modell** dargestellt wird. So lässt sich zwar die Psychoanalyse als Großtheorie nicht komplett visualisieren, aber das Strukturmodell des psychischen Apparats mit den Instanzen Es, Ich und Überich (Freud, 1923) lässt sich grafisch veranschaulichen.

Zu unterscheiden ist zwischen statischen (strukturellen) Modellen, die den Aufbau von Objekten oder Systemen darstellen, und **dynamischen (funktionalen, systemischen) Modellen**, die Prozesse und Wirkungszusammenhänge beschreiben. Ein typisches dynamisches Modell ist das Flussdiagramm, das einen zeitlichen Ablauf in seinen wichtigsten Stationen und Verzweigungen beschreibt. Pfeildiagramme (Pfadmodelle) veranschaulichen dagegen Kausalbeziehungen und zwingen dazu, sich über Wirkungsrichtungen Gedanken zu machen. Theorien mittlerer Reichweite lassen sich auch oft vollständig grafisch in Form eines **kausalen Modells** mit Ursache-Wirkungs-Relationen darstellen. Beispiel: Die Theorie der Technikakzeptanz zielt darauf ab zu erklären, welche Faktoren die Nutzung neuer Technologien (z. B. Computertechnik) am Arbeitsplatz bedingen. Ein erstes **Technology-Acceptance-Modell (TAM)** wurde Ende der 1980er-Jahre entwickelt und als Pfeildiagramm bzw. Pfadmodell visualisiert (Abb. 6.1). Parallel zur Visualisierung kann das Modell auch über ein entsprechendes **mathematisches Gleichungssystem** formalisiert werden (zu Strukturgleichungsmodellen ▶ Kap. 17).

Dieses Modell wurde anschließend zu **TAM 2** (Venkatesh & Davis, 2000) und schließlich zu **TAM 3** (Abb. 6.2) erweitert, indem die „externalen Variablen" spezifiziert wurden. Während TAM 1 die abhängige Variable (tatsächliche Nutzung des neuen technischen Systems) mit fünf Ursachenfaktoren erklärte, werden in TAM 3 insgesamt 16 Prädiktoren herangezogen. Ein komplexeres Modell hat den Vorteil, dass es die Erfahrungswirklichkeit genauer abbilden kann, es hat jedoch den Nachteil verringerter Forschungsökonomie bei der praktischen Anwendung und wirkt naturgemäß unübersichtlicher. Der angemessene Komplexitätsgrad einer Theorie wird wissenschaftstheoretisch mit dem **Kriterium der Sparsamkeit** angesprochen (▶ Abschn. 2.2.10).

Die grafische Modellbildung veranschaulicht die Relationen zwischen theoretischen Konzepten (zum Modellbegriff ▶ Stachowiak, 1992), mündet in eine stärkere Formalisierung von Theorien und erleichtert durch ihre Übersichtlichkeit zugleich die wissenschaftliche Kommunikation. Zudem regt die Anschaulichkeit des Modells zu Neuordnungen oder Ergänzungen von Elementen an, macht Lücken und Brüche ebenso wie Kombinationsmöglichkeiten sichtbar. Entsprechende theoretische Modelle lassen sich statistisch als Ganzes z. B. mittels Pfadanalyse (▶ Abschn. 12.5.2) oder Strukturgleichungsmodellierung (▶ Kap. 17) überprüfen.

Letztlich läuft Theoriearbeit darauf hinaus, die Aussagengebäude der einzelnen Theorien systematisch zu vergleichen, zu bewerten und daraus Schlussfolgerungen zu ziehen. Im Ergebnis kann sich z. B. eine Theorie als überlegen herausstellen oder es wird deutlich, dass und wie sich die Voraussagen von zwei Theorien widersprechen. Eine eigenständige Theoriearbeit kann dann wiederum späteren empirischen Arbeiten als Grundlage dienen.

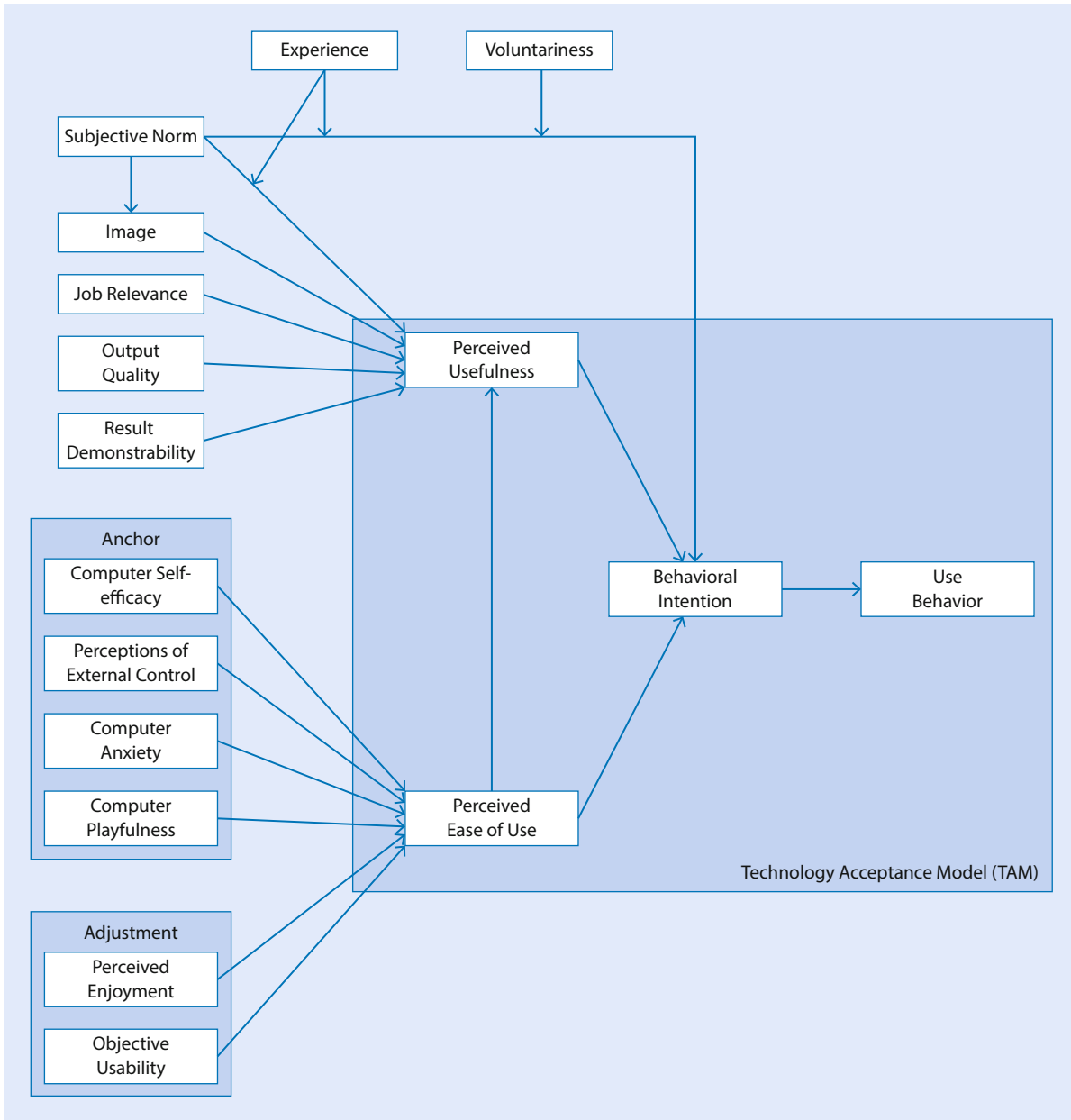


Abb. 6.2 Die dritte Fassung des Technology Acceptance Model TAM 3. (Aus Venkatesh & Bala, 2008. Mit freundlicher Genehmigung von Wiley)

6.3.2 Empirische Überprüfung und Modifikation einer einzelnen Theorie

Viele Studien nutzen als theoretischen Hintergrund letztlich eine einzige Theorie. Die entsprechende **Auswahl der Theorie** ist – mit Blick auf das gewählte Forschungsproblem – zu begründen. Hilfreich kann es dafür sein, eine Übersichtstabelle (Synopse) der wichtigsten, auf den Untersuchungsgegenstand anwendbaren Theorien zu erstellen (z. B. hinsichtlich Anzahl der Einflussfaktoren in der Theorie, Grad der empirischen Überprüfung, Vorliegen etablierter Instrumente zur Messung der zentralen Konstrukte etc.), anhand derer dann die Entscheidung für die konkrete Theorie plausibel gemacht wird. Auf der Basis von theoretischen Vorüberlegungen sowie Vorkenntnissen über den Untersuchungsgegenstand kann die Theorie vor ihrer Anwendung auch noch in begründeter Weise **modifiziert** werden, etwa indem zusätzliche Ein-

flussfaktoren ergänzt oder als unwichtig erachtete Konstrukte herausgenommen werden.

Aus der gewählten Theorie werden Forschungshypothesen und/oder Forschungsfragen abgeleitet und diese dann empirisch überprüft. Dabei können frühere Studien in identischer oder leicht modifizierter Weise wiederholt (sog. Replikationsstudie) oder auch neue Untersuchungsdesigns umgesetzt werden. Als Ergebnis der empirischen Theorieprüfung kann die Theorie vorläufig bestätigt, infrage gestellt oder modifiziert werden.

6.3.3 Formalisierung und Überprüfung einer Theorie mittels Computersimulation

Sozialwissenschaftliche Theorien sind im Unterschied zu naturwissenschaftlichen Theorien sehr viel weniger formalisiert. Sie sind oft alltagssprachlich formuliert, enthalten neben präzise definierten Fachbegriffen auch diffuse Begriffe und neben eindeutigen Ursache-Wirkungs-Pfaden auch vage Annahmen über die behandelten Wirkungszusammenhänge. Ausnahmen sind z. B. Modelle in der Lernpsychologie, der Entscheidungstheorie und der sozialpsychologischen Austauschtheorie, die funktionale Zusammenhänge mit mathematischen Gleichungen beschreiben (Bierhoff, 1997; Jungermann, Pfister, & Fischer, 2009; Kelley et al., 2003; Lefrancois, 2008). Der **geringe Formalisierungsgrad** der meisten sozialwissenschaftlichen Theorien ist nicht per se als Schwäche oder Fehler anzusehen, sondern auch der Komplexität sozialer Sachverhalte geschuldet. So konnte z. B. in jahrzehntelangen theoretischen Auseinandersetzungen und Forschungsaktivitäten bis heute in der Sozialpsychologie nicht konsensfähig geklärt werden, was eine „Situation“ ist, welche Arten von Situationen zu unterscheiden sind und wie man sie misst (Hogan, 2009). Dennoch postulieren zahlreiche Theorien, dass die „Situation“ ein bedeutsamer Einflussfaktor zur Erklärung menschlichen Erlebens und Verhaltens darstellt.

Um den Informationsgehalt von Theorien transparent zu machen, ist eine Präzisierung und Formalisierung ihrer Aussagen anzustreben. Der „epische“ Charakter mancher Theorien, deren Annahmengefüge über viele Seiten hinweg beschrieben, erörtert und begründet wird, täuscht durch Begriffsneuschöpfungen, Beispiele und Vergleiche nicht selten über Inkonsistenzen und Unklarheiten hinweg. Eine Möglichkeit, bestehende Theorien zu verbessern und neue Hypothesen zu formulieren, besteht folglich in forcierten Bemühungen, die zentralen Annahmen einer Theorie herauszuarbeiten und zu formalisieren (Gilbert & Troitzsch, 2005; Manhart, 1995; Mosler, Schwarz, Ammann, & Gutscher, 2001; Weber, 2007).

Neben den gängigen grafischen Modellen und Schaubildern zur Darstellung von Theorien werden Computermodelle bzw. **Computersimulationen** als besonders vorteilhaft empfohlen, um eine Theorie zu formalisieren und zu überprüfen (Conte, 2006; Davis, Eisenhardt, & Bingham, 2007; Dooley, 2002; Garson, 2009; Meeker & Leik, 1997; Mosler et al., 2001; Schnell, 1991). Ein Computermodell ist ein lauffähiges Programm, das die von einer Theorie postulierten Prozesse simuliert. Dabei kann man quantitative und qualitative Computermodelle unterscheiden. **Quantitative Computermodelle** beruhen in der Regel auf einem System von mathematischen Gleichungen und haben das Ziel, für unterschiedliche Anfangssituationen die entsprechenden, theoriekonformen Konsequenzen in Form von Parameterschätzungen zu berechnen. Dieses Vorgehen kann zur Theorieprüfung und zur Prognose verwendet werden. Bei **qualitativen Computermodellen** geht es nicht um korrekte Parameterschätzungen, sondern darum, ob die von einer Theorie beschriebenen Phänomene oder Effekte überhaupt nachgestellt werden können.

Die ersten sozialwissenschaftlichen Computersimulationen waren quantitativ ausgerichtet und wurden in den 1960er-Jahren vor allem unter Prognoseaspekten (Forecasting) betrachtet. In sog. Weltmodellen versuchte man, globale Entwicklungen in der Wirtschaft und Demografie abzubilden. Tatsächlich berechnete das **Weltmodell von Forrester** (1971) die Bevölkerungsentwicklung von 1900 bis 1971 genau so, wie es dem tatsächlichen Verlauf entsprach. An der Validität des Modells kamen allerdings Zweifel auf, als sich herausstellte, dass es frühere Bevölkerungszahlen nicht korrekt zurückrechnen konnte (Backcasting), sondern drastische Bevölkerungsrückgänge rekonstruierte, die historisch nicht stattgefunden hatten. Neben globalen Entwicklungen wurden auch Stadtentwicklungen modelliert (sog. Urban-Dynamics-Modelle). Beispiele für den explorativen Einsatz von System-Dynamics-Modellen sind Davis, Eisenhardt und Bingham (2007), Dooley (2002) sowie Kreutz und Bacher (1991) zu entnehmen.

Nach Schnell (1991, S. 118f.) sind qualitative Computermodelle besonders gute Katalysatoren der Theoriebildung:

> » In Simulationsprogramme übersetzte Theorien sind präziser als Alltagssprache sein kann. Andererseits sind Simulationen flexibler als es mathematisch formalisierte Theorien sein können. Die Präzision wird durch die Syntax der verwendeten Programmiersprache erzwungen: Eine ungenaue, widersprüchliche oder unvollständige Theorie läßt sich nicht ohne Präzisierung in ein funktionierendes, d. h. zunächst einmal syntaktisch korrektes,

> dann auch die gewünschte Dynamik hervorbringendes, lauffähiges Programm übersetzen . . . Der Zwang zur Präzision bei der Erstellung eines Simulationsprogrammes äußert sich vor allem in der Notwendigkeit, alle theoretischen Annahmen explizit angeben zu müssen. Diese Notwendigkeit führt bei jeder Programmierung einer Simulation zur Entdeckung von Wissenslücken.

Der – sorgfältig kommentierte – **Programmcode** bzw. Quelltext stellt nach der Programmierarbeit im Idealfall eine kompakte Kurzversion der Theorie dar und erleichtert damit Rezeption und Kritisierbarkeit des Gedankengebäudes. Erweist sich das Programm, das die Kernaussagen der Theorie zu Algorithmen verdichtet enthält, als lauffähig, ist eine **notwendige** Bedingung für die Gültigkeit der Theorie erfüllt, nicht jedoch eine hinreichende. Die Simulationsergebnisse müssen für eine Validierung des Modells auch mit empirischen Ergebnissen konfrontiert werden, wobei sich das bei allen Validierungsbemühungen unvermeidbare Problem stellt, dass unplausible Ergebnisse (hier der Simulation) sowohl auf Fehler in der Methode (hier im Programm) als auch in der Theorie zurückführbar sind.

Aber lassen sich sozialwissenschaftliche Theorien überhaupt sinnvoll in Computerprogramme umsetzen? Sind sie nicht viel zu komplex für eine simplifizierende Programmierung? Schnell (1990, S. 115) weist diese Befürchtung zurück. Die angebliche Komplexität vieler Theorien sei letztlich eher durch undefinierte oder gar zirkuläre Begriffsverwendung sowie durch implizite Zusatzannahmen verursacht. Viele Theorien seien eigentlich überraschend simpel und ließen sich in weniger als 100 Programmzeilen vollständig abbilden. Diese „verborgene Trivialität", die erst durch eine Computersimulation sichtbar wird, sei möglicherweise ein Grund dafür, dass Programmtexte so selten publiziert werden. (Weitere Hinweise zur computergestützten Exploration findet man bei Bossel, 1992; Carley, 2001; Davis et al., 2007; Garson, 2009; Gilbert & Troitzsch, 2005; Hanneman & Patrick, 1997; Squazzoni, 2008; Troitzsch, Mueller, & Gilbert, 1996).

6.3.4 Verbindung mehrerer Theorien zu einem integrativen Theoriemodell

Nicht selten werden im Rahmen der Theoriearbeit mehrere Theorien zu einem neuen Modell verknüpft (▶ Studienbeispiel „Integration mehrerer Theorien").

Eine schlüssige Integration mehrerer Theorien zu einem fruchtbaren neuen Theoriemodell ist sicherlich ein Glücksfall (für ein prominentes Beispiel s. ◘ Abb. 6.3 und ◘ Abb. 6.4). Die Bemühungen um ein übergreifendes „Rahmenmodell" sind in der Regel wenig erhellend, wenn man aus mehreren Theorien Elemente entnimmt, jedes in einem Kasten darstellt, diese Kästchen miteinander verbindet und meint, damit ein neues „Modell" geschaffen zu haben. Erst wenn aus einer additiven Zusammenfassung auch sinnvolle Querverbindungen und Kausalrelationen konstruierbar sind, hat die Modellkonstruktion einen heuristischen Wert. Der Wunsch, die Vielfalt einzufangen, führt oft zur Aufnahme einer Überzahl von Einzelaspekten, was die Unübersichtlichkeit steigert, nicht jedoch den Erklärungswert. So ist es nur in seltenen Fällen lohnend, pauschal „anthropologische", „biologische" oder „kulturelle" Einflüsse als Modellparameter aufzunehmen, wenn diese weder theoretisch ausformuliert sind noch in Empfehlungen für die Forschungspraxis münden.

Studienbeispiel

Integration mehrerer Theorien

Der Kommunikationspsychologe Schulz von Thun (1991) integrierte drei bekannte Kommunikationsmodelle und entwickelte daraus eine eigene Theorie zur Beschreibung und Erklärung sozialer Interaktionen, das auch als **Vier-Ohren-Modell** bekannt geworden ist (◘ Abb. 6.4).

Als Grundmodell für die theorieintegrierende Entwicklung des „Vier-Ohren-Modells" diente das aus der Nachrichtentechnik stammende Modell von Shannon und Weaver (1949), das Kommunikation als linearen Ablauf darstellt: Vom Sender wird die Nachricht an den Empfänger geschickt. Eine erste psychologische Ausdifferenzierung erhielt dieses Sender-Empfänger-Modell durch die Verbindung mit der Kommunikationstheorie von Watzlawick, Beavin und Jackson (2000), die darauf hinweist, dass jede interpersonale Nachricht unvermeidbar eine Botschaft sowohl auf der Sachebene als auch auf der Beziehungsebene enthält, also Auskunft darüber gibt, wie der Akteur zu seinem Kommunikationspartner steht. Eine weitere Differenzierung erreichte Schulz von Thun (1991, S. 30) durch die Integration des Sprachmodells von Bühler (1934), das Sprache drei Funktionen zuordnet: Darstellung (von Sachverhalten), Ausdruck (von Gefühlen und Gedanken) und Appell (Anweisungen an den Empfänger; ◘ Abb. 6.3).

6.3.5 Gegenstandsverankerte Theoriebildung

Das „Arbeiten" mit Theorien kann darin bestehen, die vorhandenen Theorien vergleichend darzustellen, eine ausgewählte Theorie anhand empirischer Daten oder mithilfe einer Computersimulation zu prüfen und in diesem Zuge zu modifizieren oder mehrere Theorien zu einem integrativen Modell zu vereinen und dieses dann zu prü-

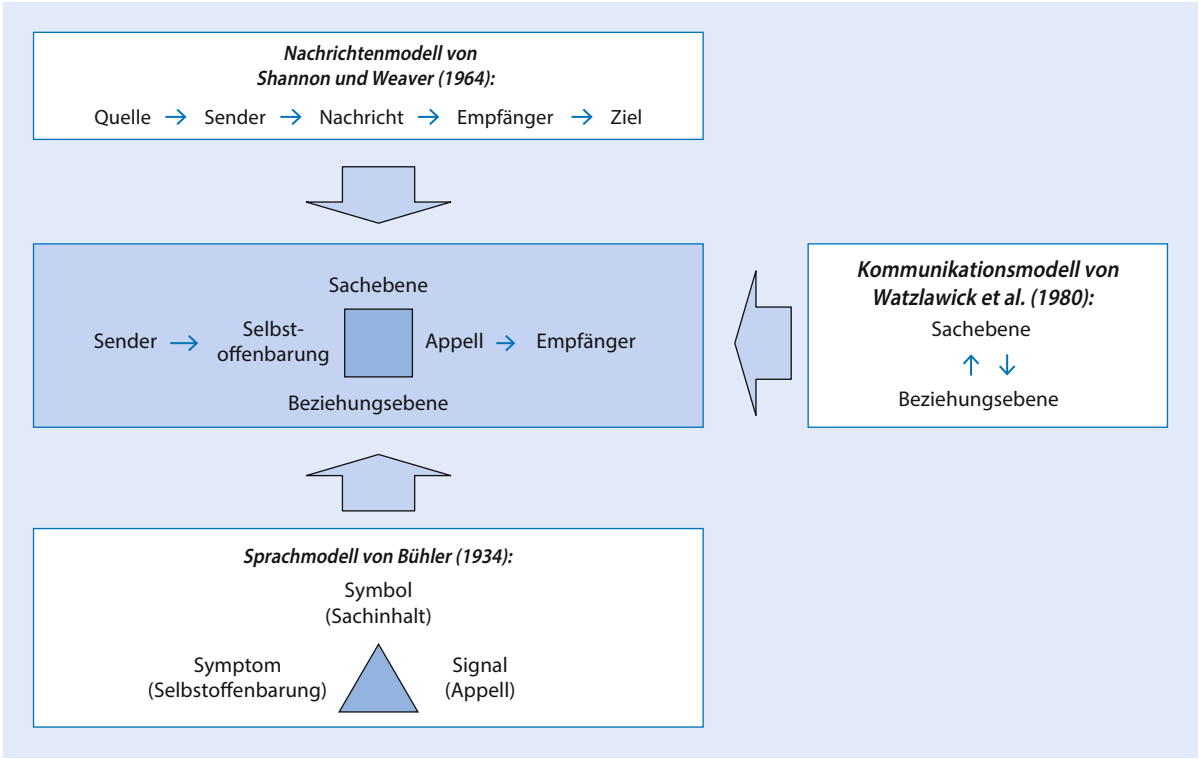


Abb. 6.3 Integration von drei Theorien: Das Kommunikationsmodell von Schulz von Thun. (Aus Schulz von Thun, 1991, S. 30. Copyright 2013 Rowohlt)

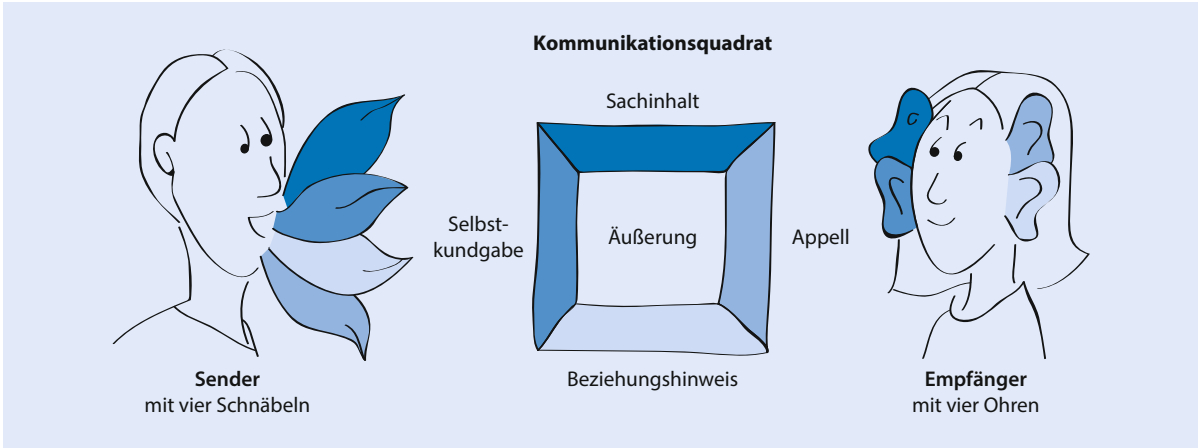


Abb. 6.4 Vier-Ohren-Modell. (Bildrechte: http://www.schulz-von-thun.de. Mit freundlicher Genehmigung von Friedemann Schulz von Thun)

fen. Insbesondere im quantitativen Paradigma der empirischen Sozialforschung wird auf der Basis der Wissenschaftstheorie des Kritischen Rationalismus die Theoriebildung als Ausgangspunkt wissenschaftlichen Erkenntnisgewinns betrachtet (▶ Abschn. 2.2). Im qualitativen Paradigma dagegen wird gemäß des **Prinzips der relativen theoretischen Offenheit** eher davor gewarnt, sich durch Rückgriff auf bestehende Theorien im Vorfeld einer Studie in der Untersuchungsperspektive zu stark einzuengen (▶ Abschn. 2.3). Möglichst unvoreingenommen zunächst den Untersuchungsgegenstand detailliert empirisch in seinem Kontext zu erkunden und dabei bislang womöglich unentdeckte Aspekte aufzudecken wird gemäß der **Grounded-Theory-Methodologie** (GTM; Gla-

ser & Strauss, 1999) als Ausgangspunkt für die qualitative Sozialforschung empfohlen. Die Forschungsstrategie der GTM ist dabei auf eine zunehmende Annäherung an den Untersuchungsgegenstand ausgerichtet: Das betrifft sowohl die schrittweise Fallauswahl nach den Prinzipien des Theoretical Sampling (▶ Abschn. 9.2.1) als auch die nichtstrukturierte Datenerhebung etwa mittels qualitativer Beobachtung (▶ Abschn. 10.1.2), qualitativem Interview (▶ Abschn. 10.2.2) oder qualitativem Fragebogen (▶ Abschn. 10.3.2) sowie die in mehreren Überarbeitungszyklen sich vollziehende Kodierung des qualitativen Datenmaterials (▶ Abschn. 10.6.2, ▶ Abschn. 12.1.1). Im Ergebnis eines solchen Forschungsprozesses soll eine sog. gegenstandsnahe oder gegenstandsverankerte Theorie entstehen, die sich ganz eng an den sozialen Phänomenen orientiert und möglichst wenig durch überkommene Vorurteile aus früheren Theorien verzerrt ist. Mit der Grounded-Theory-Methodologie steht eine Forschungsstrategie zur Verfügung, die sich ausdrücklich der Theorieneubildung widmet.

6.3.6 Unterschiedliche Strategien zur Entwicklung neuer Theorien

Während die qualitative Sozialforschung in ihrem offenen Forschungsprozess auf Theoriebildung als Ergebnis von Studien ausgerichtet (▶ Abschn. 2.3) und die Vorgehensweise dabei z. B. durch die Grounded-Theory-Methodologie detailliert vorgegeben ist, betrachtet die quantitative Sozialforschung Theorien als den Ausgangspunkt der Forschungsaktivität (▶ Abschn. 2.2). Im Zentrum der quantitativen Forschung steht oft die Überprüfung und Weiterentwicklung von Hypothesen und Theorien (sog. Begründungszusammenhang von Theorien). Doch wie man auf eine gänzlich neue und fruchtbare wissenschaftliche Theorie kommen kann oder soll (sog. **Entdeckungszusammenhang von Theorien**; ▶ Abschn. 2.2.6) wird in der quantitativen Forschungstradition nicht genau angegeben. Im englischen Sprachraum sagt man scherzhaft, dass wissenschaftliche Theorien im Kontext der drei „B's" entstehen: „Bed", „Bathroom" und „Bicycle" (Gigerenzer, 1994, S. 109).

Viele Beispiele aus der Wissenschaftsgeschichte scheinen das **irrationale Moment der Theoriebildung** zu belegen: Der Umstand, dass unvollendete Handlungen besser im Gedächtnis haften bleiben als vollendete, ist als Zeigarnik-Effekt bekannt und wurde von der Lewin-Schülerin Bluma Zeigarnik im Caféhaus entdeckt: Sie beobachtete den Kellner, der sich notorisch immer die unbezahlten (= unvollendeten) Bestellungen merkte, die bezahlten aber sofort vergaß. Der Behaviorist Burrhus F. Skinner fütterte am Bahnsteig die Tauben und stellte fest, dass er einige von ihnen durch seine Futterspenden zu absonderlichen Tänzen konditioniert hatte. Der Chemiker Dmitri I. Mendelejeff fand die Struktur des Periodensystems beim Patiencelegen. Der Psychologe Max Wertheimer erlebte das Phi-Phänomen beim Zugfahren. Der Chemiker August Kekulé hat die Ringstruktur des Benzolmoleküls entweder geträumt (ihm erschien eine Schlange) oder beim Blick in den Kamin erfasst. Der Mathematiker und Physiker Archimedes von Syrakus stieß in der Badewanne auf das Prinzip des statischen Auftriebs. Die Liste spektakulärer Einfälle und Entdeckungen ließe sich beliebig verlängern – wenngleich so manches wohl eher in den Bereich der Legendenbildung, denn der Wissenschaftsgeschichte fallen dürfte (Dörner, 1994, S. 343).

Spontan und unkontrollierbar, eben „intuitiv", erscheinen die theoriebildenden Geistesblitze bekannter Wissenschaftlerinnen und Wissenschaftler – „einer logischen Analyse weder fähig noch bedürftig", befand der Wissenschaftstheoretiker und Begründer des Kritischen Rationalismus Karl Popper (1989, S. 6). Beginnen wir mit letzterem: Bedarf es einer **Systematisierung der Theoriebildung**? Der Psychologe Dietrich Dörner (1994, S. 344) fordert sie:

> » Nicht nur der Forscher in der Psychologie, sondern auch der Praktiker ist ständig mit der Notwendigkeit konfrontiert, Theorien erfinden zu müssen. Welche Aspekte der Familienstruktur mögen wohl dafür verantwortlich sein, daß Frau X so depressiv ist? Oder liegt es gar nicht an der Familie? Die eine Theorie, die alles erklärt, gibt es in der Psychologie nicht, und so muß man sich ständig, Theorien erfindend, prüfend, revidierend, irgendwie durchwursteln. Wenn das aber so ist, darf man sich in der psychologischen Methodenlehre auf die Prüfmethoden nicht beschränken.

In letzter Zeit wird immer häufiger die Notwendigkeit betont, den Prozess der Theoriebildung transparenter zu machen und in methodologische Überlegungen mit einzubeziehen (zur Theoriebildung ▶ Kleining, 1994; Esser & Troitzsch, 1991; Strube, 1990).

Die Forderung nach reflektierter Entwicklung von Theorien setzt voraus, dass Theoriebildung einer Systematisierung fähig ist. Dafür spricht die Beobachtung, dass bereits jetzt de facto Normen bestehen; so ist z. B. ein sorgfältiges Literaturstudium im Sinne einer Aufarbeitung des Forschungsstandes eine weitgehend anerkannte Voraussetzung jeder theoriebildenden Arbeit. Einige Autoren formulieren explizit Strategien und Me-

thoden für eine in den wissenschaftlichen Forschungsprozess voll integrierte, systematische Gegenstandserkundung (Exploration) und darauf aufbauende Theoriebildung. Dazu zählen neben der Grounded-Theory-Methodologie in der qualitativen Forschung (▶ Abschn. 6.3.5) z. B. auch **hypothesenbildende Strategien der Datenauswertung in der quantitativen Forschung**, wie sie unter anderem von John Tukey (1977) mit dem Exploratory-Data-Analysis-(EDA-)Ansatz vorgeschlagen wurden (▶ Abschn. 12.3.1).

Gerade wenn man das Finden guter Theorien in die Nähe anderer kreativer und künstlerischer Schöpfungen rückt, ist anzuerkennen, dass nach einer bekannten Redewendung „Kunst von Können kommt". Wäre dies nicht so, könnte man im künstlerischen Bereich auf jegliche Ausbildung verzichten. Auch wenn es wohl keine Patentrezepte gibt, erhöhen fundierte Kenntnisse und systematisches Vorgehen die Wahrscheinlichkeit, empirisch brauchbare Hypothesen und Theorien zu finden. Aus dieser Sicht ist es auch keineswegs erforderlich, einen **Widerspruch zwischen Kreativität und Systematik** zu konstruieren. Spektakuläre Ideenfindungen wie die oben genannten ereignen sich eben nicht urplötzlich und ganz „zufällig", sondern nach langjähriger intensiver Auseinandersetzung mit einem Forschungsthema. Wer hat schließlich nicht alles schon im Café oder in der Badewanne gesessen, Tauben gefüttert, Patiencen gelegt und von Schlangen geträumt – und dabei rein gar nichts entdeckt.

Die Integration des Entdeckungszusammenhangs in den Forschungsprozess meint nicht, dass normativ Methoden der Theoriebildung festgeschrieben werden sollen. Schließlich würde strenge Reglementierung – die ohnehin nur konventionell und nicht rational zu fundieren wäre – die Gefahr in sich bergen, das Finden guter Ideen zu verhindern. Vielmehr geht es darum, die Wahrscheinlichkeit, wissenschaftlich brauchbare, innovative Ideen zur Theoriebildung zu produzieren, zu erhöhen. **Vier Strategien der Gegenstandserkundung (Exploration) und Theoriebildung** lassen sich unterscheiden:

1. **Theoriebasierte Exploration**: Diese Vorgehensweise wurde in den vergangenen Abschnitten ausführlich dargestellt und stellt ein etabliertes Vorgehen dar. Vorliegende Theorien werden als Ausgangspunkt mehr oder minder grundlegend neuer Theoriebildungen genommen.
2. **Empirisch-qualitative Exploration**: Auch diese Strategie wurde bereits erwähnt, sie ist eng verbunden mit der Grounded-Theory-Methodologie und beinhaltet Theoriebildung auf der Basis einer qualitativen Forschungsstrategie.
3. **Empirisch-quantitative Exploration**: In sehr großen quantitativen Datensätzen bislang unbeachtete oder unerkannte Muster zu entdecken und auf dieser Basis neue Theorien zu entwickeln, ist Zielsetzung der explorativen Datenanalyse EDA sowie auch des Data Mining.
4. **Methodenbasierte Exploration**: Nicht zuletzt können auch Methoden der Datenerhebung und Datenanalyse zu neuen Theorien inspirieren, etwa wenn man sie vergleicht, kombiniert oder modifiziert und dadurch unerwartete, erklärungsbedürftige Effekte erhält. Methodenkombinationen werden (nicht immer mit der ausdrücklichen Zielsetzung der Theoriebildung, aber durchaus dafür nutzbar) u. a. in der quantitativen Forschung mittels Multitrait-Multimethod-Ansatz, in der qualitativen Forschung mittels Triangulation und in der Mixed-Methods-Forschung mittels unterschiedlicher Mixed-Methods-Designs realisiert. Auch methodische Hilfsmittel wie der Computer haben sich durch Analogiebildung als fruchtbar für die Theoriebildung in den Sozialwissenschaften erwiesen. So arbeiten Theorien des menschlichen Gedächtnisses und der Kognition mit Computeranalogien, etwa wenn vom „Arbeitsspeicher" die Rede ist. Hier fungiert der Computer nicht nur als Hilfsmittel für die Analyse wissenschaftlicher Daten, auf deren Basis dann Theorien gebildet werden, sondern die Computertools selbst werden ohne „Umweg über die Daten" zur Theoriebildung herangezogen (sog. **Tools-to-Theories-Heuristik**; Gigerenzer, 1991, 1994).

6.4 Ableitung von Forschungsfragen und Forschungshypothesen

Wir haben schon betont, wie wichtig es ist, ein allgemeines Forschungsthema in ein empirisch untersuchbares Forschungsproblem bzw. eine Untersuchungsfragestellung zu transformieren (▶ Abschn. 5.1). Zu einem Forschungsproblem wiederum wird dann ein Satz von ca. ein bis zehn Forschungsfragen und/oder Forschungshypothesen formuliert (▶ Abschn. 5.2). Diese sind jeweils einzeln möglichst schlüssig aus dem Forschungsstand sowie aus dem verwendeten theoretischen Rahmenmodell abzuleiten. Üblicherweise werden sie durchnummeriert und in einer logischen Reihenfolge präsentiert oder in thematische Blöcke gruppiert.

Einige typische Konstellationen der Ableitung von Forschungsfragen und Forschungshypothesen seien im Folgenden anhand authentischer Studienbeispiele vorgestellt.

6.4.1 Ableitung von Forschungsfragen aus dem Forschungsstand

Eine explorative kommunikationswissenschaftliche Studie widmete sich dem Thema Computerspieler und formulierte sinngemäß als Forschungsproblem (Quandt, Grueninger, & Wimmer, 2009): Jugendliche galten lange als die typischen Computerspieler und werden oft untersucht, aber was kennzeichnet das Erleben und Verhalten **älterer Computerspieler**? Dieses Forschungsproblem sollte durch qualitative Leitfaden-Interviews bearbeitet werden und wurde in vier relativ allgemeine Forschungsfragen (F) aufgegliedert (▶ Abschn. 5.2; ◘ Tab. 5.1).

- F1: Warum bzw. aus **welchem Anlass** haben die älteren Spieler ursprünglich mit dem Computerspielen begonnen?
- F2: Wie haben die älteren Spieler die für Computerspiele **notwendigen Kenntnisse und Fähigkeiten** erworben?
- F3: Wie integrieren die älteren Spieler das Computerspielen in ihren **Alltag**?
- F4: Wie und durch welche Einflussfaktoren (z. B. Familie, Beruf) **verändert** sich das Spielverhalten der älteren Spieler mit der Zeit?

Bei der Darstellung des Forschungsstandes zu ihrer Arbeit betonen die Autoren, dass zwar einige Studien zu soziodemografischen Merkmalen älterer Spieler vorliegen, aber eine detailliertere Erfassung des Spielverhaltens bei der Zielgruppe „Ü35“ bislang fehlt. Ihre eigene Untersuchung differenziert das Spielverhalten in **vier Unteraspekte**, die in ähnlicher Form schon in Studien mit jüngeren Spielern auftauchen, und strebt somit eine Vergleichbarkeit der Befunde für ältere und jüngere Spieler an. Die Untersuchung beschränkt sich auf das **Nutzungsverhalten der Spieler** und thematisiert z. B. nicht Medienwirkungen. Theorien werden in dieser explorativen Arbeit nicht näher behandelt. Im Ausblick wird auf die Vielzahl weiterer offener Forschungsfragen hingewiesen.

6.4.2 Ableitung von Forschungsfragen aus einer oder mehreren Theorien

Eine ebenfalls explorativ angelegte kommunikationswissenschaftliche Studie zum Thema Computerspieler widmete sich folgendem Forschungsproblem (Krause, 2010): Männer galten lange als die typischen Computerspieler – insbesondere als die Intensivspieler – und werden oft untersucht, aber was kennzeichnet das Erleben und Verhalten **weiblicher Vielspieler**? Auch in dieser Arbeit wurden zunächst vier relativ allgemeine Forschungsleitfragen (FL) sinngemäß so formuliert:

- FL1: Welches **Nutzungsverhalten** zeigen weibliche Computervielspieler?
- FL2: Aus welchen **Nutzungsmotiven** wenden sich weibliche Vielnutzer Computerspielen zu?
- FL3: Welche **Typen von weiblichen Vielspielern** lassen sich unterscheiden?
- FL4: Wie kann die **Zugehörigkeit** zu den verschiedenen Vielspielerinnentypen erklärt werden?

Da die Daten nicht nur mittels qualitativer Leitfadeninterviews, sondern auch mittels einer ergänzenden standardisierten Online-Fragebogenerhebung gesammelt werden sollten, war es hier besonders wichtig, die Forschungsfragen zu konkretisieren. Die Autorin erarbeitete für ihre Studie deswegen einen umfangreichen theoretischen Rahmen, bei dem sie mehrere Mediennutzungstheorien miteinander verknüpfte. Als Leittheorie griff sie auf den etablierten kommunikationswissenschaftlichen **Nutzen-und-Belohnungsansatz** zurück („Uses and Gratifikations Approach“), diesen ergänzte sie um einige weitere gut etablierte Theorien, die sich auf Motive und Funktionen der Mediennutzung beziehen lassen (z. B. Mood-Management-Theorie, Theorie der sozialen Identität). Auf der Basis dieses **erweiterten Nutzen-und-Belohnungsmodells** transformierte sie die vier allgemein gehaltenen Forschungsleitfragen (s. oben) jeweils in theoretisch fundierte konkrete Forschungsfragen:

- F1: Welches Nutzungsverhalten zeigen weibliche Vielspieler im Hinblick auf Spielintensität, Spielmodi, Spielgenres und Spielcharaktere?
- F2: Welche Rolle spielen die Nutzungsmotive „Herausforderung“, „Eskapismus“, „Autonomie“, „Wettbewerb“, „Stimmungsmanagement“, „soziale Interaktion“ und „Anerkennung der Spieler-Community“ für die weiblichen Vielspieler?
- F3: Welche Typen von weiblichen Vielspielern lassen sich anhand von Nutzungsverhalten und Nutzungsmotiven unterscheiden?
- F4: Inwiefern können individuelle Eigenschaften (Sozialisation, sozialer Status, Persönlichkeitsmerkmale) und grundlegende Bedürfnisse (Geschlechtsrollenidentität, Affinität zu männlichen Gruppen) die Zugehörigkeit weiblicher Vielspieler zu einem bestimmten Spielerinnentyp erklären?

Insbesondere quantitative explorative Studien sind auf einen entsprechenden theoretischen Hintergrund angewiesen, da alle relevanten Variablen im Vorfeld identifiziert und operationalisiert werden müssen (z. B. Abfrage aller relevanten Computerspielmotive im Fragebogen). Eine rein qualitative Studie dagegen muss die verschiede-

nen Motive nicht im Vorfeld herausarbeiten, sondern kann entsprechende Daten im Rahmen offener Interviewfragen erheben.

6.4.3 Ableitung von Forschungsfragen und/oder Forschungshypothesen aus Alltagstheorien

Sozialwissenschaftliche Forschung befasst sich oft mit Sachverhalten, die auch im Alltag diskutiert werden und über die zahlreiche vorwissenschaftliche Theorien kursieren. Nicht-wissenschaftliche Theorien, die im gesellschaftlichen oder politischen Diskurs große Bedeutung gewinnen, können zum Anlass einer systematischen empirischen Prüfung genommen werden. Sei es, dass anlässlich der Überprüfung neue Primärdaten erhoben oder vorliegende Studien und Datensätze herangezogen werden (Sekundäranalyse).

Eine intensive **öffentliche Debatte** löste z. B. der Volkswirt und ehemalige Berliner Finanzsenator **Thilo Sarrazin** mit seinem Bestseller „Deutschland schafft sich ab“ aus. In diesem 2010 erschienenen Sachbuch sowie in Presseinterviews äußerte Sarrazin zahlreiche Thesen zur Zukunft Deutschlands, die durch die Kombination von Geburtenrückgang (v. a. bei deutschen Akademikern), Anwachsen der Unterschicht und Zustrom seiner Ansicht nach integrationsunwilliger muslimischer Migranten bedroht sei. In einem Interview mit der Wochenzeitung „Die Zeit“ sagte er (Ulrich & Topcu, 2010, S. 4):

> » Das kulturelle Problem ist in der Gruppe der muslimischen Migranten verankert und kann gegen deren Willen kaum verändert werden. Schwimmunterricht und Kopftücher, generell die Rolle der Frauen und Mädchen, sind dafür die Symbole. Auch der Umstand, dass sich die Türken und die Araber zu großen Teilen kaum Mühe geben, Deutsch zu lernen, ist ein Ausdruck fehlenden Interesses an der Mehrheitskultur und mangelnder Bildungsbereitschaft.

Sarrazins Meinung, „dass sich die Türken und die Araber zu großen Teilen kaum Mühe geben, Deutsch zu lernen“ (Ulrich & Topcu, 2010, S. 4) lässt sich in eine empirisch prüfbare Forschungsfrage (F) oder auch in eine gerichtete Hypothese (H) umformulieren. Hierbei wird man als Indikator vermutlich nicht das schwer greifbare „sich Mühe geben“, sondern eher das Ergebnis der vorhandenen oder nicht vorhandenen Bemühungen, nämlich die Sprachkenntnisse aufgreifen, auf die Sarrazins Argumentation auch abzielt:

- F: Über welche Deutschkenntnisse verfügen Menschen mit türkischem und arabischem Migrationshintergrund in Deutschland?
- H: Die Mehrheit der Menschen mit türkischem und arabischem Migrationshintergrund in Deutschland verfügt über keine oder sehr geringe Deutschkenntnisse.

Zur wissenschaftlichen Überprüfung müssten dann entsprechend einschlägige Studien recherchiert oder Daten neu erhoben, ausgewertet und anhand begründeter Bewertungsmaßstäbe zusammenfassend interpretiert werden. Eine Forschungsgruppe an der Humboldt Universität zu Berlin hat auf diese Weise Sarrazins Thesen systematisch auf den empirischen Prüfstand gestellt (Foroutan, Schäfer, Canan, & Schwarze, 2010; http://www.heymat.hu-berlin.de).

6.4.4 Ableitung von Forschungshypothesen aus einer Theorie

Eine medienpsychologische Untersuchung widmete sich dem Forschungsproblem, ob Computerspiele zur Erholung vom Arbeitsstress beitragen können (Reinecke, 2009). Die Grundlage der Studie bildete ein etabliertes **theoretisches Modell der Erholung**, das vier Elemente des Erholungserlebens nennt (Sonnentag & Fritz, 2007). Der Autor argumentiert, dass und wie alle vier Komponenten des Erholungserlebens beim Computerspielen prinzipiell auftreten können:

- „Detachment from Work“: Die Ablenkung beim Spielen kann eine gedankliche Ablösung vom stressbelasteten Arbeitsalltag fördern.
- „Relaxation“: Positive Gefühle beim Spielen können zur Entspannung führen.
- „Control“: Die Steuerung des Spiels vermittelt ein Gefühl von Kontrolle und Autonomie, oft im Kontrast zu den Zwängen des Arbeitslebens.
- „Mastery Experience“: Das Beherrschen des Spiels vermittelt über die eigenen Leistungen ein positives Erleben der Selbstwirksamkeit und des Selbstwertes.

Auf dieser theoretischen Basis leitete er dann seine Hypothesen (H) zum Erholungserleben im Zusammenhang mit Computerspielen ab (▶ Abschn. 5.2; ◘ Tab. 5.1).

- H1: Je erholsamer das Computerspielen empfunden wird, umso häufiger wird für Erholungszwecke auf Computerspiele zurückgegriffen.
- H2: Je mehr arbeitsbezogene Erschöpfung empfunden wird, umso häufiger wird für Erholungszwecke auf Computerspiele zurückgegriffen.

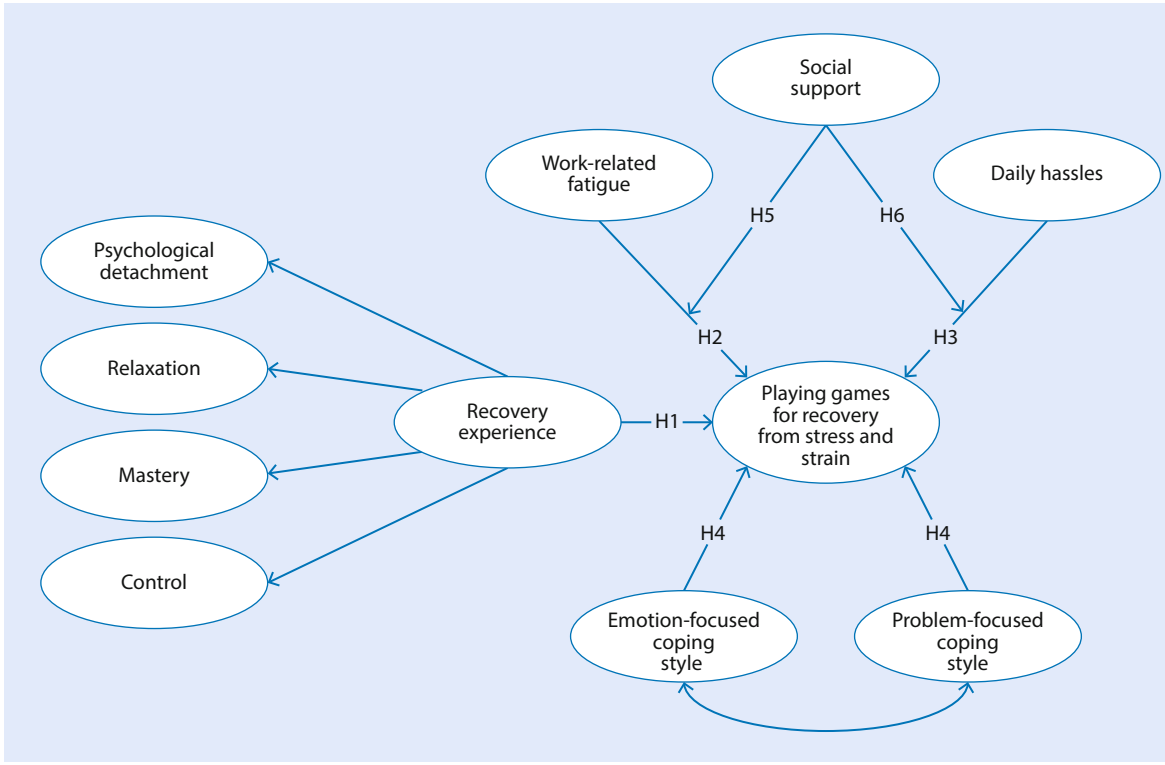


Abb. 6.5 Strukturgleichungsmodell zum Erholungserleben beim Computerspielen. (Aus Reinecke, 2009. Used by permission from *Journal of Media Psychology,* 2009, Vol 21 (3), pp. 126–142. Copyright © 2009 Hogrefe Publishing, DOI 10.1027/1864-1105.21.3.126)

- H3: Je mehr Alltagsstress erlebt wird, umso häufiger wird für Erholungszwecke auf Computerspiele zurückgegriffen.
- H4: Personen mit emotionsbezogenen Bewältigungsstrategien greifen eher zu Erholungszwecken auf Computerspiele zurück als Personen mit problemlösungsorientierten Bewältigungsstrategien.
- H5: Der Zusammenhang zwischen arbeitsbezogener Erschöpfung und dem Spielen zu Erholungszwecken ist enger, wenn weniger soziale Unterstützung zur Verfügung steht.
- H6: Der Zusammenhang zwischen Alltagsstress und dem Spielen zu Erholungszwecken ist enger, wenn weniger soziale Unterstützung zur Verfügung steht.

Die zentralen Konzepte des Modells, sowie die von den Hypothesen postulierten Effekte sind in einem Strukturgleichungsmodell dargestellt (Abb. 6.5).

6.4.5 Ableitung von Forschungshypothesen aus mehreren Theorien

Im Rahmen der wissenschaftlichen Arbeit kann man aus mehreren Theorien ein integriertes Modell entwickeln, aus dem die zu testenden Hypothesen abgeleitet werden.

Eine Studie kann auch dazu genutzt werden, **konkurrierende Theorien** im Vergleich zu testen. Zu den in der Öffentlichkeit wie in der Scientific Community kontrovers diskutierten möglichen Wirkungen gewalthaltiger Computerspiele liegen divergierende Theorien vor, aus denen konkurrierende Hypothesen ableitbar sind (Ferguson & Rueda, 2010):

Die **Theorie des sozialen Lernens** postuliert, dass gewalthaltige Medieninhalte als Vorbilder dienen und somit aggressives Denken und Handeln fördern:

- H1: Frustrierte Personen sind **aggressiver**, nachdem sie ein gewalthaltiges Computerspiel gespielt haben – im Vergleich zu einem nichtgewalthaltigen oder keinem Computerspiel.

Die **Katharsis-Theorie** postuliert, dass Mediennutzende ihre Aggressivität über gewalthaltige Medieninhalte symbolisch abreagieren und somit abbauen:

- H2: Frustrierte Personen sind **weniger aggressiv**, nachdem sie ein gewalthaltiges Computerspiel gespielt haben – im Vergleich zu einem nichtgewalthaltigen oder keinem Computerspiel.

Die **Mood-Management-Theorie** postuliert, dass Mediennutzende depressive Gefühle durch spannende und gewalthaltige Medieninhalte reduzieren können:

- H3: Frustrierte Personen sind **weniger depressiv**, nachdem sie ein gewalthaltiges Computerspiel gespielt haben – im Vergleich zu einem nichtgewalthaltigen oder keinem Computerspiel.

Lernquiz

1. Wie stellt man die Suchbegriffe für eine wissenschaftliche Literaturrecherche zusammen?
2. Auf welche Quellen stützt sich die Literaturrecherche bei wissenschaftlichen Arbeiten hauptsächlich?
3. Wie funktioniert eine Suche nach wissenschaftlicher Literatur gemäß dem Schneeballsystem?
4. Welche Typen von wissenschaftlichen Publikationen sind für die Literaturrecherche die nützlichsten?
5. Welche Bezugswege gibt es für wissenschaftliche Volltexte?
6. Welche fünf Fragen sollten bei der Darstellung des Forschungsstandes beantwortet werden?
7. Wie können Theorien gemäß der Reichweite ihrer Erklärungen eingeordnet werden?
8. Welches Ziel verfolgt die grafische Darstellung von Theorien?
9. Welche Computersimulationen werden in der Theorieentwicklung unterschieden und welche Ziele verfolgen sie?
10. Aus welchen unterschiedlichen Ausgangspunkten lassen sich Forschungsfragen und Forschungshypothesen ableiten?

Übungsaufgaben

1. Stellen Sie für das Forschungsthema „Social Networking Sites“ eine Liste primärer und sekundärer Suchbegriffe zusammen!
2. Führen Sie anhand der Suchbegriffe eine Literaturrecherche in zwei sozialwissenschaftlichen Datenbanken durch und vergleichen Sie die Ergebnisse!
3. Identifizieren Sie unter den Suchtreffern drei für die Erarbeitung des Forschungsstandes besonders einschlägige Quellen (z. B. aktuelle Review-Beiträge)!

Weiterführende Literatur

- Andermann, U., Drees, M., & Grätz, F. (2006). *Duden. Wie verfasst man wissenschaftliche Arbeiten?: Ein Leitfaden für das Studium und die Promotion* (3. Aufl.). Mannheim: Dudenverlag.
- Grix, J. & Watkins, G. (2010). *Information skills: Finding and using the right resources*. New York: Palgrave.
- Niedermair, K. (2010). *Recherchieren und Dokumentieren: Der richtige Umgang mit Literatur im Studium*. Stuttgart: UTB.

Literatur

Barlett, C. & Harris, R. (2008). The impact of body emphasizing video games on body image concerns on men and women. *Sex Roles, 59*(7), 586–601.

Bierhoff, H. W. (1997). Austauschtheorien. In D. Frey & S. Greif (Hrsg.), *Sozialpsychologie. Ein Handbuch in Schlüsselbegriffen* (4. Aufl., S. 136–140). Weinheim: Beltz.

Blumer, H. (1973). Der methodologische Standort des symbolischen Interaktionismus. In Arbeitsgruppe Bielefelder Soziologen (Hrsg.), *Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit Bd. I* (S. 80–101). Reinbek: Rowohlt.

Bossel, H. (1992). *Modellbildung und Simulation. Konzepte, Verfahren und Modelle zum Verhalten dynamischer Systeme*. Braunschweig: Vieweg.

Bühler, K. (1934). *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Frankfurt: Ullstein.

Carley, K. M. (2001). *Computational Approaches to Sociological Theorizing*. New York: Kluwer, Plenum.

Cassell, J. & Henry Jenkins, H. (Eds.). (1998). *From Barbie to Mortal Kombat: gender and computer games*. cambridge: the mit press.

Consalvo, M. (2008). *Crunched by passion: women game developers and workplace challenges*. Cambridge: MIT.

Conte, R. (2006). From simulation to theory (and backward). In F. Squazzoni (Ed.), *Epistemological aspects of computer simulation in the social sciences.* Second International Workshop, EPOS 2006, Brescia, Italy (pp. 29–47). Berlin: Springer.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science, 35*, 982–1003.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory with simulation methods. *Academy of Management Review, 32*(2), 580–599.

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829–848). London: Blackwell.

Dörner, D. (1994). Heuristik in der Theoriebildung. In T. Herrmann & W. H. Tack (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich B, Serie I, Bd. I, Methodologische Grundlagen der Psychologie* (S. 343–388). Göttingen: Hogrefe.

Esser, H. & Troitzsch, K. G. (Hrsg.). (1991). *Modellierung sozialer Prozesse. Neuere Aufsätze und Überlegungen zur soziologischen Theoriebildung*. Bonn: Informationszentrum Sozialwissenschaften.

Ferguson, C. J. & Rueda, S. M. (2010). The Hitman study: Violent video game exposure effects on aggressive behaviour, hostile feelings, and depression. *European Psychologist, 15*(2), 99–108.

Foroutan, N., Schäfer, K., Canan, C., & Schwarze, B. (2010). *Sarrazins Thesen auf dem Prüfstand – Ein empirischer Gegenentwurf zu Thilo Sarrazins Thesen zu Muslimen in Deutschland.* Abgerufen 5. September, 2013, unter http://www.heymat.hu-berlin.de/sarrazin2010.

Forrester, J. W. (1971). *World Dynamics*. Cambrige: Wright-Alben.

Freud, S. (1923). *Das Ich und das Es*. Leipzig: Internationaler Psychoanalytischer Verlag.

Garson, G. D. (2009). Computerized simulation in the social sciences: A survey and evaluation. *Simulation & Gaming, 40*(267).

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*(2), 254–267.

Gigerenzer, G. (1994). Woher kommen die Theorien über kognitive Prozesse? In A. Schorr (Hrsg.), *Die Psychologie und die Methodenfrage. Reflexionen zu einem zeitlosen Thema* (S. 109–127). Göttingen: Hogrefe.

Gilbert, N. & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd. ed.). New York: Mcgraw–Hill.

Glaser, B. G. & Strauss, A. L. (1999). *The discovery of grounded theory: Strategies for qualitative research* (8th ed.). Chicago: Aldine.

Graner-Ray, S. (2004). *Gender inclusive game design. Expanding the market*. Hingham: River.

Hanneman, R. & Patrick, S. (1997). On the uses of computer-assisted simulation modeling in the social science. *Sociological Research Online, 2*(2). Retrieved September 5, 2013, from http://www.socresonline.org.uk/2/2/5.html.

Hayes, E. (2011). The Sims as a catalyst for girls' IT learning. *International Journal of Gender, Science and Technology, 3*(1), 121-147. Retrieved September 5, 2013, from http://genderandset.open.ac.uk/index.php/genderandset/article/view/132.

Hogan, R. (2009). Much ado about nothing: The person-situation debate. *Journal of Research in Personality, 43*(2), 249.

Jansz, J., Avis, C., & Vosmeer, M. (2010). Playing The Sims 2: An exploration of gender differences in players' motivations and patterns of play. *New Media & Society, 12*(2), 235–251.

Jenson, J. & Castell, S. (2010). Gender, simulation, and gaming: Research review and redirections. *Simulation & Gaming, 41*(1), 51–71.

Jungermann, H., Pfister, H.-R., & Fischer, K. (2009). *Die Psychologie der Entscheidung: Eine Einführung*. Heidelberg: Spektrum.

Kafai, Y., Heeter, C., Denner, J., & Sun, Y. J. (Eds.). (2008). *Beyond Barbie and Mortal Kombat: New perspectives on gender and gaming*. Cambridge: Cambridge: MIT.

Kelley, H., H., Holmes, J. G., Kerr, N. L., Reis, H. T., Rusbult, C. E., & Van Lange, P. A. M. (2003). *An atlas of interpersonal situations*. Cambridge: Cambridge University Press.

Kirsh, S. J. (2010). *Media and Youth. A Developmental Perspective*: New York: Wiley-Blackwell.

Kleining, G. (1994). *Qualitativ-heuristische Sozialforschung. Schriften zur Theorie und Praxis*. Hamburg: Rolf Fechner.

Konijn, E. A., Nije Bijvank, M., & Bushman, B. J. (2007). I wish i were a warrior: The role of wishful identification in the effects of violent video games on aggression in adolescent boys. *Developmental Psychology, 43*(4), 1038–1044.

Krause, M. (2010). *Weibliche Nutzer von Computerspielen: Differenzierte Betrachtung und Erklärung Erklärung der Motive und Verhaltensweisen weiblicher Nutzer von Computerspielen*. Wiesbaden: VS Verlag.

Kreutz, H. & Bacher, J. (Hrsg.). (1991). *Disziplin und Kreativität. Sozialwissenschaftliche Computersimulation: theoretische Experimente und praktische Anwendung*. Opladen: Leske & Budrich.

Lazarro, N. (2008). Are boy games even necessary? In Y. Kafai, C. Heeter, J. Denner & J. Y. Sun (Eds.), *Beyond Barbie and Mortal Kombat: New perspectives on gender and gaming* (pp. 199–216). Cambridge: MIT.

Lefrancois, G. R. (2008). *Psychologie des Lernens* (4. Aufl.). Berlin: Springer.

Luhmann, N. (1984). *Soziale Systeme: Grundriß einer allgemeinen Theorie*. Frankfurt/M.: Suhrkamp.

Madanes, C. & Haley, J. (1977). Dimensions of family therapy. *Journal of Nervous and Mental Disease, 165*(2), 88–98.

Manhart, K. (1995). *KI-Modelle in den Sozialwissenschaften: Logische Struktur und wissensbasierte Systeme von Balancetheorien.* München: Oldenbourg.

Meeker, B. F. & Leik, R. K. (1997). Uses of computer simulation for theory construction: An evolving component of sociological research programs. In J. Szmatka, J. Skvoretz & J. Berger (Eds.), *Status, Network and Structure: Theory Construction and Theory Development* (pp. 47–70). Palo Alto: Stanford University Press.

Merton, R. K. (1949). *Social theory and social structure: Toward the codification of theory and research*. Glencoe: Free Press.

Miller, M. K. & Summers, A. (2007). Gender differences in video game characters' roles, appearances, and attire as portrayed in video game magazines. *Sex Roles, 56*(9–10), 733–742.

Mosler, H.-J., Schwarz, K., Ammann, F., & Gutscher, H. (2001). Computer simulation as a method of further developing a theory: Simulating the elaboration likelihood model (ELM). *Personality and Social Psychology Review, 5*(3), 201–215.

Peukert, P., Sieslack, S., Barth, G., & Batra, A. (2010). Internet– und Computerspielabhängigkeit: Phänomenologie, Komorbidität, ätiologie, Diagnostik und therapeutische Implikationen für Betroffene und Angehörige. *Psychiatrische Praxis, 37*(5), 219–224.

Popper, K. (1989). *Logik der Forschung* (9. Aufl.). Tübingen: Mohr.

Prescott, J. & Bogg, J. (2011). Segregation in a male-dominated industry: Women working in the computer games industry. *International Journal of Gender, Science and Technology, 3*(1). Retrieved September 5, 2013, from http://genderandset.open.ac.uk/index.php/genderandset/article/view/122.

Quandt, T., Grueninger, H., & Wimmer, J. (2009). The gray haired gaming generation: Findings from an explorative interview study on older computer games. *Games and Culture, 4*(1), 27–46.

Reinecke, L. (2009). Games and recovery: The use of video and computer games to recuperate from stress and strain. *Journal of Media Psychology: Theories, Methods, Applications, 21*(3), 126–142.

Sanford, K. & Madill, L. (2007). Understanding the power of new literacies through video game play and design. *Canadian Journal of Education, 30*(2), 432–455.

Sax, L. (2009). *Boys adrift: The five factors driving the growing epidemic of unmotivated boys and underachieving young men*. New York: Basic Books.

Schnell, R. (1990). Computersimulaltion und Theoriebildung in den Sozialwissenschaften. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42*, 109–128.

Schnell, R. (1991). Wer ist das Volk ? Zur faktischen Grundgesamtheit bei „allgemeinen Bevölkerungsumfragen": Undercoverage, Schwererreichbare und Nichtbefragbare. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43*(1), 106–137.

Schulz von Thun, F. (1991). *Miteinander reden 1: Störungen und Klärungen. Allgemeine Psychologie der Kommunikation*. Reinbeck: Rowohlt.

Shannon, C. E. & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: The University of Illinois Press.

Sonnentag, S. & Fritz, C. (2007). The recovery experience questionnaire: Development and validation of a measure for assessing recuperation and unwinding from work. *Journal of Occupational Health Psychology, 12*, 204–221.

Squazzoni, F. (2008). A (computational) social science perspective on societal transitions. *Computational & Mathematical Organization Theory, 14*(4), 266–282.

Stachowiak, F. J. (1992). Was leisten Computer in der Sprachtherapie? In W. Widdich, I. Ohlendorf, T. A. Pollow & J. P. Malin (Hrsg.), *Sprache uns Sprechen aus neurolinguistischer und medizinischer Sicht. Beiträge des ersten Rhein-Ruhr-Meetings* (S. 85–139). Bochum: Klinik Bergmannsheil.

Strube, G. (1990). Neokonnektionismus: Eine neue Basis für die Theorie und Modellierung menschlicher Kognitionen. *Psychologische Rundschau, 41*(129–143).

Trepte, S. & Reinecke, L. (2010). *Gender und games – Medienpsychologische Gender-Forschung am Beispiel Video- und Computerspiele*. Wiesbaden: VS Verlag.

Troitzsch, K. G., Mueller, U., & Gilbert, N. G. (1996). *Social science microsimulation* (Vol. 1). Berlin: Springer.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Ulrich, U. & Topcu, Ö. (2010). Sind Muslime dümmer? *DIE ZEIT ONLINE*. Abgerufen 5. September, 2013, unter http://www.zeit.de/2010/35/Sarrazin

Venkatesh, V. & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences, 39*, 273–315.

Venkatesh, V. & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46*, 186–204.

Watzlawick, P., Beavin, J. H., & Jackson, D. D. (2000). *Menschliche Kommunikation* (10. Aufl.). Bern: Hans Huber.

Weber, K. (2007). Simulationen in den Sozialwissenschaften. *Journal for General Philosophy of Science, 38*(1), 111–126.

Weis, R. & Cerankosky, B. C. (2010). Effects of video-game ownership on young boys' academic and behavioral functioning: A randomized, controlled study. *Psychological Science, 21*(4), 463–470.

Untersuchungsdesign

Nicola Döring, Jürgen Bortz

7.1 Quantitative, qualitative oder Mixed-Methods-Studie – 184

7.2 Grundlagen- oder anwendungswissenschaftliche Studie – 185

7.3 Empirische, Methoden- oder Theoriestudie – 186

7.4 Primär-, Sekundär- oder Metaanalyse – 191

7.5 Explorative, deskriptive oder explanative Studie – 192

7.6 Experimentelle, quasi-experimentelle oder nicht-experimentelle Studie – 193

7.7 Labor- oder Feldstudie – 205

7.8 Studie mit oder ohne Messwiederholungen – 208

7.9 Gruppen- oder Einzelfallstudie – 214

Literatur – 218

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_7, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Wissen, was man unter einem Untersuchungsdesign versteht.
- Verschiedene Klassifikationskriterien für Untersuchungsdesigns kennen.
- Die Vor- und Nachteile unterschiedlicher Untersuchungsdesigns abwägen können, insbesondere unter Berücksichtigung von Aufwand und Aussagekraft.
- Wissen, wie man das geeignete Untersuchungsdesign für das eigene Forschungsproblem auswählt.

Angenommen, Sie erfahren von einer Mitstudentin oder Kollegin, sie überprüfe gerade die Forschungshypothese, dass Computerspiele depressionsmindernd wirken. Sie können sich nun fragen, was der bisherige Forschungsstand dazu ergeben hat (wurde dies schon untersucht, unter welchen Bedingungen und mit welchen Ergebnissen?) und auf welche Theorie sie sich stützt (z. B. auf die Mood-Management-Theorie)? Im Hinblick auf das methodische Vorgehen der Studie ist v. a. das Untersuchungsdesign interessant:

1. Will sie rein quantitativ oder rein qualitativ forschen oder quantitative und qualitative Methoden zu einem Mixed-Methods-Design kombinieren?
2. Betreibt sie reine Grundlagenforschung oder handelt es sich um ein anwendungsbezogenes Forschungsprojekt, etwa um eine von einem Spieleentwickler in Auftrag gegebene Evaluationsstudie?
3. Ist es eine empirisch angelegte Untersuchung oder ist die Studie als reine Literatur- bzw. Theoriearbeit ausgerichtet oder vielleicht sogar als Methodenarbeit zu kennzeichnen, weil eine neue Methode der Datenerhebung oder Datenauswertung entwickelt und erprobt wird?
4. Ist die Studie so angelegt, dass sie auf die Analyse selbst erhobener Daten hinausläuft oder sollen im Rahmen einer Sekundäranalyse bereits vorhandene Datensätze neu ausgewertet oder im Rahmen einer Metaanalyse die Ergebnisse mehrerer Studien zu einem Gesamtbefund zusammengefasst werden?
5. Da sie eine Hypothese überprüfen will, scheint zumindest die Frage nach dem Erkenntnisinteresse der Studie bereits beantwortet: Die Untersuchung ist offenbar hypothesen- bzw. theorieprüfend und nicht theoriebildend oder populationsbeschreibend angelegt.
6. Soll die Hypothesenprüfung durch ein echtes Experiment erfolgen oder mittels einer quasi-experimentellen oder nicht-experimentellen Studie realisiert werden?
7. Sollen Daten im Forschungslabor oder im Feld – d. h. im natürlichen Lebensumfeld von Computerspielenden wie z. B. dem häuslichen Wohnzimmer – erhoben werden?
8. Sollen die Versuchspersonen nur zu einem einzigen Zeitpunkt, zu mehreren Zeitpunkten oder im Rahmen einer Längsschnittstudie sogar über einen längeren Zeitraum hinweg untersucht werden?
9. Ist eine Einzelfallstudie mit einer einzigen Person oder eine Gruppenuntersuchung mit einer Stichprobe geplant?

Die Wahl des Untersuchungsdesigns bzw. der Untersuchungsstrategie entscheidet ganz wesentlich darüber, welche **Aussagekraft** die wissenschaftlichen Befunde am Ende haben. Nicht nur in ihrem Erkenntniswert, auch in ihrem **forschungsökonomischen Aufwand** und in ihrer **Umsetzbarkeit** bei verschiedenen Zielgruppen und Sachverhalten unterscheiden sich Untersuchungsdesigns deutlich voneinander. Zudem differieren Designs im Hinblick auf **forschungsethische Aspekte** (▶ Abschn. 4.1): Während manche Untersuchungsdesigns kaum ethische Fragen aufwerfen (z. B. Sekundäranalyse mit öffentlich verfügbaren Daten), sind andere mit vielen ethischen Problemen behaftet (z. B. experimentelle Studien mit Kindern).

Untersuchungsdesign – Das Untersuchungsdesign (Untersuchungsart, Untersuchungsplan, Forschungsdesign, Studiendesign, „research design“) charakterisiert ganz allgemein die methodische Vorgehensweise einer Studie. Zur Kennzeichnung von Untersuchungsdesigns werden neun verschiedene Klassifikationskriterien herangezogen, die teilweise in einem hierarchischen Verhältnis zueinander stehen. Innerhalb eines konkreten Untersuchungsdesigns können unterschiedliche Stichprobenarten und Datenerhebungs- sowie Datenanalyseverfahren zum Einsatz kommen.

Ein einheitliches Klassifikationssystem für Untersuchungsdesigns fehlt, allerdings haben sich eine Reihe von **Beschreibungsdimensionen** in der Literatur etabliert, die im vorliegenden Kapitel behandelt werden (Montero & León, 2007; ◘ Tab. 7.1). Die verschiedenen Ausgestaltungsmöglichkeiten für Untersuchungsdesigns und ihre jeweiligen Vor- und Nachteile zu kennen ist wichtig, um eigene Studien planen sowie vorliegende Studien beurteilen zu können.

Hinsichtlich aller neun in ◘ Tab. 7.1 genannten Dimensionen von Untersuchungsdesigns sind im Zuge des Forschungsprozesses Entscheidungen zu treffen. Die Gründe für entsprechende Entscheidungen müssen im Sinne intersubjektiver Nachvollziehbarkeit als zentralem Kriterium von Wissenschaftlichkeit (▶ Abschn. 3.1.2) transparent gemacht werden. Bei der Bezeichnung einer Studie verzichtet man in der Regel darauf, alle Aspekte des Forschungsdesigns nennen, sondern beschränkt sich auf die wesentlichen. Wer z. B. mitteilt, die eigene Studie sei als „Laborexperiment“ angelegt, sagt damit indirekt bereits aus, dass es sich um eine explanative quantitative Primärstudie handelt.

Tabelle 7.1 Neun Klassifikationskriterien für Untersuchungsdesigns

Kennzeichen des Untersuchungsdesigns	Varianten von Untersuchungsdesigns
1. Wissenschaftstheoretischer Ansatz der Studie	– Quantitative Studie („quantitative study“) – Qualitative Studie („qualitative study“) – Mixed-Methods-Studie („mixed methods study“)
2. Erkenntnisziel der Studie	– Grundlagenwissenschaftliche Studie („basic research study“) – Anwendungswissenschaftliche Studie („applied research study“) a) Unabhängige Studie („non-commercial research study“, „independent research study“, „researcher-initiated study“) b) Auftragsstudie („contract research study“, „commercial research study“)
3. Gegenstand der Studie	– Empirische Studie („empirical study“) a) Originalstudie („original study“) b) Replikationsstudie („replication study“) – Methodenstudie („methodological study“) – Theoriestudie („theoretical study“, „research/literature review“) a) Review/Forschungsüberblick („review of research“) b) Metaanalyse („meta analysis“)
4. Datengrundlage bei empirischen Studien	– Primäranalyse („primary analysis“) – Sekundäranalyse („secondary analysis“) – Metaanalyse („meta analysis“)
5. Erkenntnisinteresse bei empirischen Studien	– Explorative (gegenstandsbeschreibende/theoriebildende) Studie („exploratory study“) – Deskriptive (populationsbeschreibende) Studie („population descriptive study“) – Explanative (hypothesenprüfende) Studie („explanatory study“)
6. Bildung und Behandlung von Untersuchungsgruppen bei explanativen Studien	– Experimentelle Studie („experimental study“, „true experiment“) bzw. randomisierte kontrollierte Studie („randomized controlled trial“ RCT) – Quasi-experimentelle Studie („quasi-experimental study“) bzw. nicht-randomisierte kontrollierte Studie („non-randomized controlled trial“ NRCT) – Nicht-experimentelle Studie („non experimental study“)
7. Untersuchungsort bei empirischen Studien	– Laborstudie („laboratory study“) – Feldstudie („field study“)
8. Anzahl der Untersuchungszeitpunkte bei empirischen Studien	– (Quasi-)Experimentelle Studien mit und ohne Messwiederholungen a) (Quasi-)Experimentelle Studie ohne Messwiederholungen („independent measures designs“, „between-subjects design“) b) (Quasi-)experimentelle Studie mit Messwiederholungen („repeated measures/measurement designs“; „within- subjects design“) – Nicht-experimentelle Studien mit und ohne Messwiederholungen a) Querschnittstudie („cross-sectional study“) b) Trendstudie („trend study“) c) Längsschnittstudie („longitudinal study“)
9. Anzahl der Untersuchungsobjekte bei empirischen Studien	– Gruppenstudie („group study“) a) Stichprobenstudie („sample study“) b) Vollerhebung („population study“) – Einzelfallstudie („single participant study“)

Bei Entscheidungen hinsichtlich des Forschungsdesigns sind die geplanten Methoden der Datenerhebung und Datenauswertung schon einzubeziehen. Denn es nutzt Studierenden mit wenig Forschungserfahrung nichts, wenn sie ein komplexes Design realisieren (z. B. mehrere Untersuchungsgruppen zu mehreren Untersuchungszeitpunkten untersuchen), später aber die erhobenen Daten überhaupt nicht sachgerecht statistisch auswerten können. Wir empfehlen bereits vor der Datenerhebung zu prüfen, welche qualitativen und/oder quantitativen Ver-

fahren der Datenanalyse zum Einsatz kommen sollen und ob diese beherrscht werden. Neben **forschungsökonomischen Fragen** der Umsetzbarkeit eines Designs durch die jeweils Forschenden sind es jedoch vordringlich **erkenntnistheoretische Fragen**, die den Ausschlag dafür geben, mit welchem Untersuchungsdesign das gewählte Forschungsproblem vor dem Hintergrund des aktuellen Forschungsstandes am besten bearbeitet werden soll.

7.1 Quantitative, qualitative oder Mixed-Methods-Studie

Ein wichtiges Klassifikationskriterium für wissenschaftliche Studien und Untersuchungsdesigns ist die Zuordnung zu einem **wissenschaftstheoretischen Ansatz bzw. Paradigma**. Lange Zeit herrschte ein erkenntnistheoretischer Streit, ob soziale Sachverhalte besser **qualitativ** in der Tradition der Geisteswissenschaften oder **quantitativ** in der Tradition der Naturwissenschaften zu untersuchen seien. Heute gilt diese Lagerbildung in den Sozialwissenschaften als weitgehend überholt: Beide wissenschaftstheoretischen Paradigmen bzw. Forschungsansätze haben ihre Vor- und Nachteile, keiner ist per se besser (▶ Abschn. 1.1.3 und ▶ Kap. 2). Eine gute Studie greift stets auf diejenige qualitative oder quantitative Forschungsstrategie zurück, die am besten zum Forschungsproblem passt. Zudem sind beide Ansätze in unterschiedlicher Weise systematisch zu **Mixed-Methods-Designs** kombinierbar (▶ Abschn. 2.4).

Qualitativer Forschungsansatz – Im qualitativen Forschungsansatz („qualitative research approach/strategy“) werden offene Forschungsfragen an wenigen Untersuchungseinheiten sehr detailliert mit unstrukturierten oder teilstrukturierten Datenerhebungsmethoden untersucht. Ziel ist eine Gegenstandsbeschreibung samt Theoriebildung. Die erhobenen qualitativen (nicht-numerischen, d. h. verbalen, visuellen) Daten werden interpretativ ausgewertet.

Quantitativer Forschungsansatz – Im quantitativen Forschungsansatz („quantitative research approach/strategy“) werden theoretisch abgeleitete Forschungshypothesen an vielen Untersuchungseinheiten mit strukturierten Datenerhebungsmethoden untersucht. Ziel ist meist die Theorieprüfung. Die erhobenen quantitativen (numerischen) Daten werden statistisch ausgewertet.

Mixed-Methods-Ansatz – Der Mixed-Methods-Ansatz („mixed methods approach“) kombiniert qualitative und quantitative Vorgehensweisen zu neuen Mixed-Methods-Designs.

Der **qualitative Forschungsansatz** hat sich aus den Geisteswissenschaften entwickelt und arbeitet primär hermeneutisch, d. h. interpretiert systematisch Textmaterial (z. B. Beobachtungsprotokolle und Interviewtranskripte) sowie auch Bildmaterial (z. B. Fotos). Typisch für den qualitativen Ansatz sind ein bewusst nicht oder wenig strukturiertes Vorgehen, das unerwartete Befunde ermöglicht. Auch konzentriert man sich meist auf eine sehr detaillierte und umfassende Analyse weniger Fälle, die im natürlichen Umfeld untersucht werden. Der qualitative Forschungsansatz versucht, offene Forschungsfragen zu beantworten, passt sich möglichst flexibel dem Gegenstand an und läuft auf Theoriebildung hinaus (▶ Abschn. 2.3). Qualitative Forschung legt eigene Gütekriterien zugrunde (z. B. Vertrauenswürdigkeit, Authentizität; ▶ Abschn. 3.3).

Die Wurzeln des **quantitativen Forschungsansatzes bzw. Paradigmas** liegen in den Naturwissenschaften (▶ Abschn. 2.2). Im quantitativen Paradigma werden die Ausprägungen von Variablen gemessen und die Messwerte statistisch ausgewertet. Typisch für den quantitativen Ansatz sind die statistische Analyse vieler Fälle und/oder die Durchführung von Experimenten im Labor. Der quantitative Forschungsansatz zielt oft auf die Überprüfung theoretisch abgeleiteter Hypothesen, arbeitet mit strukturierten Abläufen sowie standardisierten Datenerhebungsinstrumenten und läuft auf die Weiterentwicklung von Theorien hinaus. In der quantitativen Forschung werden Untersuchungsdesigns v. a. nach den Kriterien der **Objektivität** (intersubjektiven Nachvollziehbarkeit einer Studie, Unabhängigkeit von den forschenden Personen), der **Replizierbarkeit** (Wiederholbarkeit einer Studie) sowie der **Validität** (Gültigkeit der Befunde einer Studie) beurteilt. Die Validität ist dabei das wichtigste Kriterium und differenziert sich in **interne Validität** (Schlüssigkeit der nachgewiesenen Ursache-Wirkungs-Beziehungen) und **externe Validität** (Generalisierbarkeit der Untersuchungsergebnisse auf andere als genau die untersuchten Bedingungen und Personengruppen; zu Gütekriterien quantitativer Studien ▶ Abschn. 3.2).

Der **Mixed-Methods-Ansatz** („mixed methods research“) verbindet qualitatives und quantitatives Forschungsparadigma mit unterschiedlichen Integrationsstrategien (▶ Abschn. 2.4). Im Unterschied zu herkömmlichen **Mono-Method-Designs**, die entweder einen qualitativen oder einen quantitativen Forschungsprozess vorsehen, kombinieren **Mixed-Methods-Designs** qualitative und quantitative Forschungsstrategien im Rahmen einer einzigen Studie miteinander, typischerweise indem nacheinander oder gleichzeitig **qualitative und quantitative Teilstudien** durchgeführt werden. Die Teilstudien stehen dabei jedoch nicht unabhängig nebeneinander, sondern sind direkt aufeinander bezogen. Beim **Vorstudienmodell** als einem vergleichsweise einfachen Mixed-Methods-Design dient eine qualitative Studie als Vorstudie dazu, Hypothesen zu generieren (z. B. Durchführung aus-

führlicher Leitfadeninterviews mit wenigen Personen), die anschließend in einer quantitativen Studie (z. B. Verteilung eines standardisierten Fragebogens an eine große Stichprobe) überprüft werden. Die Ergebnisse der qualitativen Teilstudie fließen hier in die Hypothesenbildung und Fragebogenkonstruktion der quantitativen Teilstudie ein. Umgekehrt ist das Vorgehen beim **Vertiefungsmodell** als einem weiteren Mixed-Methods-Design: Hier folgt auf eine quantitative Studie (z. B. standardisierte Fragebogenerhebung mit großer Stichprobe) eine qualitative Studie, in deren Rahmen einige ausgewählte Respondenten aus der Fragebogenerhebung noch einmal ausführlich mit Leitfadeninterviews befragt werden (Mayring, 2001). Da in der quantitativen und qualitativen Teilstudie beim Vertiefungsmodell teilweise dieselben Personen untersucht werden, können die qualitativen und quantitativen Befunde direkt aufeinander bezogen und in ihrer Gesamtheit interpretiert werden. Darüber hinaus existieren viele weitere, teilweise sehr komplexe Varianten von Mixed-Methods-Designs (Bryman, 2008; Creswell & Plano Clark, 2010; Teddlie & Tashakkori, 2009).

Entscheidungskriterien

- Wählen Sie ein **qualitatives Untersuchungsdesign**, wenn Sie offene Forschungsfragen an wenigen Fällen hinsichtlich vieler Aspekte detailliert bearbeiten, einen neuen Gegenstand erkunden und eine Hypothese oder Theorie entwickeln wollen. Bevor eine umfangreiche Studie (z. B. Abschlussarbeit) in Angriff genommen wird, sollten Vorerfahrungen mit kleineren qualitativen Studien (z. B. Seminararbeiten) gesammelt worden sein.
- Wählen Sie ein **quantitatives Untersuchungsdesign**, wenn Sie Forschungsfragen oder Forschungshypothesen an vielen Fällen hinsichtlich weniger Aspekte bearbeiten, bereits erforschte Sachverhalte unter spezifischen Fragestellungen betrachten und etablierte Theorien überprüfen möchten. Sie sollten über das notwendige statistische Handwerkszeug verfügen, um die erhobenen Daten dann auch sachgerecht analysieren zu können.
- Wählen Sie bevorzugt ein **Mixed-Methods-Design**, wenn Sie eine besonders umfassende Studie realisieren möchten, über solide Vorkenntnisse sowohl in qualitativer als auch quantitativer sowie Mixed-Methods-Forschung verfügen und genügend zeitliche, finanzielle und personelle Ressourcen bereitstehen.

7.2 Grundlagen- oder anwendungswissenschaftliche Studie

Neben dem wissenschaftstheoretischen Ansatz ist das **Erkenntnisziel** ein weiteres zentrales Klassifikationskriterium für die Anlage wissenschaftlicher Studien. Eine Untersuchung kann darauf abzielen, **wissenschaftliche Probleme** zu lösen, indem Theorien und Methoden weiterentwickelt oder bestimmte Sachverhalte genauer analysiert werden. Eine solche **Grundlagenforschung** („basic resarch“) trägt zum Erkenntnisgewinn der jeweiligen Wissenschaftsdisziplin bei. Zwar werden mögliche praktische Konsequenzen oder Anwendungsmöglichkeiten grundlagenwissenschaftlicher Erkenntnisse in den jeweiligen Arbeiten diskutiert, ein Praxisnutzen wird jedoch nicht unmittelbar angestrebt. Grundlagenforschung wird an Universitäten, Fachhochschulen und Forschungseinrichtungen betrieben und ist zu einem großen Teil staatlich finanziert. Dabei ist die Freiheit der Forschung in Deutschland im Grundgesetz verankert, so dass eine unmittelbare politische Einflussnahme ausgeschlossen wird. Durch Forschungsförderprogramme auf der Ebene der EU, des Bundes und der Länder wird die Grundlagenforschung in ihrer Themenwahl dennoch gesteuert. Der Wert der freien Grundlagenforschung liegt in der Weiterentwicklung wissenschaftlicher Erkenntnis und wird über Publikationen und deren Zitationshäufigkeit nachgewiesen. Oft erweisen sich grundlagenwissenschaftliche Erkenntnisse im Nachhinein an vorher unerwarteter Stelle durchaus als nützlich für die Praxis (▶ Abschn. 1.1.3).

Grundlagenwissenschaftliche Studie – Die grundlagenwissenschaftliche Studie („basic research study“) dient primär dem wissenschaftlichen Erkenntnisfortschritt.

Anwendungswissenschaftliche Studie – Die anwendungswissenschaftliche Studie („applied research study“) dient primär der Lösung praktischer Probleme mithilfe wissenschaftlicher Methoden und Theorien. Anwendungsforschung findet als unabhängige Forschung („independent research“, „non-commercial research“) oder als Auftragsforschung („contract resarch“, „commercial research)“ statt.

Im Unterschied zur Grundlagenforschung zielt die Anwendungsforschung bzw. **angewandte Forschung** („applied science“) hinsichtlich ihres Erkenntnisziels auf die Lösung **praktischer Probleme** oder die Verbesserung von Maßnahmen und Technologien ab. Ihr Erkenntnisgewinn wird an praxisrelevanten Ergebnissen gemessen. Studien im Bereich der Anwendungsforschung sind weniger generalisierbar bzw. stärker auf konkrete Praxisprobleme zugeschnitten. Angewandte Forschung findet häufig im Feld (und nicht im Forschungslabor) statt und bezieht sich auf vordefinierte Zielgruppen, so dass

Studienbeispiel

Angewandte und grundlagenwissenschaftliche Studien zu kindlichen Zeugenaussagen vor Gericht

In der **angewandten Forschung zur Qualität kindlicher Zeugenaussagen** werden unter anderem Studien durchgeführt, die gerichtliche Zeugenbefragungen möglichst praxisnah simulieren (z. B. indem per Rollenspiel ein Gerichtsprozess inszeniert und das Verhalten der kindlichen Zeugen beobachtet wird). Zudem haben sich aber auch Ergebnisse **grundlagenwissenschaftlicher Studien zur Entwicklung des kindlichen Gedächtnisses** (z. B. Experimente auf der Basis der kognitiven Skripttheorie) als fruchtbar erwiesen, wenn es um das praktische Problem geht, mit welchen Interviewtechniken kindliche Zeugen optimal zu befragen sind (Tang, 2006).

die Praxisbedingungen die Spielräume bei der Gestaltung des Forschungsdesigns stärker einengen als in der Grundlagenforschung (▶ Studienbeispiel „Angewandte und grundlagenwissenschaftliche Studien zu kindlichen Zeugenaussagen vor Gericht“).

Anwendungsforschung wird ebenso wie Grundlagenforschung vonForschungsfördereinrichtungen finanziert und findet auch an staatlichen Hochschulen statt. Dann spricht man von akademischer, nichtkommerzieller bzw. **unabhängiger Forschung** („independent research“, „non-commercial research“). Darüber hinaus handelt es sich bei Anwendungsforschung oft um direkte **Auftragsforschung** („contract research“, „commercial research“), etwa wenn ein Unternehmen als Anwendungsstudie eine wissenschaftliche Evaluation für ein Produkt in Auftrag gibt. Bei der Auftragsforschung sind die Auftraggebenden als Mitentscheider mehr oder minder intensiv in den Forschungsprozess einzubeziehen, insbesondere bei der Präzisierung des Forschungsproblems und der einzelnen Forschungsfragen bzw. Forschungshypothesen, aber auch bei der Untersuchungsplanung. Dies kann Freiheitsgrade bei der Ausgestaltung des Forschungsdesigns einschränken. Ergebnisse der Auftragsforschung sind wegen Geheimhaltungsverpflichtungen oft nicht oder erst nach einer Sperrfrist publizierbar. Ihr Wert bemisst sich primär an der Zufriedenheit des Auftraggebers.

Generell besteht bei der human- und sozialwissenschaftlichen Auftragsforschung die Herausforderung darin, einerseits mit den Auftraggebenden zu kooperieren, andererseits aber auch wissenschaftliche Unabhängigkeit zu wahren. Im Sinne der Transparenz ist eine **Nennung der Studienauftraggeber** in Publikationen notwendig (zur Wissenschaftsethik ▶ Abschn. 4.2). Somit lassen sich dann z. B. unabhängige Medikamententests von denjenigen abgrenzen, die von der Pharmaindustrie in Auftrag gegeben bzw. finanziert werden. Darüber hinaus wird v. a. in der medizinischen Forschung gefordert, dass Forschende mögliche **Interessenskonflikte offenlegen**, die indirekt die Befunde sowie deren Publikation beeinflussen könnten (z. B. wenn Forschende ein untersuchtes Produkt selbst vermarkten oder an einem entsprechenden Unternehmen wirtschaftlich beteiligt sind).

Entscheidungskriterien

- Wählen Sie bevorzugt eine **grundlagenwissenschaftliche Studie**, wenn Sie im wissenschaftlichen Bereich weiterarbeiten und/oder bewusst unabhängig von Praxisbedingungen und Auftraggebenden bleiben möchten.
- Wählen Sie bevorzugt eine **anwendungswissenschaftliche Studie**, wenn Sie über entsprechende Praxiskontakte verfügen und sich in ihrer Forschung eher berufsfeldorientiert ausrichten wollen.
- Sofern Sie eine anwendungswissenschaftliche Studie als **Auftragsstudie** durchführen können, wägen Sie die Vorteile (z. B. finanzielle Unterstützung, mögliche Weiterbeschäftigung beim Auftraggeber) und die Nachteile ab (z. B. größerer Aufwand für Abstimmungsprozesse mit dem Auftraggeber; unterschiedliche Bewertungskriterien für eine Qualifikationsarbeit vonseiten der Auftraggebenden in der Praxis und der Betreuenden in der Wissenschaft).

7.3 Empirische, Methoden- oder Theoriestudie

Auf einer übergeordneten Ebene lassen sich Studien danach unterscheiden, welchem **Gegenstand** sie sich widmen: Wird ausschließlich der bisherige Theorie- und Forschungsstand auf der Basis von Publikationen bearbeitet, so handelt es sich um eine Theoriestudie. Werden wissenschaftliche Forschungsmethoden zum Gegenstand einer Untersuchung, so liegt eine Methodenstudie vor. Werden anhand von neu erhobenen Daten inhaltliche Forschungsfragen und Forschungshypothesen bearbeitet, so spricht man von einer empirischen Studie.

Theoriestudie – Die Theoriestudie („theoretical study“, „research/literature review“) präsentiert und bewertet auf der Basis einer Literaturrecherche den Forschungsstand. Der Forschungsstand wird entweder in einem **Review bzw. Überblicksartikel** zusammengefasst („research review“) oder **mittels** Metaanalyse **(„meta analysis“) werden die zentralen statistischen** Ergebnisse bisheriger Studien zu einem Gesamtbefund aggregiert.

Methodenstudie – Die Methodenstudie („methodological study“) dient dem Vergleich und der Weiterentwicklung von Forschungsmethoden.

Empirische Studie – Die empirische Studie („empirical study“) dient der Lösung von inhaltlichen Forschungsproblemen auf der Basis systematischer eigener Datenerhebung und/oder Datenanalyse, wobei es sich um eine Originalstudie („original study“) oder um eine Replikationsstudie („replication study“) handeln kann.

Jede wissenschaftliche Arbeit ist darauf angewiesen, den bisherigen Forschungsstand zum Thema aufzuarbeiten und einen theoretischen Hintergrund für das eigene Forschungsproblem zu entwickeln (▶ Kap. 6). Studien können sich auch gänzlich auf diese Aufgabe konzentrieren und auf die Erhebung und Auswertung eigener Daten verzichten. Entsprechende Studien werden als **Theoriearbeiten** („theoretical studies“) bezeichnet. Theoriearbeiten sind für den Erkenntnisfortschritt sehr wichtig, indem sie Einzelstudien zusammenfassend darstellen und bewerten. Die Durchführung einer Theoriearbeit hat forschungsökomisch den Vorteil, dass sie als reine **Schreibtischforschung** („desk research“) absolviert werden kann und keine Abhängigkeit von anderen Personen besteht (z. B. entfällt die Rekrutierung von Untersuchungspersonen). Eine gute Theoriearbeit erfordert indessen sehr systematische Recherchen, um zunächst alle relevanten Quellen zu identifizieren. Zudem müssen diese gründlich durchgearbeitet, verstanden, sinnvoll strukturiert und kritisch beurteilt werden. Für Neulinge in einem wissenschaftlichen Gebiet können sich theoretische Übersichtsarbeiten deswegen als zu anspruchsvoll erweisen. **Zwei Varianten von Theoriearbeiten** sind zu differenzieren:

- **Review/Forschungsüberblick** („research review“, „literature review“). Ein Forschungsüberblick geht auf den aktuellen Stand an Theorien, Methoden und v. a. auch Befunden zu einem Thema ein und beschreibt und bewertet diese (z. B. aktuelle Review-Artikel zu Sexualität im Internet: Döring, 2009; zu intravenösem Drogengebrauch: Wagner, Unger, Bluthenthal, Andreeva, & Pentz, 2010; oder zu Diversität im Topmanagement: Nielsen, 2010). Ein Sonderfall des Forschungsüberblicks ist der **Theorieüberblick** im engeren Sinne; er konzentriert sich auf den systematischen Vergleich von theoretischen Ansätzen in einem bestimmten Themenfeld („review of theories“; z. B. ein Vergleich von verschiedenen Emotionstheorien: Moors, 2009). Es existieren auch reine Theoriearbeiten, die unterschiedliche Theorien und Konzepte argumentativ in ein **neues theoretisches Modell** zusammenfügen, dessen empirische Prüfung dann zukünftigen Studien obliegt (z. B. integratives Modell kognitiver Komplexität von Psychotherapeuten: Owen & Lindley, 2010). Entsprechende rein literaturbasierte Übersichtsartikel findet man in den üblichen Fachzeitschriften sowie in speziellen **Review-Zeitschriften** (z. B. Zeitschriften wie „Review of Educational Research“, „Psychological Review“) und auch in disziplinären Jahrbüchern (kommunikationswissenschaftliches „Communication Yearbook“ der International Communication Association). Das Review als wissenschaftlicher Überblicksbeitrag ist nicht zu verwechseln mit dem Review im Sinne einer Buchrezension.
- **Metaanalyse** („meta analysis“). Metaanalysen fassen die qualitativen oder quantitativen Ergebnisse direkt vergleichbarer Studien zum selben Sachverhalt durch eine qualitative oder quantitative Metaauswertung zu einem Gesamtergebnis zusammen. Metaanalysen werden teilweise den Theoriearbeiten zugeordnet, da sie keine eigene Datenerhebung beinhalten (wie das bei Primärstudien der Fall ist) und auch keine Fremddatensätze auswerten (wie dies Sekundäranalysen tun), sondern auf publizierte Studienergebnisse zurückgreifen. Was sie allerdings von sonstigen Theoriearbeiten unterscheidet, ist die Anwendung statistischer Auswertungsmethoden auf die Studienergebnisse, was entsprechendes Know-how in der Analysemethodik verlangt (zur Metaanalyse ▶ Kap. 16).

Methodenstudien („methodological studies“) zielen auf die Überprüfung und Weiterentwicklung von qualitativen und/oder quantitativen Forschungsmethoden ab. Dabei kann es sich um Innovationen bei Forschungsdesigns (z. B. Vergleich von Labor- und Online-Experiment), um die Konstruktion neuer Skalen und Testverfahren (z. B. Entwicklung einer neuen Skala zur Messung eines bestimmten Merkmals) oder um Varianten der Datenerhebung (z. B. Umgang mit sensiblen Fragen im Interview ▶ Studienbeispiel „Methodenstudie ‚Sensible Fragen‘“; Auswirkung von unterschiedlichen Belohnungen auf die Teilnahmebereitschaft von Versuchspersonen) und nicht zuletzt auch um eine Überprüfung und Verbesserung von Analysemethoden gehen (z. B. Entwicklung neuer statistischer Verfahren). Methodenstudien gelten als besonders anspruchsvoll und werden von Studierenden selten in Angriff genommen. Tatsächlich würde es die meisten Studierenden überfordern, neue statistische oder interpretative Methoden der Datenanalyse zu entwickeln und zu testen.

Eine Reihe von Methodenstudien lassen sich jedoch mit relativ einfachen empirischen Methoden umsetzen, insbesondere **Methodenvergleichsstudien** sind prinzipiell auch von weniger erfahrenen Forschenden gut zu bewältigen. Eine empirische Methodenstudie ist nicht komplizierter, sondern teilweise sogar einfacher als eine inhaltliche empirische Studie. Methodenstudien kommen oft ohne einen umfassenden theoretischen Hintergrund aus, da eine praktische methodische Frage im Zentrum

Studienbeispiel

Methodenstudie „Sensible Fragen"

Mündliche Befragungen, in denen heikle oder sensible Themen angesprochen werden sollen (z. B. sexuelle oder kriminelle Verhaltensweisen, extreme Einstellungen), können sowohl von Befragten als auch von Interviewenden als unangenehm erlebt werden. Dies kann dazu führen, dass Interviewpersonen die Antwort verweigern oder das ganze Gespräch abbrechen, weil ihnen die Fragen zu persönlich werden. Es kann auch sein, dass verfälschte – z. B. sozial erwünschte – Antworten gegeben werden, um vor dem Interviewer das Gesicht zu wahren. Auch kommt es vor, dass Interviewende die heiklen Fragen in einem standardisierten Interview überspringen, um peinliche Situationen zu umgehen, und nachher fiktive Daten eintragen. Um derartige Probleme zu vermeiden, werden in der Methodenliteratur drei unterschiedliche Strategien vorgeschlagen, wie ein konfrontatives „direktes Abfragen" heikler Sachverhalte durch ein subtileres Vorgehen ersetzt werden kann:

- **Wording-/Framing-Techniken** („wording/framing technique"): Ein als heikel wahrnehmbarer Sachverhalt wird in der Formulierung entschärft, etwa indem man darauf hinweist, dass es sich um ein verbreitetes Phänomen handelt.
- **Technik des vertraulichen Kuverts** („sealed envelope technique"): Die heiklen Fragen werden nicht mündlich gestellt, sondern am Ende des Interviews von den Befragten schriftlich beantwortet. Der entsprechende Kurzfragebogen wird in einen anonymen Umschlag gesteckt und verschlossen. Er darf nicht vom Interviewer, sondern nur vom Forschungsinstitut geöffnet werden.
- **Randomized-Response-Technik** („randomized response technique"): Die Befragten beantworten nach einem von ihnen selbst gesteuerten Zufallsprinzip (z. B. Kartenspiel, Würfel, Münzwurf, dessen Ergebnis der Interviewer nicht sieht) entweder ehrlich die heikle Frage (z. B. ob man schon einmal illegale Drogen konsumiert hat) oder geben eine vordefinierte andere Antwort. Beim **Unrelated-Question-Ansatz** wäre z. B. bei Münzwurf „Kopf" die heikle Frage ehrlich zu beantworten, bei Münzwurf „Zahl" eine ehrliche Antwort auf eine inhaltlich andere Frage zu geben (z. B. ob die eigene Mutter an einem geraden oder ungeraden Tag geboren wurde). Beim **Forced-Response-Ansatz** wäre z. B. beim Wurf von zwei Münzen bei 1-mal Zahl und 1-mal Kopf die heikle Frage ehrlich zu beantworten, bei 2-mal Zahl ist einfach mit „Ja", bei 2-mal Kopf mit „Nein" zu antworten. Der Interviewer weiß also im konkreten Einzelfall nicht, was die gegebene Ja-/Nein-Antwort zu bedeuten hat. Bei der Auswertung der Daten lässt sich jedoch der wahre Wert für die interessierende Merkmalsausprägung der befragten Stichprobe bestimmen (z. B. Prozentwert der Drogenkonsumenten), indem die Zufallsantworten herausgerechnet werden.

Im Rahmen einer Lehrveranstaltung wurde gemeinsam mit Studierenden eine Methodenstudie durchgeführt, die diese drei indirekten Abfragetechniken einem direkten empirischen Vergleich mit der direkten Abfrage unterzog (Preisendörfer, 2008): Aus dem Immatrikulationsverzeichnis der Universität Mainz wurde eine echte Zufallsstichprobe von Studierenden gezogen. Die Interviewpersonen wurden auf die studentischen Mitglieder des Forschungsteams verteilt. Es konnten halbstündige persönliche Interviews mit $n = 578$ Studierenden realisiert werden. Bei 37 % der Interviews wurden heikle Themen direkt abgefragt, bei 63 % der Interviews kam eine der drei indirekten Strategien zum Einsatz. Die Ergebnisse zeigten zwar bei einzelnen Fragen höhere Antworthäufigkeiten bei den indirekten Techniken, über mehrere Themen hinweg ergab sich jedoch ein eher ernüchterndes Bild: Die Methodeneffekte waren meist sehr klein oder sogar inkonsistent, d. h. zuweilen wurde bei direkter Abfrage mehr normverletzendes Verhalten zugegeben als bei indirekter Abfrage. Dies unterstreicht jedoch umso mehr die Bedeutung von Methodenvergleichsstudien, um die Wirksamkeit methodischer Varianten zuverlässig einschätzen zu können (Tab. 7.2).

steht. Sie sind jedoch auf eine sorgfältige methodische Umsetzung angewiesen. Der Vorteil von Methodenstudien liegt darin, dass ihre Ergebnisse oft von vielen anderen Forschenden – unabhängig von deren inhaltlichen Schwerpunkten – genutzt werden können. Die **meistzitierten wissenschaftlichen Publikationen** in den Sozialwissenschaften sind Methodenbeiträge.

Empirische Studien („empirical studies"), die sich der Lösung inhaltlicher Forschungsprobleme widmen, sind der am weitesten verbreitete Untersuchungstyp. Aufwand und Anspruch empirischer Studien mit einem qualitativen, quantitativen oder Mixed-Methods-Design variieren in Abhängigkeit von den konkreten Design-Entscheidungen auf den hier behandelten neun Beschreibungsdimensionen (z. B. Einzelfall- oder Gruppenstudie; Studie mit oder ohne Messwiederholungen etc.). Darüber hinaus hängt der Aufwand einer empirischen Studie auch von Art und Umfang der Stichprobe (▶ Kap. 9) sowie den gewählten Datenerhebungs- sowie Datenanalysemethoden ab. Bei empirischen Studien unterscheidet man zwischen **Originalstudien** („original studies"), deren Design selbst entwickelt wurde, und **Replikationsstudien** bzw. Wiederholungsstudien („replication studies"), deren Design sich an dem einer bereits durchgeführten Studie orientiert.

Replikationsstudien durchzuführen ist besonders empfehlenswert, wenn die eigene Forschungserfahrung noch überschaubar ist. Denn bei einer Wiederholungsstudie kann man sich an einem bereits entwickelten Forschungsdesign orientieren. Voraussetzung ist jedoch, dass die entsprechende Publikation der Studie sehr detailliert auf das methodische Vorgehen eingeht und auch die verwendeten Datenerhebungsinstrumente (z. B. Interviewleitfaden, Fragebogen) enthält bzw. diese über die Autoren zugänglich sind (für ein Beispiel einer US-amerikanischen Replikationsstudie, die untersucht, wie Eltern ihre Kinder über das Thema Kindesmissbrauch aufklären; Deblinger, Thakkar-Kolar, Berry, & Schroeder, 2010). Zwei Varianten von Replikationsstudien sind zu unterscheiden:

Tabelle 7.2 Methodenvergleich von direkten und indirekten Abfragemethoden bei heiklen Themen. (Aus Preisendörfer, 2008, S. 18, 20, 21)

Heikle Themen	Direkte Abfrage	Indirekte Abfrage	Signifikanztestergebnis (Irrtumswahrscheinlichkeit)
Illegaler Drogenkonsum	**Direkte Abfrage**: „Hast du jemals illegale Drogen genommen?“	**Wording/Framing-Technik**: „Auch VIPs, also sehr bedeutsame Personen aus dem öffentlichen Leben berichten, in ihrer Jugend sogenannte illegale Drogen (angefangen von Marihuana bis hin zu härtere Sachen) probiert und genommen zu haben. Hast du jemals illegale Drogen genommen?“	
	54.8 % ($n = 211$)	60.5 % ($n = 367$)	$p = .18$ nicht signifikant
Homosexuelle Kontakte	**Direkte Abfrage**: „Hattest du in deinem bisherigen Leben jemals homosexuelle Kontakte?“	**Vertrauliches Kuvert**: „Hattest du in deinem bisherigen Leben jemals homosexuelle Kontakte?“	
	14.8 % ($n = 211$)	18.5 % ($n = 367$)	$p = .26$ nicht signifikant
Diebstahl aus der Unibibliothek	**Direkte Abfrage**: „Hast du an der Uni schon jemals in der Bibliothek ein Buch entwendet?“	**Randomized-Response-Technik**: „Hast du an der Uni schon jemals in der Bibliothek ein Buch entwendet?“ kombiniert im Forced-Response-Ansatz mit einem Kartenspiel	
	0 % ($n = 211$)	6.7 % ($n = 367$)	$p < .05$ signifkant

1. **Direkte/exakte Replikationsstudien** („direct/exact replication studies“), die eine Originalstudie bis ins Detail möglichst genau nachstellen, dienen dazu, bisherige Befunde zu überprüfen und Zufallsergebnisse oder gar Täuschungen in der früheren Forschung aufzudecken. Dies begründet den oft unterschätzten großen Wert von Replikationsstudien.
2. **Systematische Replikationsstudien** („systematic replication studies“) variieren die Originalstudie, d. h. sie weichen in der einen oder anderen Designentscheidung bewusst systematisch von ihr ab: So kann ein Laborexperiment als Feldexperiment repliziert werden (Variation des Untersuchungsortes). Die interessierenden Variablen können im Vergleich zur ursprünglichen Studie mit genaueren Datenerhebungsmethoden gemessen, die Untersuchungsgruppen vergrößert oder zusätzliche Untersuchungsgruppen in das Design einbezogen werden etc. Systematische Replikationen tragen genau wie exakte Replikationen dazu bei, bisherige Befunde abzusichern. Zudem können sie zu neuen Erkenntnissen verhelfen (▶ Studienbeispiel „Der Mozart-Effekt und seine Replikation“).

Entscheidungskriterien

- Wählen Sie eine **Theoriestudie**, wenn zu einem Thema bereits eine kritische Masse an Studien vorliegt, wenn Sie große Mengen an Literatur gut bearbeiten und strukturieren können und wenn Sie ganz bewusst in ihrer Forschungsarbeit nicht von Untersuchungsteilnehmenden abhängig sein wollen.
- Wählen Sie eine **Methodenstudie**, insbesondere eine empirische Methodenvergleichsstudie, wenn Sie ein methodisches Problem lösen wollen und Ihr Schwerpunkt weniger in der Theoriearbeit, sondern stärker im Bereich der methodischen Umsetzung liegt.
- In den Human- und Sozialwissenschaften fällt bei Qualifikationsarbeiten die Wahl meist auf eine inhaltliche **empirische Arbeit** – und zwar auf eine Originalstudie mit qualitativem, quantitativem oder Mixed-Methods-Design. Bei einer Originalstudie können Sie selbst kreativ sein, dies führt jedoch bei begrenzter eigener Forschungserfahrung oft zu langwierigen und wenig ertragreichen Forschungsprozessen. Ziehen Sie auch eine **Replikationsstudie** in Erwägung, die sowohl wissenschaftlich von großem Wert ist als auch für Neulinge gut realisierbar ist, weil man sich auf ein erprobtes Studiendesign stützen kann.

Studienbeispiel

Der Mozart-Effekt und seine Replikation

Am 13. Januar 1998 beschloss Zell Miller, der damalige Gouverneur von Georgia, dass jährlich 105 000 US-Dollar des Staatshaushaltes investiert werden sollen, um jedes im US-Bundesstaat neu geborene Kind mit klassischer Musik auf einer CD auszustatten. In Florida wurde gesetzlich festgelegt, dass **in Kindergärten jeden Tag eine Stunde klassische Musik** abgespielt werden sollte. Hintergrund dieser Maßnahmen war die Hoffnung, durch das Anhören klassischer Musik könne die Intelligenz von Kindern gesteigert werden.

Diese Hoffnung stützte sich auf eine in der renommierten Zeitschrift **Nature** publizierte wissenschaftliche Studie, die gezeigt hatte, dass $n = 36$ studentische Versuchspersonen nach dem 10-minütigen Anhören von **Mozarts Sonate für Klavier in D-Dur** (Köchelverzeichnis 448) Intelligenztestaufgaben zum räumlichen Denken (aus dem Stanford-Binet-Intelligenztest) deutlich besser lösen konnten (durchschnittlicher räumlicher IQ-Wert von 119) als Versuchspersonen, die vorher 10 min lang Entspannungsmusik gehört (IQ-Wert von 111) oder der Stille ausgesetzt waren (IQ-Wert von 110; Rauscher, Shaw, & KY, 1993). Es handelte sich also um eine experimentelle Studie mit der dreifach gestuften unabhängigen Variable „Art der Musik" (Mozart, Entspannung, keine) und der intervallskalierten abhängigen Variable „räumlicher Intelligenzquotient".

Vergleicht man die Studie mit der **Medienberichterstattung und den politischen Entscheidungen**, die auf der These beruhen, dass klassische Musik die Intelligenz steigere, so fällt auf, dass es hier zu einer **gravierenden Überinterpretation** gekommen war: So ist unklar, ob der gefundene Mozart-Effekt für die Klaviersonate in D-Dur auch für andere Arten klassischer Musik gilt. Unzulässig ist zudem die Verallgemeinerung von räumlicher Intelligenzleistung auf Intelligenzleistung allgemein. Zudem ist zu beachten, dass die Originalstudie betont, die Leistungssteigerung beim Lösen räumlicher Intelligenztestaufgaben halte nur ca. eine Viertelstunde lang an. Da sich die Überinterpretation des Mozart-Effektes jedoch international stark verbreitete, gab am Ende sogar das deutsche Bundesministerium für Bildung und Forschung einen Sammelband heraus (BMBF, 2006), der 300 Studien zum Mozart-Effekt auswertete und zu dem Schluss kommt, die empirische Forschung zu Musikwirkungen belege keine allgemeine Intelligenzsteigerung durch Mozart oder andere klassische Musikstücke. Der von Don Campbell inzwischen patentierte „Mozart-Effekt" (http://www.mozarteffect.com) wird trotzdem sehr erfolgreich vermarktet (z. B. Verkauf angeblich intelligenzförderlicher Musik-CDs für Kinder).

Neben der Frage der öffentlichen Wahrnehmung und Überinterpretation der Originalstudie stellt sich auch die Frage nach ihrer empirischen **Replizierbarkeit.**

Eine **exakte Replikationsstudie** würde die Originalbedingungen nachstellen und als Laborstudie mit studentischen Probanden realisiert werden. Da der Originalartikel extrem kurz gehalten ist (eine Seite), sind ihm keine Detailinformationen zum Versuchsablauf zu entnehmen, etwa zur Gestaltung des Versuchslabors oder zur Reihenfolge der Testdurchgänge (alle 36 Versuchspersonen durchliefen in verschiedenen Reihenfolgen alle drei Bedingungen). Auch ist unbekannt, welche Informationen bzw. Instruktionen die Versuchspersonen vorab bekamen. Wurde ihnen direkt gesagt, dass es um den Test von Musikwirkungen geht oder erhielten sie eine Coverstory? Derartige Detailfragen stellt man sich meist nur, wenn man versucht, eine Studie zu replizieren. Fehlen entsprechende Hinweise in der Publikation, so kann man die Autoren um Auskunft bitten oder muss im Zweifelsfall selbst entsprechende Entscheidungen treffen, die dem Original möglichst nahe kommen sollen.

Eine **systematische Replikationsstudie** könnte – im Unterschied zur exakten Replikation – nun darauf hinauslaufen, bestimmte Randbedingungen des Originalexperiments systematisch zu variieren (Abb. 7.1):

- Unabhängige Variable: Es werden andere Kontroll- und/oder Experimentalbedingungen ergänzt (z. B. Rockmusik; andere Musikstücke von Mozart).
- Abhängige Variable: Anstelle des Stanford-Binet-Tests könnten Testaufgaben zum räumlichen Denken aus anderen Intelligenztests herangezogen werden. Oder es wird anstelle der räumlichen Intelligenz der Effekt auf die verbale Intelligenz geprüft.
- Versuchspersonen: Anstelle von Studierenden werden Kinder untersucht.
- Versuchsumgebung: Anstelle einer Laborstudie wird eine Feldstudie im Kindergarten durchgeführt.
- Versuchsablauf: Anstelle der unmittelbaren Intelligenztestung wird diese mit zeitlicher Verzögerung durchgeführt (z. B. eine halbe Stunde oder zwei Stunden nach dem Musikhören), um die Nachhaltigkeit des Effekts zu prüfen.

Zwei Beispiele für systematische Replikationsstudien zum Mozart-Effekt:

- **Wilson & Brown (1997)**: $n = 22$ Studierende durchliefen in unterschiedlicher Reihenfolge die drei experimentellen Bedingungen (jeweils 10 min lang Musik von Mozart, Entspannungsmusik und Stille) und lösten anschließend Aufgaben zum räumlichen Denken. Variiert wurde hierbei eine Ausprägung der unabhängigen Variablen (anstelle des Klavierkonzerts in D-Dur KV 448 wurde das Klavierkonzert Nr. 23 KV 488 verwendet) sowie die Messung der abhängigen Variablen (anstelle von Intelligenztestaufgaben wurden schriftliche Labyrinth-Aufgaben verwendet). Es zeigte sich **kein Mozart-Effekt**: Zwar löste die Mozart-Gruppe die Aufgaben etwas schneller, machte jedoch auch mehr Fehler.
- **Thompson, Husain, & Schellenberg (2001)**: $n = 24$ Studierende hörten sich im Rahmen eines Laborexperiments jeweils 10 min lang Mozarts Sonate in D-Dur (KV 448) und Albionis Adagio in g-Moll an und bearbeiteten anschließend jeweils die Intelligenztestaufgaben zum räumlichen Denken, die auch in der Originalstudie verwendet worden waren. Diese Replikationsstudie variierte also die unabhängige Variable, nicht jedoch die abhängige Variable. Zudem wurde als Mediatorvariable die Stimmung der Probanden erfasst. Es zeigte sich, dass ein **signifikanter Mozart-Effekt** replizierbar war: Nach dem Hören der Mozart-Sonate erreichten die Versuchspersonen signifikant bessere Werte im räumlichen Denken als nach dem Adagio. Dieser Effekt war jedoch vollständig durch die **Mediatorvariable Stimmung** erklärbar: So hatten die Probanden nach der heiteren Mozart-Sonate bessere Stimmung und fühlten sich energiegeladener als nach dem langsamen und traurigen Adagio. Wurde der Einfluss der Stimmung in die Datenanalyse einbezogen, so unterschieden sich die beiden Gruppen nicht mehr in den erreichten Leistungspunkten im Intelligenztest.



Abb. 7.1 Systematische Replikationsstudien zum Mozart-Effekt können z. B. die unabhängige Variable verändern, indem anstelle der Klaviersonate in D-Dur ein anderes Mozartstück gewählt wird. (© Marty Bucella/Catalog number: 33232112, Rights Available from Jantoo.com)

7.4 Primär-, Sekundär- oder Metaanalyse

Gemäß ihrem Gegenstand wurden bereits Theorie- und Methodenstudien von empirischen Studien abgegrenzt (► Abschn. 7.3). Kennzeichen empirischer Studien ist, dass sie auf einer eigenen Datenerhebung und/oder Datenanalyse basieren. Bei der Primärstudie ist die Datengrundlage ein selbst erhobener Datensatz. Bei der Sekundärstudie werden bereits vorhandene Datensätze erneut analysiert. Bei der Metaanalyse greift man aus einer mehr oder minder großen Menge vergleichbarer Studien zum selben Sachverhalt die berichteten statistischen Ergebnisse heraus und rechnet mit diesen weiter. Da die Metaanalyse empirische Befunde verrechnet, wird sie hier als Sondertyp einer empirischen Studie eingeordnet. Da sie sich aber ausschließlich auf Literaturquellen stützt, kann sie auch als Theorie- bzw. Literaturstudie (► Abschn. 7.3) aufgefasst werden – je nach Betrachtungsperspektive.

Primäranalyse – Bei der Primärstudie bzw. Primäranalyse („primary analysis") als dem typischen Fall einer empirischen Studie werden die empirischen Daten selbst erhoben und anschließend analysiert.

Sekundäranalyse – Bei der Sekundärstudie bzw. Sekundäranalyse („secondary analysis") werden vorhandene Datensätze neu ausgewertet. Voraussetzung dafür ist die Beschaffung der Originaldatensätze.

Metaanalyse – Bei der Metastudie bzw. Metaanalyse („meta analysis") werden die Ergebnisse direkt vergleichbarer Studien zum selben Thema zu einem Gesamtergebnis statistisch zusammengefasst. Da die Metaanalyse mit empirischen Ergebnissen weiterrechnet und somit eine statistische Datenanalyse enthält, kann sie als Sonderform einer empirischen Studie eingeordnet werden. Allerdings arbeitet sie nicht mit Orgininaldatensätzen, sondern stützt sich ausschließlich auf Ergebnisberichte von Studien und wird deswegen auch als Theorie- oder Literaturarbeit eingestuft (► Kap. 16).

Die meisten qualitativen wie quantitativen human- und sozialwissenschaftlichen Studien sind als **Primäranalysen** angelegt, d. h., sie laufen auf eine erstmalige Auswertung selbst erhobener Daten hinaus (Glass, 1976). Dies hat den Vorteil, dass die Details des Forschungsdesigns, die Art der Stichprobe und auch die Datenerhebungsmethoden selbst festgelegt und somit die Beschaffenheit des Datensatzes genau auf das selbst gewählte Forschungsproblem zugeschnitten werden kann. Gleichzeitig haben **Primärstudien** den Nachteil, dass aus forschungsökonomischen Gründen oft nur relativ kleine Datensätze erzeugt werden können.

Bei **Sekundäranalysen** greift man auf einen oder mehrere vorhandene Datensätze zurück, um diese mit verbesserten Auswertungsmethoden oder unter neuen Fragestellungen noch einmal zu analysieren. **Sekundärstudien** werden sowohl mit quantitativen als auch mit qualitativen Datensätzen durchgeführt, die über entsprechende Datenbanken oder direkt von den Forschenden zu beziehen sind. Der Vorteil von Sekundäranalysen besteht darin, dass der Aufwand für die eigene Datenerhebung entfällt, man auf sehr umfassende Datensätze zurückgreifen und sich stark auf die Datenauswertung konzentrieren kann. Nachteilig ist dagegen, dass die vorgefundenen Datensätze oft Variablen oder Aspekte nicht enthalten, die für das eigene Forschungsproblem relevant wären.

Die **Metaanalyse** schließlich fasst die Ergebnisse mehrerer vergleichbarer Studien, die denselben Effekt untersuchen, zusammen. **Metastudien** sind nur durchführbar, wenn zu dem interessierenden Phänomen bereits intensive Forschung betrieben wurde und eine Mindestzahl direkt vergleichbarer Studien vorliegt. Die Metaanalyse hat sich primär als **quantitatives Verfahren** etabliert, das die statistischen Effekte mehrerer hochwertiger und inhaltlich vergleichbarer Studien zu einem Gesamteffekt aggregiert. Metaanalysen sind z. B. von Interesse, wenn es um die Wirksamkeit und Wirkungen von bestimmten Medien, Medikamenten, Psychotherapien oder Lehr- und Lernmethoden geht (zu Details der quantitativen Metaanalyse ► Kap. 16). Vereinzelt werden auch methodische Ansätze für **qualitative Metaanalysen** bzw. Metastudien vorgeschlagen, die auf einer systematischen qualitativen Auswertung einschlägiger qualitativer Studienergebnisse

basieren (Stall-Meadows & Hyle, 2010; Timulak, 2009) und auch die Grenzen der Synthetisierung von Metabefunden diskutieren (Nicholas, Globerman, Antle, McNeill, & Lach, 2006; Thorne et al., 2002).

Entscheidungskriterien

- Wählen Sie eine **Primäranalyse**, wenn Sie den empirischen Forschungsprozess inklusive Planung und Durchführung einer eigenen Datenerhebung vollständig durchlaufen können und wollen.
- Entscheiden Sie sich für eine **Sekundäranalyse**, wenn Ihr Schwerpunkt auf der qualitativen und/oder quantitativen Datenanalyse liegen soll und sie mit mehreren oder größeren Datensätzen operieren können und wollen. Voraussetzung für Sekundärstudien ist die Verfügbarkeit entsprechender Datensätze.
- Avisieren Sie eine **Metaanalyse**, wenn zum interessierenden Forschungsproblem bereits eine Reihe hochwertiger und vergleichbarer quantitativ-empirischer Studien durchgeführt wurden, auf deren Studienberichte Sie sich stützen können. Eine quantitative Metaanalyse ist methodisch aufwändig und komplex, d. h. als Einstiegsstudie für Neulinge eher ungeeignet.

7.5 Explorative, deskriptive oder explanative Studie

Empirische Studien lassen sich nach dem zugrunde liegenden **Erkenntnisinteresse** in drei große Gruppen einteilen: explorative, deskriptive und explanative Studien.

Explorative Studie – Die explorative Studie („exploratory study") dient der genauen Erkundung und Beschreibung eines Sachverhaltes mit dem Ziel, wissenschaftliche Forschungsfragen, Hypothesen und Theorien zu entwickeln.

Explanative Studie – Die explanative Studie („explanatory study dient der Überprüfung vorher aufgestellter Hypothesen und somit auch der Theorien, aus denen sie abgeleitet wurden (▶ Abschn. 2.2.8).

Deskriptive Studie – Die deskriptive Studie („descriptive study of populations") dient der Feststellung der Verbreitung von Merkmalen und Effekten in großen Grundgesamtheiten, etwa der Bevölkerung eines Landes.

Explorative Studien erkunden einen Gegenstand, indem sie Daten über ihn sammeln. Auf der Basis von offenen Forschungsfragen werden verschiedene Aspekte eines Sachverhaltes beleuchtet und anschließend differenziert beschrieben. Explorative (gegenstandserkundende, theoriebildende) Studien beantworten offene Forschungsfragen und helfen dabei, neue Hypothesen und Theorien zu generieren. Explorative Studien werden oft als nicht oder wenig strukturierte **qualitative Studien** durchgeführt, etwa als ethnografische Feldstudien oder als Interviewstudien, da dieses Vorgehen offen ist für unerwartete Befunde. Beispiel: Da sich die Computerspielforschung bislang auf Jugendliche und junge Erwachsene konzentriert hat, wurden Spielverhalten und Spielerleben von älteren Spielerinnen und Spielern im Rahmen einer explorativen Studie auf der Basis von Leitfadeninterviews untersucht (Quandt, Grueninger, & Wimmer, 2009; ▶ Abschn. 5.2).

Explanative Studien überprüfen vorher aufgestellte (aus Theorien abgeleitete) Hypothesen auf ihre Gültigkeit. Insbesondere Ursache-Wirkungs-Relationen sollen bestätigt bzw. verworfen werden. Bei explanativen (hypothesen-/theorieprüfenden) Studien handelt es sich meist um vollstrukturierte **quantitative Studien**, die genau die postulierten Effekte testen. Die höchste Erklärungskraft im Hinblick auf den Nachweis von Kausalität haben **experimentelle Studien** (▶ Abschn. 7.6.1). Beispiel: Um die beiden konkurrierenden Medienwirkungsthesen zu testen, dass gewalthaltige Inhalte von Computerspielen a) zu schlechteren oder b) zu besseren Spielleistungen führen, wurde ein Laborexperiment durchgeführt (Bösche, 2009; ▶ Abschn. 5.2). Auch bei der Originalstudie und den Replikationsstudien zum Mozart-Effekt handelt es sich um explanative Studien (▶ Abschn. 7.3). Bei explanativen Studien steht der Nachweis von Effekten im Vordergrund, wobei sie in Form von experimentellen und quasi-experimentellen Studien auch zur Klärung von Ursache-Wirkungs-Mechanismen beitragen. Entsprechende Studien arbeiten – entgegen landläufiger Meinung – meist gerade nicht mit bevölkerungsrepräsentativen Stichproben, sondern mit hinsichtlich soziodemografischer Variablen eher ‚homogenen Untersuchungsgruppen (z. B. Studierende), um personengebundene Störvariablen auszuschalten und den Ursache-Wirkungs-Mechanismus mit hoher interner Validität zu belegen. Dafür sind Einbußen in der externen Validität (Generalisierbarkeit der Befunde auf andere Personengruppen) hinzunehmen, was bei der Ergebnisinterpretation zu beachten ist.

Deskriptive Studien beschreiben ganze Populationen möglichst präzise im Hinblick auf die Verbreitung ausgewählter Merkmale, seien es Aspekte des Gesundheits-, Konsum-, Politik- oder Sozialverhaltens. Derartige populationsbeschreibende Untersuchungen werden als quantitative Studien realisiert. Sie sind auf repräsentative Stichproben angewiesen (▶ Abschn. 9.1.3) sowie auf zuverlässige Messinstrumente für die interessierenden Va-

riablen. Deskriptive Studien spielen unter anderem in der Meinungsforschung **(Demoskopie)** und in der Forschung zur Verbreitung von Gesundheit und Krankheiten in der Bevölkerung **(Epidemiologie)** eine wichtige Rolle. Sie basieren auf strukturierten mündlichen oder schriftlichen Befragungen. Beispiel: Eine deskriptive Studie aus Norwegen zeigte – basierend auf der Befragung einer bevölkerungsrepräsentativen Zufallsstichprobe ($n = 2\,500$ Personen im Alter zwischen 15 und 40 Jahren) – dass die Mehrheit von 56 % der norwegischen Bevölkerung in der betrachteten Altersspanne Computerspiele nutzt und bei 0.6 % eine suchtartige Extremnutzung auftritt (Mentzoni et al., 2011).

Entscheidungskriterien

- Wählen Sie eine **explorative Studie**, wenn Sie einen bislang wenig untersuchten Gegenstand erkunden oder theoriebildend tätig sein wollen. Typisch ist bei diesem Studientyp eine qualitative Forschungsstrategie. Es existieren aber auch quantitative explorative Studien mit speziellen Techniken der explorativen statistischen Datenanalyse (▶ Abschn. 12.3).
- Planen Sie eine **explanative Studie**, wenn Sie hypothesen- und theorieprüfend vorgehen möchten. Typisch für diese Studienart ist eine quantitative Forschungsstrategie. Besonders empfehlenswert für den Einstieg sind explanative Studien in Form von Replikationsstudien oder in Form von Methodenvergleichsstudien, die z. B. die erwartete Überlegenheit einer neuen Methodenvariante gegenüber dem herkömmlichen Vorgehen prüfen. Für explanative quantitative Studien sind Vorkenntnisse in statistischen Verfahren der Hypothesenprüfung notwendig (▶ Abschn. 12.5).
- **Deskriptive Studien** können im Rahmen von Qualifikationsarbeiten meist nicht als Primärstudien durchgeführt werden, weil Zufallsstichproben für relevante Populationen zu aufwändig wären. Sie können allerdings mit deutlich geringerem Aufwand als Sekundäranalysen realisiert werden, sofern einschlägige Originaldatensätze verfügbar sind. Voraussetzung für populationsbeschreibende Studien sind nicht zuletzt solide Kenntnisse der statistischen Parameterschätzung (▶ Abschn. 12.4).

7.6 Experimentelle, quasi-experimentelle oder nicht-experimentelle Studie

Wir haben die **explanative Studie** bereits als hypothesenprüfenden quantitativen Studientyp kennengelernt (▶ Abschn. 7.5). Besonderen Stellenwert bei hypothesenprüfenden Studien nehmen **Kausalhypothesen** ein, die im Kontext des jeweiligen Forschungsthemas angeben, welche Ursachen zu welchen Wirkungen führen (Ursache-Wirkungs-Relationen werden als Kausalität bezeichnet). Damit eine explanative Studie einen eindeutigen **Kausalschluss** zulässt, müssen mindestens zwei Untersuchungsgruppen gebildet und unterschiedlich behandelt sowie schließlich hinsichtlich der interessierenden Wirkungen verglichen werden (Experimentalgruppe vs. Kontrollgruppe). Bei einem **Experiment** wird dabei mit exakt vergleichbaren Gruppen gearbeitet (durch zufällige Zuordnung der Untersuchungsobjekte zu den Gruppen: **Randomisierung**). Kann keine Randomisierung umgesetzt werden, so handelt es sich um ein Quasi-Experiment, sofern Untersuchungs- und Kontrollgruppen im Sinne des untersuchten Ursachenfaktors gezielt unterschiedlich behandelt werden. Entfällt auch die experimentelle Variation und werden lediglich vorgefundene Gruppen verglichen, die sich im Ursachenfaktor unterscheiden, so handelt es sich um eine **nicht-experimentelle Studie**, die nur äußerst bedingt zur Prüfung von Kausalhypothesen geeignet ist. Bei nicht-experimentellen, quasi-experimentellen oder echten experimentellen Untersuchungsdesigns kommt es also darauf an, wie man die für den Kausalschluss zu vergleichenden **Untersuchungsgruppen bildet und behandelt**.

Experimentelle Studie/randomisierte kontrollierte Studie – Die experimentelle Studie/randomisierte kontrollierte Studie („experimental study") bildet für die Prüfung einer Kausalhypothese zunächst künstlich nach dem Zufallsprinzip mindestens zwei Gruppen (Randomisierung; „randomization", „random assignment to groups"), behandelt diese systematisch unterschiedlich (experimentelle Manipulation der unabhängigen Variable/n; „experimental manipulation", „manipulated independent variable/s"), und misst die in den Experimental- und Kontrollgruppen resultierenden Effekte auf die abhängige/n Variable/n. Man spricht auch von einem echten Experiment (im Unterschied zum Quasi-Experiment) oder von einer randomisierten Kontrollgruppenstudie bzw. einer randomisierten kontrollierten Studie („randomized controlled trial", RCT).

Quasi-experimentelle Studie/nicht-randomisierte kontrollierte Studie – Die quasi-experimentelle Studie/nicht-randomisierte kontrollierte Studie („quasi-experimental study", „non-randomized study") greift zur Prüfung einer Kausalhypothese auf Gruppen zurück, die nicht zufällig zusammengestellt, sondern oft einfach vorgefunden oder anderweitig gebildet wurden (keine Randomisierung), behandelt diese jedoch ebenso wie im echten Experiment systematisch unterschiedlich (experimentelle Variation der unabhängigen Variable/n) und misst die in den Experimental- und Kontrollgruppen resultierenden Effekte auf die ab-

hängige/n Variable/n. Man spricht auch von einer nicht-randomisierten kontrollierten Studie („non-randomized controlled trial", NRCT).

Nicht-experimentelle Studie – Eine nicht-experimentelle Studie („non-experimental study", „descriptive study") greift auf vorgefundene Gruppen zurück (keine Randomisierung) und betrachtet deren vorgefundene Unterschiede (keine experimentelle Variation der unabhängigen Variable/n, sondern vorgefundene Variation). Sie ist zur Prüfung von Kausalhypothesen nur sehr bedingt geeignet, aus forschungsökonomischen sowie forschungsethischen Gründen aber bei manchen Forschungsthemen die einzige Option.

Das Experiment gilt im Hinblick auf die Prüfung von Kausalhypothesen als „**Königsweg der Erkenntnis**". Denn in einer **experimentellen Studie** werden die laut Theorie bzw. Hypothese postulierten Ursache-Wirkungs-Relationen unter Ausschaltung von personenbezogenen und untersuchungsbedingten Störeinflüssen **aktiv hergestellt**, d. h. es wird mindestens eine Experimentalgruppe mit einer Kontrollgruppe verglichen. Beim **Quasi-Experiment** werden bei der experimentellen Manipulation die Störeinflüsse nur teilweise kontrolliert. Eine **nicht-experimentelle Studie** weist keine aktive experimentelle Manipulation der unabhängigen Variable/n im Sinne einer unterschiedlichen Behandlung der Versuchsgruppen auf, stattdessen werden vorgefundene Gruppenunterschiede verglichen (z. B. Computerspieler vs. Nichtspieler). Man spricht auch von „passiven" oder „korrelativen" Studien. Die kausale Aussagekraft nicht-experimenteller Untersuchungen ist somit am geringsten. Aufgrund praktischer und ethischer Rahmenbedingungen muss jedoch oft auf ein experimentelles oder quasi-experimentelles Untersuchungsdesign verzichtet und auf eine nicht-experimentelle Studie zurückgegriffen werden. Dabei ist es dann wichtig, die Grenzen ihrer kausalen Aussagekraft zu kennen und transparent zu machen. Gleichzeitig sind auch die Stärken nicht-experimenteller Studien zu beachten, die meist ein breiteres Spektrum an Variablen und Untersuchungseinheiten berücksichtigen, weil sie sich nicht auf den Nachweis weniger Kausaleffekte beschränken. Die Besonderheiten experimenteller, quasi-experimenteller und nicht-experimenteller Designs werden im Folgenden genauer betrachtet.

7.6.1 Die experimentelle Studie

Ziel eines Experiments („experimental study") ist die theoriegestützte Überprüfung des kausalen Einflusses einer oder mehrerer **unabhängiger Variablen** (UV„independent variable", IV) auf die Ausprägungen einer oder mehrerer **abhängiger Variablen** (AV, „dependent variable", DV). Die betrachteten Ursachenfaktoren sind die unabhängigen Variablen, die Wirkungen oder Effekte werden anhand der abhängigen Variablen gemessen. Die einfachste Form der experimentellen Versuchsanordnung arbeitet mit einer zweifach gestuften unabhängigen Variablen, so dass ein **Zwei-Gruppen-Plan** resultiert (z. B. Experimentalgruppe vs. Kontrollgruppe). Die Wirkungen des Ursachenfaktors werden auf mindestens einer kardinalskalierten abhängigen Variablen gemessen (zu Skalenniveaus ▶ Abschn. 8.4). Soll z. B. die Hypothese geprüft werden, dass Kinder Englisch-Vokabeln mit einem Handy-Lernprogramm besser lernen als mit dem herkömmlichen Lehrbuch (unabhängige Variable: Lernform mit den Ausprägungen Handy-Lernprogramm vs. Lehrbuch), so müsste als abhängige Variable der Lernerfolg gemessen werden (z. B. Anzahl der nach einer Woche gelernten Vokabeln).

Im Experiment wird die unabhängige Variable als experimentelle Variable grundsätzlich von den Forschenden variiert bzw. manipuliert. Mit **experimenteller „Manipulation"** ist keine Verfälschung der Ergebnisse gemeint, sondern ein aktives Herstellen von Verursachungsbedingungen. Würde man unter Schülern herumfragen, wer ein Handy-Lernprogramm nutzt und die vorgefundenen Nutzer mit Nichtnutzern vergleichen, so wäre dies eine nicht-experimentelle Studie, denn es fand keine experimentelle Manipulation statt. Zum Gegenstand eines Experiments wird das Vokabellernen mit dem Handy erst dadurch, dass zunächst eine soziodemografisch homogene Gruppe von Versuchspersonen rekrutiert wird (z. B. Schülerinnen und Schüler derselben Klassenstufe, die bislang keine Erfahrung mit Handylernen haben) und dass diese Versuchspersonen dann per Zufallsprinzip entweder der Kontroll- oder der Experimentalgruppe zugeordnet und dementsprechend mit Buch oder Handy-Lernprogramm ausgestattet werden. Der Lernerfolg wird z. B. anhand der Anzahl der nach einer Woche gelernten Englisch-Vokabeln gemessen, wobei das Stichprobenergebnis in dieser fiktiven Studie auf einen besseren durchschnittlichen Lernerfolg der Experimentalgruppe (rund 22 gelernte Vokabeln) als der Kontrollgruppe (durchschnittlich rund 13 gelernte Vokabeln) hinweist (◘ Tab. 7.3).

Das Design wird komplexer, wenn eine UV mit mehr als zwei Stufen verwendet wird. So kann die unabhängige Variable „Lernform" vierfach gestuft untersucht werden (z. B. Lernen mit einem Buch, mit einem Hörbuch, mit einer Website oder mit einem Handy-Lernprogramm). Auch kann mehr als eine unabhängige Variable einbezogen werden (z. B. UV1: Lernform mit 4 Ausprägungen, UV2: Lerninhalt mit 2 Ausprägungen: Vokabeln, Formeln). Man spricht bei mehreren UV von einem **faktoriellen Design**. Im hier skizzierten zweifaktoriellen Beispiel (Lernform und Lerninhalt) würde ein 4×2-Design resultieren, wenn alle vier Ausprägungen von UV1 mit allen zwei Ausprägungen von UV2 kombiniert werden (◘ Tab. 7.4). Die $4 \times 2 = 8$ Zellen bzw. Versuchsbedingungen im Design (z. B. mit dem Buch Vokabeln lernen, mit dem Buch Formeln lernen etc.) sind jeweils mit einer

Tabelle 7.3 Ein einfacher experimenteller Zwei-Gruppen-Plan mit einer zweifach gestuften unabhängigen Variablen (Lernform) und einer kardinalskalierten abhängigen Variablen (Lernerfolg gemessen über die Anzahl der nach einer Woche gelernten Vokabeln, fiktive Daten)

Experimentalgruppe Vokabellernen mit einem Handy-Lernprogramm		Kontrollgruppe Vokabellernen mit dem Buch	
Vp1	11	Vp1	12
Vp2	26	Vp2	7
Vp3	22	Vp3	6
Vp4	27	Vp4	25
Vp5	26	Vp5	9
Vp6	13	Vp6	14
Vp7	17	Vp7	16
Vp8	32	Vp8	19
Vp9	28	Vp9	11
Vp10	22	Vp10	7
$M_{Experimentalgruppe} = 22.40$ (SD = 6.82)		$M_{Kontrollgruppe} = 12.60$ (SD = 6.06)	

Vp Versuchsperson

Tabelle 7.4 Ein zweifaktorieller Versuchsplan: 4 × 2-Plan mit 8 Zellen bzw. Untersuchungsbedingungen

		UV1: Lernform			
		Buch	Hörbuch	Website	Handy
UV2: Lerninhalt	Vokabeln				
	Formeln				

UV unabhängige Variable

Mindestzahl von ca. 10–30 Versuchspersonen zu besetzen. Man müsste also in einem **vollständig besetzten Versuchsplan** ca. 80–240 Versuchspersonen rekrutieren. Je mehr Versuchspersonen pro Zelle untersucht werden, umso größer ist bei der statistischen Auswertung die **Teststärke**, d. h., umso eher kann ein vorhandener Ursache-Wirkungs-Effekt tatsächlich als statistisch überzufälliger Effekt im Experiment nachgewiesen werden. Die Festlegung der Versuchspersonenzahl pro Zelle sollte im Vorfeld schon bei der Planung des Untersuchungsdesigns anhand des sog. **optimalen Stichprobenumfangs** erfolgen (▶ Abschn. 14.3).

Je mehr Untersuchungsbedingungen berücksichtigt werden, umso mehr Versuchspersonen werden benötigt. Typischerweise werden in der Experimentalforschung ein bis fünf unabhängige Variablen berücksichtigt. Zudem sind auch **unvollständige Versuchspläne** möglich, bei denen nicht alle Zellen im Design mit Untersuchungspersonen besetzt werden (▶ Abschn. 12.5.3). Der Vorteil eines faktoriellen Designs besteht darin, dass bei einem solchen Plan nicht nur die separate Wirkung der einzelnen UV bzw. Ursachenfaktoren auf die AV geprüft werden kann, sondern auch das spezifische Zusammenwirken verschiedener Ursachenfaktoren (sog. **Interaktionseffekte**, bei denen Faktoren nicht einfach additiv, sondern in spezifischer überadditiver Weise in ihren Effekten auf die abhängige/n Variable/n zusammenwirken). So könnte es sein, dass Formeln generell schlechter gelernt werden als Vokabeln. Würde jedoch in der Versuchsbedingung Formellernen mit dem Handy plötzlich ein besseres Lernergebnis entstehen als bei den anderen Lernformen, würde dies auf einen Interaktionseffekt hindeuten.

Auch hinsichtlich der abhängigen Variablen kann ein experimenteller Versuchsplan flexibel gestaltet werden. Statt einer einzigen abhängigen Variablen **(univariater Versuchsplan)** können mehrere abhängige Variablen **(multivariater Versuchsplan)** gemessen werden. Mit steigender Komplexität des Designs wächst die benötigte Datenmenge. Entsprechend sind komplexere statistische Auswertungsverfahren indiziert. Typischerweise werden in experimentellen Studien **Unterschiedshypothesen** geprüft und statistisch mit *t*-Tests (zwei Gruppen) oder **Varianzanalysen** (mehr als zwei Gruppen) ausgewertet (▶ Abschn. 12.5.3). Wie nach einem „**Baukastenprinzip**" können experimentelle Designs unterschiedlichster Art durch Hinzunahme weiterer unabhängiger und abhängiger Variablen gestaltet werden.

Interne Validität – („internal validity") Eine Untersuchung ist intern valide, wenn ihre Ergebnisse kausal eindeutig interpretierbar sind in dem Sinne, dass Effekte in den abhängigen Variablen zweifelsfrei auf die Wirkung der unabhängigen Variablen zurückzuführen sind. Die interne Validität steigt durch die sorgfältige Kontrolle von personenbezogenen Störvariablen (v. a. durch Randomisierung) sowie von untersuchungsbedingten Störvariablen (z. B. Registrieren, Ausschalten oder Konstanthalten von Störeinflüssen durch standardisierten Ablauf des Experimentes; Verblindung von Versuchspersonen und Versuchsleitenden etc.).

Externe Validität – („external validity") Eine Untersuchung ist extern valide, wenn ihre Ergebnisse über die Bedingungen der Untersuchungssituation und über die untersuchten Personen hinausgehend generalisierbar sind. Die externe Validität sinkt meist mit wachsender Unnatürlichkeit der Versuchsbedingungen im Hinblick auf die alltägliche Lebenswelt und mit abnehmender Repräsentativität der Versuchspersonen im Hinblick auf die Zielpopulation.

Ein gutes Experiment ist so beschaffen, dass es einen eindeutigen Rückschluss von der unabhängigen Variablen als Ursachenfaktor auf die gemessenen Effekte (Ausprägungen der abhängigen Variablen als Wirkungen) erlaubt

(**interne Validität**). **Alternativerklärungen** sollen möglichst vollständig ausgeschlossen werden. Man spricht auch von der **Ceteris-Paribus-Klausel** (lat. unter sonst gleichen Bedingungen), um zum Ausdruck zu bringen, dass eine intern valide Studie die Wirkungsweise der betrachteten unabhängigen Variablen isoliert, indem **alle anderen Einflussfaktoren ausgeschaltet bzw. konstant gehalten** werden.

Im obigen Beispiel will man den besseren Lernerfolg in der Experimentalgruppe tatsächlich kausal auf die Lernmethode Handy zurückführen. Alternativerklärungen, etwa dass die Versuchspersonen in der Experimentalgruppe von vorne herein besser Englisch konnten, motivierter waren, stärker von den Eltern beim Lernen gefördert wurden usw. müssen also ausgeschlossen werden. Mit Alternativerklärungen sind Einflussfaktoren auf die abhängige Variable gemeint, die nichts mit der UV zu tun haben. Solche Einflussfaktoren werden im Experiment und im Quasi-Experiment als Bedrohung der internen Validität und somit als **Störvariablen** oder **konfundierende Variablen** („confounding variables", „confounder") betrachtet. Grundsätzlich unterscheidet man zwei Arten von Störvariablen:

1. **Personenbezogene Störvariablen**: Systematische Unterschiede in den persönlichen Merkmalen und Voraussetzungen der Versuchspersonen in den verschiedenen Untersuchungsgruppen, z. B. wenn sich in der Experimentalgruppe motiviertere oder intelligentere Versuchspersonen befinden würden als in der Kontrollgruppe.
2. **Umwelt- bzw. untersuchungsbedingte Störvariablen**: Unterschiede in der Behandlung der verschiedenen Untersuchungsgruppen, die nichts mit der unabhängigen Variablen zu tun haben, z. B. wenn die Experimentalgruppe von den Forschenden intensiver betreut würde und mehr Aufmerksamkeit und Zuspruch bekäme als die Kontrollgruppe.

Um personenbezogene Störfaktoren auszuschließen, arbeitet man in experimentellen Studien mit Versuchspersonen, die einander hinsichtlich soziodemografischer Merkmale ähnlich sind, sowie in jedem Fall auch mit der Technik der **Randomisierung** („randomization"), d. h. der zufälligen Zuordnung der Versuchspersonen zu den Untersuchungsbedingungen. Die Randomisierung darf dabei nicht mit dem Konzept der Zufallsstichprobe („random sample") verwechselt werden, das sich auf die Auswahl von Untersuchungsobjekten aus einer Population bezieht (► Abschn. 9.3.2). In der Experimentalforschung wird üblicherweise **nicht** mit Zufallsstichproben, sondern mit nicht-probabilistischen Stichproben (z. B. Gelegenheitsstichproben) gearbeitet (► Abschn. 9.3.1). Ziel der Randomisierung – also der zufälligen Zuordnung von Versuchspersonen auf Experimental- und Kontrollgruppen im Experiment – ist es, dass sich die verschiedenen Gruppen hinsichtlich der Ausprägung aller möglichen psychologischen und sozialen Merkmale (die als personengebundene Störvariablen wirken könnten) nicht systematisch unterscheiden, sondern weitgehend äquivalent sind. Denn es ist – bei genügender Gruppengröße – sehr unwahrscheinlich, dass sich z. B. nach einer Zufallsaufteilung in der einen Gruppe überwiegend die hochmotivierten und in der anderen Gruppe die weniger motivierten Schüler befinden. Im Durchschnitt sollten bei ausreichender Gruppengröße und Zufallsaufteilung der Versuchspersonen alle für die Untersuchung potenziell relevanten personenbezogenen Störvariablen in den verschiedenen Untersuchungsgruppen annähernd gleich ausgeprägt sein, d. h., mögliche Gruppenunterschiede in Bezug auf die abhängige Variable (d. h. im Beispiel in Bezug auf die Lernleistung) gehen mit hoher Wahrscheinlichkeit auf die unabhängige Variable (Lernform) zurück. Ein solches Untersuchungsergebnis wäre (relativ) eindeutig interpretierbar: Die Untersuchung verfügt über eine **hohe interne Validität bzgl. der personenbezogenen Störvariablen**. (Überlegungen zur Kalkulation der Gruppengröße, die erforderlich ist, um Äquivalenz der zu vergleichenden Gruppen herzustellen, findet man bei Mittring & Hussy, 2004).

Umgang mit Störvariablen beim Experiment

Personengebundene Störvariablen werden im Experiment kontrolliert, indem man von vorne herein auf soziodemografisch recht ähnliche Versuchspersonen zurückgreift und diese per Zufallsprinzip (Randomisierung) auf Experimental- und Kontrollgruppen verteilt.

Untersuchungsbedingte Störvariablen werden im Experiment nach Möglichkeit ausgeschaltet, konstant gehalten oder im Zuge der Datenerhebung miterfasst. Zur Kontrolle der untersuchungsbedingten Störvariablen finden Experimente oft im Labor statt (also in einer kontrollierten räumlichen Umgebung). Zudem wird möglichst mit Verblindung gearbeitet, d. h. weder Versuchspersonen noch Versuchsleitende oder Versuchsauswertende wissen, welche Gruppe jeweils die Experimental- oder Kontrollgruppe ist.

Der Umgang mit Störvariablen, der im Zuge von Entscheidungen zum Untersuchungsdesign festgelegt wird, ist bestimmend für die Aussagekraft eines Experimentes und damit für seine wissenschaftliche Qualität (zur Validität ► Abschn. 3.2.2).

Zudem müssen für ein aussagekräftiges Ergebnis beim Experiment auch die **untersuchungsbedingten Störvariablen** möglichst vollständig ausgeschaltet oder konstant gehalten werden. Ein gravierender untersuchungsbedingter Störfaktor, der die interne Validität der hier beispielhaft beschriebenen Studie gefährden würde, käme z. B. zustande, wenn man die Handy-Lerngruppe anlässlich des Experiments leihweise mit leistungsfähigen neuen Mobiltelefonen ausstatten würde, die Kontrollgruppe aber nicht. Beide Gruppen würden sich dann nicht nur hinsichtlich der Lernform unterscheiden, sondern die Experimentalgruppe würde darüber hinaus durch das Leihgerät eine Bevorzugung erleben und dadurch möglicherweise generell motivierter und engagierter reagieren. Entsprechende Effekte sind durch eine geschickte Versuchsplanung auszuschließen, etwa indem man das Handy-Lernprogramm auf den vorhandenen Handys installiert (**Ausschaltung des untersuchungsbedingten Störfaktors** Leihgerät) oder sowohl Experimental- als auch Kontrollgruppe mit einem Leihgerät ausstattet (**Konstanthalten des untersuchungsbedingten Störfaktors** Leihgerät).

Ein nicht zu vernachlässigender untersuchungsbedingter Störfaktor sind **Erwartungseffekte**, die sowohl auf Seiten der Untersuchsteilnehmenden als auch der Versuchsleitenden auftreten können. Als Versuchsperson an einem Experiment teilzunehmen veranlasst Menschen in der Regel, über Sinn und Zweck der Studie nachzudenken. Sie entwickeln in Abhängigkeit von den Untersuchungsbedingungen entsprechende Erwartungen. Wäre etwa in einem Medikamententest den Versuchspersonen bekannt, ob sie das neue Medikament, das alte Medikament oder ein Placebo erhalten, so könnte dieses Wissen allein die untersuchten Effekte beeinflussen. Man arbeitet deswegen wenn möglich mit **„blinden" Versuchspersonen**, die keinen Einblick in die Details der Versuchsgestaltung haben. Alle Versuchspersonen erhalten also unter exakt denselben Umständen eine genau gleich aussehende Tablette und wissen nicht, in welcher Versuchsbedingung sie sich befinden bzw. ob sie den Wirkstoff oder das Placebo erhalten. Allerdings müssen sie sich gemäß Richtlinien der Forschungsethik zuvor ausdrücklich mit den Versuchsbedingungen einverstanden erklärt haben, d. h. sie werden zuvor aufgeklärt, dass sie entweder ein bestimmtes Medikament oder ein Placebo erhalten und welche Wirkungen dies haben könnte („informed consent"; ▶ Abschn. 4.1.1).

Um Versuchspersonen vom eigentlichen Untersuchungsziel abzulenken, muss im Rahmen von Verblindung teilweise auch mit **Täuschung** („deception") gearbeitet werden: Angenommen, man möchte auf der Basis kognitiver Identitätstheorien experimentell die Hypothese überprüfen, dass sich der aktuelle Kleidungsstil (z. B. legere vs. formale Kleidung) auf die Selbstwahrnehmung auswirkt, man sich entsprechend der legeren Kleidung selbst als „entspannter", „lässiger" und „zwangloser" einschätzt. Diese Zielsetzung würde man den Probanden jedoch nicht vorab mitteilen, da man an spontanen, unverzerrten Reaktionen interessiert ist. Stattdessen würde man den blinden Versuchspersonen zur Ablenkung eine andere Zielsetzung der Studie mitteilen **(Coverstory)**. In einem entsprechend verblindeten sozialpsychologischen Experiment wurde den Versuchspersonen als Coverstory erzählt, es handele sich um eine Studie zu Fehlurteilen bei polizeilichen Gegenüberstellungen. Dementsprechend wurden sie aufgefordert, sich für das Experiment entsprechend der Beschreibung eines Verdächtigen zu kleiden (Hannover & Kühnen, 2002), so dass dann die Gegenüberstellung simuliert werden könne. Um die Effekte des Kleidungsstils auf die Selbstwahrnehmung zu prüfen, sollten die Probanden sich anschließend auf einer Reihe von Eigenschaften selbst einschätzen. Diese Selbsteinschätzungen wurden im Rahmen der Coverstory als kleine Vorstudie zur Überprüfung des Instruments einer Diplomarbeit angekündigt, so dass aus Sicht der Probanden zwei unabhängige Studien stattfanden. Täuschung zum Zweck der Verblindung bzw. zur Kontrolle untersuchungsbedingter Störfaktoren sollte mit Augenmaß eingesetzt werden und erfordert aus forschungsethischen Gründen in jedem Fall eine nachträgliche Aufklärung (**Debriefing**; zur Forschungsethik ▶ Abschn. 4.1).

Welche Hypothesen Versuchspersonen angesichts der Versuchsbedingungen über das Untersuchungsziel entwickeln, lässt sich im Zweifelsfall im Rahmen von methodologischen **Pretests** vorab prüfen. Hier werden Probanden gefragt, welche Zielsetzung sie aus welchen Gründen vermuten. Entsprechende Hinweise können dazu genutzt werden, die Abläufe gegebenenfalls noch umzugestalten. Auch am Ende der **Hauptuntersuchung** können die Vorstellungen der Probanden zum Untersuchungsziel als Kontrollvariable erfasst werden.

Erwartungseffekte können nicht nur auf Seiten der Versuchspersonen, sondern auch der Versuchsleitenden auftreten: So könnte ein Versuchsleiter, der weiß, dass er gerade der Experimentalgruppe das neue Medikament verabreicht, dabei unbewusst enthusiastischer vorgehen als bei der Vergabe des Placebos. Nachweislich können minimale Änderungen in der Körpersprache der Versuchsleitenden dazu führen, dass Versuchspersonen wiederum ihr Verhalten ändern, sich z. B. wie „gute Versuchspersonen" verhalten und unbewusst aktiv zur Bestätigung der Hypothesen beitragen. Man spricht hier von **Versuchsleitereffekten** (Rosenthal, 2002). Um diese auszuschließen wird in einem idealen Experiment nicht nur mit blinden Versuchspersonen (◘ Abb. 7.2), sondern auch mit **blinden Versuchsleitenden** gearbeitet, die ihrerseits nicht wissen, ob sie gerade ein Medikament an die Experimentalgruppe oder ein Placebo an die Kontrollgruppe vergeben bzw. ob sie gerade Mitglieder der Experimental-



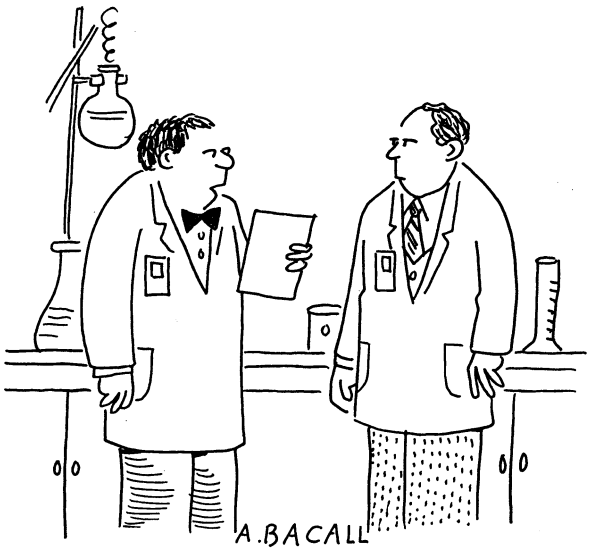
Abb. 7.2 Blinde Versuchspersonen wissen nicht, ob sie zur Experimental- oder zur Kontrollgruppe gehören und welches Ziel mit der Studie verfolgt wird. (© Jack Corbett/Search ID: jco0117, Rights Available from CartoonStock.com)

oder Kontrollgruppe nach ihrem Befinden befragen. Man spricht bei einer Verblindung des Vorgehens sowohl auf Seiten der Versuchspersonen als auch der Versuchsleitenden von einem **Doppelblind-Versuch** („double-blind trial", „double-blind experiment"). Bei einem **Dreifachblind-Versuch** würde man zusätzlich dafür sorgen, dass auch die Datenanalyse von Personen durchgeführt wird, die nicht wissen, welche Gruppe die Experimental- oder Kontrollgruppe ist, um auch hier unbewusste hypothesenkonforme Verzerrungen – z. B. bei der Datenbereinigung (▶ Kap. 11) – zu vermeiden.

So wünschenswert **Verblindung** ist, weil sie untersuchungsbedingte Störfaktoren reduziert und somit die interne Validität von Experimenten steigert, sie stößt auch an ihre Grenzen:

- **Forschungspraktische Grenzen**: Während die Medikamenten-/Placebovergabe forschungspraktisch problemlos verblindet werden kann, ist dies bei anderen experimentellen unabhängigen Variablen nicht möglich. So muss bei einem experimentellen Vergleich von Operationsmethoden die Operateurin bzw. der Operateur natürlich wissen, welcher Gruppe der aktuelle Patient gerade zuzuordnen ist. Ebenso kann in bestimmten Fällen den Versuchspersonen nicht verheimlicht werden, dass sie sich in der Experimentalgruppe befinden und welche Effekte angezielt werden. So ist davon auszugehen, dass Teilnehmende eines Anti-Rauch-Trainings wissen, dass es um eine Reduktion ihres Nikotinkonsums geht.
- **Forschungsökonomischer Aufwand**: Die Verblindung ist mit zusätzlichem organisatorischen und personellen Aufwand verbunden, weshalb in der Sozialforschung nicht selten aus forschungsökonomischen Gründen auf blinde Versuchsleiter und blinde Auswerter verzichtet wird. Insbesondere bei Qualifikationsarbeiten liegen Planung, Durchführung und Auswertung des Experiments meist in einer Hand und Studierende haben kaum Möglichkeiten, blinde Versuchsleiter für sich arbeiten zu lassen. In der Medizinforschung wird Verblindung dagegen fast immer realisiert. Wie bei anderen Methodenentscheidungen auch, muss hier die Kosten-Nutzen-Relation reflektiert und das Vorgehen begründet werden.
- **Forschungsethik**: Da die Teilnahme an wissenschaftlichen Studien grundsätzlich freiwillig erfolgen und auf informierter Einwilligung basieren soll, stellen sich durch Verblindung auch ethische Fragen, wenn Versuchspersonen vorab nicht darüber aufgeklärt werden, was im Rahmen des Experiments denn nun genau auf sie zukommt. Die Einwilligungserklärung muss deswegen auf die verschiedenen Behandlungsmöglichkeiten hinweisen. In jedem Fall sollten Versuchspersonen – insbesondere wenn sie über den wahren Hintergrund der Studie zunächst getäuscht werden mussten – im Nachgang genau aufgeklärt werden **(Debriefing)**. Auch muss den Probanden dann nachträglich noch die Möglichkeit gegeben werden, der Verwendung ihrer Daten zu widersprechen.

Bislang haben wir Maßnahmen zur Steigerung der internen Validität betrachtet. Sie stehen leider in einem Spannungsverhältnis zur externen Validität. Denn intern valide ist ein Experiment besonders dann, wenn mit sehr **homogenen Versuchspersonengruppen** unter **kontrollierten Laborbedingungen** gearbeitet wird. Beides schränkt die Generalisierbarkeit ein (Abb. 7.3). Befunde, die in Deutschland mit Psychologiestudierenden des ersten Semesters gewonnen werden, lassen sich nicht ohne weiteres auf andere Kulturen, andere Altersgruppen, andere Bildungsschichten etc. verallgemeinern. Allerdings macht es im Rahmen der Experimentalforschung überhaupt keinen Sinn, die kleinen Untersuchungsgruppen bewusst heterogen oder vermeintlich „repräsentativ" zu besetzen, denn dies würde die interne Validität der Studie infrage stellen. Die **externe Validität im Hinblick auf die Zielpopulation** lässt sich jedoch steigern durch **systematische Replikationsstudien**, bei denen die z. B. an Studierenden nachgewiesenen Effekte mit homogenen Gruppen von Schülern oder mit homogenen Gruppen von Senioren repliziert werden. Die **externe Validität im Hinblick auf die alltägliche Lebenswelt** lässt sich steigern, indem Laborbedingungen möglichst alltagsnah gestaltet werden **(ökologische Validität)** oder indem das Experiment statt als Laborexperiment als **Feldexperiment** durchgeführt wird (zu Feldstudien ▶ Abschn. 7.7). Zur Bewertung und Steigerung der Aussagekraft experimenteller Studien und insbesondere ihrer Validität



"There's a flaw in your experimental design. All the mice are libras."

Abb. 7.3 Homogene Untersuchungsgruppen sind im Experiment im Sinne interner Validität wünschenswert, allerdings können sie die externe Validität einschränken. (© Aaron Bacall/Search ID: aba0006, Rights Available from CartoonStock.com)

► Abschn. 3.2.2, der die vier Typen der Validität gemäß Campbell-Tradition vorstellt.

Entscheidungskriterien

Wählen Sie eine experimentelle Studie, wenn Sie gut begründete, theoretisch abgeleitete Kausalhypothesen prüfen wollen. Experimente müssen methodisch sehr gründlich geplant werden, um interne und externe Validität zu sichern. Dabei sind methodische Pretests unverzichtbar, um vor der Hauptuntersuchung sicherzustellen, dass der Versuchsablauf reibungslos funktioniert. Der Aufwand für ein Experiment hängt stark vom konkreten Design ab: Er kann relativ gering (z. B. Online-Experiment), aber auch sehr groß sein (z. B. Laborexperiment mit großer Probandenzahl und einstündigem Versuchsdurchgang pro Versuchsperson). Typischerweise unterschätzen Studierende den Aufwand für die Versuchsdurchführung und überschätzen die Wahrscheinlichkeit, tatsächlich statistisch und praktisch bedeutsame Effekte nachweisen zu können. Als Einstieg in die Experimentalforschung eignen sich ein- oder zweifaktorielle Designs mit überschaubarem Aufwand und dabei v. a. Replikationsstudien (► Abschn. 7.3).

7.6.2 Die quasi-experimentelle Studie

Ebenso wie ein echtes Experiment dient ein Quasi-Experiment („quasi-experimental study") der Überprüfung des kausalen Einflusses einer oder mehrerer unabhängiger Variablen (UV) auf die Ausprägungen einer oder mehrerer abhängiger Variablen (AV). Auch im Quasi-Experiment wird die unabhängige Variable aktiv variiert, indem den Untersuchungsgruppen jeweils eine Behandlung (Treatment) zukommt, während die Kontrollgruppe unbehandelt bleibt. Allerdings findet beim Quasi-Experiment **keine Randomisierung** statt: Die Versuchspersonen werden nicht zufällig den Untersuchungsgruppen zugeordnet, sondern man arbeitet mit **vorgefundenen Gruppen**. Deswegen besteht im Quasi-Experiment die Möglichkeit, dass sich die Vergleichsgruppen nicht nur hinsichtlich der unabhängigen Variablen, sondern zusätzlich hinsichtlich weiterer Merkmale (personengebundene Störvariablen bzw. konfundierende Variablen) systematisch unterscheiden. Ergeben sich in einer quasi-experimentellen Untersuchung Gruppenunterschiede in Bezug auf die abhängige Variable, so sind diese nicht eindeutig auf die unabhängige Variable zurückzuführen: Die quasi-experimentelle verfügt im Vergleich zu einer experimentellen Untersuchung über eine **geringere interne Validität**.

Umgang mit Störvariablen beim Quasi-Experiment

Bei **quasi-experimentellen Studien** („quasi-experimental studies", „quasi-experiments") findet im Unterschied zum echten Experiment keine Randomisierung statt, wohl aber eine experimentelle Variation der unabhängige/n Variablen. Die interne Validität des Quasi-Experiments hinsichtlich personenbezogener Störvariablen ist wegen fehlender Randomisierung geringer als beim echten Experiment. Es müssen verstärkt versuchsplanerische und auch statistische Verfahren der Kontrolle von **personengebundenen Störvariablen** zum Einsatz kommen, um die interne Validität quasi-experimenteller Studien zu steigern (z. B. Bildung der Untersuchungsgruppen nach den Prinzipien der Parallelisierung oder des Matching von Stichproben). Auch die Kontrolle von **untersuchungsbedingten Störvariablen** ist bei Quasi-Experimenten erschwert, etwa wenn sie als Feldexperimente stattfinden.

Auf quasi-experimentelle Designs wird v. a. dann zurückgegriffen, wenn anstelle von Laborstudien **Feldstudien**

durchgeführt werden (z. B. Evaluationsstudien im Feld). Im natürlichen Umfeld ist es nämlich in der Regel kaum möglich, festgefügte Personengruppen (z. B. Patientengruppen, Abteilungen in einer Firma, Schulklassen etc.) auseinanderzureißen und zum Zweck des Experiments per Zufallsprinzip neu zusammenzusetzen. Will man etwa überprüfen, ob sich der Einsatz von Laptops im Schulunterricht positiv auf den Lernerfolg auswirkt, so wird man einige Klassen mit Laptops ausstatten, andere nicht, und am Ende die Leistungen der Schülerinnen und Schüler in der Laptop-Bedingung mit denen der Kontrollbedingung vergleichen. Wenn die Laptop-Klassen besser abschneiden ist aufgrund des quasi-experimentellen Designs jedoch unklar, ob dies wirklich auf den Laptop als neuem Lehr-Lern-Medium zurückzuführen ist oder nicht vielmehr einen Lehrer- bzw. Klasseneffekt darstellt, weil vielleicht nur besonders engagierte und innovative Lehrkräfte mit ihren ohnehin besonders leistungsfähigen Klassen sich zum Laptop-Einsatz bereit erklärt haben. Um diese Einbußen der internen Validität auszugleichen, werden in quasi-experimentellen Studien folgende **Techniken zur Kontrolle personenbezogener Störvariablen** eingesetzt (► Abschn. 3.2.2):

1. **Konstanthalten**: Die potenzielle personengebundene Störvariable wird konstant gehalten, ist also in allen Untersuchungsgruppen gleich ausgeprägt. So wird bei unserem Beispiel eines Quasi-Experiments zur Wirkung von Laptop-Unterricht vermutlich das Alter der Schülerinnen und Schüler wesentlich mitbestimmen, ob und wie sie vom Laptop-Einsatz beim Lernen profitieren. Würde man also eine Laptop-Klasse mit älteren Schülern mit einer herkömmlichen Klasse mit jüngeren Schülern vergleichen, so wäre die interne Validität gefährdet. Stattdessen wird man diesen Störfaktor zunächst ausschalten, indem man nur Klassen mit gleichaltrigen Schülern vergleicht. Auch die Störvariable Klassengröße wird man am besten konstant halten.
2. **Parallelisierung von Stichproben**: Ist das Konstanthalten nicht möglich oder sinnvoll, so sollten die Gruppen so ausgewählt werden, dass die verschiedenen Ausprägungen der potenziellen Störvariablen zumindest vergleichbar unterschiedlich ausgeprägt sind (z. B. gleicher Anteil an Muttersprachlern vs. Nichtmuttersprachlern in allen Untersuchungsgruppen; gleicher Mittelwert und gleiche Streuung hinsichtlich häuslicher Computernutzungszeit in allen Untersuchungsgruppen). Für die Laptop-Klasse wird also eine hinsichtlich **Aggregatwerten** der relevanten Störvariablen möglichst vergleichbare Kontrollklasse gesucht.
3. **Matching von Stichproben**: Bei kleinen Gruppen ($n < 20$ pro Versuchsbedingung) verwendet man anstelle der Parallelisierung gemäß Aggregatwerten (z. B. Prozent- oder Mittelwerten) lieber sog. Matched Samples (verbundene Stichproben) oder Paired Samples (gepaarte Stichproben). Hierbei wird für jeden einzelnen Schüler der Laptop-Klasse ein hinsichtlich aller relevanten Störvariablen (z. B. Schulnoten, Sozialstatus der Eltern, Computererfahrung etc.) möglichst ähnlicher „Kontrollschüler" aus einer anderen Klasse gesucht. Auf diese Weise wird eine gematchte bzw. gepaarte Kontrollgruppe zusammengestellt (das Matching-Verfahren ist auch auf mehr als zwei Gruppen erweiterbar). Durch Parallelisierung und Matching entstehen **abhängige Stichproben**, so dass bei der statistischen Auswertung auf entsprechende Verfahren zurückzugreifen ist (z. B. t-Test für abhängige Stichproben, Varianzanalyse mit Messwiederholungen).
4. **Störvariable als unabhängige Variable**: Möchte man dem Einfluss einer nominalen bzw. kategorialen Störvariable genauer nachgehen, so kann man diese auch als weitere unabhängige Variable in das Design aufnehmen. Somit kann dann deren Interaktion mit der ursprünglichen unabhängigen Variable getestet werden. So könnte für das Quasi-Experiment zum Laptop-Unterricht das Geschlecht als weitere unabhängige Variable in das Design aufgenommen werden, sofern vermutet wird, dass Schülerinnen und Schüler geschlechtsrollenbedingt die Laptops unterschiedlich nutzen und somit ggf. andere Lernergebnisse resultieren.
5. **Störvariable als Kontrollvariable**: Ob und wie die Eltern die Kinder hinsichtlich der schulischen Laptop-Nutzung unterstützen und welche Einstellungen die Eltern dem Projekt gegenüber haben, könnte ebenfalls einen Einfluss auf die untersuchten Lerneffekte haben. Entsprechende Störvariablen ließen sich durch einen Elternfragebogen erfassen und bei der Auswertung der Daten statistisch als Kontrollvariablen mitberücksichtigen (z. B. durch Herauspartialisieren ihres Einflusses aus den Ergebnissen). Neben der Kovarianzanalyse kann zur Kontrolle des Einflusses von Störvariablen u. a. auch die **Propensity-Score-Analyse** eingesetzt werden (Guo & Fraser, 2010; Harder, Stuart, & Anthony, 2010; Pruzek, 2011).
6. **Messwiederholungen**: Um den Einfluss einer unabhängigen Variable auf eine interessierende abhängige Variable zu bestimmen, sind Messwiederholungen einschlägig. Im einfachsten Fall wird die abhängige Variable vor und nach der Intervention gemessen. Sind Treatment- und Kontrollgruppe vergleichbar, so sollten beide Gruppe ähnliche durchschnittliche Vorher-Messungen aufweisen (Prä-/Pretestwerte). In der Treatmentgruppe (auch: Untersuchungs- oder Expe-

rimentalgruppe) sollte es zu einer Veränderung der abhängigen Variable kommen, so dass hier die Posttestwerte von den Prätestwerten abweichen. In der Kontrollgruppe dagegen sollten kaum Prä-Post-Unterschiede erkennbar sein. Wird die Messwiederholung als zusätzliche unabhängige Variable bzw. **Messwiederholungsfaktor** in das Design einbezogen, sollte sich der Treatmenteffekt in einem signifikanten Interaktionseffekt der mehrfaktoriellen Varianzanalyse widerspiegeln. Eine andere Auswertungsvariante besteht darin, die **Prätestwerte als Kontrollvariablen** zu behandeln und die Posttestwerte entsprechend im Zuge einer Kovarianzanalyse zu bereinigen. Damit können Baseline-Unterschiede (d. h. Vorher-Unterschiede) zwischen den Gruppen statistisch kontrolliert werden (Senn, 2006).

Zudem müssen (genau wie im echten Experiment) auch im Quasi-Experiment **Techniken zur Kontrolle der untersuchungsbedingten Störvariablen** eingesetzt werden. Dies setzt jedoch voraus, dass im Einzelnen bekannt ist, welche untersuchungsbedingten Störvariablen wirksam werden könnten (► Abschn. 3.2.2):

1. **Ausschalten**: Störungen, die den Ablauf in einzelnen Gruppen unterschiedlich beeinträchtigen könnten, sollten nach Möglichkeit ausgeschlossen werden. So sollte z. B. für das geschilderte Quasi-Experiment zum Laptop-Unterricht sichergestellt werden, dass es während des Versuchszeitraums in keiner Gruppe zu Unterrichtsausfall kommt (Vertretungslösung) und dass der Unterricht in der Experimentalgruppe nicht durch Geräteausfall behindert wird (Ersatzgeräte).
2. **Konstanthalten**: Untersuchungsbedingte Störvariablen können z. B. räumliche Umgebungsfaktoren sein. Wenn der Laptop-Unterricht im Computerraum (höherer Geräuschpegel, schlechtere Luft etc.), der herkömmliche Unterricht aber im normalen Klassenzimmer stattfindet, so kann dies die interne Validität einschränken. Der Unterricht sollte also zur Vergleichbarkeit der Gruppen in vergleichbaren Räumlichkeiten stattfinden.
3. **Registrieren**: Relevante Störeinflüsse, die sich weder ausschalten noch konstant halten lassen, müssen zumindest registriert werden, so dass sie nachher bei der statistischen Auswertung zur Bereinigung der Daten oder zumindest zur Interpretation der Ergebnisse genutzt werden können. So könnte sich das Pausenverhalten der Schülerinnen und Schüler der Laptop-Bedingung verändern. Wie lange und wozu die Geräte in den Pausen genutzt werden, könnte z. B. mittels Befragung, Beobachtung oder nonreaktiver Datenerhebung (z. B. Logfiles der Computer) registriert werden.

Entscheidungskriterien

Wählen Sie eine **quasi-experimentelle Studie**, wenn Sie Kausalhypothesen prüfen und Bedingungen variieren wollen, aber die Untersuchungsgruppen nicht randomisiert zusammenstellen können (z. B. bei einem Feldexperiment). Treffen Sie bei einer quasi-experimentellen Studie geeignete Maßnahmen, um trotz fehlender Randomisierung die personengebundenen Störvariablen zu kontrollieren (z. B. Parallelisierung oder Matching der Stichproben; Erhebung und statistische Kontrolle der Störvariablen). Wenn das Quasi-Experiment im Feld stattfindet, sind zudem besondere Vorkehrungen zu treffen, um die untersuchungsbedingten Störvariablen zu kontrollieren. Die Versuchsplanung ist hier besonders gründlich zu durchdenken und durch Pretests abzusichern, um einerseits die Aussagekraft des Quasi-Experiments zu optimieren, andererseits aber auch den Bedingungen im Untersuchungsfeld gerecht zu werden.

7.6.3 Die nicht-experimentelle Studie

Viele Wirkungshypothesen werden weder in einer klassischen experimentellen Laborstudie noch in einer quasi-experimentellen Feldstudie überprüft, sondern im Rahmen **nicht-experimenteller Studien** („non-experimental studies"). Nicht-experimentelle Studien vergleichen vorgefundene Gruppen hinsichtlich vorgefundener Unterschiede auf den abhängigen Variablen. Es findet nicht nur **keine Randomisierung**, sondern auch **keine aktive experimentelle Manipulation** statt. Nicht-experimentelle Studien werden deswegen auch als **Ex-post-facto-Studien** („ex post facto studies") bezeichnet, denn Effekte werden nur „im Nachhinein" konstatiert (► Studienbeispiel „Hatte das Training für Führungskräfte einen Effekt? Diskussion verschiedener Untersuchungspläne"). Um zum Ausdruck zu bringen, dass der Ursachenfaktor nicht von den Forschenden hergestellt, sondern nur vorgefunden und beobachtet wird, werden nicht-experimentelle Studien – etwas missverständlich – auch als **Beobachtungsstudien** („observational studies"; Rosenbaum, 2010) oder als **korrelative Studien** („correlational studies") bezeichnet, obwohl sie keinesfalls immer mit Beobachtungsmethoden Daten erheben oder mit Korrelationsanalysen Daten auswerten. Man spricht bei nicht-experimentellen Studien des weiteren von **natürlichen Experimenten** („natural experiment"), um zum Ausdruck zu bringen, dass bestimmte Ursachenfaktoren auf manche Menschen

Studienbeispiel

Hatte das Training für Führungskräfte einen Effekt? Diskussion verschiedener Untersuchungspläne

Die Firma K. beabsichtigt, die Führungsqualitäten ihres mittleren Managements durch ein Trainingsprogramm zu verbessern. Nachdem Frau W. als Abteilungsleiterin das Trainingsprogramm absolviert hat, überprüft die Firmenleitung das Betriebsklima, die Arbeitszufriedenheit und die Produktivität ihrer Abteilung. (Empirische Untersuchungen, die sich mit der Wirksamkeit von Trainingsprogrammen bzw. Interventionen befassen, bezeichnet man als „Evaluationsstudien"; ► Kap. 18.) Die Auswertung der Fragebögen führt zu dem Resultat, dass es in dieser Abteilung keine Gründe für Beanstandungen gibt.

Formal lässt sich dieses Untersuchungsdesign folgendermaßen beschreiben:

$$T \rightarrow M$$

Mit T ist die Schulungsmaßnahme gemeint. Der Buchstabe kürzt die Bezeichnung „Treatment" (Behandlung) ab, die üblicherweise für experimentelle Eingriffe, Manipulationen oder Maßnahmen verwendet wird. M steht für „Messung" und symbolisiert in diesem Beispiel die Befragung der Mitarbeitenden der Abteilung von Frau W. nach dem Treatment.

Diese **One-Shot Case Study** (Cook & Campbell, 1976) ist kausal nicht interpretierbar, d. h., die Tatsache, dass es in der Abteilung nach der Schulung nichts zu beanstanden gibt, kann nicht zwingend auf die Schulung der Abteilungsleiterin als Ursache zurückgeführt werden, denn vielleicht gab es ja vorher schon nichts zu beklagen. Um Veränderungen in der Abteilung registrieren zu können, hätte die Abteilung nicht nur nach, sondern auch vor der Schulungsmaßnahme befragt werden müssen. Für dieses **Ein-Gruppen-Prätest-Posttest-Design** wird die folgende Formalisierung verwendet:

$$M_1 \rightarrow T \rightarrow M_2$$

Nach einer Prätestmessung (M_1) erfolgt das Treatment und danach eine erneute Messung, die Posttestmessung (M_2). Ein Vergleich dieser beiden Messungen liefert Hinweise über mögliche, zwischenzeitlich eingetretene Veränderungen.

Aber auch dieser Plan lässt nicht den zwingenden Schluss zu, die Veränderungen seien ursächlich auf das Training bzw. das Treatment zurückzuführen. Generell muss bei Untersuchungen von diesem Typus damit gerechnet werden, dass eine Veränderung auftritt, weil

- zwischenzeitliche Einflüsse unabhängig vom Treatment wirksam werden (z. B. eine Lohnerhöhung),
- sich die Untersuchungsteilnehmenden unabhängig vom Treatment weiter entwickelten (sie werden z. B. mit ihren Aufgaben besser vertraut),
- allein die Prätestmessung das Verhalten veränderte (die Untersuchungsteilnehmenden werden z. B. durch die Befragung auf bestimmte Probleme aufmerksam gemacht),
- das gemessene Verhalten ohnehin einer starken Variabilität unterliegt (z. B. könnten die Arbeitsanforderungen saisonalen Schwankungen unterliegen, die den Effekt des Trainings überlagern), oder
- sich die Messungen aus formal-statistischen Gründen verändern können (diese „Regressionseffekte" betreffen vorzugsweise Extremwerte, die bei wiederholten Messungen zur Mitte tendieren).

Auch dieses Untersuchungsdesign lässt also keine eindeutige Interpretation zu.

Ein dritter Plan könnte die „behandelte" Gruppe mit einer unbehandelten, nicht-äquivalenten Kontrollgruppe vergleichen (nicht-äquivalent deshalb, weil die Kontrollgruppe, anders als in rein experimentellen Untersuchungen, natürlich im Feld angetroffen wird und nicht per Randomisierung zustande kommt). Diese könnte z. B. aus einer anderen Abteilung bestehen, deren Leiter keine Schulung erhielt.

$$\frac{T \rightarrow M_1}{M_2}$$

Man bezeichnet diesen Plan als **Ex-post-facto-Plan**, d. h., die vergleichende Messung wird erst nach erfolgtem Treatment vorgenommen. Auch dieser Plan leidet an schlechter Interpretierbarkeit. Unterschiede zwischen den Vergleichsgruppen sind uneindeutig, da man nicht ausschließen kann, dass sie bereits vor Behandlung der Experimentalgruppe bestanden.

Zuverlässigere Interpretationen ließe ein Plan zu, der wiederholte Messungen bei beiden Gruppen vorsieht, also bei einem **Kontrollgruppenplan mit Prä- und Posttest**-Messung:

$$\begin{array}{lcl} M_{11} \rightarrow & T & \rightarrow M_{12} \\ M_{21} & \rightarrow & M_{22} \end{array}$$

Mit M_{11} und M_{12} werden die Prätest- und Posttestmessungen in der Experimentalgruppe (Gruppe 1 mit Führungskräftetraining) verglichen. Besteht hier ein Unterschied, informiert der Vergleich M_{21} und M_{22} in der Kontrollgruppe (Gruppe 2 ohne Führungskräftetraining) darüber, ob die Differenz $M_{11} - M_{12}$ für einen Treatmenteffekt spricht oder ob andere Ursachen für die Differenz verantwortlich sind, was zuträfe, wenn die gleiche Veränderung auch in der Kontrollgruppe registriert wird.

Zeigen sich nun in der Experimentalgruppe andere Veränderungen als in der Kontrollgruppe, ist dies noch immer kein sicherer Beleg für die kausale Wirksamkeit des Treatments (hier: Führungskräftetraining). Es könnte sein, dass der Effekt darauf zurückzuführen ist, dass die trainierte Abteilungsleiterin hauptsächlich jüngere Mitarbeitende anleitet, die den neuen Führungsstil positiv aufnehmen. Ältere Mitarbeitende hätten auf den neuen Führungsstil möglicherweise völlig anders reagiert. Das Alter der Mitarbeitenden übt damit einen Einfluss auf die abhängige Variable aus; die Wirkung des Treatments richtet sich danach, mit welcher Altersstufe es kombiniert wird.

Eine höhere interne Validität hätte eine experimentelle Studie, bei der per Zufall entschieden wird, welche Mitarbeitenden zur Experimentalgruppe und welche zur Kontrollgruppe gehören sollen (Randomisierung). Ob jedoch die Geschäftsführung der Firma K. diese Studie genehmigen würde, muss bezweifelt werden. (Für Details zur Beurteilung der Gültigkeit von Schlussfolgerungen, die aus unterschiedlichen experimentellen, quasi-experimentellen und nicht-experimentellen Untersuchungsdesigns gezogen werden sollen sowie einer weiteren Variante der ► Versuchsplan-Formalisierung ► Abschn. 3.2.2).

einwirken („Experimentalgruppe“) und auf andere nicht („Kontrollgruppe“), dass diese Umstände jedoch nicht von den Forschenden zu beeinflussen sondern sozusagen vom „natürlichen Lauf der Dinge“ vorgegeben sind: Vergleiche zwischen Rauchern und Nichtrauchern, Vegetariern und Nichtvegetariern oder Computerspielern und Nichtspielern wären Beispiele für natürliche Experimente bzw. nicht-experimentelle Studien, bei denen der Einfluss eines vorgefundenen Treatments (Rauchen, Ernährungsstil, Mediennutzungsverhalten) auf bestimmte abhängige Variablen von Interesse ist. Die interne Validität nicht-experimenteller Studien ist noch geringer als bei quasi-experimentellen Studien.

Angenommen, man möchte die z. B. aus der Theorie des sozialen Lernens (Lernen am Modell) abgeleitete Hypothese prüfen, dass Jugendliche, die häufig gewalthaltige Computerspiele nutzen, aufgrund der Vorbildwirkung der Spielfiguren häufiger aggressive Gedanken, Gefühle und Verhaltensweisen zeigen als Jugendliche, die seltener oder gar nicht auf gewalthaltige Computerspiele zurückgreifen. Die unabhängige Variable wäre hier die **Nutzungsintensität gewalthaltiger Computerspiele** mit drei Ausprägungen (intensive Nutzung, moderate Nutzung, keine Nutzung). Die abhängige Variable wäre die Intensität der **Aggressivität**, die über verschiedene kognitive, emotionale und verhaltensbezogene Aggressivitätsmaße mittels entsprechender Fragebögen (oder anderer Datenerhebungsverfahren) gemessen wird. Zeigt sich nun hypothesenkonform, dass die Intensivnutzer von gewalthaltigen Computerspielen im Durchschnitt tatsächlich signifikant aggressiver sind als die Wenig- und Nichtnutzergruppen, so kann dies – aufgrund mangelnder interner Validität nicht-experimenteller Studien – nicht eindeutig auf eine Wirkung der unabhängigen Variablen (hier Nutzungsintensität gewalthaltiger Computerspiele) zurückgeführt werden. Man beachte: Diese **Unterschiedshypothese** kann analog auch als **Zusammenhangshypothese** geprüft werden, indem die Computerspielintensität direkt gemessen wird (z. B. in Stunden und Minuten pro Tag) anstatt die Spieler in Gruppen einzuteilen. Dementsprechend würde man einen positiven Zusammenhang zwischen der Intensität der Nutzung gewalthaltiger Computerspiele und der Intensität der Aggressivität postulieren.

Wenn eine nicht-experimentelle Studie einen hypothesenkonformen Gruppenunterschied bzw. Variablenzusammenhang belegt, ist damit kein Ursache-Wirkungs-Mechanismus nachgewiesen. Denn die vorgefundenen regelmäßigen Gewaltspieler können sich in vielen weiteren aggressionsrelevanten Faktoren als nur ihrem Computerspielverhalten von den anderen beiden Untersuchungsgruppen unterscheiden: Sie stammen vielleicht häufiger aus sozial unterprivilegierten Elternhäusern, haben schlechtere Schulnoten, sind sozial isolierter, haben häufiger häusliche Gewalt erlebt, haben eine generell erhöhte Aggressionsneigung etc. – all dies wären mögliche Ursachenfaktoren der beobachteten höheren Aggressivitätswerte in der Gruppe. Wird auf eine experimentelle Manipulation der unabhängigen Variablen bzw. des Ursachenfaktors verzichtet und werden Gruppen verglichen, bei denen sich die verschiedenen Ausprägungen der unabhängigen Variablen bereits in der Vergangenheit entwickelt haben, so ist von einer **Konfundierung der unabhängigen Variablen mit anderen Merkmalen (Störvariablen)** auszugehen. Diese kann in gewissen Grenzen statistisch aufgelöst werden, indem bekannte Störvariablen (z. B. elterlicher Erziehungsstil, Persönlichkeitsmerkmale) als Kontrollvariablen erfasst und statistisch herausgerechnet werden. Die Erhöhung der internen Validität durch die Erfassung von **Kontrollvariablen** stößt jedoch an zwei Grenzen: Zum einen müssen alle relevanten Störvariablen bekannt sein, was meist nicht der Fall ist. Zum anderen müssen die potenziell relevanten Kontrollvariablen alle gemessen werden, was aufgrund des damit verbundenen Aufwandes für die Untersuchungsteilnehmenden oft nicht möglich ist.

Es sind v. a. drei Umstände, unter denen zur Prüfung kausaler Hypothesen auf Experimente und Quasi-Experimente verzichtet werden und eine **nicht-experimentelle Studie** durchgeführt werden muss:

- Die interessierende unabhängige Variable ist eine personen- oder umweltgebundene Variable, die **forschungspraktisch gar nicht variierbar** ist: Studien, in denen Singles und Verheiratete, Linkshänder und Rechtshänder, Inländer und Ausländer, Einzelkinder und Geschwisterkinder etc. verglichen werden, lassen sich nur nicht-experimentell realisieren. Denn Personen bringen entsprechende personengebundene Variablen wie Familienstand, Händigkeit oder Nationalität bereits mit und man kann ihnen diese nicht experimentell zuweisen. Auch interkulturelle Studien, in denen einzelne Erlebens- und Verhaltensweisen der Bevölkerungen unterschiedlicher Länder miteinander verglichen werden, verwenden das Herkunftsland als nicht-experimentellen Faktor. Fehlende Variierbarkeit kommt zudem bei umweltgebundenen Variablen vor: So können kulturelle oder naturbedingte Großereignisse nicht künstlich hergestellt werden, um deren Effekte (quasi-)experimentell zu prüfen (z. B. gesellschaftliche Auswirkungen einer Scheidung im Königshaus, eines Jahrhundertsommers oder eines Wirtschaftsaufschwungs).
- Die interessierende unabhängige Variable lässt sich **nur mit sehr großem organisatorischen oder finanziellen Aufwand variieren**, was forschungsökonomisch dann nicht möglich ist: Innovative Bildungs- oder Therapiekonzepte lassen sich nur empirisch prü-

fen, wenn sie entsprechend in die Praxis umgesetzt werden, was oft an zahlreichen Hürden scheitert.

- Die interessierende unabhängige Variable ist eine prinzipiell gut variierbare Variable, deren experimentelle Variation jedoch **forschungsethisch nicht vertretbar** wäre (► Abschn. 4.1): Eine experimentelle Überprüfung der Wirkungen gewalthaltiger Computerspiele oder pornografischer Medienangebote auf Minderjährige ist nicht möglich, da man Kinder und Jugendliche zu wissenschaftlichen Zwecken nicht mit potenziell beeinträchtigenden Inhalten konfrontieren darf.

Die vergleichsweise beste Möglichkeit, Kausalhypothesen für Variablen zu prüfen, die nicht experimentell variiert werden können oder dürfen, ist die Durchführung von **Längsschnittstudien**, bei denen Veränderungen verfolgt werden können (► Abschn. 7.8). Sie haben im Vergleich zur Querschnittstudie eine höhere interne Validität hinsichtlich kausaler Interpretationen von vorgefundenen Gruppenunterschieden. Als nicht-experimentelles Design zur Analyse von Kausaleffekten ist u. a. die **prospektive Kohortenstudie** („prospective cohort study") als Spezialfall der Längsschnittstudie einsetzbar. Hierbei werden eine Gruppe, die aktuelle einem **potenziellen Ursachen- oder Risikofaktor ausgesetzt ist**, sowie eine vergleichbare nicht-exponierte Gruppe über die Zeit hinweg in ihrer Entwicklung verfolgt (z. B. daraufhin, ob eine bestimmte Krankheit bzw. Problematik als Wirkung auftritt oder nicht). Beispiele wären Studien, in denen man jeweils eine Stichprobe von Rauchern und Nichtrauchern oder von Vielspielern, Wenigspielern und Nichtspielern zieht und diese wiederholt hinsichtlich relevanter Wirkungen untersucht (z. B. Auftreten von Bronchitis und Lungenkrebs oder von Aggressivität, Bildungs- und Berufsmisserfolg).

Muss man den mit Messwiederholungen verbundenen hohen Aufwand einer Längsschnittstudie umgehen, will aber dennoch Kausaleffekte klären, bietet sich bei nicht-experimentellen Untersuchungen das Design der **Fall-Kontroll-Studie** („case control study") an. Hierbei zieht man eine Stichprobe von Fällen, in denen die **interessierende Wirkung bereits eingetroffen ist** (z. B. Erkrankung an Lungenkrebs oder hohe Aggressivität). Diesen Fällen stellt man eine möglichst genau passende (z. B. ähnliche sozialstatistische und gesundheitsbezogene Merkmale) Stichprobe von Personen gegenüber, die von der Wirkung nicht betroffen sind und die als Kontrollpersonen die Kontrollgruppe bilden. Nun erhebt man rückwirkend (retrospektiv) inwiefern die Gruppen früher dem vermuteten Ursachen- oder Risikofaktor ausgesetzt waren: Wie viele der Lungenkrebs-Patienten haben früher geraucht gegenüber den nicht an Lungenkrebs erkrankten Kontrollpersonen? Wie viele von den hochaggressiven Personen haben früher gewalthaltige Medien genutzt im Vergleich zu den durchschnittlich oder unterdurchschnittlich aggressiven Kontrollpersonen? (Zur statistischen Auswertung von Fall-Kontroll-Studien über die Effektgröße der **Odds-Ratio** ► Abschn. 14.2.1).

Einfache nicht-experimentelle Studien, die nicht auf ein längsschnittliches Kohorten- oder zumindest ein Fall-Kontroll-Design zurückgreifen, haben insgesamt eine so geringe interne Validität, dass sie zur Prüfung von Kausalhypothesen als ungeeignet anzusehen sind. Entsprechende Studien werden oft **überinterpretiert**, indem statistisch signifikante vorgefundene Zusammenhänge oder Unterschiede fälschlich als Nachweise von Kausalrelationen ausgelegt werden. Dabei sind bei Ex-Post-Facto-Studien oft die **Existenz einer Kausalbeziehung** und die **kausale Wirkungsrichtung** unklar: Stellt die nominelle unabhängige Variable (z. B. Pornografienutzung) tatsächlich die Ursache und die nominelle abhängige Variable tatsächlich die Wirkung (z. B. freizügiges Sexualverhalten) dar? Plausibel ist oft auch der umgekehrte Kausalpfad, demgemäß im Rahmen freizügigen Sexualverhaltens dann auch mehr Pornografie genutzt wird. Beide Variablen könnten sich auch wechselseitig bedingen und verstärken. Darüber hinaus sind noch komplexere Kausalpfade möglich und mit der empirischen Korrelation vereinbar. So könnte eine Drittvariable wie z. B. das sexuelle Interesse, Sensation Seeking, Neugier oder Neurotizismus dafür verantwortlich sein, dass Menschen sowohl ein freizügigeres Sexualverhalten als auch mehr Pornografienutzung an den Tag legen.

Bei nicht-experimentellen Studien ist nicht nur die kausale Wirkungsrichtung oft unklar, sondern meist auch das **kausale Wirkprinzip**. Der in diversen nicht-experimentellen Studien gut belegte Befund, dass Männer im Durchschnitt mehr Pornografie nutzen als Frauen, ist nicht einfach dadurch kausal erklärbar, dass das eben „am Geschlecht" der Personen liege (zur inhaltlichen Bedeutung und Messung der Variable Geschlecht ► Abschn. 8.5.1). Denn mit dem Geschlecht sind zahlreiche andere Variablen konfundiert, die kausal für den Unterschied verantwortlich sein können (sexuelle Einstellungen, Masturbationsgewohnheiten, soziale Identität, Genrekenntnisse etc.). Um kausalen Wirkprinzipien auf die Spur zu kommen, ist eine gründliche Theoriearbeit, eine breite Erfassung möglicher kausaler Einflussvariablen sowie eine entsprechend komplexe statistische Auswertung notwendig (zu multiplen Regressionsanalysen ► Abschn. 12.5.2 oder zu Strukturgleichungsmodellen ► Kap. 17).

Umgang mit Störvariablen bei nicht-experimentellen Studien

Bei nicht-experimentellen Studien („non-experimental studies“) findet keine experimentelle Manipulation steht, deswegen entfallen auch **untersuchungsbedingte Störvariablen**, die mit der Bildung und Behandlung der Untersuchungsgruppen verbunden sein könnten. Da eine nicht-experimentelle Studie ausschließlich vorgefundene Gruppen vergleicht, die sich nicht nur hinsichtlich der interessierenden unabhängigen Variablen, sondern gleichzeitig zahlreicher weiterer damit verbundener (konfundierter) Variablen unterscheiden, sind **personengebundene Störvariablen** kaum so gut kontrollierbar, dass eindeutige (intern valide) kausale Schlussfolgerungen über Ursache-Wirkungs-Relationen (deren Existenz, Richtung und Wirkmechanismus) getroffen werden können. Damit nicht-experimentelle Studien erste Hinweise auf Kausaleffekte geben können, sind Störvariablen möglichst vollständig als Kontrollvariablen zu erfassen und statistisch zu kontrollieren. Auch eine Erweiterung des nicht-experimentellen Querschnittsdesigns zum Längsschnittdesign, insbesondere zur **prospektiven Kohortenstudie** ist zu erwägen (► Abschn. 7.8). Sofern dies wegen des hohen Aufwandes nicht möglich ist, kann ein **Fall-Kontroll-Design** genutzt werden.

Bis hierher wurden nicht-experimentelle Studien v. a. hinsichtlich ihrer **Grenzen kausaler Aussagekraft** im Vergleich zu Quasi-Experimenten und echten Experimenten diskutiert. Es wäre allerdings zu einseitig, würde man nicht-experimentelle Studien nur als mangelbehaftete Designs beschreiben. In vielen Fällen geht es in der sozialwissenschaftlichen Forschung schließlich darum, Unterschieds-, Zusammenhangs- oder Veränderungshypothesen zu prüfen, bei denen der Nachweis von Kausalität gar nicht im Zentrum steht: So kann es – unabhängig von den Verursachungsverhältnissen – relevant sein herauszufinden, ob und wie sich z. B. Fernsehnachrichten in den letzten Dekaden verändert haben, welche Unterschiede und Gemeinsamkeiten in der deutschen und japanischen Unternehmenskultur bestehen, welche Verhaltensweisen besonders erfolgreiche Führungspersönlichkeiten zeigen oder ob es Zusammenhänge zwischen Persönlichkeitsmerkmalen und Drogenkonsum gibt. Generell sind nicht-experimentelle Studien gut geeignet, um **nicht-kausale Forschungsfragen** zu beantworten und **nicht-kausale Forschungshypothesen** zu prüfen, die sich z. B. darauf beziehen, wer, wo, was, wie oft, wie lange, wozu etc. tut oder erlebt. Anstelle der für Experimente und Quasi-Experimente gemäß der Ceteris-Paribus-Klausel gebotenen Beschränkung auf wenige Kausalfaktoren und relativ kleine und homogene Untersuchungsgruppen, kann bei nicht-experimentellen Studien ein **viel breiteres Spektrum von Variablen und Untersuchungseinheiten** einbezogen werden.

Entscheidungskriterien

Planen Sie eine **nicht-experimentelle Studie**, wenn Sie an nicht-kausalen Forschungsfragen oder Forschungshypothesen interessiert sind. Greifen Sie zudem auf eine nicht-experimentelle Studie zurück, wenn eine experimentelle Manipulation forschungsethisch nicht gerechtfertigt oder forschungspraktisch nicht zu leisten oder unmöglich wäre. Sind Sie hier dennoch an der Klärung von Ursache-Wirkungs-Relationen interessiert, streben Sie die Umsetzung der nicht-experimentellen Studie als Längsschnittstudie (z. B. **prospektive Kohortenstudie**; ► Abschn. 7.8) oder zumindest als querschnittliche **Fall-Kontroll-Studie** an. Kausale Interpretationen von vorgefundenen Gruppenunterschieden hinsichtlich des interessierenden Ursachenfaktors sind nur mit großer Vorsicht möglich und verlangen, dass möglichst viele relevante personengebundene Störvariablen als **Kontrollvariablen erfasst und statistisch kontrolliert** werden. Dies erfordert umfassende theoretische Vorarbeit (► Kap. 6), um die einschlägigen Störvariablen zu kennen und angemessen im Untersuchungsinstrument (z. B. Fragebogen) zu operationalisieren (► Kap. 8). Zudem sind spezielle statistische Auswertungsverfahren notwendig, um den Einfluss von Kontrollvariablen aus dem eigentlich interessierenden Gruppenunterschied oder Variablenzusammenhang herauszurechnen (► Abschn. 12.5.2).

7.7 Labor- oder Feldstudie

Bei der Planung von empirischen Studien ist der **Untersuchungsort** ein wesentlicher Faktor. Grob unterscheidet man zwischen den Untersuchungsorten **Labor** (künstliche bzw. kontrollierte/kontrollierbare Umgebung) und **Feld** (natürliche bzw. wenig kontrollierte/kontrollierbare Umgebung). Zwischen beiden Extremen von „Labor“ und

„Feld“ (z. B. perfekt akustisch abgeschirmter und klimatisierter Laborraum in einem Forschungsinstitut vs. heimisches Wohnzimmer der Untersuchungsperson) existiert ein breites **Spektrum** mehr oder minder umfassend durch die Forschenden kontrollierbarer Untersuchungsorte (z. B. als Untersuchungsraum genutzter Pausenraum eines Betriebes oder einer Kindertagesstätte, der in begrenztem Rahmen für die Studiendurchführung umgestaltet werden konnte).

Laborstudie – Die Laborstudie („laboratory study“) findet in einer kontrollierten Umgebung statt, um den Einfluss umwelt- bzw. untersuchungsbedingter Störvariablen möglichst auszuschließen. Erkauft wird die Kontrolle der Störvariablen im Labor meist durch eine gewisse Künstlichkeit des Untersuchungsortes, was die Übertragbarkeit der Befunde auf den Alltag erschwert.

Feldstudie – Die Feldstudie („field study“) findet im natürlichen Umfeld statt, so dass die Untersuchungsbedingungen den Alltagsbedingungen ähneln und auf diese möglichst gut übertragbar sind. Erkauft wird die Natürlichkeit der Bedingungen durch eine verminderte Kontrolle von Störvariablen, was die kausale Interpretierbarkeit der Befunde erschwert.

Das **Forschungslabor** als eine kontrollierte Untersuchungsumgebung hat den Vorteil, dass umgebungs- bzw. untersuchungsbezogene **Störvariablen** beeinflussbar sind (z. B. Raumtemperatur, Möblierung, anwesende Personen, Geräuschpegel etc.), was die **interne Validität steigert**. Zudem sind viele Forschungslabore mit einem **Einwegspiegel** ausgestattet, so dass das Verhalten der Probanden unbemerkt von außen beobachtet werden kann. Auch **Videokameras** sind in der Regel vorhanden. Der Nachteil von Laborstudien besteht darin, dass die kontrollierte Umgebung die Versuchspersonen in der Weise beeinflussen kann, dass sie alltagsuntypisches Verhalten zeigen. Unter diesem Gesichtspunkt würde z. B. ein im Laborexperiment beobachteter Effekt (z. B. Flirtverhalten im Labor) möglicherweise nicht auf das natürliche Umfeld (z. B. Flirten in der Kneipe) übertragbar sein. Dies entspricht einer **Einschränkung der externen Validität**. Ein weiterer Nachteil der Laborstudie besteht im **Aufwand für die Probanden**, die extra ins Forschungslabor kommen müssen, sowie für die Forschenden, die einen entsprechenden Laborraum und Laborpersonal benötigen (Abb. 7.4).



Abb. 7.4 Nicht nur experimentelle Untersuchungen mit menschlichen Probanden, sondern auch Tierversuche finden meist im Labor statt, was mit entsprechendem Aufwand verbunden ist. (© Stik -Bill Greenhead-/Search ID: bgrn980, Rights Available from CartoonStock.com)

Quantitative wie qualitative Studien finden sowohl im Feld als auch im Labor statt:

- **Quantitative Laborstudien:** Bei Laborstudien, die in einem Forschungslabor stattfinden, handelt es sich meist um **quantitative Studien** – v. a. um **Experimente**, bei denen die Kontrolle der Störvariablen und damit die Steigerung in der internen Validität von großer Bedeutung ist. Der Vorteil der hohen internen Validität von Laborexperimenten wird jedoch erkauft durch reduzierte externe Validität: Je künstlicher die Laborsituation, desto schwieriger lassen sich die Befunde auf Alltagsbedingungen generalisieren. Man spricht auch von **ökologischer Validität**, um auszudrücken, inwiefern eine Laborsituation dem natürlichen Lebensumfeld ähnelt. Dabei ist zu beachten, dass Alltagsnähe einer Untersuchungssituation nicht automatisch die Generalisierbarkeit auf andere Situationen garantiert. So kann die Untersuchung des Vokabellernens im natürlichen Schulkontext so viele Störfaktoren beinhalten (Lehrereffekt, Klassendynamik, Lehrbuch etc.), dass keine auf andere Klassen oder Schulen verallgemeinerbaren Schlussfolgerungen möglich sind. Demgegenüber kann ein gedächtnispsychologisches Experiment, das unter künstlichen Bedingungen im Labor stattfindet und sogar mit sinnlosen Silben arbeitet, verallgemeinerbare Befunde zur Arbeit des Gedächtnisses hervorbringen.

Quantitative Laborstudien sind nicht nur zur Kontrolle von Störvariablen bzw. zur Sicherstellung der internen Validität von Experimenten notwendig, sondern auch immer dann indiziert, wenn die Messung der abhängigen Variablen entsprechende **Apparatu-**

ren und Geräte erfordert (z. B. physiologische Messungen ▶ Abschn. 10.5). **Usability-Tests**, mit denen die Gebrauchstauglichkeit technischer Geräte geprüft wird, finden typischerweise im Labor statt.

- **Qualitative Laborstudien**: Neben quantitativen Laborexperimenten finden manche qualitativen Studien im Labor statt, etwa Fokusgruppen-Diskussionen, deren Verlauf man durch die Einwegscheibe beobachten oder über Videokameras aufzeichnen möchte (zur Fokusgruppen-Diskussion ▶ Abschn. 10.2.2).
- **Qualitative Feldstudien**: Feldstudien sind im qualitativen Forschungsansatz typisch, etwa wenn man Befragungspersonen im häuslichen Umfeld interviewt oder in der ethnografischen Feldforschung über längere Zeit am Alltagsleben der interessierenden Personen und Gruppen teilnimmt und dieses beobachtet (zur ethnografischen Feldforschung ▶ Abschn. 10.1.2).
- **Quantitative Feldstudien**: Auch quantitative Studien mit standardisierten Interviewleitfäden oder Fragebögen finden teilweise im Feld statt, etwa wenn Interviewer von Meinungsforschungsinstituten an der Tür klingeln oder zu Hause anrufen. Wenn Internetnutzende zu Hause am Computer per Online-Fragebogen befragt werden, handelt es sich ebenfalls um eine Feldstudie. Hier möchte man den Befragungspersonen den Aufwand ersparen, das Forschungslabor aufzusuchen. Es sind jedoch neben nicht-experimentellen reinen Befragungsstudien auch experimentelle Studien im Feld realisierbar. Am ehesten umsetzbar sind Feldexperimente als Quasi-Experimente, in deren Rahmen natürliche Gruppen (z. B. Schulklassen) unterschiedlich behandelt werden. Optimal im Sinne externer und interner Validität ist indessen das echte Feldexperiment, das mit randomisierter Gruppenzuteilung arbeitet. Während Feldexperimente, die als Quasi-Experimente durchgeführt werden, oft in authentischen Umgebungen stattfinden und mit den dort vorfindbaren Gruppen arbeiten (in der Schule, im Unternehmen etc.), werden bei Feldexperimenten, die echte Experimente darstellen, die Feldbedingungen oft simuliert. Ein bekanntes Beispiel ist das Standford-Prison-Experiment, bei dem die Gefängnisumgebung in einem Institutskeller der Stanford University simuliert wurde (▶ Studienbeispiel „Das Stanford-Prison-Experiment“). Dass nur beim echten Feldexperiment interne und externe Validität besonders stark ausgeprägt sind, verdeutlicht noch einmal ◘ Tab. 7.5.

Entscheidungskriterien

- Wählen Sie eine **Laborstudie**, wenn Sie a) quantitativ forschen und ein **echtes Experiment** oder einen **Usability-Test** realisieren wollen, die im Feld nicht durchführbar sind (z. B. wegen notwendiger Messgeräte; ▶ Abschn. 7.6.1) oder wenn Sie b) qualitativ forschen und eine **Gruppendiskussion oder Fokusgruppe** durchführen wollen, bei der eine Videoaufzeichnung und Live-Beobachtung durch eine Spiegelscheibe erwünscht sind (zu qualitativen Gruppeninterviews ▶ Abschn. 10.2.2). Soll eine Laborstudie durchgeführt werden, so muss rechtzeitig im Vorfeld ein passendes Labor gefunden und reserviert werden. Zudem muss realistisch geplant werden, wie genügend Untersuchungspersonen angeworben und zum zeit- und kostenintensiven Aufsuchen des Forschungslabors motiviert werden können. Nicht zuletzt wird ausreichendes Personal benötigt, um die Untersuchungsdurchgänge im Labor zu betreuen, wobei mit einem nennenswerten Anteil von Ausfällen und Terminverschiebungen zu rechnen ist.
- Wählen Sie eine **Feldstudie**, wenn Sie a) quantitativ forschen und ein **Quasi-Experiment** unter möglichst natürlichen Bedingungen durchführen wollen (▶ Abschn. 7.6.2) oder wenn Sie b) qualitativ forschen und natürliche Lebensumwelten beobachten wollen (zur **qualitativen Beobachtung** ▶ Abschn. 10.1.2). Je nach dem, welches Untersuchungsfeld wie oft und wie lange aufgesucht werden muss, ist mit entsprechendem Zeit- und Kostenaufwand (Fahrtkosten) für die Forschenden zu rechnen. Sofern es sich beim Feld nicht um einen öffentlich frei zugänglichen Ort handelt (z. B. Verkehrsbeobachtung auf der Straße), sondern um halböffentliche oder geschlossene Orte, sind mit Verantwortlichen vor Ort besondere Absprachen und Vorkehrungen zu treffen, um im Feld unter Beachtung der Forschungsethik sowie wissenschaftlicher Kriterien systematisch empirische Daten erheben zu können (z. B. Unterrichtsbeobachtung in Schule oder Hochschule).

Studienbeispiel

Das Stanford-Prison-Experiment – Ein Feldexperiment

1971 führte der Sozialpsychologe Philip Zimbardo das berühmte **Stanford-Prison-Experiment** durch. Im Rahmen des Experiments wurden $n = 24$ Studenten in zwei Gruppen eingeteilt (die zweifach gestufte unabhängige Variable war die Rolle im Gefängnis: Gefängnisaufseher oder Gefängnisinsasse). Die Versuchspersonen verbrachten mehrere Tage in dem zu einem Gefängnis umgebauten Keller des Forschungsinstituts. Eine genaue Hypothese über die erwarteten Wirkungen lag nicht vor, deswegen wurde das Verhalten der Aufseher und Insassen während des mehrtätigen Versuchsablaufs durchgängig auf Video aufgezeichnet. Dem Videomaterial ließen sich dann nachträglich Wirkungen auf unterschiedliche abhängige Variablen entnehmen.

Das simulierte Gefängnis mit vergitterten Zellen, Gefängniskleidung, Uniformen für die Aufseher etc. macht das Experiment zum **Feldexperiment**. Im Sinne der ökologischen Validität wurden die Versuchspersonen in der Gefängnisinsassenbedingung zu Beginn des Experiments sogar unter den Blicken der Nachbarn von echten Polizisten zu Hause abgeführt. Obwohl alle Beteiligten wussten, dass es sich um ein freiwilliges psychologisches Experiment und keinen Ernstfall handelte, entwickelte sich eine so starke Eigendynamik, dass das eigentlich auf zwei Wochen angelegte Experiment schon nach sechs Tagen abgebrochen werden musste. Ein Teil der Aufseher hatte sadistische Verhaltensweisen an den Tag gelegt, ein Teil der Insassen zeigte extreme Stressbelastung. Dass ganz normale Studenten in der Rolle von Aufsehern nach kurzer Zeit sadistisches Verhalten zeigten, wurde von Zimbardo v. a. dahingehend interpretiert, dass Situationsfaktoren (Gefängnis-Setting) ausreichen, um inhumanes Verhalten heraufzubeschwören und dass dieses eben nicht auf die vermeintlich sadistische Persönlichkeit der Aufseher zurückzuführen ist.

Die Validität dieses Feldexperiments bzw. der daraus abgeleiteten Schlussfolgerungen wurde in verschiedener Hinsicht kritisch diskutiert. So kann man sich fragen, ob ein an jungen US-amerikanischen Männern gezeigter Effekt tatsächlich auf die gesamte Menschheit generalisierbar ist (**Einschränkung der externen Validität**). Es besteht der Verdacht, dass Zimbardo, der selbst die Rolle des Gefängnisdirektors spielte und die Aufseher über mögliche Disziplinierungsmaßnahmen instruierte, dabei möglicherweise aktiv zu einer Eskalation beitrug – er agierte jedenfalls nicht als blinder Versuchsleiter (**Einschränkung der internen Validität**). Ohnehin ist die interne Validität schwer zu beurteilen, da vor dem Experiment keine klare Wirkungshypothese formuliert und die relevanten abhängigen Variablen nicht definiert worden waren. Auch **forschungsethische Bedenken** wegen der starken Belastung für die Versuchspersonen wurden im Nachhinein angemeldet, obwohl das Experiment im Vorfeld von einer Ethikkommission genehmigt worden war.

Dreißig Jahre nach dem Stanford-Prison-Experiment führten die britischen Sozialpsychologen Alex Haslan und Steve Reicher in Kooperation mit der BBC die „**BBC Prison Study**“ durch – ein Feldexperiment in Anlehnung an die Originalstudie, aber ohne deren ethische Probleme, mit theoretischer Basis und mit anderen Befunden.

- **Stanford-Prison-Experiment (Zimbardo, 2001):** http://www.prisonexp.org
- **The BBC Prison Study (Reicher & Haslam, 2006):** http://www.bbcprisonstudy.org

Tabelle 7.5 Interne und externe Validität für (quasi-)experimentelle Labor- und Feldstudien

	Externe Validität: gering	Externe Validität: hoch
Interne Validität: hoch	Laborexperiment	Feldexperiment
Interne Validität: gering	Quasi-experimentelle Laborstudie	Quasi-experimentelle Feldstudie

7.8 Studie mit oder ohne Messwiederholungen

Die meisten empirischen Untersuchungen beschränken sich aus forschungsökonomischen Gründen auf einen einzigen Untersuchungszeitpunkt; dies ist sowohl für Forschende als auch für Untersuchungsteilnehmende mit dem geringsten Aufwand verbunden. Doch Untersuchungsdesigns mit Messwiederholungen, bei denen **dieselben Studienteilnehmenden über mehr oder minder lange Zeiträume hinweg wiederholt untersucht** werden, sind unter bestimmten Umständen vorteilhaft oder sogar notwendig, um das Forschungsproblem zu lösen oder das Untersuchungsdesign effizient zu gestalten (Ellis, 1999).

Wir behandeln im Folgenden zunächst **(quasi-)experimentelle Studien mit und ohne Messwiederholungen**, bei denen es um die Prüfung kurz- oder mittelfristiger Ursache-Wirkungs-Relationen geht. Anschließend geht es um **nicht-experimentelle Studien mit und ohne Messwiederholungen**, bei denen eher langfristige individuelle und soziale Veränderungen in den Blick genommen werden: Querschnitt-, Trend- und Längsschnittstudie lassen sich bei den nicht-experimentellen Studien hinsichtlich des Umgangs mit Messwiederholungen voneinander abgrenzen, wobei Längsschnittstudien noch Spezialformen wie die prospektive Kohortenstudie und die Zeitreihenanalyse aufweisen.

Messwiederholungen sind nicht zuletzt nützlich, um die **Genauigkeit gemessener Werte** zu erhöhen. So erhält man präzisere Angaben über den von Jugendlichen konsumierten Alkohol, wenn man die typische tägliche Alkoholmenge nicht an einem einzigen Messzeitpunkt summarisch erfragt, sondern über ein oder zwei Wochen hinweg wiederholt den täglichen Alkoholkonsum erfasst und daraus den Durchschnittswert bildet (zu **Tage-**

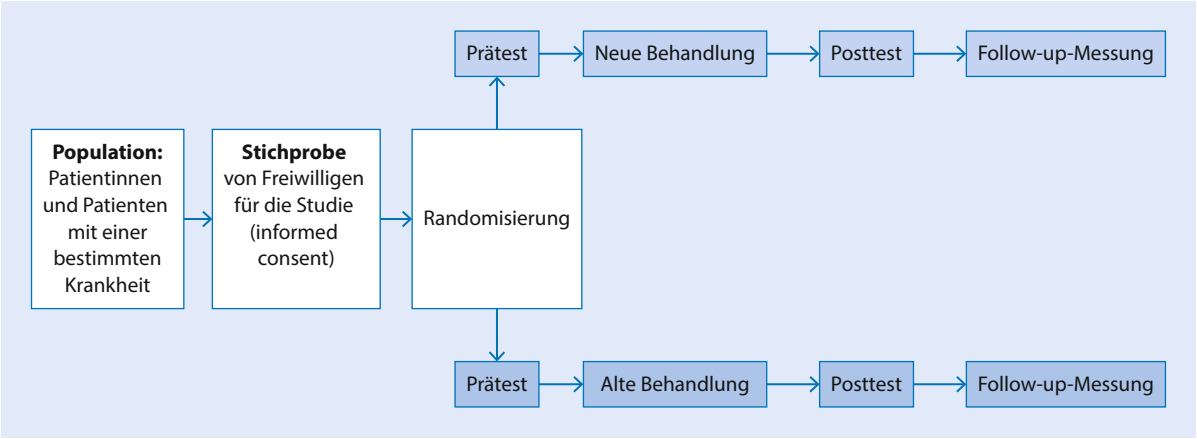


Abb. 7.5 Randomisierte klinische Kontrollgruppenstudie mit Messwiederholungen

buchmethode und ambulantem Assessment ▶ Abschn. 10.3.3). Derartige Vorgehensweisen betreffen aber weniger das gesamte Untersuchungsdesign, sondern eher die **Operationalisierung und Datenerhebung**.

7.8.1 (Quasi-)Experimentelle Studien mit und ohne Messwiederholungen

Wir haben experimentelle und quasi-experimentelle Studien schon als Designs kennengelernt, bei denen die Ursachenfaktoren bewusst durch die Forschenden beeinflusst werden, um das Auftreten der erwarteten Wirkungen zu prüfen (▶ Abschn. 7.6.1 und ▶ Abschn. 7.6.2). Sehr häufig werden (quasi-)experimentelle bzw. (nicht-)randomisierte kontrollierte Studien mit Messwiederholungen im Sinne einer **Prä- und einer Post-Messung** durchgeführt, um die Wirkung einer Intervention zu prüfen. Der Faktor Zeit mit den beiden Ausprägungen „vor und nach der Intervention" fungiert dann im Vorher-Nachher-Design als weitere unabhängige Variable.

Nicht selten wird eine Prätest-Posttest-Messung noch um eine oder mehrere deutlich spätere **Follow-Up-Messung(en)** ergänzt (Abb. 7.5). So kann eine Studie zur Überprüfung der Wirksamkeit einer Depressionstherapie die Depressivität der Patientinnen und Patienten zu Beginn der stationären Behandlung (Prätest), am Ende der stationären Behandlung (Posttest) sowie ein Jahr (1. Follow-Up) und dann noch einmal fünf Jahre später (2. Follow-Up) erheben. Die Ergebnisse vergleicht man mit Messungen an einer unbehandelten Kontrollgruppe. Die statistischen Verfahren zur Auswertung von Versuchsplänen mit Messwiederholungen (▶ Abschn. 12.5.4) weichen von den Auswertungsverfahren für Studien ohne Messwiederholungen leicht ab, da ein Teil der Fehlervarianz entfällt, wenn anstelle von Werten unterschiedlicher Personen Werte derselben Personen in die Berechnung einfließen (dies führt zu einer Steigerung der Teststärke).

Neben Messwiederholungen im Sinne von Prä-Post-Messungen ist es bei manchen Forschungsproblemen darüber hinaus möglich, (quasi-)experimentelle Designs so anzulegen, dass **dieselben Versuchspersonen nacheinander die verschiedenen Experimental- und Kontrollbedingungen** durchlaufen. Sofern dies nicht zu Überforderung, Ermüdung oder verfälschenden Lerneffekten führt, können durch derartige Messwiederholungsdesigns Versuchspersonen eingespart werden. Zudem werden personenbezogene Störvariablen ausgeschlossen, wenn sich in den verschiedenen Untersuchungsgruppen nicht unterschiedliche, sondern jeweils dieselben Personen befinden. Für die statistische Auswertung von Messwiederholungsdesigns sind spezifische Verfahren indiziert, etwa die Varianzanalyse mit Messwiederholungen (▶ Abschn. 12.5.4).

(Quasi-)experimentelle Studie ohne Messwiederholungen – Eine (quasi-)experimentelle Studie ohne Messwiederholungen („independent measures designs", „between-subjects design") verzichtet auf eine Messwiederholung im Sinne einer Prä-Post-Messung und beschränkt sich auf die Nachher-Messung. Dies schränkt die interne Validität der Studie ein, wenn die Vorher-Werte nicht in allen Untersuchungsgruppen identisch waren. Zudem werden in einem Design ohne Messwiederholungen in allen Untersuchungs- und Kontrollgruppen jeweils andere Versuchspersonen untersucht.

(Quasi-)experimentelle Studie mit Messwiederholungen – Eine (quasi-)experimentelle Studie mit Messwiederholungen („repeated measures/measurement designs"; „within-subjects design") e beinhaltet a) eine Prä-Post-Messung zur Prüfung individueller Veränderungen durch die Intervention (was die interne Validität der Studie erhöht) und/oder es werden b) ganz oder teilweise dieselben Versuchspersonen in den unterschiedlichen Experimental- und Kontrollgruppen eingesetzt (was sowohl die Zahl der benötigten Probanden als auch den Einfluss personengebundener Störvariablen reduziert). Die Vorteile von Messwiederholungen sind jedoch mit gewissen Nachteilen und Risiken verbunden, so dass die Einsatzmöglichkeiten von Messwiederholungen in Abhängigkeit von

Forschungsproblem und Rahmenbedingungen bei der Planung des Untersuchungsdesigns abzuwägen sind.

Nachteilig ist, dass Messwiederholungen ihrerseits die Ergebnisse beeinflussen und somit die **interne Validität einer Studie gefährden** können, sei es durch untersuchungsbedingte **Testübung oder Testmüdigkeit** (z. B. wiederholter Einsatz derselben oder ähnlicher Messinstrumente verändert das Ergebnis etwa im Sinne weniger sorgfältig ausgefüllter Depressionsfragebögen) oder durch Faktoren jenseits der Studie wie z. B. **Reifung** (z. B. im Laufe der Zeit bessert sich Depressivität teilweise von allein) oder äußerer **historischer Ereignisse** (z. B. ein Wirtschaftsaufschwung und/oder ein positives nationales Großereignis wie ein WM-Sieg beeinflussen die gemessene Depressivität). Um diese Einflussfaktoren auszuschließen, werden teilweise alternative Designs mit reduzierten Messwiederholungen und dafür zusätzlichen Untersuchungsgruppen eingesetzt (z. B. **Solomon-Vier-gruppen-Plan** anstelle eines einfachen Vorher-Nachher-Designs; ▶ Abschn. 12.5.4), zudem ist eine Kontrollgruppe unabdingbar, um Reifungs- und historische Effekte von den Treatmenteffekten abgrenzen zu können.

Wenn zur Messung von Effekten **Differenzwerte** (z. B. Differenz zwischen Vorher-Nachher-Messung) verwendet werden, ist zu beachten, dass die Differenzwerte stärker von Messfehlern betroffen sein können, wenn sowohl die Vorher- als auch die Nachher-Werte messfehlerbehaftet sind. Problematisch ist die Interpretation von Differenzwerten aber v. a. dann, wenn mit **Extremgruppen** gearbeitet wird. Denn greift man bei der Vorher-Messung selektiv Personen mit besonders hohen oder besonders niedrigen Werten heraus, so verändern sich diese bei der zweiten Messung naturgemäß in Richtung der üblichen Werte – dabei handelt es sich um die sog. **Regression zur Mitte** (bzw. zur Dichte der Verteilung). Bietet man also Personen, die unter besonders starken Rückenschmerzen leiden, ein Rückentraining an und stellt bei der Nachher-Messung eine deutliche Verbesserung der Beschwerden fest, so ist damit zu rechnen, dass zumindest ein Teil der Besserung nicht durch das Training bewirkt wurde, sondern einfach der Regression zur Mitte geschuldet ist (zu „Herausforderungen der Veränderungsmessung in quasi-experimentellen Studien" ▶ Abschn. 12.5.4). Eine unbehandelte und parallelisierte oder gematchte Kontrollgruppe, bei der idealerweise nur die Regression zur Mitte greift, kann dazu dienen, den Treatmenteffekt realistisch einzuschätzen (quasi-experimenteller Zwei-Gruppen-Vorher-Nachher-Plan; nicht-randomisiertes Prä-Post-Kontrollgruppen-Design).

Interventionswirkungen können nicht nur in Gruppenuntersuchungen, sondern auch in **Einzelfallstudien** durch Messwiederholungen geprüft werden (▶ Abschn. 12.5.5).

7.8.2 Nicht-experimentelle Studien mit und ohne Messwiederholungen

Während in experimentellen und quasi-experimentellen Designs die Messwiederholungen dazu dienen, die – meist kurz- oder mittelfristigen – Effekte der experimentellen Manipulation (d. h. des Einflusses der unabhängigen Variablen) nachzuweisen, geht es in **Veränderungsstudien** darum, **mittel- bis langfristigen individuellen und sozialen Wandel** zu untersuchen. Dabei werden meist nicht-experimentelle Designs genutzt, weil die Effekte von 1. **Alter** („age"), 2. **Geburtskohorte**/Geburtsjahrgang/Generation („cohort") oder 3. historischem Erhebungszeitraum/Periode/**Epoche** („period") nicht experimentell variierbar sind. Zudem sind diese drei Erklärungsfaktoren miteinander verknüpft: Zu einem bestimmten Zeitpunkt gehören Personen derselben Altersgruppe automatisch einer bestimmten Geburtskohorte an. Im Zusammenhang mit der Untersuchung mittel- und langfristiger Veränderungen unterscheiden wir Längsschnitt-, Trend- und Querschnittstudien (zur statistischen Prüfung von Veränderungshypothesen ▶ Abschn. 12.5.4).

Querschnittstudie – Bei einer Querschnittstudie („cross-sectional study") wird eine Stichprobe zu einem Zeitpunkt untersucht. Umfasst die Stichprobe mehrere Altersgruppen, so sind Alterseffekte immer nur im Zusammenhang mit Kohorteneffekten interpretierbar.

Trendstudie – Eine Trendstudie („trend study") besteht aus mehreren, in zeitlichem Abstand durchgeführten Querschnittstudien, in denen jeweils (zumindest teilweise) dieselben Variablen erhoben werden. Trendstudien dienen der Untersuchung von gesellschaftlichem Wandel.

Längsschnittstudie – Bei einer Längsschnittstudie bzw. Panelstudie („longitudinal study") wird eine Stichprobe (Panel) über längere Zeit hinweg wiederholt untersucht. Längsschnittstudien dienen der Untersuchung individueller Veränderungen über die Lebensspanne. Diese Alter(n)seffekte sind jedoch mit Kohorten- und Epocheneffekten konfundiert, was man durch komplexere längsschnittliche Designs teilweise auflösen kann.

Querschnittstudie

Eine Stichprobe wird zu **einem Messzeitpunkt** untersucht, dabei können Personen unterschiedlichen Alters untersucht und Altersgruppen verglichen werden. Altersdifferenzen, die sich hierbei zeigen (z. B. Jüngere berichten sehr viel intensivere Internetnutzung als Ältere; ◘ Abb. 7.6; auch Busemann & Gscheidle, 2010), lassen sich kausal nicht eindeutig interpretieren. Die Schlussfolgerung, ältere Menschen seien aufgrund ihres Alters generell weniger technikinteressiert, ignoriert den Kohorteneffekt: Die zum Untersuchungszeitpunkt Älteren gehören einer Geburtskohorte bzw. Generation an, die ohne Internet aufgewachsen ist und somit eine größere Einstiegshürde zu überwinden hatte. Querschnittstudien sind aufgrund ihres geringen forschungsökonomischen

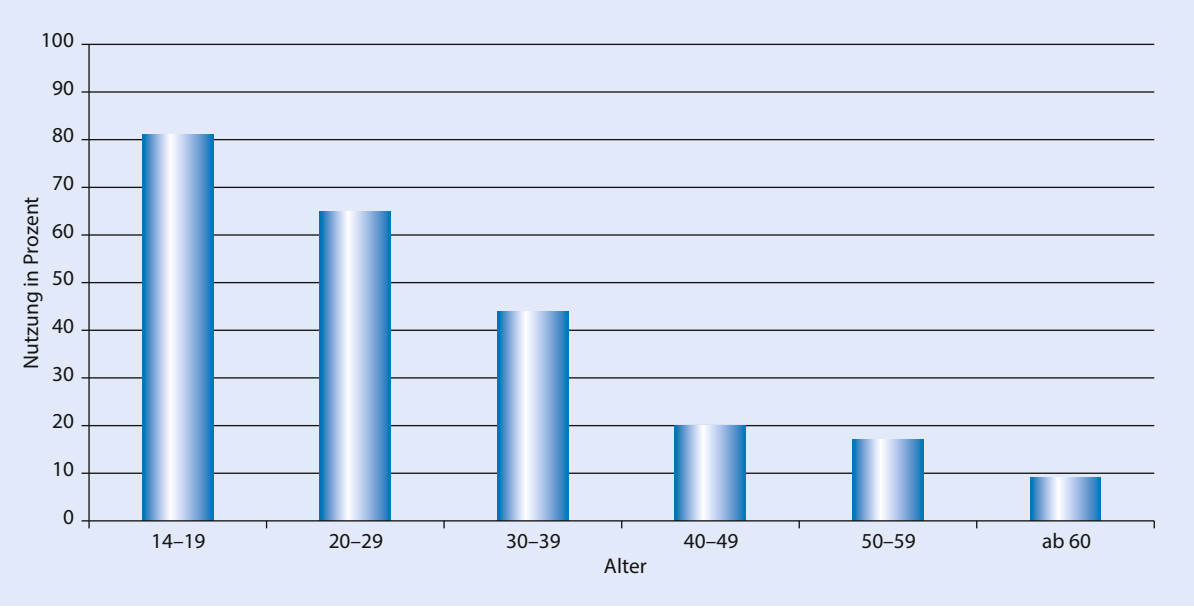


Abb. 7.6 Querschnittstudie im Jahr 2010: Beteiligung *(in Prozent)* an sozialen Netzwerkplattformen und Online-Communitys im Internet in unterschiedlichen Altersgruppen. (Nach Busemann & Gscheidle, 2010, S. 364)

Aufwandes sehr verbreitet, weisen aber eine **Konfundierung von Alters- und Kohorteneffekten** auf und geben im Sinne einer **Momentaufnahme** nur Auskunft über die jeweilige historische Situation zum Zeitpunkt der Datenerhebung. So kann der Anteil von Senioren, die sich an sozialen Netzwerkplattformen im Internet beteiligen, schon zehn Jahre nach der Datenerhebung stark vergrößert haben, oder aber die Nutzung ist in allen Altersgruppen rückläufig, weil soziale Netzwerkplattformen aus der Mode gekommen sind.

Trendstudie

Eine Trendstudie besteht aus mehreren Querschnittstudien, d. h. **mehrere Stichproben** werden in zeitlichem Abstand jeweils zu **einem Messzeitpunkt** mit **demselben Instrument** untersucht. Interessiert man sich für soziale und gesellschaftliche Veränderungen, so kann eine einzelne Querschnittstudie zur Trendstudie erweitert werden, indem sie wiederholt durchgeführt wird; es handelt sich also um **Replikationen der ursprünglichen Erhebung**. Eine Trendstudie erfasst in regelmäßigen oder unregelmäßigen Abständen dieselben Variablen an jeweils einer neuen Stichprobe aus der Grundgesamtheit. Der Fragebogen kann auch minimal verändert oder ergänzt werden, um historisch neue Fragen oder Themen einzubeziehen. Beispiel: In der Öffentlichkeit recht verbreitet ist die Hypothese, dass Jugendliche – v. a. unter dem Einfluss sexualisierter Medienangebote – heutzutage verstärkt schon in sehr jungem Alter sexuell aktiv werden. Eine repräsentative Trendstudie der Bundeszentrale für gesundheitliche Aufklärung (BZgA, 2010) hat in bislang sieben Erhebungen über 30 Jahre hinweg Merkmale der Jugendsexualität erfasst. Betrachtet man die Koituserfahrung in der Gruppe der 14-Jährigen, so zeigt sich, dass diese bei den Jungen in den letzten 30 Jahren bei maximal 10 % lag und zum letzten Erfassungsjahr 2009 wieder sinkt, während sie bei Mädchen bei maximal 12 % lag und ebenfalls rückläufig ist (Abb. 7.7). Für eine Gesamtbetrachtung der Jugendlichen sind dann noch die anderen Altersgruppen einzubeziehen.

Längsschnittstudie

Dieselbe Stichprobe (das Panel) wird in zeitlichem Abstand zu **mehreren Messzeitpunkten** mit **demselben Instrument** untersucht. Nur auf diese Weise können individuelle Veränderungen nachvollzogen werden. Die Prozent- oder Mittelwerte der interessierenden Variablen werden in der grafischen Darstellung bei Längsschnittstudien durch Linien miteinander verbunden, um den Veränderungsverlauf darzustellen (Abb. 7.9). Bei Querschnitt- und Trendstudien dagegen sollten anstelle von Liniendiagrammen lieber Balkendiagramme verwendet werden, um zu unterstreichen, dass es sich um unabhängige Messungen bzw. um Messungen an unterschiedlichen Personen handelt (Abb. 7.6 und Abb. 7.7).

Ein Beispiel für die **unzulässige und irreführende Interpretation von Querschnittdaten** als Veränderungswerten stammt aus der Intelligenzforschung. So wurden die durchschnittlichen Intelligenzwerte unterschiedlicher

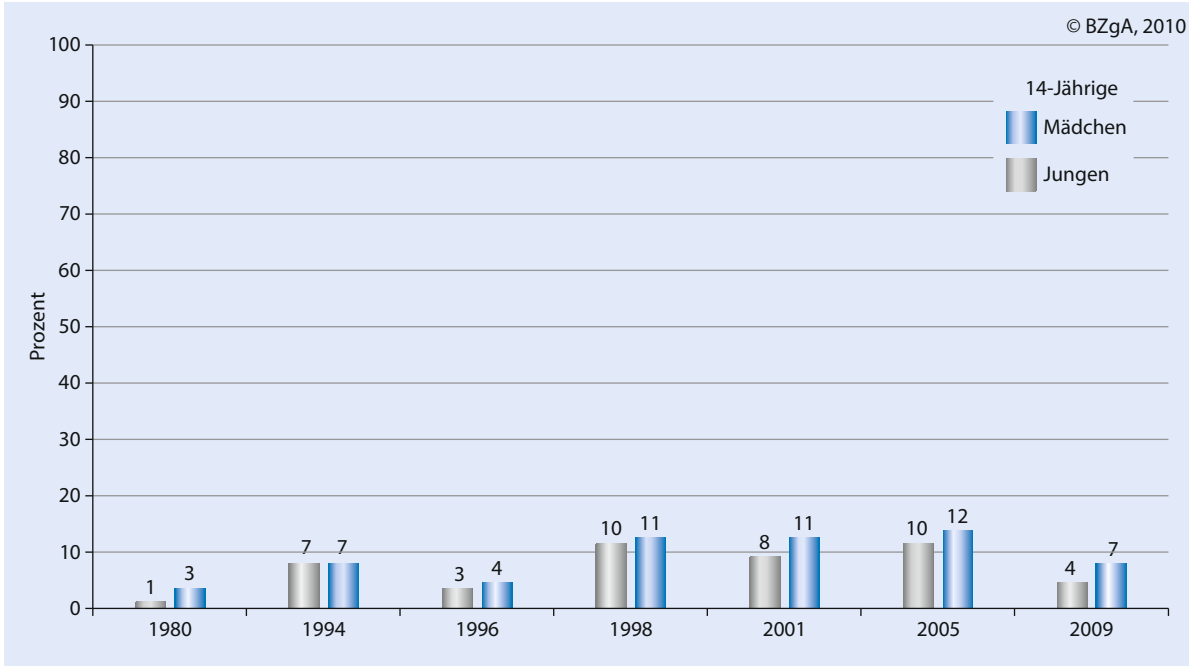


Abb. 7.7 Trendstudie über 30 Jahre: Koituserfahrung in der Altersgruppe der 14-Jährigen von 1980–2009 in Deutschland. (Nach BZgA, 2010, S. 113. Mit freundlicher Genehmigung)

Altersgruppen im Liniendiagramm miteinander verbunden und als individuelle Veränderung interpretiert, dergemäß es im Alter zu deutlichem Intelligenzabbau kommt (Abb. 7.8). Die Querschnittdaten lassen diesen direkten Rückschluss durch die Konfundierung mit dem Kohorteneffekt jedoch nicht zu. Denn bei einer Intelligenztestung in den 1950er-Jahren waren die Jüngeren eben nicht nur jünger, sondern gehörten auch einer Generation an, die oft eine bessere Schulbildung genossen hatte als älteren Studienteilnehmer (zu Theorie und Empirie der Intelligenzentwicklung über die Lebensspanne: Baltes, 1984).

Echte Längsschnittstudien sind aufwändiger als Trendstudien, denn das **Panel** muss sorgfältig gepflegt werden, damit es wirklich gelingt, dieselben Personen nach Jahren und Jahrzehnten immer wieder zu erreichen (Aktualisierung der Adresskartei) sowie immer wieder zur Teilnahme zu motivieren (z. B. Versendung von Briefen zu Feiertagen und Geburtstagen, Incentivierung). Welche Zeitspanne für eine Längsschnittstudie von Interesse ist, hängt von der Forschungsfrage ab. Besonders einschlägig sind Veränderungsmessungen, wenn es um Lernerfolge, Krankheitsverläufe, Intelligenz- oder Persönlichkeitsentwicklung geht. Längsschnittstudien über ein Jahr hinweg existieren ebenso wie Längsschnittstudien, deren Teilnehmerinnen und Teilnehmer über zehn (z. B. Entwicklung von Depressionen bei Jugendlichen: Hankin et al., 1998; Abb. 7.9), zwanzig (z. B. Veränderung von maskulinen und femininen Eigenschaften bei Frauen: Kasen, Chen,

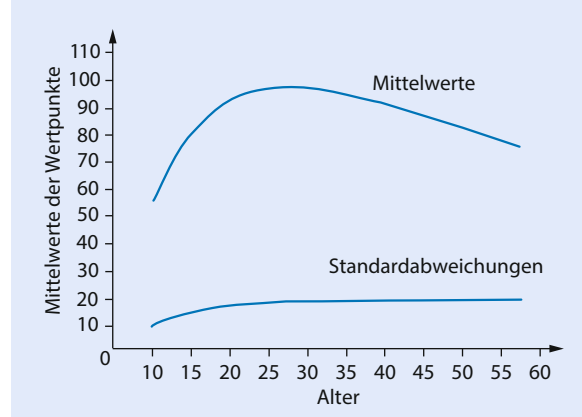


Abb. 7.8 Kurve des Intelligenzanstiegs und -abfalls nach Wechsler (1964), wobei das Liniendiagramm eine Veränderung über die Lebensspanne suggeriert. Tatsächlich handelt es sich aber nicht um Längsschnitt-, sondern um Querschnittdaten, also um Unterschiede zwischen Altersgruppen bzw. Geburtsjahrgängen. (Aus Wechsler, 1964. Mit freundlicher Genehmigung von Huber, Bern)

Sneed, Crawford, & Cohen, 2006), dreißig (Veränderung und Stabilität von Persönlichkeitseigenschaften: Block & Block, 2006) oder vierzig Jahre (Veränderung von Persönlichkeitsmerkmalen: Helson, Jones, & Kwan, 2002) hinweg regelmäßig untersucht werden. Längsschnittstudien bergen nicht nur besondere logistische Herausforderungen (Block & Block, 2006), sondern erfordern auch eine

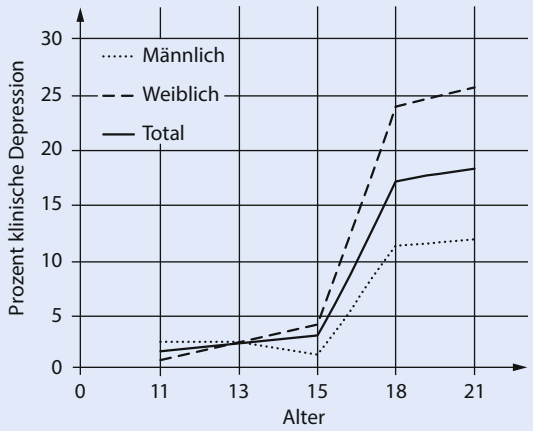


Abb. 7.9 Längsschnittstudie über zehn Jahre: Erstdiagnose klinischer Depression in der Generation 1972 zwischen dem 11. und 21. Lebensjahr. (Aus Hankin et al., 1998, S. 133. Copyright © 2012 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA)

spezielle statistische Behandlung von Veränderungswerten, etwa Varianzanalysen mit Messwiederholungen oder alternative Verfahren (Davis, 2002; Helson et al., 2002; ▶ Abschn. 12.5.4 „Veränderungshypothesen für Entwicklungen: Alters-, Kohorten- und Epocheneffekte“).

Längsschnittstudien sind in der Regel **prospektive (vorausschauende) Studien**, die ausgehend vom ersten Messzeitpunkt die zukünftige Entwicklung erfassen: Wie verläuft die berufliche Entwicklung von Abiturientinnen und Abiturienten mit unterschiedlichen Abschlussnoten? Wie verbessert oder verschlechtert sich der Zustand von Patienten nach einem Klinikaufenthalt? Wie entwickeln sich Kinder nach der Scheidung der Eltern? Will man den Kausaleinfluss eines entsprechenden Ursachenfaktors längsschnittlich untersuchen, so kann eine **prospektive Kohortenstudie** („prospective cohort study“) genutzt werden, bei der eine Gruppe von Exponierten, die dem Ursachenfaktor ausgesetzt sind (z. B. Scheidungskinder) sowie eine vergleichbare Gruppe von Nicht-Exponierten (d. h. hier: nicht von Scheidung betroffene Kinder: Kontrollgruppe) in ihrer weiteren Entwicklung durch Messwiederholungen begleitet werden. In der prospektiven Kohortenstudie interessiert man sich dann dafür, ob, wann und in welcher Weise die erwarteten Wirkungen (hier Scheidungswirkungen) bei den von Scheidung Betroffenen in der Zukunft dann stärker auftreten als bei den Nicht-Betroffenen.

Da prospektive Längsschnittstudien sehr aufwändig sind, weil sich ein Forschungsteam über Jahre oder Jahrzehnte mit der Studie befassen muss, wird zuweilen auch mit der weniger validen, aber forschungsökonomischeren Variante der **retrospektiven (rückschauenden) Längsschnittstudie** gearbeitet. Sie setzt bei Personen bzw. Patienten an, bei denen sich die interessierenden Wirkungen oder Probleme bereits zeigen. Auf der Basis einer Rekonstruktion der Vergangenheit soll eruiert werden, ob, wann und wie die Person früher mit dem vermuteten Ursachenfaktor in Berührung gekommen ist: Lungenkrebspatienten werden nach ihrem früheren Lebensstil und Rauchverhalten befragt, ggf. werden auch ärztliche Dokumente über ihre Krankengeschichte hinzugezogen. Straffällige oder beruflich besonders Erfolgreiche werden zu ihrer Vorgeschichte befragt. Jugendliche, die selbstverletzendes Verhalten zeigen, werden rückblickend um Angaben zu ihren Kindheitserfahrungen und ihrem Umgang mit Selbstverletzungen im Freundeskreis gebeten (Heath, Ross, Toste, Charlebois, & Nedecheva, 2009). Retrospektive Studien haben den Nachteil, dass Gedächtniseffekte, selektive Wahrnehmung und selektives Berichten die Daten verzerren können. Die Aussagekraft retrospektiver Studien steigt, wenn eine Kontrollgruppe einbezogen wird, also z. B. gleichaltrige Jugendliche, die sich nicht selbst verletzen. Sogenannte **Fall-Kontroll-Studien** („case control studies“) folgen einem entsprechenden Design (▶ Abschn. 7.6.3).

Längsschnittstudien werden nicht nur als **quantitative Gruppenstudien** (▶ Studienbeispiel „Quantitative Längsschnittstudie (Gruppenstudie)“), sondern auch als **qualitative Einzelfallstudien** (▶ Studienbeispiel „Qualitative Längsschnittstudie (Einzelfallstudie)“) durchgeführt. In beiden Fällen ist es erforderlich, dass das Forschungsteam sich über Jahre oder gar Jahrzehnte hinweg einer Studie widmet – dieser große forschungsökonomische und logistische Aufwand erklärt, warum Längsschnittstudien trotz ihres hohen Erkenntniswertes in der Praxis vergleichsweise selten realisiert werden.

Studienbeispiel

Quantitative Längsschnittstudie (Gruppenstudie)

Das **sozioökonomische Panel** (SOEP) ist eine groß angelegte bevölkerungsrepräsentative Panelstudie, die in Deutschland bereits **seit 25 Jahren** läuft. Im Auftrag des **Deutschen Instituts für Wirtschaftsforschung** (DIW) werden jedes Jahr über 20 000 Personen aus rund 11 000 bundesdeutschen Haushalten von **TNS Infratest Sozialforschung** befragt. Die Daten geben Auskunft über Einkommen, Erwerbstätigkeit, Bildung und Gesundheit. Weil jedes Jahr **dieselben Personen** befragt werden, können langfristige soziale und gesellschaftliche Veränderungen besonders gut verfolgt werden. Diese von einem großen Forschungsteam betreute Längsschnittstudie erfordert gute Panelpflege inklusive einer regelmäßigen Nachrekrutierung, um ausgeschiedene Panelisten (Umzug ins Ausland, Tod) zu ersetzen. Die Daten werden der Wissenschaft kostenlos für Sekundäranalysen zur Verfügung gestellt: **http://www.diw.de/soep**

Studienbeispiel

Qualitative Längsschnittstudie (Einzelfallstudie)

Im Jahr 2002 nahm der damals 12-jährige „Steven" aus einem sozialen Brennpunktviertel in den USA an einer nachmittäglichen Arbeitsgemeinschaft seiner Schule teil, in der Filme produziert wurden. Sein fertiger Film mit dem Titel **„LEMONADE!!"** stand unter dem Motto „Wenn dir das Leben Zitronen gibt, mach' Limonade daraus" und erzählte von seiner familiären Herkunft: Aufgrund des Drogenkonsums seiner Mutter während der Schwangerschaft kam er als sog. Crackbaby auf die Welt, hatte gesundheitliche Beeinträchtigungen und wurde zur Adoption freigegeben. Doch dank seiner liebevollen Adoptivmutter entwickelte er sich vom kränklichen Baby zum fröhlichen Jungen mit vielen Freunden und Hobbys sowie guten Schulnoten – so zeigt es der von Steven in dem pädagogischen Medienprojekt produzierte autobiografische Film.

Hat der sozial benachteiligte Steven von diesem **pädagogischen Medienprojekt** profitiert? Konnte er seine eigene Geschichte erzählen und wurde gehört? Hat ihn die aktive Medienarbeit in seiner Identität und in seinem Selbstwertgefühl bestärkt? Auf den ersten Blick könnte es so wirken. Doch eine Einzelfallstudie im Längsschnittdesign erbrachte ein ganz anderes Ergebnis.

Die qualitative Studie rekonstruierte, wie der Film im Einzelnen zustande gekommen war, wie er im sozialen Umfeld aufgenommen wurde und welche Bedeutung er fünf Jahre später für den inzwischen 17-jährigen Steven hatte (Nelson, Hull, & Roche-Smith, 2008). Die Datenerhebung umfasste Dokumentenanalysen, Beobachtungen und Interviews. Es zeigte sich unter anderem, dass Steven eigentlich ein Drehbuch für einen viel weniger optimistischen Film geschrieben hatte, die Betreuer jedoch befürchteten, dieser würde in den Augen Außenstehender negative Stereotype über Afro-Amerikaner bestätigen. Also produzierte Steven den von den Betreuern gewünschten optimistischen Film. Dieser wurde auf Elternabenden und von Lehrern sehr positiv aufgenommen. Allerdings machten sich die anderen Schüler anlässlich des Films über Steven als „Crackbaby" sowie als „Streber" lustig. Trotz eines späteren Schulwechsels kursierte der Film bald auch unter seinen neuen Mitschülern und wirkte sich für Steven negativ aus. Diese Langzeitfolgen wären unentdeckt geblieben, hätte man die Studie nicht als Längsschnittuntersuchung angelegt.

Nicht zuletzt können Längsschnittstudien auch mit komplexeren Designs durchgeführt werden, indem z. B. die Entwicklung mehrerer Gruppen aus unterschiedlichen Generationen parallel mitverfolgt wird (sog. Zeitwandelstudien ► Abschn. 12.5.4).

Einen Sonderfall der Längsschnittstudie stellt schließlich noch die **Zeitreihenanalyse** dar. Eine Zeitreihe entsteht, wenn dieselbe Variable sehr oft in gleichen oder unterschiedlichen Zeitabständen gemessen wird. Eine regelmäßige reaktive Datenerhebung mit Untersuchungspersonen ist in den meisten Fällen zu aufwändig und belastend. Zahlreiche **nonreaktive Datenerhebungsmethoden** liefern jedoch Zeitreihendaten: z. B. stündliche Zugriffszahlen auf eine Website registriert über Logfiles, tägliche Börsenkurse eines Unternehmens, jährliche Suizidraten eines Landes gemäß amtlicher Statistik. Zu verschiedenen Zeitreihenhypothesen und dem statistischen Instrumentarium der Zeitreihenanalyse ► Abschn. 12.5.4.

Entscheidungskriterien

- Wählen Sie eine **(quasi-)experimentelle Studie mit Messwiederholungen** im Sinne von Prä-Post-Messungen sowie im Sinne wiederholten Einsatzes derselben Versuchspersonen in verschiedenen Untersuchungsbedingungen, wenn Sie kurzfristige Ursache-Wirkungs-Relationen prüfen und dabei Versuchspersonen einsparen sowie personengebundene Störvariablen noch umfassender ausschließen wollen. Zulässig und sinnvoll ist ein Messwiederholungsdesign allerdings nur, wenn es die Ergebnisse nicht verfälscht (z. B. durch Lerneffekte) und wenn es die Versuchspersonen nicht überfordert. Andernfalls müssen Sie auf ein **(quasi-)experimentelles Design mit wenigen oder keinen Messwiederholungen** ausweichen.
- Planen Sie eine **nicht-experimentelle Studie mit Messwiederholungen**, wenn Sie sich speziell für mittel- und langfristige Veränderungen interessieren. Da Sie in der Regel im Rahmen einer Qualifikationsarbeit nicht über die Ressourcen für eine ggf. jahrelang dauernde prospektive Längsschnittstudie verfügen, können Sie folgende Alternativen erwägen: Sie können anstelle einer Primäranalyse eine Sekundäranalyse vorliegender Längsschnittdaten oder Zeitreihen durchführen. Zudem können Sie eine retrospektive Fall-Kontroll-Studie in Erwägung ziehen. Wenn Sie lediglich eine **nicht-experimentelle Studie ohne Messwiederholungen** – also etwa eine Trendstudie oder eine Querschnittstudie – durchführen, bei denen keinerlei Veränderungen an denselben Personen erfasst werden, können Veränderungshypothesen auch nicht geprüft werden.

7.9 Gruppen- oder Einzelfallstudie

Bei den meisten sozial- und humanwissenschaftlichen empirischen Untersuchungen handelt es sich um **Gruppenstudien**: Dabei wird üblicherweise eine Stichprobe von Untersuchungseinheiten aus der interessierenden Grundgesamtheit bzw. dem interessierenden Forschungs-

feld gebildet und dann mit der Stichprobe gearbeitet (**Stichprobenerhebung**; zur Stichprobenziehung Kap. 9). In Abhängigkeit von Art und Größe der jeweiligen Stichprobe können anhand der Stichprobenergebnisse dann mehr oder minder genaue Rückschlüsse auf die Grundgesamtheit gezogen werden. Handelt es sich bei der interessierenden Grundgesamtheit um eine relativ kleine und heterogene Population (z. B. alle Beschäftigten eines mittelständischen Unternehmens; alle Folgen einer abgeschlossenen Fernsehserie; alle Mitglieder einer Online-Selbsthilfegruppe), so ist es sinnvoll und möglich, sämtliche Populationsmitglieder zu untersuchen (**Vollerhebung**; ▶ Abschn. 9.1.1).

Gruppenstudien haben den Vorteil, dass sie mehrere bis sehr viele Fälle zusammenfassend betrachten und somit von den Eigenheiten eines Einzelfalles abstrahieren. Auf diese Weise werden generalisierbare Erkenntnisse über die Grundgesamtheit bzw. das Untersuchungsfeld gewonnen. Zuweilen ist es jedoch sinnvoll, eben nicht mit statistischen Aggregatwerten wie Mittelwerten oder mit qualitativ zusammengefassten Typen zu arbeiten, sondern tatsächlich einen Einzelfall individuell zu beleuchten. Dann ist eine **Einfallstudie bzw. Fallstudie** („case study“) indiziert (Yin, 2008).

Gruppenstudie – Bei einer Gruppenstudie („group design“) wird eine Stichprobe von Objekten aus der Grundgesamtheit untersucht und zusammenfassend ausgewertet – dies ist sowohl in der qualitativen als auch in der quantitativen Forschung der häufigste Fall (z. B. quantitative Aggregatwerte wie Mittelwerte, Prozentwerte, Korrelationskoeffizienten oder qualitative Aggregate wie Kategorien und Typen). Eine Gruppenstudie kann nicht nur als **Stichprobenstudie**, sondern auch als **Vollerhebung** realisiert werden, indem die gesamte Population untersucht wird, was nur bei kleinen Populationen möglich ist.

Einzelfallstudie – Bei einer Einzelfallstudie bzw. Fallstudie („case study“) wird ein typischer oder untypischer Einzelfall (d. h. eine Person, ein Ereignis, eine Organisation etc.) umfassend untersucht, wobei unterschiedliche Datenerhebungsmethoden zum Einsatz kommen (Beobachtung, Interview, Fragebogen, Dokumentenanalyse etc.). Fallstudien können qualitativ, quantitativ oder nach einem Mixed-Methods-Design angelegt sein. Einen Sonderfall stellen quantitative **Einzelfallstudien mit quasi-experimentellen Versuchsplänen** dar, die durch Messwiederholungen bei einer Versuchsperson umgesetzt werden („single subject design“, „single participant design“).

Bei dem **Einzelfall** kann es sich um eine einzelne Person, aber auch um eine Organisation, eine Veranstaltung, ein politisches Ereignis, eine Freundschaft, eine Unterrichtseinheit etc. handeln. Die Merkmale und Funktionsweisen des individuellen Falles gilt es – auf der Basis spezifischer Forschungsfragen oder Forschungshypothesen – detailliert zu untersuchen. Dabei können unterschiedliche Arten der Datenerhebung (z. B. Beobachtung, Interview, Dokumentenanalyse) und Datenauswertung kombiniert werden. Fallstudien lassen sich **qualitativ, quantitativ oder in einem Mixed-Methods-Design** anlegen (Yin, 2008). Es sind auch Fallstudien möglich, die mehrere separate Einzelfälle analysieren und nebeneinander stellen.

Fallstudien als wissenschaftliche empirische Studien sind abzugrenzen vom didaktischen Einsatz von Fallstudien im Unterricht. Qualitative Fallstudien haben vorwiegend theoriegenerierende, quantitative Fallstudien theorieprüfende Funktion. Darüber hinaus werden Fallstudien auch als Praxistests einer etablierten Theorie oder Methode verstanden (für Hinweise zur Gestaltung von Fallstudien in der Beratungspsychologie oder Managementforschung Cepeda & Martin, 2005; Lowman & Kilburg, 2011). Für Fallstudien kommen sowohl besonders **typische als auch untypische Fälle** infrage (z. B. Fallstudie einer Familie mit vier Suiziden innerhalb von zwei Jahren: Pompili, Lester De Pisa, Del Casale, Tatarelli, & Girardi, 2008; Fallstudie einer Familie mit zwei HIV-positiven Elternteilen: Lowe, 2007)

Einen Sonderfall der quantitativen Fallstudie stellen **quasi-experimentelle Einzelfallstudien** dar, die auch als **„single-participant study“** oder **„single subject study“** bezeichnet werden (Morgan & Morgan, 2001). Bei der quasi-experimentellen Einzelfallstudie geht es darum, den Einfluss einer experimentell variierten unabhängigen Variable auf die abhängige(n) Variable(n) zu prüfen. Dazu wird die abhängige Variable wiederholt unter unterschiedlichen Bedingungen (verschiedene Ausprägungen der unabhängigen Variablen) gemessen. Reagiert die Versuchsperson immer wieder konsistent auf einen bestimmten Stimulus und reagiert bei dessen Eliminierung oder bei einem anderen Stimulus konsistent anders, so kann hier eine Ursache-Wirkungs-Relation postuliert werden. Da dieselbe Person unterschiedlichen Bedingungen ausgesetzt wird, fungiert sie sozusagen als **ihre eigene Kontrollbedingung**. Da bei einer einzelnen Versuchsperson keine Randomisierung vorgenommen werden kann, sind Einzelfallstudien mit experimenteller Manipulation als quasi-experimentell zu kennzeichnen. Typischerweise werden die Ergebnisse von quasi-experimentellen Einzelfallstudien grafisch über Verlaufskurven dargestellt, denen zu entnehmen ist, wie sich über die Zeit und verschiedene experimentelle Variationen hinweg (x-Achse) die Messwerte der abhängigen Variablen verändern (y-Achse). Auch eine statistische Auswertung über Signifikanztests ist möglich, dabei werden die wiederholten Messungen als Stichprobe einer Population von Verhaltensweisen des betreffenden Individuums aufgefasst (zu statistischen Analyseverfahren für Einzelfallhypothesen ▶ Abschn. 12.5.5).

Dass anhand systematischer Versuche mit einzelnen Versuchspersonen oder Versuchstieren allgemeine Ursache-Wirkungs-Prinzipien nachweisbar sind, belegen be-

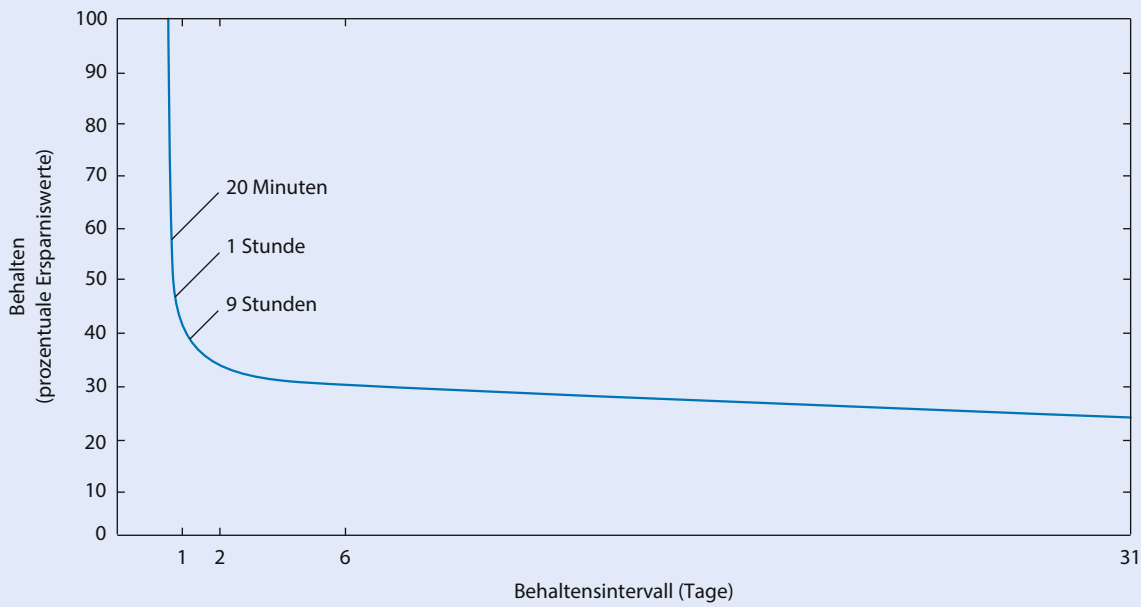
Studienbeispiel

Quasi-experimentelle Einzelfallstudie: Die Ebbinghaus'sche Behaltenskurve

Der Psychologe Hermann Ebbinghaus publizierte 1885 die erste bedeutende Untersuchung, die eine quantitative Messung der Gedächtnisleistung darstellte. Er benutzte sinnlose Silben, die jeweils aus einem Vokal zwischen zwei Konsonanten bestanden (z. B. „MUB"; ◘ Tab. 7.6). Er entschied sich bewusst für **sinnlose Silben**, um eine unverfälschte Messung der Gedächtnisleistung zu erlangen. Denn kurze Wörter, die eine Bedeutung haben (z. B. „ROT"), sind mit Assoziationen behaftet oder bereits gelernt worden.

Ebbinghaus war in dem geschilderten Versuch seine eigene Versuchsperson. Er gab sich selbst eine Liste mit sinnlosen Silben vor und las sich diese zum Lernen wiederholt durch (Wiederholungen). Die Liste galt als gelernt, wenn er alle Silben in der richtigen Abfolge aufsagen konnte. Während des **Behaltensintervalls** lenkte er sich von der ursprünglich gelernten Liste ab, indem er andere Listen lernte. Als Behaltensleistung dokumentierte er, wie viele Versuche er brauchte, um die ursprüngliche Liste nach einer bestimmten Zeit (z. B. nach 2 Minuten, 1 Stunde, 9 Stunden oder 24 Stunden) erneut zu lernen. Mit zunehmenden Durchgängen brauchte er immer weniger Wiederholungen, weil er Informationen aus den vorherigen Lerndurchgängen bereits gespeichert hatte. Diese Verminderung der Lerndurchgänge lässt sich durch sog. **Ersparniswerte** ausdrücken. Diese berechnen sich aus der Differenz der Wiederholungen, geteilt durch die anfänglichen Wiederholungen. Angenommen, Ebbinghaus brauchte im ersten Versuch 20 Wiederholungen, um eine Liste zu lernen, und nach 24 Stunden nur noch 15 Wiederholungen. Die Ersparnis beträgt dann $(20 - 15)/20 = 25\,\%$.

Mit der Ersparnismethode prüfte Ebbinghaus, wie viel des gespeicherten Materials nach unterschiedlichen Behaltensintervallen vergessen wird und ermittelte daraus eine Kurve. Die **Behaltenskurve** nach Ebbinghaus verdeutlicht, dass gleich zu Beginn der Durchgänge sehr deutliche Verluste des zu erinnernden Materials gemessen werden, diese Vergessensrate aber mit zunehmender Anzahl der Durchgänge abflacht (◘ Abb. 7.10).



◘ **Abb. 7.10** Ebbinghaus'sche Behaltenskurve beim Erinnern sinnloser Silben. (Nach Ebbinghaus, 1885, zit. nach Zimbardo & Gerrig, 2004, S. 252)

kannte Beispiele aus der **Kognitions- bzw. Gedächtnispsychologie** (▶ Studienbeispiel „Quasi-experimentelle Einzelfallstudie: Die Ebbinghaus'sche Behaltenskurve") sowie aus der Lerntheorie: **Klassische behavioristische Studien** mit dem **Pavlov'schen Hund** oder mit **Skinners Tauben** basierten jeweils auf zahlreichen Versuchsdurchgängen bei einem einzelnen Versuchstier. Nicht zuletzt sind in der **klinischen Forschung** oder in der **Sportwissenschaft** quantitative Einzelfallstudien verbreitet, um die Wirkung von Therapie- oder Trainingsmaßnahmen auf einzelne Individuen systematisch zu prüfen. Hier sind Gruppenstudien oft gar nicht möglich, da bestimmte Krankheitsbilder singulär auftreten oder bestimmte Trainingspläne individuell auf eine konkrete Spitzensportlerin oder einen bestimmten Spitzensportler zugeschnitten sind.

Tabelle 7.6 Listen sinnloser Silben zur Messung der Gedächtnisleitung. (Aus Hoppe-Graff, 2006, S. 5. *Skript zur Vorlesung Entwicklung und Lernen II*. Unveröff. Manuskript, Leipzig, in Anlehnung an Ebbinghaus, 1885, 1992)

Liste 1	Liste 2
MUB	KUD
VEF	VEP
TUZ	GOF
PEV	FEK
GUB	JAL
JIS	TIZ
FAB	HIF
ZAS	WEF
ROF	SAB
SEG	MOZ
NUF	VEN
WAM	LOP

Entscheidungskriterien

- Wählen Sie eine **Gruppenstudie**, wenn Sie mehrere bis sehr viele Fälle übergreifend analysieren möchten. Gruppenstudien in Form von Stichprobenerhebungen sind in der qualitativen wie quantitativen und Mixed-Methods-Forschung der Normalfall. Ihr Aufwand bemisst sich unter anderem an der Art und Größe der untersuchten Stichproben. Vollerhebungen sind selten und nur bei kleinen Populationen möglich.
- Entscheiden Sie sich für eine **Einzelfallstudie**, wenn Sie Forschungsfragen oder Forschungshypothesen untersuchen möchten, die sich speziell auf einen individuellen Einzelfall beziehen. Einzelfallstudien lassen sich im Rahmen der qualitativen, quantitativen oder Mixed-Methods-Forschung umsetzen. Quantitative Einzelfallstudien in Form von quasi-experimentellen Designs mit Messwiederholungen erfordern spezielle statistische Auswertungsmethoden (► Abschn. 12.5.5).

Lernquiz

1. Kennzeichnen Sie Labor- und Feldstudie hinsichtlich ihrer Validität!
2. Grenzen Sie nicht-experimentelle, quasi-experimentelle und experimentelle Designs voneinander ab!
3. Kennzeichnen Sie Studientypen nach der Anzahl der Untersuchungsteilnehmenden!
4. Was versteht man unter Replikationsstudien, wozu dienen sie und welche Varianten werden unterschieden?
5. Was charakterisiert angewandte Forschung und welche Varianten gibt es?
6. Handelt es sich bei der Metaanalyse um eine theoretische Studie oder eine empirische Studie? Begründen Sie!
7. Was versteht man unter einem Mixed-Methods-Design?
8. Worin unterscheiden sich Trendstudie und Panelstudie?
9. Worin unterscheiden sich explorative, deskriptive und explanative Studien?
10. Was versteht man unter Störvariablen und welche Typen von Störvariablen sind voneinander abzugrenzen?

Übungsaufgaben

1. Das Stanford-Prison-Experiment variierte die unabhängige Variable „Rolle im Gefängnis“ (Gefangener, Aufseher). Die „BBC Prison Study“ (http://www.bbcprisonstudy.org) stellt eine systematische Replikation dieses klassischen Experiments dar. 1.) Wurden in der Replikation weitere unabhängige Variablen berücksichtigt? Wenn ja, welche und warum? 2.) Welches waren die zentralen abhängigen Variablen und wie wurden sie erfasst? 3.) Wie wurde mit Blick auf die Forschungsethik sichergestellt, dass die Versuchspersonen keiner zu großen Beeinträchtigung oder gar Schädigung durch das Experiment ausgesetzt waren?
2. Angenommen, Sie möchten in einer empirischen Studie die Wirkung des Memory-Spielens auf die Gedächtnisleistung von Kindern untersuchen. Skizzieren Sie ein nicht-experimentelles, ein quasi-experimentelles und ein experimentelles Design und entscheiden Sie jeweils, ob Sie eine Querschnitt- oder Längsschnitt-, eine Feld- oder Laborstudie durchführen! Oder wäre am besten eine Fallstudie indiziert? Mit Begründung!

3. Führen Sie ein Experiment zur Behaltensleistung nach Ebbinghaus im Sinne einer quantitativen Einzelfallstudie mit sich selbst als Versuchsperson durch:
   a) Lesen Sie dazu die Silbenliste 1 in Tab. 7.6 in der vorgegebenen Reihenfolge durch, bis Sie alle Silben der Liste in der richtigen Reihenfolge aufsagen können. Stellen Sie für 20 Minuten später den Wecker.
   b) Lernen Sie dann die Liste 2 nach dem gleichen Prinzip bis zum vollständigen Behalten auswendig.
   c) Wenn nach 20 Minuten der Wecker klingelt: Lernen Sie Liste 1 und protokollieren Sie die Anzahl der Lerndurchgänge, die Sie benötigen, bevor Sie alle Silben der ersten Liste wieder in der richtigen Reihenfolge aufsagen können. Stellen Sie für eine Stunde später den Wecker.
   d) Lernen Sie dann erneut die Liste 2 nach dem gleichen Prinzip bis zum vollständigen Behalten auswendig.
   e) Wenn nach einer Stunde der Wecker klingelt: Lernen Sie Liste 1 und protokollieren Sie die Anzahl der Lerndurchgänge, die Sie benötigen, bevor Sie alle Silben der ersten Liste wieder in der richtigen Reihenfolge aufsagen können.
   f) Berechnen Sie die Ersparniswerte für den zweiten und dritten Durchgang.

Weiterführende Literatur

- Creswell, J. W. (2003). *Research design: qualitative, quantitative, and mixed method approaches* (2., illustrierte und kommentierte Aufl.). Thousand Oaks: Sage.
- Kuckartz, U. & Creswell, J. W. (2012). *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: Verlag für Sozialwissenschaften.
- Mayring, P. (2010). Design. In G. Mey & K. Mruck (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (S. 225–237). Wiesbaden: Verlag für Sozialwissenschaften.
- Mitchell, M. & Jolley, J. (2010). *Research Design Explained*. Belmont, CA: Wadsworth.
- Montero, I. & León, O. G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology*, *7*(3), 847–862.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental design for generalized causal inference*. Belmont, CA: Wadsworth.
- Yin, R. K. (2008). *Case study research: design and methods* (4. Aufl.). Thousand Oaks: Sage.

Literatur

Baltes, P. B. (1984). Intelligenz im Alter. *Spektrum der Wissenschaft*, 4, 46–60.

Block, J. & Block, J. H. (2006). Venturing a 30-year longitudinal study. *American Psychologist*, *61*(4), 315–327.

BMBF (Bundesministerium für Bildung und Forschung). (2006). *Macht Mozart schlau? Die Förderung kognitiver Kompetenzen durch Musik*. Abgerufen 8. August, 2013, unter: http://www.bmbf.de/pub/macht_mozart_schlau.pdf

Bösche, W. (2009). Violent content enhances video game performance. *Journal of Media Psychology: Theories, Methods, Applications*, *21*(4), 145–150.

Bryman, A. (2008). *Social research methods* (3rd. ed.). Oxford: Oxford University Press.

Busemann, K. & Gscheidle, C. (2010). Web 2.0: Nutzung steigt – Interesse an aktiver Teilhabe sinkt. Ergebnisse der ARD/ZDF–Onlinestudie. *Media Perspektiven, 7–8*, 359–368.

BZgA (Bundeszentrale für gesundheitliche Aufklärung). (2010). JugendsexualitätRepräsentative Wiederholungsbefragung von 14- bis 17-Jährigen und ihren Eltern – Aktueller Schwerpunkt Migration. Abgerufen 16. August, 2013, unter: http://www.bzga.de/infomaterialien/studien/?uid=0338b2d793e248a3d438fbf95da61d4d&idx=1789.

Cepeda, G. & Martin, D. (2005). A review of case studies publishing in Management Decision 2003–2004: Guidelines and criteria for achieving quality in qualitative research. *Management Decision*, *43*(6), 851–876.

Cook, T. D. & Campbell, D. T. (1976). The design and conduct of true experiments and quasi-experiments in field settings. In M. D. Dunnette (Eds.), *Handbook of industrial and organizational psychology* (pp. 223–326). Skokie, IL: Rand McNally.

Creswell, J. W. & Plano Clark, V. L. (2010). *Designing and conducting mixed methods research*. Thousand Oaks: Sage.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York: Springer.

Deblinger, E., Thakkar–Kolar, R. R., Berry, E. J., & Schroeder, C. M. (2010). Caregivers' efforts to educate their children about child sexual abuse: A replication study. *Child Maltreatment*, *15*(1), 91–100.

Döring, N. (2009). Psychological aspects of interviewing by cellular telephone. In E. Maxl, N. Döring, & A. Wallisch (Eds.), *Mobile market research* (pp. 40–80). Köln: von Halem.

Ebbinghaus, H. (1885). Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie. Leipzig: Duncker & Humblot.

Ebbinghaus, H. (1885, 1992). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie. Neue, unveränd. und ungek. Ausgabe nach der 1. Aufl. 1885*. Darmstadt: WBG.

Ellis, M. V. (1999). Repeated measures designs. *The Counseling Psychologist*, *27*(4), 552–578.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(3), 3–8.

Guo, S. & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications. Advanced quantitative techniques in the social sciences*. Thousands Oaks, CA: Sage.

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R., & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology*, *107*(1), 128–140.

Hannover, B. & Kühnen, U. (2002). „The clothing makes the self" via knowledge activation. *Journal of Applied Social Psychology*, *32*(12), 2513–2525.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234–249.

Heath, N. L., Ross, S., Toste, J. R., Charlebois, A., & Nedecheva, T. (2009). Retrospective analysis of social factors and nonsuicidal self-injury among young adults. *Canadian Journal of Behavioural Science*, *41*(3), 180–186.

Helson, R., Jones, C., & Kwan, V. S. Y. (2002). Personality change over 40 years of adulthood: Hierarchical linear modeling analyses of two longitudinal samples. *Journal of Personality and Social Psychology*, *83*(3), 752–766.

Hoppe-Graff, S. (2006). *Skript zur Vorlesung Entwicklung und Lernen II*.Unpublished manuscript, Leipzig.

Kasen, S., Chen, H., Sneed, J., Crawford, T., & Cohen, P. (2006). Social role and birth cohort influences on gender–linked personality traits in women: A 20-year longitudinal analysis. *Journal of Personality and Social Psychology*, *91*(5), 944–958.

Lowe, W. (2007). „I finally got real parents, and now they're gonna die" – A case study of an adolescent with two HIV-positive parents. *Families, Systems and Health*, *25*(2), 227–233.

Lowman, R. L. & Kilburg, R. R. (2011). Guidelines for case study submissions to Consulting Psychology Journal: Practice and Research. *Consulting Psychology Journal: Practice and Research*, *63*(1), 1–5.

Mayring, P. (2001). Kombination und Integration qualitativer und quantitativer Analyse. *Forum: Qualitative Social Research*, *2*(1), Art. 6. Abgerufen 25. Juni, 2013, unter http://www.qualitativeresearch.net/index.php/fqs/article/viewArticle/967/2110.

Mentzoni, R. A., Brunborg, G. S., Molde, H., Myrseth, H., Mar Skouverøe, K. J., Hetland, J., et al. (2011). Problematic video game use: Estimated prevalence and associations with mental and physical health. Cyberpsychology, Behavior, and Social Networking, 14(10), 591–596.

Mittring, G. & Hussy, W. (2004). Die Ermittlung der kleinsten hinreichend großen Stichprobe bei wissenschaftlichen Experimenten mit Randomisierung. *Kölner Psychologische Studien. Beiträge zur natur–, kultur–, sozialwissenschaftlichen Psychologie, IX*(1), 1–57.

Montero, I. & León, O. G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology*, *7*(3), 847–862.

Moors, A. (2009). Theories of emotion causation: A review. *Cognition and Emotion*, *23*(4), 625–662.

Morgan, D. L. & Morgan, R. K. (2001). Single-participant research design. Bringing science to managed care. *The American Psychologist*, *56*(2), 119–127.

Nelson, M. E., Hull, G., & Roche–Smith, J. (2008). Challenges of Multimedia Self–Presentation: Taking, and Mistaking, the Show on the Road. *Written Communication*, *25*(4), 415–440.

Nicholas, D. B., Globerman, J., Antle, B. J., McNeill, T., & Lach, L. M. (2006). Processes of Metastudy: A Study of Psychosocial Adaptation to Childhood Chronic Health Conditions. *International Journal of Qualitative Methods*, *5*(1). Retrieved February 3, 2012, from http://www.ualberta.ca/~iiqm/backissues/5_1/PDF/NICHOLAS.PDF

Nielsen, S. (2010). Top management team diversity: A review of theories and methodologies. *International Journal of Management Reviews*, *12*(3), 301–316.

Owen, J. & Lindley, L. D. (2010). Therapists' cognitive complexity: Review of theoretical models and development of an integrated approach for training. *Training and Education in Professional Psychology*, *4*(2), 128–137.

Pompili, M., Lester, D., De Pisa, E., Del Casale, A., Tatarelli, R., & Girardi, P. (2008). Surviving the suicides of significant others: A case study. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *29*(1), 45–48.

Preisendörfer, P. (2008). Heikle Fragen in mündlichen Interviews: Ergebnisse einer Methodenstudie im studentischen Milieu. Abgerufen 3. Februar, 2012, unter http://repec.ethz.ch/ets/papers/preisendoerfer_sensitive_questions.pdf

Pruzek, R. M. (2011). Introduction to the special issue on propensity score methods in behavioral research. *Multivariate Behavioral Research*, *46*(3), 389–398.

Quandt, T., Grueninger, H., & Wimmer, J. (2009). The gray haired gaming generation: Findings from an explorative interview study on older computer games. *Games and Culture*, *4*(1), 27–46.

Rauscher, F. H., Shaw, G. L., & KY, K. N. (1993). Music and spatial task performance. *Nature*, *365*(14), 611.

Reicher, S. & Haslam, S. A. (2006). Rethinking the psychology of tyranny: The BBC prison study. *British Journal of Social Psychology*, *45*(1), 1–40.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

Rosenthal, G. (2002). Biographisch-narrative Gesprächsführung: Zu den Bedingungen heilsamen Erzählens im Forschungs- und Beratungskontext. *Psychotherapie und Sozialwissenschaften*, *4*, 204–227.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, *25*(24), 4334–4344.

Stall-Meadows, C. & Hyle, A. (2010). Procedural methodology for a grounded meta-analysis of qualitative case studies. *International Journal of Consumer Studies*, *34*(4), 412–418.

Tang, C. M. (2006). Developmentally sensitive forensic interviewing of preschool children: Some guidelines drawn from basic psychological research. *Criminal Justice Review*, *31*(2), 132–145.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks: Sage.

Thompson, W. F., Husain, G., & Schellenberg, E. G. (2001). Arousal, mood, and the Mozart effect. *Psychological Science*, *12*(3), 248–251.

Thorne, S., Paterson, B., Acorn, S., Canam, C., Joachim, G., & Jillings, C. (2002). Chronic Illness Experience: Insights from a Metastudy. *Qualitative Health Research*, *12*(4), 437–452.

Timulak, L. (2009). Meta-analysis of qualitative studies: A tool for reviewing qualitative research findings in psychotherapy. *Psychotherapy Research*, *19*(4–5), 591–600.

Wagner, K. D., Unger, J. B., Bluthenthal, R. N., Andreeva, V. A., & Pentz, M. A. (2010). Cognitive behavioral theories used to explain injection risk behavior among injection drug users: A review and suggestions for the integration of cognitive and environmental models. *Health Education and Behavior*, *37*(4), 504–532.

Wechsler, D. (1964). *Die Messung der Intelligenz Erwachsener. Textband zum Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE)*. Bern: Huber.

Wilson, T. & Brown, T. (1997). Reexamination of the effect of Mozart's music on spatial task performance. *Journal of Psychology*, *131*(4), 365–370.

Yin, R. K. (2008). *Case study research: Design and methods* (4th ed.). Thousand Oaks: Sage.

Zimbardo, P. G. (2001). The pathology of imprisonment. In J. M. Henslin (Ed.), *Down to earth sociology: Introductory readings* (11th ed.; pp. 272–277): New York, NY, US: Free Press.

Zimbardo, P. G. & Gerrig, R. J. (2004). *Psychologie* (16. Aufl.). München: Pearson Studium.

Operationalisierung

Nicola Döring, Jürgen Bortz

8.1 Theoretische Konzepte in der qualitativen und quantitativen Forschung – 222

8.2 Deduktive Konzeptspezifikation und Variablen – 223

8.3 Grundlagen zur Durchführung einer Operationalisierung – 228

8.4 Messung und die vier Skalenniveaus – 232

8.5 Messung mit Einzelindikatoren – 264

8.6 Messung mit psychometrischen Skalen und Skalenbildung – 267

8.7 Messung mit Indizes und Indexbildung – 277

Literatur – 283

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_8, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Die unterschiedlichen Vorgehensweisen bei der Bildung theoretischer Konzepte in der qualitativen Forschung und der Operationalisierung theoretischer Konzepte in der quantitativen Forschung kennen.
- Wissen, was man unter der Konzeptspezifikation versteht und wie man dabei vorgeht.
- Latente und manifeste Variablen voneinander abgrenzen können.
- Die vier Skalenniveaus definieren und an Beispielen erläutern können.
- Die Ratingskala als Antwortformat von der psychometrischen Skala als Messinstrument unterscheiden können.
- Die Messung theoretischer Konzepte mit Einzelindikatoren, psychometrischen Skalen und Indizes als drei zentralen Operationalisierungsvarianten in ihren Vor- und Nachteilen abwägen können.

Angenommen, wir möchten empirisch die Forschungshypothese prüfen, dass die Nutzung gewalthaltiger Computerspiele die Aggressivität steigert. Als Untersuchungsdesign kommt hierfür eine explanative Studie infrage, also eine theorieprüfende quantitative empirische Primärstudie, die (sofern kein Auftraggeber existiert) der Grundlagenforschung zuzurechnen wäre (zum Untersuchungsdesign ▶ Kap. 7). Die explanative Studie hat dabei den größten kausalen Erklärungswert, wenn sie als echtes Experiment realisiert wird, also wenn die Nutzung gewalthaltiger Computerspiele experimentell variiert wird und die Versuchspersonen den experimentellen Bedingungen zufällig zugeordnet werden, ehe am Ende der Effekt auf die Aggressivität gemessen wird. Entsprechende Laborexperimente sind in der Computerspielforschung recht verbreitet. Sie fokussieren dabei kurzfristige Effekte (zum kausalen Nachweis langfristiger Effekte wären Längsschnittstudien wie z. B. prospektive Kohortenstudien indiziert, ▶ Abschn. 7.8.2).

Bevor wir nun eine Stichprobe von Versuchspersonen ziehen und die Daten im Labor tatsächlich erheben können, müssen wir uns noch ganz genau darüber klar werden, was wir mit den in der Forschungshypothese verwendeten **theoretischen Konzepten** bzw. **Variablen** „Nutzung gewalthaltiger Computerspiele“ und „Aggressivität“ im Einzelnen meinen bzw. wie wir diese Konzepte definieren. In der **quantitativen Forschung** erfolgt die Definition von theoretischen Konzepten grundlegend anders als in der qualitativen Forschung (▶ Abschn. 8.1). Entsprechende inhaltliche Festlegungen theoretischer Konzepte werden als **Konzeptspezifikation** bezeichnet (▶ Abschn. 8.2). An die Konzeptspezifikation schließt sich in der quantitativen Forschung die **Operationalisierung** an, die durch die Auswahl von Indikatoren und Datenerhebungsinstrumenten vorgibt, wie die Konzepte zu messen sind, d. h., wie Objekten mit unterschiedlicher Variablenausprägung aussagekräftige quantitative Messwerte zuzuordnen sind (▶ Abschn. 8.3). Die Messung kann generell auf vier unterschiedlichen **Skalenniveaus** erfolgen und erzeugt dabei unterschiedlich aussagekräftige numerische Daten (▶ Abschn. 8.4). Theoretische Konzepte können über **Einzelindikatoren** gemessen werden (▶ Abschn. 8.5), häufiger werden komplexe Konzepte jedoch durch eine Zusammenfassung mehrerer Indikatoren erfasst, nämlich über sog. psychometrische **Skalen** (▶ Abschn. 8.6) und **Indizes** (▶ Abschn. 8.7). Von standardisierten bzw. psychometrischen Messinstrumenten wird in der quantitativen Forschung erwartet, dass sie die **Gütekriterien** der Objektivität, Reliabilität und Validität erfüllen (zum Qualitätskriterium der methodischen Strenge in der Wissenschaft ▶ Abschn. 3.1.2; zu Gütekriterien von psychometrischen Skalen ▶ Abschn. 10.4.1).

8.1 Theoretische Konzepte in der qualitativen und quantitativen Forschung

Jede Forschungsfrage oder Forschungshypothese enthält **theoretische Konzepte**, **Konstrukte** bzw. **Begriffe** („theoretical construct“, „theoretical concept“), die im Rahmen empirischer Forschung mit der beobachtbaren Erfahrungswirklichkeit verknüpft werden müssen. Dabei geht man in der qualitativen Forschung induktiv, in der quantitativen Forschung dagegen deduktiv vor (Neumann 2003, S. 177f.).

8.1.1 Induktive Bildung theoretischer Konzepte

Auf der Basis relativ offener Forschungsfragen und nichtstandardisierter Erhebungsinstrumente wird in der qualitativen Forschung reichhaltiges nicht-numerisches (z. B. textuelles, audiovisuelles, multimediales) Datenmaterial erhoben und dieses im Rahmen der qualitativen Datenanalyse und weiteren Datenerhebung schrittweise zu theoretischen Konzepten verdichtet. Die Spezifikation der theoretischen Konzepte sowie Angaben zu ihren Relationen untereinander sind somit das **Ergebnis** einer qualitativen Studie (◘ Tab. 8.1). Die Konzeptbildung erfolgt primär induktiv bzw. „bottom-up“, d. h. auf der Basis von empirischen Daten werden durch Abstraktion übergeordnete theoretische Begriffe gebildet (zur induktiven Konzeptbildung in der qualitativen Forschung, speziell im Grounded-Theory-Ansatz Glaser 2002). In der qualitativen Sozialforschung entfällt somit die Arbeitsphase

Tabelle 8.1 Induktive Bildung theoretischer Konzepte in der qualitativen Forschung

3. Theoretische Ebene	Detaillierte Darstellung der theoretischen Konzepte **(Konzeptspezifikation)** und ihrer Relationen untereinander **(Theoriebildung)** als **Ergebnis** einer **qualitativen** Studie
	↑
2. Operationalisierung	**Interpretation und Verdichtung** des reichhaltigen nicht-numerischen Datenmaterials im Rahmen der qualitativen Datenanalyse
	↑
1. Empirische Ebene	Datenerhebung mithilfe **nicht-standardisierter Erhebungsinstrumente** anhand von Forschungsfragen

Tabelle 8.2 Deduktive Definition und Messung theoretischer Konzepte in der quantitativen Forschung

1. Theoretische Ebene	Präzise Definition der theoretischen Konzepte **(Konzeptspezifikation)** und Angaben zu ihren Relationen untereinander in Form von theoretisch abgeleiteten Hypothesen als **Ausgangspunkt** einer **quantitativen** Studie
	↓
2. Operationalisierung	Auswahl vorhandener und/oder Entwicklung eigener **standardisierter Messinstrumente** (z. B. Einzelindikatoren, psychometrische Skalen und Indizes), um die theoretischen Konzepte in messbare Variablen zu überführen
	↓
3. Empirische Ebene	Erhebung von Messwerten für alle Objekte in der Stichprobe und statistische Datenanalyse zu den Relationen der Variablen, auf dieser Basis **Rückschlüsse auf die Gültigkeit der Hypothesen** als **Ergebnis** der **quantitativen** Studie

der Operationalisierung vor der Datengewinnung. Vielmehr verlangt es das qualitative Wissenschaftsverständnis ausdrücklich, **theoretische Offenheit** zu bewahren, damit sich die Bedeutung von theoretischen Konzepten im Zuge der Datenerhebung und Datenanalyse noch verändern bzw. dem Untersuchungsgegenstand annähern kann. Präzise Definitionen von theoretischen Konstrukten und deren Festschreibung in Messinstrumenten durch Operationalisierung werden im qualitativen Paradigma abgelehnt (► Abschn. 2.3.2).

8.1.2 Deduktive Definition theoretischer Konzepte

In der quantitativen Forschung werden die theoretischen Konzepte im Vorfeld der Datenerhebung ganz genau definiert, so dass mithilfe standardisierter Messinstrumente Merkmalsausprägungen in Form von sinnvoll interpretierbaren numerischen Messwerten erhoben werden können. Es ist also vor allem in der quantitativen Forschung von großer Bedeutung, bereits **vor** der Datenerhebung und Datenanalyse alle theoretischen Konzepte präzise auszuarbeiten und die standardisierten Messverfahren für die einzelnen Variablen detailliert festzulegen (Tab. 8.2). Die Operationalisierung in der quantitativen Forschung umfasst sowohl die bewusste Auswahl vorhandener als auch die Entwicklung eigener standardisierter Messinstrumente (z. B. Beobachtungspläne, Fragebögen). Die Konzeptbildung erfolgt deduktiv bzw. „top-down“: Aus Theorien werden Hinweise darauf entnommen, welche Konzepte wie zu verstehen und zu messen sind, um mithilfe der entsprechenden Messinstrumente empirische Daten zu gewinnen. Das vorliegende Kapitel zur Operationalisierung ist somit nur für die quantitative sowie die Mixed-Methods-Forschung relevant.

8.2 Deduktive Konzeptspezifikation und Variablen

Quantitative empirische Sozialforschung strebt auf der Basis der Wissenschaftstheorie des Kritischen Rationalismus (► Abschn. 2.2) eine Analyse der sozialen Wirklichkeit an, bei der **einzelne Merkmale** der Erfahrungswirklichkeit und deren Relationen untereinander genau betrachtet werden. Es wird also nicht der „ganze Mensch“, das „komplette Unternehmen“ oder eine „vollständige Gesellschaft“ untersucht. Denn dies wäre für eine einzelne empirische Studie viel zu komplex und letztlich

nicht leistbar. Indem man theoriegestützt und systematisch jeweils Ausschnitte der Erfahrungswirklichkeit analysiert, soll gemäß kritischem Rationalismus ein wissenschaftlicher Erkenntnisfortschritt möglich sein. Dabei interessieren generell nur Merkmale, die unterschiedliche Ausprägungen annehmen können **(Variablen)**. Merkmale, die bei allen Untersuchungsobjekten unveränderlich sind bzw. völlig identisch ausgeprägt sind **(Konstanten)**, bieten keine Grundlage für Forschungsaktivitäten. Hinsichtlich der unterschiedlichen Ausprägungen von Variablen unterscheiden wir in der quantitativen Forschung

- **stetige** bzw. **kontinuierliche Variablen** („continuous variable“) mit unendlich vielen, beliebig fein abgestuften Ausprägungen (z. B. Reaktionszeit, Intelligenz) und
- **diskrete** bzw. **diskontinuierliche Variablen** („discrete variables“, „discontinuous variables“), die nur wenige verschiedene Werte annehmen können (z. B. Kinderzahl, Automarke, Parteipräferenz).

Variable – Die Variable („variable“) ist ein Merkmal, das verschiedene Ausprägungen annehmen kann. Man unterscheidet **stetige Variablen** („continuous variable“) mit unendlich vielen, fein abgestuften Ausprägungen von **diskreten Variablen** („discrete variable“) mit einer kleinen und begrenzten Zahl an Ausprägungen.

Welche Merkmale bzw. Variablen im Rahmen einer empirischen Studie betrachtet werden, hängt vom Forschungsproblem ab. In dem aus dem Forschungsproblem abgeleiteten Satz an Forschungsfragen bzw. Forschungshypothesen (► Abschn. 5.2) werden jeweils alle relevanten Merkmale genannt. Dabei lassen sich hinsichtlich Messbarkeit und theoretischem Gehalt zwei Typen von Merkmalen bzw. Variablen differenzieren:

1. **Manifeste Merkmale** („manifest variables“, „observable variables“) sind in ihrer Ausprägung unmittelbar beobachtbar. So muss ein Interviewpartner im persönlichen Gespräch nicht erst ausdrücklich danach gefragt werden, ob er im Rollstuhl sitzt oder eine weiße Hautfarbe hat – entsprechende Merkmale sind meistens offensichtlich. Auch Merkmale wie Körpergröße oder Lebensalter, die z. B. im Ausweis registriert sind, lassen sich den manifesten Variablen zuordnen. Bei manifesten Variablen ist es nicht nur sehr einfach, die Ausprägung festzustellen bzw. zu messen, sondern sie sind in der Regel auch theoretisch nicht sehr kompliziert. In einer Studie, in der manifeste Variablen wie „Alter“, „Familienstand“ oder „Wohnort“ erfasst werden, wird man kaum ausführliche theoretische Erläuterungen dazu finden, was mit „Alter“ oder „Wohnort“ gemeint ist, denn die Bedeutung dieser einfachen manifesten Variablen kann in der Regel als bekannt und unstrittig vorausgesetzt werden.
2. **Latente Merkmale** („latent variables“) sind im Gegensatz zu manifesten Variablen nicht direkt beobachtbar und zudem in ihrer theoretischen Bedeutung oft kompliziert und erklärungsbedürftig. Bevor man überhaupt messen kann, welche Ausprägung von „Intelligenz“, „Aggressivität“, „Religiosität“, „Ausländerfeindlichkeit“, „Eifersucht“ oder „Ehrgeiz“ eine Person aufweist, muss erst theoretisch geklärt werden, was im Einzelnen unter diesen Konzepten verstanden werden soll.

Manifestes Merkmal bzw. manifeste Variable – Bei einem **manifesten Merkmal** bzw. einer **manifesten Variable** („manifest variable“, „observable variable“) sind die Ausprägungen direkt beobachtbar bzw. leicht feststellbar und die theoretische Bedeutung gilt als eindeutig und bekannt.

Latentes Merkmal bzw. latente Variable – Bei einem **latenten Merkmal** bzw. einer **latenten Variablen** („latent variable“) bzw. einem **theoretischen Konzept/Konstrukt** („theoretical concept“, „theoretical construct“) sind die Ausprägungen nicht direkt beobachtbar und die theoretische Bedeutung ist erklärungsbedürftig.

Die besonderen Anstrengungen der präzisen Konzeptspezifikation und anschließenden Operationalisierung beziehen sich in der quantitativen Forschung primär auf latente Merkmale bzw. theoretische Konzepte. Zuweilen kann es aber auch sinnvoll und notwendig sein, vermeintlich einfache manifeste Variablen (z. B. „Familienstand“ oder „Geschlecht“) theoretisch genauer zu betrachten und diese theoretischen Überlegungen bei der Operationalisierung zu berücksichtigen (► Abschn. 8.5.1).

8.2.1 Nominaldefinition von theoretischen Konzepten

In einer quantitativen empirischen Studie müssen alle interessierenden theoretischen Konzepte spezifiziert bzw. alle latenten Merkmale präzise definiert werden. Dabei handelt es sich jeweils um sog. Nominaldefinitionen (Kromrey 2000a, S. 163; Schnell, Hill, & Esser, 2008, S. 128). Bei einer **Nominaldefinition** wird der zu definierende Begriff (Definiendum) durch einen oder mehrere bereits bekannte Begriffe (Definiens) festgelegt. Beispiel: Organisationale Gerechtigkeit (Definiendum) meint die Wahrnehmung der Beschäftigten von der Fairness am Arbeitsplatz (Definiens; zur Entwicklung von Definitionen der organisationalen Gerechtigkeit Greenberg 1990).

Voraussetzung für Nominaldefinitionen ist also ein Bestand an bereits eingeführten bekannten Begriffen, die als nicht weiter erklärungsbedürftig gelten und somit zur Präzisierung des zu definierenden Konzeptes verwendbar sind. Nominaldefinitionen sind nichts anderes als begriffliche Umformulierungen. Sie sollen die intersubjektive

Nachvollziehbarkeit der verwendeten Konzepte sicherstellen. Die Präzisierung eines theoretisches Begriffs kann auch als **Explikation** bezeichnet und von der Definition insofern abgegrenzt werden, als die Definition im engeren Sinne auf die Klärung des Sprachgebrauchs abzielt, während die Explikation theoretische Festlegungen vornimmt.

Welche unterschiedlichen Arten von Definitionen neben der Nominaldefinition existieren (z. B. Realdefinition) oder wie Begriffsdefinitionen und Explikationen formalisiert und analysiert werden können, wird im Rahmen der Sprachphilosophie, der Logik sowie der Wissenschaftstheorie detaillierter behandelt. Entsprechende Details sind für die empirische Forschungspraxis jedoch von untergeordneter Bedeutung und werden hier nicht vertieft (zur Wissenschaftstheorie ► Kap. 2).

Konzeptspezifikation – Unter Konzeptspezifikation („conzeptualization", „concept specification") versteht man die Nominaldefinition bzw. Explikation eines theoretischen Konzeptes, Konstruktes bzw. latenten Merkmals. Bei komplexen Merkmalen nennt die Definition theoriebasiert in der Regel noch verschiedene Subdimensionen oder Aspekte des Konzeptes.

Man unterscheidet bei theoretischen Konzepten die intensionale Bedeutung **(Begriffsinhalt)** und die extensionale Bedeutung **(Begriffsumfang)** und somit **zwei Varianten von Nominaldefinitionen**:

1. Soll definiert werden, was unter dem Konzept der „Gewalthaltigkeit" eines Computerspiels zu verstehen ist, so sind im Rahmen einer **intensionalen Nominaldefinition** eine Reihe von Eigenschaften zu nennen, anhand derer man gewalthaltige von weniger oder gar nicht gewalthaltigen Spielen abgrenzen kann. Entsprechende Merkmale könnten z. B. sein: a) dass die Spielerin oder der Spieler im Rahmen des Spiels selbst Gewalt ausüben kann, b) dass man im Zuge der Gewalthandlung im Spiel andere Spieler töten kann und c) dass die Tötungshandlungen im Spiel für den Spielerfolg notwendig sind. Die begründete Auswahl dieser Merkmale entspricht dem Begriffsinhalt. Sofern ein Spiel genau diese Merkmale erfüllt, gilt es als gewalthaltiges Spiel. Bei der Konzeptspezifikation greift man meistens auf intensionale Nominaldefinitionen zurück.
2. In manchen Fällen ist es jedoch schwierig, den Begriffsinhalt vollständig anzugeben. Dann wird im Sinne einer **extensionalen Nominaldefinition** eine Liste von Beispielen angegeben, um das interessierende Konzept zu definieren. So würde man gewalthaltige Computerspiele laut extensionaler Nominaldefinition spezifizieren als Spiele in der Art wie „Counterstrike", „Battlefield", oder „Call of Duty". Ein Spiel, das den genannten Beispielen ähnelt, würde somit als gewalthaltiges Spiel gelten. Extensionale Definitionen haben den Vorteil, dass sie leicht generierbar und sehr anschaulich sind, sie haben im Vergleich zu intensionalen Definitionen jedoch den Nachteil geringerer Präzision.

Studienbeispiel

Nominaldefinition von Mobbing („bullying")

In einem von der American Psychological Association (APA) herausgegebenen Sammelband zur Gewaltprävention befindet sich auch ein Beitrag zum Mobbing an Schulen. Dabei wird auf der Basis vorliegender Literatur das Mobbing – im Englischen „bullying" – als eine Spezialform der Aggression eingeordnet mit folgender intensionaler Nominaldefinition (Orpinas & Horne, 2006):

„*Bullying is a type of aggression in which the bully (or the aggressor) is stronger or more powerful than the victim, and the aggressive behaviors are committed intentionally, repeatedly, and over time.*" (Orpinas & Horne, 2006, S. 148)

Zur Erläuterung dieser Definition wird ein Beispiel für Nichtmobbing sowie eines für Mobbing beigefügt:

„*Therefore, a fight among gang members of equal power would be aggressive but not bullying, whereas pushing, teasing, or threatening a child who is younger, smaller, or somehow weaker would be bullying if the behavior is repeated over time.*" (Orpinas & Horne, 2006, S. 148)

Nominaldefinitionen sind Festlegungen, sie können nicht als wahr oder falsch überprüft werden, sondern sind daran zu messen, ob sie für die Forschung **fruchtbar** sind. Dies setzt voraus, dass verwendete Definitionen sich am bisherigen Stand der Forschung und Theorieentwicklung orientieren. Im einfachsten Fall greift man eine in der Literatur häufig verwendete und brauchbar erscheinende Definition heraus und stützt sich auf diese (► Studienbeispiel „Nominaldefinition von Mobbing").

Vielfach finden sich in der Literatur auch Studien, die auf eine theoretische Konzeptspezifikation ganz verzichten und unmittelbar die für die relevanten latenten Merkmale verwendeten Messinstrumente nennen (für Studien zur kurzfristigen Wirkung von Computerspielen auf Aggression ohne Konzeptspezifikation von Aggression siehe z. B. Anderson et al., 2010; Ferguson & Rueda, 2010). Dieses Vorgehen ist insofern nachteilig, als ohne ausdrückliche Konzeptspezifikation der Bezug zwischen den mit den genannten Messinstrumenten erhobenen Daten einerseits und ihrer übergreifendem theoretischen und praktischen Bedeutung andererseits nicht ausdrücklich hergestellt wird. Liest man etwa, dass eine Studie Aggression mittels der Lautstärke von Bestrafungstönen misst, die eine Versuchsperson einem vermeintlichen Spielpartner im Labor appliziert, so muss man selbst den Rück-

Studienbeispiel

Bedeutungsanalyse von Aggression

In einem Forschungsüberblick sollen die vorliegenden Befunde zu den wechselseitigen Beziehungen zwischen Aggression, Feindseligkeit und Ärger dargestellt werden (Ramírez & Andreu, 2009). Die Autoren referieren dazu über drei Seiten hinweg verschiedene Definitionen von Aggression. Dabei wird die „Absicht, einem anderen Lebewesen zu schaden", als **allgemein konsensfähige Definition von Aggression** herausgearbeitet und mit diversen Literaturquellen belegt. Des Weiteren werden diverse Typen von Aggression auf unterschiedlichen Dimensionen unterschieden:

- Art der Absicht bzw. Zielsetzung: **feindselige Aggression**, bei der die unmittelbare Schädigung Selbstzweck ist (z. B. Schläge im Rahmen von Hass), vs. **instrumentelle Aggression**, bei der die unmittelbare Schädigung Mittel für ein übergeordnetes Ziel ist (z. B. Schläge im Rahmen eines Raubüberfalls).
- Mittel der Schädigung: **physische Aggression**, bei der eine andere Person körperlich angegriffen wird, vs. **verbale/soziale/relationale Aggression**, bei der eine andere Person verbal bzw. im Kontext des Sozialverhaltens angegriffen wird (z. B. Beschimpfungen).
- Aktivität des Aggressors: **aktive Aggression**, bei der die Schädigung durch aktives Handeln erzeugt wird, vs. **passive Aggression**, bei der die Schädigung durch Nichthandeln hervorgerufen wird (z. B. nicht grüßen, nicht zuhören).
- Unmittelbarkeit des schädigenden Verhaltens: **direkte Aggression**, bei der die Schädigung im direkten Face-to-Face-Kontakt auftritt, vs. **indirekte Aggression**, bei der die Schädigung nicht im direkten Kontakt erfolgt (z. B. heimlich Eigentum des Opfers zerstören, hinter dem Rücken der betroffenen Person Gerüchte verbreiten).

Die Autoren konzentrieren sich auf die erstgenannte Dimension (Art der Absicht) und führen genauer aus, warum diese Dimension für das Verständnis von Aggression so wichtig ist, was feindselige vs. instrumentelle Aggression laut vorliegenden Definitionen im Einzelnen kennzeichnet und welche synonymen Bezeichnungen (z. B. reaktive vs. proaktive Aggression) dafür in der Literatur verwendet werden.

schluss ziehen, welche Art von Aggression hier eigentlich gemessen wird und inwiefern diese überhaupt alltagsrelevant ist (► Studienbeispiel „Bedeutungsanalyse von Aggression").

8.2.2 Bedeutungsanalyse von etablierten theoretischen Konzepten

Bei theoretisch umstrittenen Konzepten, für die mehrere unterschiedliche Definitionen vorliegen, muss eine **Bedeutungsanalyse** bzw. **semantische Analyse** der vorliegenden Begriffsverwendungen und Definitionen vorgenommen und eine begründete Auswahl getroffen werden (für eine ausführliche Darstellung siehe Kromrey 2000b, S. 130ff.). Hierfür greift man auf die existierende Literatur zurück, die im jeweiligen Theorieteil meistens Konzeptualisierungsfragen diskutiert (► Studienbeispiel „Bedeutungsanalyse von Aggression"). Zudem existieren Publikationen, die sich im Sinne einer Bedeutungsanalyse ausschließlich der Darstellung, dem Vergleich und der Integration unterschiedlicher Konzeptualisierungen eines Phänomens in der wissenschaftlichen Literatur widmen (z. B. Konzeptualisierung von „Mobbing am Arbeitsplatz": Crawshaw 2009; Konzeptualisierung von „Agression im Straßenverkehr": Galovski, Malta, & Blanchard, 2006).

Deskriptive und explanative quantitative Studien sind nur sinnvoll, wenn eine entsprechende Forschungs- und Theoriebasis existiert, so dass man auf vorhandene Definitionen zurückgreifen kann. Für die eigene Studie wird man entweder die einschlägigste Definition herausgreifen und übernehmen/anpassen oder zunächst mittels Bedeutungsanalyse alle vorliegenden Definitionsvorschläge analysieren und auf dieser Basis eine Auswahl treffen oder eine Modifikation vornehmen.

8.2.3 Dimensionale Analyse von neuen theoretischen Konzepten

Es sind vor allem explorative Studien, die sich neuen oder bislang wenig untersuchten Gegenständen zuwenden, die es erforderlich machen, eine eigene Konzeptspezifikation zu entwickeln und zu diesem Zweck eine **Dimensionsanalyse** bzw. **dimensionale Analyse** durchzuführen. Während die Bedeutungsanalyse bei den bereits vorliegenden Definitionen und Begriffsverwendungen in der Literatur ansetzt und diese vergleichend beleuchtet, betrachtet die Dimensionsanalyse direkt den **Gegenstand** selbst und versucht anhand von konkreten Beispielen dessen zentrale Merkmale und Aspekte möglichst systematisch herauszuarbeiten.

Diese angestrebte begriffliche Strukturierung erfolgt im Rahmen der dimensionalen Analyse in vier Schritten (Kromrey 2000b, S. 116ff.):

1. **Ideen- und Materialsammlung**: Für eine möglichst breite Sammlung von Ideen und Materialien zum Gegenstand ist auf verschiedene Quellen zurückzugreifen: eigene Alltagserfahrung, Konsultation von Experten, Presseberichte, Fachliteratur zu verwandten Themen, Brainstorming mit Kollegen und vor allem die Betrachtung von Beispielen. Will man z. B. ein Konzept wie „Handykompetenz von Kindern" einer dimensionalen Analyse unterziehen, um zu einer genauen Konzeptspezifikation mit allen relevanten Unteraspekten zu kommen, so kann man auf allgemeine Definitionen von „ Medienkompetenz" zurückgreifen,

Studienbeispiel

Konzeptspezifikation von „Globalisierung" mittels dimensionaler Analyse

Eine politikwissenschaftliche Methodenstudie widmet sich der Konzeptspezifikation von „**Globalisierung**", einem Begriff, der bekanntlich unser Zeitalter prägt, dessen genaue Bedeutung jedoch schwer greifbar ist. Auf der Basis vorliegender Theorien und Definitionsvorschläge **(Bedeutungsanalyse)** präsentiert der Autor zunächst eine **allgemeine Nominaldefinition**: „Globalisierung bezeichnet Prozesse der Zunahme sowie der geographischen Ausdehnung grenzüberschreitender anthropogener/sozialer Interaktion." (Kessler 2009).

Das Konzept der Globalisierung wird dabei als Variable eines Nationalstaates betrachtet, d. h., verschiedene Länder bzw. Staaten unterscheiden sich im Grad ihrer Globalisierung. Die auf der Basis der Bedeutungsanalyse herausgearbeitete Nominaldefinition ist für eine Messung noch zu allgemein, so dass mittels einer **dimensionalen Analyse** die **Unterdimensionen des Konstruktes** zu spezifizieren sind. Im Zuge der Dimensionsanalyse wurde hierbei die „grenzüberschreitende soziale Interaktion" in den Mittelpunkt gerückt, deren Menge und Ausdehnung zunehmen soll. Zur Systematisierung wird ein einfaches Kausalmodell zugrunde gelegt, dementsprechend werden Voraussetzungen, Arten und Folgen grenzüberschreitender Interaktionen differenziert. Die gesammelten Beispiele für Unterdimensionen werden in dieses Schema eingeordnet. Eine solche Konzeptspezifikation durch dimensionale Analyse hilft der weiteren Theoriebildung sowie der Planung empirischer Untersuchungen, weil nun ganz gezielt einzelne Globalisierungsaspekte herausgegriffen und gemessen werden können und man gleichzeitig darstellen kann, welche Globalisierungsaspekte in der jeweiligen Studie nicht untersucht werden (Tab. 8.3).

Für eine konkrete **Messung des Globalisierungsgrades** verschiedener Staaten kommt die Bildung eines **Globalisierungsindex** infrage. Dazu müssen auf der Basis der dimensionalen Analyse Anzahl und Art der relevanten Indikatoren ausgewählt werden. Zudem muss inhaltlich begründet und entschieden werden, welcher Indextyp (z. B. gewichteter additiver Index, ungewichteter multiplikativer Index) zum Einsatz kommen soll (zur Indexbildung ► Abschn. 8.7).

Tabelle 8.3 Ergebnis einer dimensionalen Analyse des Konzepts „ Globalisierung". (Mod. nach Kessler 2009, S. 38)

Voraussetzungen für grenzüberschreitende soziale Interaktionen	→	**Arten von grenzüberschreitenden sozialen Interaktionen** (über nationalstaatliche Grenzen und wachsende Entfernung hinweg)	→	**Folgen** von grenzüberschreitenden sozialen Interaktionen
Beispiele:		Beispiele:		Beispiele:
– Abbau von Handelsbeschränkungen – Abbau von Kapitalverkehrsbeschränkungen – Liberale politische Systeme – Fremdsprachenkenntnisse – Abbau von Einreise- und Aufenthaltsbeschränkungen – Verfügbarkeit von Transportmitteln, Infrastruktur und Know-how – Verfügbarkeit von Computer-Hard- und Software, Internet, PC-Kenntnissen – Verfügbarkeit von Telefon, Telefax, Fernsehen, Radio	→ ←	– Zunahme der Handelsströme – Zunahme der Direkt- und Portfolioinvestitionen – Wachsender Austausch von internationalen Organisationen, Politikern, Wissenschaftlern – Zunahme von Touristenströmen, des Personenverkehrs – Zunahme des Gütertransports – Zunahme der übermittelten Informationen in Form von Datenströmen – Zunahme der empfangenen Fernseh- oder Radiosendungen, der Telefongespräche	→ ←	– Wettbewerbsintensivierung – Entwicklungsimpulse – Höhere externe Verwundbarkeit – Bedeutungszuwachs transnationaler Akteure, internationaler Organisationen – Beschleunigte Ausbreitung von Krankheiten – Ökologische Probleme – Verbreitung extremistischer, krimineller Information – Kulturelle Durchdringung

Presseartikel zu Handygefahren für Kinder sichten oder Eltern und Lehrer dazu befragen, welche nützlichen und schädlichen Umgangsweisen sie bei Kindern mit dem Handy beobachten.

2. **Systematisierung**: Die gesammelten Beispiele und Aspekte werden in eine gegenstandsbezogene Ordnung gebracht: Dabei bietet sich eine Systematisierung nach Ursache-Wirkungs-Relationen, nach zeitlichen Abläufen oder logischen Strukturen (z. B. vom Allgemeinen zum Speziellen) an. Eine Visualisierung in Form von Pfad- oder Ablaufdiagrammen kann dabei helfen, die konzeptuelle Struktur zu verdeutlichen und hinsichtlich möglicher Lücken, Redundanzen oder Inkonsistenzen zu bearbeiten. Ziel ist eine möglichst erschöpfende Auffächerung der verschiedenen relevanten Aspekte des Gegenstandes. So lässt sich das Konzept Handykompetenz entlang üblicher Dimensionen von Medienkompetenz gliedern (z. B. Medienkunde, Medienkritik, Mediennutzung, Mediengestaltung), möglich ist aber auch eine Strukturierung nach unterschiedlichen Anwendungen und Diensten des Mobiltelefons (z. B. Mobiltelefonie, Handyspiele, Handykamera, Mobile Apps, mobiles Internet) und/oder nach Nutzungskontexten (z. B. Handynutzung zu Hause, in der Schule, unterwegs).

3. **Auswahl der untersuchungsrelevanten Aspekte**: Da die dimensionale Analyse als Hilfsmittel der Konzeptspezifikation bzw. der Gegenstandsstrukturierung eine empirische Studie vorbereiten soll, müssen schließlich auch theoretische und forschungspraktische Erwägungen herangezogen werden: Welche Aspekte lassen sich mit den vorhandenen zeitlichen, finanziellen und personellen Mitteln überhaupt untersuchen? Welche Aspekte sind besonders interessant und vielversprechend hinsichtlich ihres theoretischen oder praktischen Erkenntniswertes? So mag man sich z. B. dafür entscheiden, den Aspekt der Videotelefonie per Handy nicht in die Studie einzubeziehen, weil zum Untersuchungszeitpunkt wenige Kinder diese Option nutzen. Oder man wird insbesondere jene Aspekte der Handykompetenz umfassend untersuchen, deren praktische Förderung – z. B. im Rahmen des Schulunterrichts – als besonders gut möglich erscheint.
4. **Entwicklung eines deskriptiven Begriffsschemas**: Nachdem die in die empirische Untersuchung einzubeziehenden Subdimensionen des Konzepts identifiziert wurden, müssen für diese jeweils schlüssige Nominaldefinitionen formuliert werden. Soll z. B. im Bereich Handywissen der Aspekt „Kenntnis über die Kosten der Mobilkommunikation" erfasst werden, so muss genau definiert werden, ob damit z. B. gemeint ist, dass Kinder wissen, wie viel ihr Handy als Endgerät kostet, wie teuer Inlands- und Auslandsgespräche mit dem Handy sind oder welche verschiedenen Bezahlmodelle existieren und welchen Tarif sie selbst nutzen.

Neben explorativen empirischen Studien sind es vor allem **Theorie- und Methodenstudien** (▶ Abschn. 7.3), die komplexe Konzepte einer systematischen dimensionalen Analyse unterziehen, um deren Messung vorzubereiten (▶ Studienbeispiel „Konzeptspezifikation von ‚Globalisierung' mittels dimensionaler Analyse").

8.3 Grundlagen zur Durchführung einer Operationalisierung

Auf die Konzeptspezifikation mittels Nominaldefinition, die mithilfe von Bedeutungsanalyse sowie dimensionaler Analyse zu erarbeiten ist, folgt die Operationalisierung, die in eine operationale Definition mündet. Im Folgenden wird die operationale Definition eingeführt. Dabei werden Besonderheiten bei abhängigen und unabhängigen Variablen hervorgehoben und verbreitete Fehlinterpretationen problematisiert.

8.3.1 Operationale Definition von theoretischen Konzepten

Für ein und dasselbe theoretische Konzept existieren meistens verschiedene Möglichkeiten der Operationalisierung. So kann das **theoretische Konstrukt „Intelligenz"** mit unterschiedlichen **Messinstrumenten** (Intelligenztests) operationalisiert werden, die sich jeweils aus verschiedenen **Indikatoren** (Intelligenztestaufgaben) zusammensetzen, welche einzeln und gemeinsam dazu dienen, die Intelligenzleistung einer Person zu erfassen. Eine **operationale Definition** wie „Intelligenz ist, was der Intelligenztest misst" bzw. genauer: „Intelligenz ist, was der Hamburg-Wechsler-Intelligenztest misst" oder „Intelligenz ist, was der Raven-Test misst" erscheint für sich genommen meistens unbefriedigend. Deswegen sollten operationale Definitionen nicht die Konzeptspezifikation ersetzen. Hat man zuerst im Rahmen der Konzeptspezifikation theoretisch festgelegt, welches Intelligenzkonzept man zugrunde legen möchte und die entsprechende Nominaldefinition angegeben, dann wird man auf dieser Basis gezielt die passende Operationalisierung bzw. den passenden Intelligenztest auswählen. Bei dieser Vorgehensweise ist die verwendete operationale Definition nicht zirkulär, sondern **in einem theoretischen Verständnis von Intelligenz verankert**.

Operationalisierung – Die Operationalisierung („operationalization") eines theoretisches Konzepts bzw. einer latenten Variable legt fest, anhand welcher beobachtbaren Variablen **(Indikatoren)** die Ausprägung des theoretischen Konzepts bei den Untersuchungsobjekten festgestellt werden soll. Neben der Auswahl der Indikatoren gehört zur Operationalisierung auch die **Festlegung der Messinstrumente**, mittels derer den Ausprägungen der einzelnen Indikatoren jeweils entsprechende numerische Werte zugeordnet und zu einem Gesamtmesswert für das Konstrukt verrechnet werden. Komplexe theoretische Konstrukte werden selten mit einem einzigen Indikator (**Einzelindikator** als Messinstrument) operationalisiert, sondern meist über einen Satz von Indikatoren (d. h. über eine **psychometrische Skala** oder einen **Index**). Mit der Festlegung der Operationalisierung wird für ein theoretisches Konzept (dargelegt über seine Nominaldefinition) eine konkretisierende **operationale Definition („operational definition")** vorgenommen.

Angenommen, die **Intensität des Mobbings von Schulkindern** wurde im Zuge der Konzeptspezifikation über die Dimensionen a) Anzahl der Aggressoren, b) Dauer des Mobbings und c) Schwere der aggressiven Handlungen definiert. Indikatoren für diese theoretischen Aspekte können nun entsprechend detaillierte Fragen in einem Elternfragebogen oder in einem Schülerinterview sein. Welche Operationalisierungsvariante und damit auch **Datenerhebungsmethode** (▶ Kap. 10) gewählt wird, hängt u. a. von forschungspraktischen sowie von theoretischen Erwägungen ab. So mag es zwar weniger aufwändig sein,

Eltern kollektiv beim Elternabend einen standardisierten Fragebogen zur Mobbingbetroffenheit ihres Kindes ausfüllen zu lassen als alle Kinder einzeln in kindgerechter Weise zu interviewen. Da jedoch nicht davon auszugehen ist, dass Eltern über alle Mobbingvorfälle ihrer Kinder genau informiert sind, wären vermutlich Indikatorvariablen, die direkt an den Kindern selbst erhoben werden, zu bevorzugen.

Die zur Konzeptspezifikation der **Globalisierung unterschiedlicher Länder** angegebenen theoretischen Dimensionen (z. B. „Touristenströme"; ▶ Studienbeispiel „Konzeptspezifikation von ‚Globalisierung' mittels dimensionaler Analyse") sind im Zuge der Operationalisierung in konkret messbare Indikatoren zu übersetzen (z. B. Anzahl der Ankünfte internationaler Touristen pro 1 000 Einwohner pro Jahr), die z. B. den amtlichen Statistiken zu entnehmen wären. Zuweilen muss aus **forschungspraktischen bzw. forschungsökonomischen Gründen** auf Indikatoren zurückgegriffen werden, die das theoretische Konzept nur teilweise oder nur ungenau abbilden (etwa weil relevante Kennwerte nicht schnell oder kostengünstig genug beschaffbar sind). Abweichungen zwischen theoretischem Konstrukt und den zur Operationalisierung genutzten Indikatoren sind zu begründen und bei der Ergebnisinterpretation zu berücksichtigen (z. B. wenn die interessierenden internationalen „Touristenströme" über die Zahl der Hotelübernachtungen operationalisiert würden, wobei dann auch nationale Touristen sowie Geschäftsreisende enthalten wären).

Multiple Indikatoren tragen dazu bei, dass die verschiedenen Aspekte eines komplexen theoretischen Konstruktes möglichst vollständig abgebildet werden und somit wirklich das erfasst wird, was gemessen werden soll (Kriterium der Gültigkeit bzw. **Validität** des Messinstrumentes). Zudem können durch den Einsatz multipler Indikatoren auch Messfehler reduziert werden, so dass die Messgenauigkeit bzw. **Reliabilität** des Instruments steigt (zu psychometrischen Gütekriterien im Überblick ▶ Abschn. 10.4.1). Dementsprechend wird eine groß angelegte bevölkerungsrepräsentative Studie zur Verbreitung von Depressionen mit einem etablierten Depressionsmessinstrument arbeiten, welches das Konstrukt möglichst differenziert mithilfe einer Reihe sorgfältig entwickelter und aufeinander abgestimmter Selbstauskunftsfragen bzw. Indikatoren erfasst. Demgegenüber wird eine Studie, die sich dem Essverhalten widmet, das Konstrukt Depression – wenn es denn am Rande auch erhoben werden soll – möglicherweise nur mit einem **Einzelindikator** (z. B. mit einer einzigen Frage im Fragebogen) erheben, weil alles andere das ohnehin umfangreiche Erhebungsinstrument sonst zu sehr aufblähen, die Untersuchungspersonen demotivieren und somit letztlich den Erfolg der gesamten Studie einschränken könnte (zur Messung von latenten Merkmalen wie z. B. Depression mittels Einzelindikatoren ▶ Abschn. 8.5.2).

Wenn anstelle eines Einzelindikators ein aus mehreren Indikatoren bestehendes Messinstrument zur Operationalisierung eines theoretischen Konstruktes genutzt wird, so stellt sich die Frage, welche Indikatoren auf welche Weise zu einem Messinstrument zusammengefasst werden sollen. Die Auswahl, Modifikation oder Neuentwicklung von Messinstrumenten und Indikatoren ist die Kernaufgabe bei der Operationalisierung. Wissenschaftlich unbrauchbar sind **Ad-hoc-Instrumente**, die aus einer mehr oder minder willkürlichen und ungeprüften Menge an Indikatoren bestehen. Gefordert sind stattdessen wissenschaftliche Messinstrumente, deren Aufbau theoretisch begründet und deren Gütekriterien empirisch geprüft sind, so dass man sichergehen kann, dass sie tatsächlich genau das Zielkonstrukt erfassen (Validität) und wenig durch Messfehler verzerrt sind (Reliabilität). Die Forschung rund um die Voraussetzungen und Gütekriterien der Messung psychologischer Merkmale wird auch als **Psychometrie** („psychometrics") bezeichnet. Eine wichtige Grundlage für die Konstruktion und Bewertung von Messinstrumenten ist die **Testtheorie** (▶ Abschn. 10.4.4). In Abgrenzung von Ad-hoc-Instrumenten werden überprüfte Messinstrumente auch als **psychometrische Skalen** und **psychometrische Tests** bezeichnet. Typische Gegenstände psychometrischer Messung sind Einstellungen, aber auch Wissen und kognitive Leistungen, Persönlichkeitsmerkmale oder psychologische Störungen, zu denen eine Fülle von Messinstrumenten vorliegen (für eine Auswahl gebräuchlicher Testverfahren ▶ Abschn. 10.4.3). Auf die Konstruktionsprinzipien von unterschiedlichen psychometrischen Skalen (▶ Abschn. 8.6) sowie von Indizes (▶ Abschn. 8.7) wird im Verlauf dieses Kapitels noch genauer eingegangen.

Wichtig ist es im Hinblick auf das **Verhältnis zwischen theoretischem Konstrukt und Indikator**, zwischen zwei Typen von Indikatoren zu unterscheiden: Den reflektiven und den formativen Indikatoren (auch Bühner 2011, S. 37).

- Beim **reflektiven Messmodell**, das psychometrischen Skalen (▶ Abschn. 8.6) zugrunde liegt, wird das **theoretische Konstrukt als Ursache** und die **Indikatoren** werden **als Wirkungen** betrachtet. Beispiel: Dadurch, dass Menschen sich im Grad ihrer Schüchternheit unterscheiden (theoretisches Konstrukt als Ursache), beantworten sie Indikatorvariablen bzw. Skalen-Items wie „Mir fällt es schwer, mit Fremden ins Gespräch zu kommen" oder „Wenn ich mit Fremden spreche, fühle ich mich gehemmt" in systematischer Weise unterschiedlich (reflektive Indikatoren als Wirkungen).

In den Ausprägungen dieser Indikatoren „reflektiert" sich also die Ausprägung des latenten Merkmals: Weil eine Person schüchtern ist, stimmt sie entsprechenden Aussagen über schüchternes Verhalten und Erleben tendenziell zu. Dasselbe Ursache-Wirkungs-Prinzip zwischen Konstrukt und Indikatoren gilt für Skalen, mit denen z. B. durch mehrere Skalen-Items (Fragen oder Aufgaben) Konstrukte wie Depression, Intelligenz, Aggressivität oder Religiosität erfasst werden. Die in einer psychometrischen Skala enthaltenen reflektiven Indikatoren sind einander **formal und inhaltlich ähnlich** (z. B. bedeutungsähnliche Selbsteinschätzungen zur Schüchternheit) und relativ hoch miteinander korreliert.

- Demgegenüber betrachtet das **formative Messmodell**, welches den Indizes (▶ Abschn. 8.7) zugrunde liegt, die **Indikatoren als Ursachen** bzw. Determinanten des im Zuge der Konzeptspezifikation definierten theoretischen Konstruktes. Die Ausprägung des **Konstruktes ist eine Wirkung** der Indikatoren. Beispiel: Erst wenn nachweisbar ist, dass ein bestimmtes Land über zunehmende grenzüberschreitende Touristen-, Daten-, Handels-, Finanzströme etc. verfügt, dann wird diesem Land gemäß der Nominaldefinition und Konzeptspezifikation von Globalisierung ein hoher Globalisierungsgrad zugeschrieben. Also nicht weil ein Land globalisiert ist, entstehen dadurch Touristen- und Finanzströme, Wirtschafts- und Wissenschaftsaustausch etc., sondern das Auftreten der inhaltlich ganz verschiedenen grenzüberschreitenden sozialen Interaktionen erzeugt bzw. „formt" den Globalisierungsgrad eines Landes. Die in einem Index enthaltenen formativen Indikatoren können einander **formal und inhaltlich sehr unähnlich** sein und müssen auch nicht miteinander korrelieren. Ein weiteres Beispiel: Wenn eine Person einen geringen Bildungsstand hat, einen Beruf mit geringem Prestige ausübt und über ein unterdurchschnittliches Einkommen verfügt (drei formative Indikatoren als Determinanten), dann wird ihr definitionsgemäß ein niedriger sozioökonomischer Status zugeschrieben (theoretisches Konstrukt als Wirkung).

Neben Einzelindikatoren werden zur Operationalisierung theoretischer Konstrukte oft **Messinstrumente** verwendet, die aus mehreren Indikatoren bestehen und auf zwei grundlegend verschiedenen Messmodellen basieren.

Reflektives Messmodell – Bei einem reflektiven Messmodell geht man davon aus, dass das zu messende Konstrukt die Ursache für die Merkmalsausprägungen auf den gewählten Indikatoren ist. Als Messinstrument wird eine **psychometrische Skala** genutzt, die aus homogenen bzw. inhaltsähnlichen Fragen, Aussagen oder Aufgaben besteht (sog. **reflektive Indikatoren**, in denen sich das Konstrukt widerspiegelt).

Formatives Messmodell – Bei einem formativen Messmodell geht man davon aus, dass das zu messende Konstrukt die Wirkung oder Folge der Merkmalsausprägungen der Indikatoren ist. Als Messinstrument wird ein **Index** gebildet, in den heterogene Kennwerte eingehen (sog. **formative Indikatoren**, durch die das Konstrukt ursächlich gebildet wird).

Bereits bei der Konzeptspezifikation (▶ Abschn. 8.2), spätestens im Zuge der operationalen Definition, also bei der Auswahl oder Konstruktion von Indikatoren und Messinstrumenten, sollte man sich darüber klar geworden sein, ob man ein reflektives oder ein formatives Messmodell anzulegen hat, also ob eine psychometrische Skala oder ein Index zu verwenden ist.

8.3.2 Operationalisierung von abhängigen Variablen

In der quantitativen Sozialforschung messen wir Variablen vor allem, um im Kontext explanativer Studien Hypothesen über Variablenzusammenhänge, Gruppenunterschiede oder Veränderungen über die Zeit zu prüfen. Besonders aussagekräftig ist dabei die Prüfung von Kausalhypothesen, wie sie in experimentellen und quasi-experimentellen Designs angestrebt wird (▶ Abschn. 7.6). In diesen Designs wird mindestens eine unabhängige Variable systematisch variiert (z. B. unterschiedliche Behandlungen, Interventionen, Stimuli), um die Auswirkung auf die abhängige(n) Variable(n) zu prüfen. Dabei versucht man meistens, die Ausprägungen der abhängigen Variablen in **möglichst feinen Abstufungen** zu erfassen. Aber auch in nicht-experimentellen Studien ist eine differenzierte Messung der Variablenausprägungen nützlich. Zur Messung feiner Merkmalsabstufungen sind folgende sechs **Operationalisierungsvarianten** (modifiziert nach Conrad & Maul, 1981, S. 151) besonders geeignet:

1. **Häufigkeit:** Wie oft tritt ein bestimmtes Verhalten auf? (Beispiele: Anzahl der Fehler in einem Diktat, Häufigkeit der Blickkontakte beim Flirt, Häufigkeit von Sprechpausen in einer Vernehmungssituation, Häufigkeit von Ehestreits vor und nach einer Paartherapie)
2. **Reaktionszeit:** Wie viel Zeit vergeht, bis eine Person nach Auftreten eines Stimulus reagiert? (Beispiele: Reaktionslatenz nach Auftreten eines unerwarteten Verkehrshindernisses, Reaktionszeit bis zur Identifikation eines Wortes). Die Reaktionszeitmessung ist in der Regel experimentellen Laborstudien vorbehalten.
3. **Reaktionsdauer:** Wie lange reagiert eine Person auf einen Stimulus bzw. auf eine Intervention? (Beispiele: Lösungszeit für eine Mathematikaufgabe, Verweildauer des Auges auf einem bestimmten Bildausschnitt, Dauer des Nichtrauchens nach einem Anti-Rauch-Training)

4. **Reaktionsstärke:** Wie intensiv reagiert eine Person auf einen Stimulus bzw. auf eine Intervention? (Beispiele: Stärke der Muskelanspannung als Indikator für Aggressivität, geäußerte Stärke von Meinungen auf Ratingskalen, Höhe des Blutdruckanstiegs, Intensität der Zustimmung zu einer Partei vor und nach der Rezeption von Pressebeiträgen über einen Parteiskandal). Die Reaktionsstärke kann im Rahmen von Messwiederholungsdesigns bzw. Längsschnittstudien auch mehrfach erhoben und in ihrem Verlauf betrachtet werden (▶ Abschn. 7.8).
5. **Reaktionsqualität**: Welche Wertigkeit (Valenz) hat eine Reaktion auf einen Stimulus: Ist sie eher positiv oder negativ bzw. beinhaltet sie Zuwendung oder Abwendung? (Beispiele: Bewertung von Lebensmitteln hinsichtlich Geruch und Geschmack auf Schulnotenskalen; Einstufung der Sympathie oder Antipathie gegenüber einer Person auf einer Ratingskala; Einstufung der eigenen Stimmungslagen – angespannt, aufmerksam, fröhlich etc. – auf einer psychometrischen Skala.)
6. **Wahlreaktion:** Welche Wahl trifft eine Person angesichts mehrerer Wahlmöglichkeiten? (Beispiele: Bevorzugung eines von zwei Kunstwerken als ästhetischer im Paarvergleichsurteil; Bevorzugung eines von mehreren Reisezielen bei einer Mehrfachwahlaufgabe; Nennung der Lieblingsmarke aus einem Spektrum an Marken bei einem Präferenzurteil.)

Bei den Operationalisierungsvarianten für abhängige Variablen ist der **Zeitpunkt der Messung** zu beachten: Die Messung kann **nach Abschluss der Intervention** (z. B. Stimuluspräsentation) erfolgen, etwa wenn nach dem Betrachten eines Films, nach einem Bewerbungsinterview oder nach einer Unterrichtsstunde jeweils die Ausprägungen der interessierenden abhängigen Variablen erfasst werden (z. B. mit einer psychometrischen Skala, mit einem psychologischen Testverfahren oder per Expertenurteil). Es besteht aber auch die Möglichkeit, die Messung **prozessbegleitend** durchzuführen (d. h. **während** die Untersuchungsperson auf den Stimulus bzw. die unabhängige Variable reagiert wird z. B. die Reaktionsdauer per Stoppuhr gemessen und die Reaktionsqualität durch Beobachtung von Mimik und Gestik erfasst). Besonders gut geeignet für prozessbegleitende Messungen sind **physiologische Messverfahren** (▶ Abschn. 10.5), mit denen kontinuierlich während des gesamten Versuchsdurchlaufs z. B. der Blutdruck oder die Blickbewegungen aufgezeichnet werden. Es besteht auch die Möglichkeit, subjektive Bewertungen prozessbegleitend zu erheben. Hierfür werden als Operationalisierungsvarianten das sog. **Real Time Response (RTR) Measurement** bzw. das **Continous Response Measurement (CRM)** genutzt (zum Überblick Maier, Maier, Maurer, Reinemann, & Meyer, 2009). Bei diesen Verfahren erhält die Untersuchungsperson als Messinstrument einen **Dreh- oder Schieberegler**, über den sie fortlaufend stufenlos angeben kann, ob und wie stark sie einen präsentierten Reiz gerade positiv oder negativ bewertet (auch andere Reaktionsqualitäten wie interessant vs. langweilig etc. können erfasst werden). Die RTR-Measurement- bzw. CRM-Methode wurde bereits in den 1930er-Jahren entwickelt und wird bis heute u. a. zur prozessbegleitenden Messung von Publikumsreaktionen auf Medienangebote wie Radio- oder TV-Sendungen, Werbespots und Kinofilme genutzt (Maier, Maurer, Reinemann, & Faas, 2006; Reinemann, Maier, Faas, & Maurer, 2005; Schmeisser, Bente, & Isenbart, 2004; Schneider et al., 2011).

Je komplexer ein theoretisches Konstrukt und je wichtiger seine genaue Messung für die Studie ist, umso eher wird man zu seiner Operationalisierung statt auf einen **Einzelindikator** (▶ Abschn. 8.5) auf ein **Messinstrument mit mehreren Indikatoren** zurückgreifen. Sehr verbreitet zur Operationalisierung abhängiger Variablen in der empirischen Sozialforschung sind die **psychometrische Skala** (▶ Abschn. 8.6) sowie der **Index** (▶ Abschn. 8.7). Sie messen Qualitäten sowie mehr oder minder fein abgestuft die Intensitäten von Merkmalsausprägungen.

8.3.3 Operationalisierung von unabhängigen Variablen

Während abhängige Variablen oft stetige Merkmale sind, die mit möglichst vielen Abstufungen gemessen werden, handelt es sich bei unabhängigen Variablen meistens um diskrete Variablen mit wenigen Ausprägungen. Die unabhängige Variable bzw. der Ursachenfaktor fungiert in experimentellen und quasi-experimentellen Studien (▶ Abschn. 7.6) als **Gruppierungsvariable**, wobei der Vergleich von zwei bis ca. zehn Gruppen typisch ist. Bei nicht-experimentellen Studien werden diese Gruppen durch eine vorgefundene Variable gebildet, die z. B. mittels Beobachtung oder Befragung operationalisiert wird (z. B. Vergleiche zwischen Altersgruppen, Geschlechtern, Nationalitäten auf der Basis entsprechender soziodemografischer Angaben im Fragebogen).

Die Operationalisierung von experimentellen und quasi-experimentellen unabhängigen Variablen läuft oft auf die Produktion von unterschiedlichem **Stimulusmaterial** oder die Konzeption von unterschiedlichen **Behandlungsformen** bzw. Interventionen oder Treatments hinaus. Soll etwa der Grad der Gewalthaltigkeit von Computerspielen als experimentelle unabhängige Variable variiert werden, um den Effekt auf die abhängige Variable

Aggressivität zu erfassen, so müssen mehrere Varianten eines Computerspiels mit unterschiedlicher Gewalthaltigkeit a) gezielt ausgewählt oder b) selbst produziert werden. Dabei kommt es darauf an, dass sich die Spiele möglichst nur in ihrer Gewalthaltigkeit und nicht in anderen Merkmalen unterscheiden. Würde man im Experiment z. B. die Aggressivität nach einem Denkspiel mit der Aggressivität nach dem Spielen eines Ego-Shooters vergleichen und in der Shooter-Gruppe tatsächlich erhöhte Aggressionswerte messen können, so wäre damit nicht belegt, dass es sich um einen Effekt der Gewalthaltigkeit des Spiels handelt. Denn Denkspiel und Shooter unterscheiden sich auch in vielen anderen Aspekten – von der Hintergrundmusik über die Farbgebung bis zur Spieleraktivität – deutlich voneinander. Die verschiedenen Untersuchungsbedingungen einschließlich der Stimulusmaterialien müssen idealerweise so gestaltet sein, dass sie sich exakt nur hinsichtlich der unabhängigen Variable voneinander unterscheiden und **alle anderen Aspekte der Untersuchungsbedingungen gleich** sind.

Zu beachten ist zudem, dass die **Dosierung der unabhängigen Variable** maßgeblich darüber entscheidet, welche **Effektstärke** sich in einer Studie zeigt (z. B. Messung der Aggressivität nach 10 Minuten, nach 3 Stunden oder nach 6 Monaten Nutzungszeit eines gewalthaltigen Medienangebots; ► Abschn. 14.2).

Eine theoretisch fundierte Operationalisierung der unabhängigen Variablen ist in der Experimentalforschung (► Abschn. 7.6) eine besondere Herausforderung und gleichzeitig entscheidende Voraussetzung für die Aussagekraft eines Experiments oder Quasi-Experiments.

8.3.4 Fehlinterpretation von Operationalisierungen

Die Auswahl der Indikatoren und die Wahl des Messinstrumentes sind bei **latenten Merkmalen** erklärungsbedürftig, weil eine Brücke zwischen den beobachtbaren Sachverhalten einerseits und der theoretischen Konzeptualisierung andererseits geschlagen werden muss. Bei **manifesten Variablen**, die praktisch unmittelbar als beobachtbare Indikatoren vorliegen, besteht kein besonderes konzeptuelles Überbrückungsproblem. Dennoch müssen strenggenommen auch manifeste Variablen operationalisiert werden. So etwa, wenn zur Erfassung von Alter, Geschlecht oder Wohnort entsprechende Fragen in einem Fragebogen gestellt werden (zur Messung soziodemografischer Variablen ► Abschn. 8.5.1).

Indem die Operationalisierung angibt, über welche Indikatoren und mit welchem standardisierten Messinstrument (z. B. einem standardisierten Fragebogen oder psychologischen Test) ein theoretisches Konstrukt empirisch zu erfassen ist, wird dieses greifbarer. Dabei ist im Auge zu behalten, dass die beobachteten Merkmale keine voraussetzungslosen Tatsachen, sondern immer Ergebnis eines **theoretischen Konstruktionsprozesses** sind. Der Umstand, dass eine Person auf einem Messinstrument für „Internetsucht“ eine hohe Punktzahl erreicht, bedeutet nicht, dass die Person tatsächlich internetsüchtig „ist“. Es bedeutet, dass ihr Verhalten und Erleben auf der Basis bestimmter theoretischer Vorannahmen mit dem Konzept der Sucht beschrieben und erklärt wird. Internetsucht als reale Tatsache – anstatt als theoretische Konstruktion – aufzufassen, käme einer unzulässigen Verdinglichung bzw. **Reifizierung** („reification“) gleich. Eine andere Theorie könnte dasselbe Verhalten nicht als „Sucht“, sondern als „Zwang“ oder auch als „Gewohnheit“ auffassen, woraus sich dann andere Schlussfolgerungen hinsichtlich Entstehung oder Behandlung ergeben würden. Empirische Forschung, die gemessene Variablen als Tatsachen auffasst, mündet in einen **naiven Empirismus bzw. Positivismus**.

Deswegen ist die **theoretische Konstruiertheit aller wissenschaftlichen Messungen** bei der Diskussion von empirischen Forschungsprozessen und ihren Ergebnissen stets zu berücksichtigen. Dies wird im quantitativen Paradigma der empirischen Sozialforschung im Rahmen der Wissenschaftstheorie des Kritischen Rationalismus ausdrücklich betont (► Abschn. 2.2.3). Theoretische Konzepte zu operationalisieren läuft somit keineswegs auf ein datengläubiges „empiristisches“ oder „positivistisches“ Vorgehen hinaus, vielmehr verlangt eine seriöse wissenschaftliche Operationalisierung transparente und fundierte theoretische Argumente sowohl bei der Auswahl und Konstruktion von Indikatoren und Messinstrumenten als auch bei der Interpretation der so gewonnenen quantitativen Daten.

8.4 Messung und die vier Skalenniveaus

Wurden im Zuge der Operationalisierung für ein latentes Merkmal die manifesten Indikatorvariablen ausgewählt und die Art des Messinstruments und damit auch der Datenerhebungsmethode festgelegt (z. B. Beobachtungsschema, standardisierter Fragebogen, psychologischer Test), so steht noch die Messung im engeren Sinne an, d. h., die aussagekräftige Zuordnung von numerischen Messwerten zu den beobachteten Ausprägungen der Untersuchungseinheiten auf den einzelnen Indikatorvariablen. In Abhängigkeit von der Art der Messung unterscheiden wir vier **verschiedene Messniveaus bzw. Skalenniveaus**. Je höher das Skalenniveau der Messung, umso informationshaltiger sind die erzeugten Messwerte und umso vielfältiger die Möglichkeiten der statistischen

Tabelle 8.4 Die drei respektive vier wichtigsten Skalenarten bzw. Skalenniveaus

Drei Skalenarten bzw. Skalenniveaus	Vier Skalenarten bzw. Skalenniveaus	Zulässige Transformationen	Mögliche Aussagen	Beispiele
1. Nominalskala	1. Nominalskala	Eindeutigkeitstransformation	Gleichheit, Verschiedenheit	Automarken, Krankheitsklassifikationen, Familienstand
2. Ordinalskala	2. Ordinalskala	Monotone Transformation	Größer-Kleiner-Relationen	Militärische Ränge, Windstärken
3. Kardinalskala = metrische Skala	3. Intervallskala	Lineare Transformation	Gleichheit von Differenzen	Temperatur (z. B. Celsius), Kalenderzeit, Intensität von Einstellungen
	4. Verhältnisskala	Ähnlichkeitstransformation	Gleichheit von Verhältnissen	Längenmessung, Gewichtsmessung, Häufigkeiten pro Person

Datenanalyse. Generell ist es empfehlenswert, sich im Vorfeld der Datenerhebung über den Informationsgehalt der Messwerte und die damit verbundenen statistischen Auswertungsmöglichkeiten Gedanken zu machen.

In der Sozialforschung werden die vier Skalenarten bzw. Skalenniveaus zuweilen auch zu drei Varianten zusammengefasst (Tab. 8.4). Die Daten eines bestimmten Skalenniveaus ermöglichen unterschiedliche inhaltliche Aussagen über die Variablenausprägungen. Sie sind gegenüber den jeweils zulässigen skalenspezifischen Transformationen invariant, wodurch die Möglichkeiten und Grenzen einer sinnvollen statistischen Auswertung abgesteckt werden: Im Rahmen der zulässigen Transformationen können Messwerte verrechnet werden, ohne dass sich ihre inhaltliche Aussage verändert. Jede Skalenart ist durch spezifische messtheoretische Voraussetzungen definiert, auf die wir in den folgenden Abschnitten genauer eingehen.

Ein Vergleich der vier Skalenniveaus zeigt, dass die Messungen mit zunehmender Ordnungsziffer des Skalenniveaus genauer werden. Während eine Nominalskala lediglich Äquivalenzklassen von Objekten numerisch beziffert, informieren die numerischen Werte einer Ordinalskala zusätzlich darüber, bei welchen Objekten das Merkmal stärker bzw. weniger stark ausgeprägt ist. Eine Intervallskala ist der Ordinalskala überlegen, weil hier die Größe eines Merkmalsunterschiedes bei zwei Objekten genau quantifiziert wird. Eine Verhältnisskala schließlich gestattet zusätzlich Aussagen, die die Merkmalsausprägungen verschiedener Objekte zueinander ins Verhältnis setzen.

Messungen auf den vier Skalenniveaus werden in Tab. 8.5 durch Fragebogenitems verdeutlicht.

Empirische Sachverhalte werden durch die vier Skalenarten bzw. Skalenniveaus unterschiedlich genau abgebildet. Die hieraus ableitbare Konsequenz für die Planung empirischer Untersuchungen liegt auf der Hand: Bieten sich bei einer Quantifizierung mehrere Skalenarten an, sollte diejenige mit dem **höchsten Skalenniveau** gewählt werden (Bortz & Schuster, 2010. S. 22f.). Erweist sich im Nachhinein, dass die erhobenen Daten dem angestrebten Skalenniveau letztlich nicht genügen, besteht die Möglichkeit, die erhobenen Daten auf ein niedrigeres Skalenniveau zu transformieren. (Beispiel: Zur Operationalisierung des Merkmals „Schulische Reife“ sollten Experten intervallskalierte Punkte vergeben. Im Nachhinein stellte sich heraus, dass die Experten mit dieser Aufgabe überfordert waren, so dass man beschließt, für weitere Auswertungen nur die aus den Punktzahlen ableitbare Rangfolge der Kinder zu verwenden.) Eine nachträgliche Transformation auf ein höheres Skalenniveau ist hingegen nur im Ausnahmefall möglich (▶ Abschn. 8.4.3 „Indirekte Rangordnungen“).

Wie jedoch – so lautet die zentrale Frage – wird in der Forschungspraxis entschieden, auf welchem Skalenniveau ein bestimmtes Merkmal gemessen wird? Ist es erforderlich bzw. üblich, bei jedem Merkmal die gesamte Axiomatik der mit einer Skalenart verbundenen Messstruktur empirisch zu überprüfen? Kann man – um im oben genannten Beispiel zu bleiben – wirklich guten Gewissens behaupten, die Punktzahlen zur „Schulischen Reife“ seien, wenn schon nicht intervallskaliert, so doch zumindest ordinalskaliert?

Sucht man in der Literatur nach einer Antwort auf diese Frage, so wird man feststellen, dass hierzu unterschiedliche Auffassungen vertreten werden (Hand 1996; King, Rosopa, & Minium, 2010). Unproblematisch und im Allgemeinen ungeprüft ist die Annahme, ein Merkmal sei nominalskaliert. Wohnort, Parteizugehörigkeit, Studienfach etc. sind einfache manifeste Merkmale, deren Nominalskalenqualität unstrittig ist.

Weniger eindeutig fällt die Antwort jedoch aus, wenn es darum geht zu entscheiden, ob Schulnoten, Testwerte oder auf Ratingskalen abgegebene Einstellungsmessun-

Tabelle 8.5 Messungen auf allen vier Skalenniveaus am Beispiel von Operationalisierungen des Merkmals „Rauchen“

Fragebogenitem	Messwerte für die Antwortalternativen	Skalenniveau der Variable
Sind Sie Raucher/-in? Ja ☐ Nein ☐	Ja (1) Nein (2) Ja (1) Nein (0)	**2-fach gestufte nominalskalierte (binäre, dichotome) Variable** ➔ welche Zahlen zugeordnet werden, ist egal, es müssen nur zwei unterschiedliche Zahlen sein, üblich sind 0, 1 oder 1, 2
Was rauchen Sie hauptsächlich? (bitte nur *eine* Antwort ankreuzen) Zigaretten mit Filter ☐ Zigaretten ohne Filter ☐ Cigarillos ☐ Zigarren ☐ Pfeife ☐ Anderes ☐	 Zigaretten mit Filter (1) Zigaretten ohne Filter (2) Cigarillos (3) Zigarren (4) Pfeife (5) Anderes (6)	Mehrfach bzw. **6-fach gestufte nominalskalierte (polytome) Variable** ➔ bei Einfachauswahl („forced choice“) entsteht eine polytome Variable ➔ die Messwerte repräsentieren unterschiedliche Qualitäten des Rauchens
Was rauchen Sie? (Mehrfachauswahl möglich) Zigaretten mit Filter ☐ Zigaretten ohne Filter ☐ Cigarillos ☐ Zigarren ☐ Pfeife ☐ Anderes ☐	 Zigaretten mit Filter (0/1) Zigaretten ohne Filter (0/1) Cigarillos (0/1) Zigarren (0/1) Pfeife (0/1) Anderes (0/1)	**6 Variablen**, die **jeweils 2-fach gestuft nominalskaliert** sind ➔ bei Mehrfachauswahl („multiple choice“) bildet jede einzelne Antwortkategorie eine neue binäre Variable
Welcher Rauchertyp sind Sie? Kettenraucher/-in ☐ Regelmäßiger Raucher/-in ☐ Gelegenheitsraucher/-in ☐ Nichtraucher/-in ☐	 Kettenraucher/-in (4) Regelmäßiger Raucher/-in (3) Gelegenheitsraucher/-in (2) Nichtraucher/-in (1)	**Ordinalskalierte Variable** ➔ die Messwerte repräsentieren eine eindeutige Rangreihe der Intensität des Rauchens: 4 (Kettenraucher) > 3 (regelmäßiger Raucher) > 2 (Gelegenheitsraucher) etc. ➔ die Messwerte können auch in umgekehrter Reihenfolge von Kettenraucher (1) bis Nichtraucher (4) vergeben werden, intuitiv am besten erfassbar ist es, wenn für starke Merkmalsausprägungen hohe Werte vergeben werden
Wie oft rauchen Sie? Nie ☐ Gelegentlich ☐ Oft ☐	 Nie (1) Gelegentlich (2) Oft (3)	**Ordinalskalierte Variable** ➔ vergebene Messwerte sollten intuitiv verständlich sein: höhere Messwerte = stärkere Ausprägung
Wie oft rauchen Sie? Nie ☐ Sehr selten ☐ Selten ☐ Gelegentlich ☐ Oft ☐ Sehr oft ☐ Fast immer ☐	 Nie (1) Sehr selten (2) Selten (3) Gelegentlich (4) Oft (5) Sehr oft (6) Fast immer (7)	**Intervallskalierte Variable** (7stufige, annähernd gleichabständige Häufigkeits-Ratingskala als Antwortformat) ➔ intuitiv verständliche Vergabe der Messwerte: höhere Messwerte = stärkere Ausprägung
Ich werde mir das Rauchen nächstes Jahr abgewöhnen. Keinesfalls ☐ Wahrscheinlich nicht ☐ Vielleicht ☐ Ziemlich wahrscheinlich ☐ Ganz sicher ☐	 Keinesfalls (1) Wahrscheinlich nicht (2) Vielleicht (3) Ziemlich wahrscheinlich (4) Ganz sicher (5)	**Intervallskalierte Variable** (5-stufige, annähernd gleichabständige Wahrscheinlichkeits-Ratingskala als Antwortformat)
Wie viele Zigaretten haben Sie gestern geraucht? ____ Zigaretten	0 Zigaretten (0) 1 Zigarette (1) 2 Zigaretten (2) 3 Zigaretten (3) ... 50 Zigaretten (50) ...	**Verhältnisskalierte Variable** (die Messwerte sind gleichabständig und haben einen absoluten Nullpunkt)

gen ordinal- oder intervallskaliert sind (zu messtheoretischen Problemen ▶ Abschn. 8.4.4 „Messtheoretische Probleme bei Ratingskalen“). Eine richtige Entscheidung ist insoweit von Bedeutung, als die Berechnung von sinnvoll interpretierbaren Mittelwerten und anderen wichtigen statistischen Maßen nur bei intervallskalierten Merkmalen zu rechtfertigen ist. Das heißt, dass für ordinalskalierte Daten andere statistische Verfahren einzusetzen sind als für intervallskalierte Daten.

Die übliche Forschungspraxis verzichtet auf eine empirische Überprüfung der jeweiligen Skalenaxiomatik. Die meisten Messungen sind **Per-fiat-Messungen** (Messungen „durch Vertrauen“), die auf Erhebungsinstrumenten (Fragebögen, Tests, Ratingskalen etc.) basieren, von denen man annimmt, sie würden das jeweilige Merkmal auf einer Intervallskala messen. Es kann so der gesamte statistische „Apparat“ für Intervallskalen eingesetzt werden, der erheblich differenziertere Auswertungen ermöglicht als die Verfahren für Ordinal- oder Nominaldaten (Rasmussen 1989; Zumbo & Zimmerman, 1993). Hinter dieser „liberalen“ Auffassung steht die Überzeugung, dass die Bestätigung einer Forschungshypothese durch die Annahme eines falschen Skalenniveaus eher **erschwert** wird. Strengere Auffassungen fordern jedoch eine ausdrückliche Überprüfung der messtheoretischen Annahmen, dies ist z. B. im Rahmen der probabilistischen Testtheorie möglich (▶ Abschn. 10.4.4).

Im Folgenden werden wir das Konzept der „**Messung**“ von sozialwissenschaftlichen Sachverhalten etwas vertiefen (▶ Abschn. 8.4.1) und anschließend jedes einzelne **Skalenniveau** noch einmal detailliert mit seinen messtheoretischen Eigenschaften, Problemen und Operationalisierungsvarianten erörtern (▶ Abschn. 8.4.5). Von besonderem praktischem Interesse ist dabei die Intervallskala. Intervallskalierte Daten werden sehr oft erhoben, indem man z. B. im Interview oder Fragebogen **Ratingskalen** (z. B. „stimmt gar nicht – wenig – ziemlich – völlig“) als Antwortvorgaben präsentiert. Die Konstruktion derartiger Ratingskalen muss methodischen Standards folgen, damit die so gewonnenen Daten Intervallskalencharakter beanspruchen können. Schließlich wenden wir uns noch der **Skalentransformation** zu, also der Umwandlung von Daten eines Skalenniveaus auf ein niedrigeres oder höheres Niveau (▶ Abschn. 8.4.6).

8.4.1 Messung

Das „Messen“ wird in der Alltagssprache meistens mit physikalischen Vorstellungen in Verbindung gebracht. Dabei bezeichnet man als **fundamentale Messung** das Bestimmen einer (Maß-)Zahl als das Vielfache einer Einheit (z. B. Messungen mit einem Zollstock oder einer Balkenwaage). Für derartige Messungen ist der Begriff „**Einheit**“ zentral. Man wählt hierfür eine in der Natur vorgegebene Größe (wie z. B. die Ladung eines Elektrons als Einheit des Merkmals „elektrische Ladung“) oder man legt aus Gründen der Zweckmäßigkeit willkürlich eine Größe als Normeinheit fest (z. B. der in Paris niedergelegte „Archivmeter“ bzw. „Urmeter“). Eine physikalische Messung besteht darin, möglichst genau zu erfassen, wie oft die gewählte Merkmalseinheit in dem zu messenden Objekt enthalten ist.

Eine Übertragung dieser Messvorstellung auf die Sozialwissenschaften scheitert daran, dass „Einheiten“ in diesem Sinne in den Sozialwissenschaften bislang fehlen. Dennoch sind auch hier – allerdings mit einer weiter gefassten Messkonzeption – Messoperationen möglich.

Allgemein formuliert besteht eine Messoperation im aussagekräftigen Zuordnen von Zahlen zu Objekten. Die logisch-mathematische Analyse dieser Zuordnungen und die Spezifizierung von Zuordnungsregeln sind Aufgaben der **Messtheorie** (▶ Messtheorie).

Messung – Eine Messung („measurement“) meint in der quantitativen Sozialforschung eine Zuordnung von Zahlen zu Objekten oder Ereignissen, sofern diese Zuordnung eine homomorphe (strukturerhaltende) Abbildung eines empirischen Relativs in ein numerisches Relativ ist (Orth 1983, S. 138).

Diese Definition sei kurz erläutert: Ein **empirisches Relativ** ist eine Menge an Objekten, z. B. an Personen, Ereignissen, Medienangeboten. In einem empirischen Relativ gibt es eine oder mehrere Relationen, die die Beziehung zwischen den Objekten charakterisieren (größer als; kleiner als; gleich etc.).

Ein **numerisches Relativ** ist eine Menge an Zahlen (z. B. 1; 2; 15; 17.5 …), die eine mathematische Relation aufweisen ($>$; $<$; $=$ etc.).

Eine **homomorphe (strukturerhaltende) Abbildung** ist eine Funktion. Diese ordnet jedem Objekt des empirischen Relativs (z. B. jedem Kind einer Schulklasse) genau eine Zahl des numerischen Relativs zu (z. B. die jeweilige Körpergröße in cm). Voraussetzung für Homomorphismus ist, dass für jedes Objekt im empirischen Relativ genau ein Element im numerischen Relativ gefunden wird (d. h. für dasselbe Kind dürfen nicht mehrere verschiedene Körpergrößen als Messwerte erscheinen). Außerdem müssen die Relationen im empirischen Relativ auch im numerischen Relativ gültig sein (z. B. wenn Kind A größer ist als Kind B: empirisches Relativ, dann muss auch der Körpergrößenmesswert von Kind A größer sein als der Körpergrößenmesswert von Kind B: numerisches Relativ). Sind diese beiden Voraussetzungen erfüllt, spricht man von einer eindeutigen Zuordnung.

Messtheorie

Wir haben die „Messung“ bereits als strukturerhaltende Abbildung eines empirischen Relativs in ein numerisches Relativ definiert und an dem Beispiel der Messung der Fehler in Aufsätzen (empirisches Relativ) mittels der Fehlerzahl (numerisches Relativ) illustriert. Um messtheoretische Überlegungen zu vertiefen, soll hier ein weiteres Bespiel erläutert werden, nämlich die Messung der **Spielstärke von Tennisspielerinnen** auf einer Rang- bzw. Ordinalskala.

Ein empirisches Relativ möge aus 10 Tennisspielerinnen sowie der zweistelligen Relation „spielerische Überlegenheit“ bestehen. Die spielerische Überlegenheit wird in einem Turnier „Jede gegen Jede“ ermittelt. Eine Spielerin i sei einer Spielerin j überlegen, wenn sie diese schlägt. Dieser Sachverhalt wird durch $i > j$ (i schlägt j) zum Ausdruck gebracht. Den 10 Spielerinnen sind nun wie folgt Zahlen zuzuordnen: $\phi(i)$, $\phi(j)$, $\phi(k)$, ... Für jedes Spielerinnenpaar mit $i > j$ muss die Zahlenrelation $\phi(i) > \phi(j)$ gelten. Die so resultierende Skala heißt **Rang- bzw. Ordinalskala**.

Wenn man unterstellt, dass das Merkmal „Spielstärke“ kontinuierlich ist, die 10 Spielerinnen auf diesem Kontinuum unterschiedliche Positionen einnehmen und diese „wahre“ Spielstärke allein über den Ausgang eines jeden Spiels entscheidet, wären die Rangzahlen 1 (schlechteste Spielerin) bis 10 (beste Spielerin) geeignet, das empirische Relativ homomorph bzw. strukturerhaltend abzubilden. Im Unterschied zur Verhältnisskala kann bei dieser Ordinalskala nicht behauptet werden, dass die Spielerin auf Rangplatz 10 doppelt so gut spielt wie die Spielerin auf Rangplatz 5, denn es gibt bei dieser Messung der Spielstärke keinen absoluten Nullpunkt und die Rangplätze sind auch nicht gleichabständig.

Man bedenke zudem, dass aus $i > j$ und $j > k$ nicht unbedingt $i > k$ folgen muss, denn eine der Spielerin j unterlegene Spielerin k könnte durchaus Spielerin i schlagen ($k > i$), auch wenn Spielerin i ihrerseits Spielerin j besiegt hat. Die Abbildung der Objekte i, j und k mit $\phi(i) = 3$, $\phi(j) = 2$ und $\phi(k) = 1$ wäre in diesem Falle nicht strukturerhaltend, weil die empirische Relation $k > i$ der numerischen Relation $\phi(k) < \phi(i)$ widerspricht.

Die Messbarkeit eines Merkmals ist also an Bedingungen (Axiome) geknüpft, die im empirischen Relativ erfüllt sein müssen. Diese Bedingungen werden in einem **Repräsentationstheorem** zusammengefasst, das die Existenz einer Skala behauptet, wenn diese Bedingungen erfüllt sind. In unserem Beispiel wäre das sog. Transitivitätsaxiom verletzt, wenn für eine beliebige Dreiergruppe von Spielerinnen $i > j$ und $j > k$, aber nicht $i > k$ gilt. (Auf die Möglichkeit äquivalenter Spielstärken gehen wir in ▶ Abschn. 8.4.3 ein.)

Eindeutigkeitsproblem

Mit dem Eindeutigkeitsproblem verbindet sich die Frage, ob sich die Abbildungsfunktion ϕ so in eine andere Abbildungsfunktion ϕ' transformieren lässt, dass die Eigenschaften der Skala erhalten bleiben. Die Lösung des Eindeutigkeitsproblems besteht dann in der Angabe von Transformationen, gegenüber denen die Skaleneigenschaften invariant sind. Man sagt, eine Messung sei eindeutig bis auf die in diesem Sinne zulässigen Transformationen der ursprünglichen Skala. Im Zuge der statistischen Auswertung der Daten dürfen nur die jeweils zulässigen Transformationen durchgeführt werden, damit die inhaltliche Aussage der Messwerte erhalten bleibt.

Im Beispiel wurden den 10 Tennisspielerinnen die Rangzahlen 1–10 zugeordnet. Sind die Bedingungen für eine Ordinalskala erfüllt, ist davon auszugehen, dass eine Spielerin mit einer höheren Zahl eine Spielerin mit einer niedrigeren Zahl besiegt. Dieser Informationsgehalt bliebe erhalten, wenn man zu den Rangzahlen 1–10 jeweils eine konstante Zahl addiert, wenn man sie mit einer konstanten Zahl c ($c > 0$) multipliziert oder wenn man sie so verändert, dass die Größer-Kleiner-Relationen zwischen den ursprünglichen Rangzahlen nicht verändert werden. Transformationen mit dieser Eigenschaft bezeichnet man allgemein als **monotone Transformationen**, so dass wir formulieren können: Messungen auf einer Rang- oder Ordinalskala sind eindeutig bis auf hier zulässige monotone Transformationen.

Bedeutsamkeitsproblem

Unter dem Stichwort Bedeutsamkeit wird gefragt, welche mathematischen Operationen mit den erhobenen Messungen sinnvoll sind. Dass die Beantwortung dieser Frage von der Lösung des Eindeutigkeitsproblems abhängt, lässt sich an unserem Beispiel leicht verdeutlichen: Weder die Aussage: „Spielerin i ist doppelt so spielstark wie Spielerin j“ noch die Aussage: „Spielerin i und j unterscheiden sich in ihrer Spielstärke in gleicher Weise wie die Spielerin k und l“ sind wegen der für Rangskalen zulässigen monotonen Transformation sinnvoll. Addieren wir zu den Rangzahlen 1 und 2 z. B. jeweils den Wert 100. bleibt die Größer-Kleiner-Relation zwar erhalten ($101 < 102$); das Verhältnis der Zahlen zueinander hat sich jedoch drastisch verändert (Tab. 8.6). Dass der Vergleich von Spielstärkeunterschieden keinen Sinn ergibt, verdeutlichen folgenden Zahlen: Die Messungen $\phi(i) = 1$, $\phi(j) = 3$, $\phi(k) = 7$, $\phi(l) = 9$ könnten vermuten lassen, dass der Unterschied zwischen i und j genauso groß sei wie der Unterschied zwischen k und l. Da es sich hierbei jedoch um Messungen auf einer Rangskala handelt, bei der die Größe der Messwertdifferenzen nicht inhaltlich interpretierbar ist, sind monotone Transformationen zulässig (Tab. 8.6).

Bei den Transformationen B und C sind die Größer-Kleiner-Relationen unverändert; die Spielstärkeunterschiede variieren jedoch beträchtlich: Bei Transformation B wäre der Unterschied zwischen i und j gleich und bei Transformation C größer als der Unterschied zwischen k und l. Messwertdifferenzen (oder auch Summen oder Mittelwerte) ergeben also bei Rangskalen **keinen Sinn**. Gleichwohl sind Mittelwerte mit ordinalskalierten Messwerten wie den Rangplätzen 1–10 der Tennisspielerinnen natürlich problemlos manuell oder mit Statistikprogrammen zu berechnen, denn die Messwerte selbst „wehren“ sich gegen solche statistischen Auswertungen nicht. Grundkenntnisse der Messtheorie sind deswegen wichtig, um eben keine sinnlosen Berechnungen anzustellen. Kenntnisse der Messtheorie sind zudem notwendig, um Berechnungen, die das Messniveau missachten und somit (trotz möglicherweise korrekter Rechendurchführung) zu nicht sinnvoll interpretierbaren Ergebnissen führen, als solche zu erkennen.

Allgemein sagen wir, dass eine numerische Aussage dann „bedeutsam“ ist, wenn sie sich unter den für eine Skala **zulässigen Transformationen** nicht verändert. Bei Rangzahlen sind nur diejenigen statistischen Verfahren zulässig, die lediglich die Größer-Kleiner-Relation der Messungen nutzen. Eine kritische Analyse der Bedeutsamkeitsproblematik, auch im Hinblick auf die im Folgenden zu behandelnden Skalenarten, findet man bei (Niederée & Mausfeld, 1996a, 1996b).

Tabelle 8.6 Drei mögliche monotone Transformationen einer ursprünglichen Rangreihe (Rangplätze von Tennisspielerinnen)

Ursprüngliche Messwerte (Rangplätze)	Transformation A (+100)	Transformation B (×5)	Transformation C
$\phi(i) = 1$	$\phi(i) = 101$	$\phi(i) = 5$	$\phi(i) = 1.2$
$\phi(j) = 3$	$\phi(j) = 103$	$\phi(j) = 15$	$\phi(j) = 3.8$
$\phi(k) = 7$	$\phi(k) = 107$	$\phi(k) = 35$	$\phi(k) = 7.9$
$\phi(l) = 9$	$\phi(l) = 109$	$\phi(l) = 45$	$\phi(l) = 8.1$

Ein empirisches Relativ, ein numerisches Relativ sowie eine die beiden Relative homomorph verknüpfende Abbildungsfunktion konstituieren eine **Skala**.

Skala – Unter einer Skala („scale“) versteht man **im Kontext der Messtheorie** ein empirisches Relativ, ein numerisches Relativ und eine die beiden Relative verknüpfende, homomorphe Abbildungsfunktion.

Die Messbarkeit eines Merkmals bzw. die Konstruierbarkeit einer Skala ist an Bedingungen (Axiome) geknüpft (diese werden in ▶ Messtheorie weiter ausgeführt).

Sehr einfach veranschaulichen lässt sich der Messvorgang, wenn es um einfache manifeste quantitative Merkmale geht: So wird man das Merkmal „Fehlerzahl im Aufsatz“ ganz einfach in der Weise „messen“, dass man Messwerte gemäß der Fehlerzahl zuweist (z. B. zwei Fehler → Messwert 2; fünf Fehler → Messwert 5 etc.). Zwei Schüler mit der gleichen Fehlerzahl erhalten dann auch den gleichen Messwert. Schüler, die wenige Fehler gemacht haben, erhalten entsprechend niedrigere Messwerte als Schüler, denen mehr Fehler unterlaufen sind. Die durch homomorphe Abbildung erzeugten Messwerte sind somit sinnvoll hinsichtlich des betrachteten Merkmals interpretierbar. Die in diesem Beispiel resultierende Skala heißt **Verhältnisskala**, weil sie gleichabständig ist und einen **absoluten Nullpunkt** besitzt (ein fehlerfreier Aufsatz hat 0 Fehler, negative Werte sind nicht möglich), so dass Verhältnisaussagen zu treffen sind: Schüler E hat einen doppelt so hohen Messwert bei der Fehlerzahl (z. B. Messwert 8) wie Schüler B (z. B. Messwert 4) und somit auch doppelt so viele Fehler gemacht.

Im Zusammenhang mit Messung und Messtheorie wurde verdeutlicht, dass in Abhängigkeit von der Beschaffenheit des empirischen Relativs bei der homomorphen Abbildung in ein numerisches Relativ Messwerte auf einer Skala entstehen, die unterschiedliche Aussagekraft haben. Man spricht hier auch vom **Messniveau bzw. Skalenniveau** („level of measurement“, „scales of measure“).

Mess- bzw. Skalenniveaus – Messungen in der quantitativen Sozialforschung führen in Abhängigkeit von der Beschaffenheit des empirischen Relativs zu unterschiedlich informationsreichen Messwerten. Man unterscheidet **vier Mess- bzw. Skalenniveaus** („levels of measurement“, „scales of measurement“) nach aufsteigendem Informationsgehalt: Nominalskala („nominal scale“), Ordinal-/Rangskala („ordinal scale“), Intervallskala („interval scale“) und Verhältnis-/Ratioskala („ratio scale“). Im Ergebnis spricht man von nominal-, ordinal-, intervall- oder verhältnisskalierten Messwerten bzw. Daten. Intervall- und Verhältnisskala werden häufig unter den Begriffen „Kardinalskala“ bzw. „metrische Skala“ („metric scale“) zusammengefasst, da sie mit denselben statistischen Verfahren ausgewertet werden können, während für ordinalskalierte und nominalskalierte Daten jeweils spezifische statistische Auswertungsverfahren indiziert sind.

Es werden nun die vier wichtigsten Skalenarten vorgestellt, die bereits in den 1940er-Jahren von dem **Experimentalpsychologen Stanley Smith Stevens** beschrieben wurden (Stevens 1946). Dabei werden die für eine Skalenart jeweils gebräuchlichste Messstruktur sowie die Art ihrer Repräsentation im numerischen Relativ kurz erläutert. Auf eine Behandlung der Axiomatik der Skalen wird hier unter Verweis auf die Spezialliteratur zur Messtheorie verzichtet (z. B. Krantz, Luce, Suppes, & Tversky, 2006a, 2006b; Mari 2005; Michell 2005; Orth 1983; Steyer & Eid, 1993; Suppes, Krantz, Luce, & Tversky, 2006; Wade Savage & Ehrlich, 1991). Die Behandlung der Skalen erfolgt hierarchisch, beginnend mit einfachen, relativ ungenauen Messungen auf dem Niveau der Nominalskala bis hin zu exakten, vor allem in den Naturwissenschaften gebräuchlichen – aber eben auch in den Sozialwissenschaften eingesetzten – Messungen auf Verhältnisskalenniveau (z. B. physiologische Messverfahren).

8.4.2 Nominalskala

Beispiele für nominalskalierte Merkmale wurden bereits erwähnt: Bei der Parteimitgliedschaft (Parteimitglied, kein Parteimitglied), dem Autobesitz (Auto, kein Auto), der Automarke (VW, BMW, Audi, Mercedes-Benz, Opel, Ford, Renault etc.), dem Familienstand (ledig, verheiratet, geschieden, verwitwet), der Händigkeit (Linkshänder, Rechtshänder) oder dem aktuellen Raucherstatus (Raucher, Nichtraucher) einer Person handelt es sich um zweifach oder mehrfach gestufte nominalskalierte Variablen. Eine Nominalskala bzw. kategoriale Skala setzt ein empirisches Relativ mit einer gültigen **Äquivalenzrelation** voraus. Äquivalente Objekte bzw. Objekte mit identischen Merkmalsausprägungen erhalten identische Zahlen, und Objekte mit verschiedenen Merkmalsausprägungen erhalten verschiedene Zahlen. Die Zuweisung von Zahlen entspricht bei der Nominalskala der Zuweisung von Namen zu den unterschiedlichen Merkmalsausprägungen.

Nominalskala – Eine Nominalskala („nominal scale", „categorial scale") ordnet den Objekten eines empirischen Relativs Zahlen zu, die so geartet sind, dass Objekte mit gleicher Merkmalsausprägung gleiche Zahlen und Objekte mit verschiedener Merkmalsausprägung verschiedene Zahlen erhalten.

Ein empirisches Relativ mit einer gültigen Äquivalenzrelation bezeichnet man als eine **klassifikatorische Messstruktur**. Die Auswahl der Zahlen, die den Objektklassen zugeordnet werden, ist für eine Nominalskala unerheblich, solange gewährleistet ist, dass äquivalente Objekte durch identische und nichtäquivalente Objekte durch verschiedene Zahlen abgebildet werden. Vier verschiedenen Parteien könnten also die Zahlen 1, 2, 3 und 4 zugeordnet werden oder auch andere Zahlen wie z. B. 2, 66, 5 und 1 005. Man sagt: Die quantitativen Aussagen einer Nominalskala sind gegenüber beliebigen eindeutigen Transformationen invariant.

Für Nominalskalen sind nur Aussagen über die Besetzungszahlen bzw. Häufigkeiten für Objektklassen bedeutsam. Dementsprechend beschränken sich mathematisch-statistische Operationen für Nominaldaten auf die Analyse von Häufigkeitsverteilungen (Bortz & Schuster, 2010, Kap. 9). Klassifikatorische Begriffe spielen in der qualitativen Forschung eine zentrale Rolle, zuweilen werden nominalskalierte Variablen auch als „**qualitative Variablen**" bezeichnet. Sie sind jedoch nicht mit qualitativen Daten im Sinne von nicht-numerischem Datenmaterial zu verwechseln. Eine zweifach gestufte nominalskalierte Variable heißt auch **binäre oder dichotome Variable** (z. B. Händigkeit: links- vs. rechtshändig, Serostatus: HIV positiv vs. HIV negativ), eine mehrfach gestufte nominale Variable wird als **polytome Variable** bezeichnet (z. B. Familienstand: ledig, verheiratet, geschieden, verwitwet, Studiengang: Psychologie, Soziologie, Erziehungswissenschaft, Kommunikationswissenschaft, Medizin). Messungen auf Nominalskalenniveau sind in der empirischen Human- und Sozialforschung sehr verbreitet und laufen meist auf die Darstellung und Analyse von absoluten Häufigkeiten sowie relativen Häufigkeiten bzw. Prozentzahlen hinaus.

Wenn von **nominalen, kategorialen oder qualitativen Merkmalen bzw. Kategorien** die Rede ist, so müssen diese hinsichtlich ihrer **Ausprägungen** drei Kriterien erfüllen.

1. **Genauigkeitskriterium**: Die Ausprägungen des Merkmals müssen exakt definiert sein. Für jedes Untersuchungsobjekt muss genau feststellbar sein, welche Merkmalsausprägung vorliegt. Dies erfolgt vorzugsweise durch eine operationale Definition, die Indikatoren und Messverfahren präzise festlegt. Ein nominales Merkmal wie „Kleidungsstil" wäre ungenügend operationalisiert, wenn man dafür die Ausprägungen „geschmackvoll gekleidet" vs. „geschmacklos gekleidet" vorsehen würde, aber eben nicht genau definierbar ist, worin sich geschmackvolle und geschmacklose Bekleidung unterscheiden. Eine Operationalisierung des Merkmals „Kleidungsstil" würde dagegen das Genauigkeitskriterium erfüllen, wenn die Ausprägungen „formale Kleidung" vs. „nichtformale Kleidung" erfasst werden und genau definiert ist, welche Merkmale die Kleidung erfüllen muss, um als formal zu gelten (z. B. Anzug oder Kostüm).
2. **Exklusivitätskriterium**: Die Ausprägungen müssen sich gegenseitig ausschließen. Ein Untersuchungsobjekt darf nicht gleichzeitig mehrere verschiedene Ausprägungen der Variable erfüllen. Ein nominales Merkmale wie „Berufstätigkeit" wäre ungenügend operationalisiert, wenn man dafür Ausprägungen wie „Freiberufler", „Angestellter", „Beamter" und „Manager" vorsieht, weil eine Person, die ein eigenes Unternehmen führt, dann sowohl als „Freiberufler" als auch als „Manager" eingeordnet werden kann. Das Exklusivitätskriterium wird meist dadurch verletzt, dass in einem Merkmal mehrere inhaltliche Dimensionen vermischt werden (z. B. Art des Beschäftigungsverhältnisses und Art der Tätigkeit). Eine Operationalisierung des Merkmals „Art des Beschäftigungsverhältnisses" mit den Ausprägungen „selbstständig", „angestellt", „verbeamtet" würde dem Exklusivitätskriterium genügen.
3. **Exhaustivitätskriterium**: Die Ausprägungen müssen das Merkmal erschöpfend beschreiben. Die Ausprägungen müssen so beschaffen sein, dass jedes Untersuchungsobjekt einer Merkmalskategorie zugeordnet werden kann. Eine Operationalisierung des Merkmals „Art des Beschäftigungsverhältnisses" mit den Ausprägungen „selbstständig", „angestellt", „verbeamtet" würde dem Exhaustivitätskriterium nicht genügen, falls auch Personen untersucht werden sollen, die aktuell über kein Beschäftigungsverhältnis verfügen. Dann müsste die Ausprägung „aktuell kein Beschäftigungsverhältnis" hinzugefügt werden. Um das Exhaustivitätskriterium zu erfüllen wird in der Praxis oftmals mit einer Sammelausprägung wie „Anderes" bzw. „Sonstiges" gearbeitet. Sei es, weil bestimmte Sonderfälle nicht im Vorfeld antizipiert werden können oder auch weil aus Platzgründen nur die wichtigsten Ausprägungen ausdrücklich genannt werden sollen (z. B. Beziehungsstatus: Single, feste Beziehung, Ehe, anderes; sexuelle Orientierung: heterosexuell, homosexuell, bisexuell, anderes). Die Gruppe „anderes" oder „sonstiges" sollte in der Regel nicht sehr stark besetzt sein, andernfalls wäre dies ein Hinweis darauf, dass weitere inhaltliche Ausprägungen definiert werden müssen.

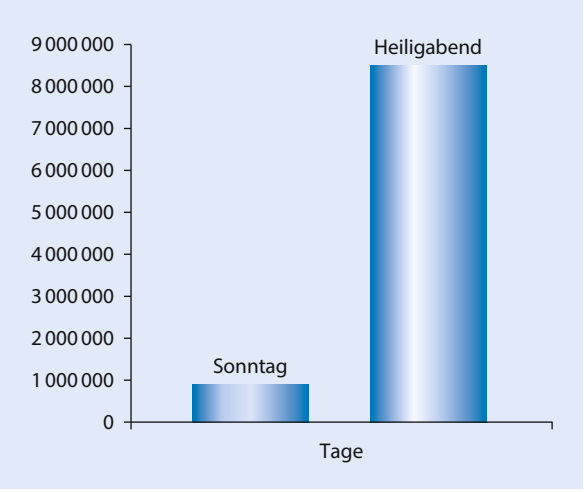


Abb. 8.1 Absolute Häufigkeitsangaben zur Besucherzahl evangelischer Gottesdienste in Deutschland laut Kirchenstatistik 2011. (Mod. nach EKD 2013, S. 14. Mit freundlicher Genehmigung der Evangelischen Kirche in Deutschland)

Die Messung nominaler Variablen lässt sich nicht nur auf Personen, sondern auch auf andere Untersuchungsobjekte anwenden (z. B. könnte im Rahmen der quantitativen Inhaltsanalyse von Medienangeboten erfasst werden, wie oft in unterschiedlichen Fernsehprogrammen „Schleichwerbung" vorkommt und welche Produkte und Marken dabei beworben werden, ▶ Abschn. 10.6.3). Die Messung auf Nominalskalenniveau läuft im Wesentlichen auf das **Zählen der Untersuchungsobjekte** hinaus, die in die jeweiligen Ausprägungen des Merkmals fallen. Bei einem dichotomen Merkmal, dessen Ausprägungen exhaustiv und exklusiv sind, reicht die Zählung einer Ausprägung, da alle anderen Untersuchungsobjekte dann automatisch in die andere Ausprägung fallen. Erfasst man etwa, wie viele Personen an einem bestimmten Sonn- oder Feiertag einen evangelischen Gottesdienst besuchen, so weiß man umgekehrt, dass der Rest der Bevölkerung dann offenbar keinen evangelischen Gottesdienst besucht hat. Die evangelische Kirchenstatistik für 2011 meldet rund 900 000 Gottesdienstbesucherinnen und -besucher für einen normalen Sonntag und rund 8.5 Millionen für Heiligabend (EKD 2013; Abb. 8.1).

Legt man für 2011 eine Gesamtbevölkerungszahl von rund 70.8 Mio. Deutschen über 14 Jahre zugrunde, so ergibt sich eine relative Häufigkeit von $0.9/70.8 = 1.3\,\%$ Besuchern des evangelischen Gottesdienstes vs. 98.7 % Nichtbesuchern an einem normalen Sonntag. An Heiligabend steigt der Besucheranteil in evangelischen Gottesdiensten auf bis zu 12 %, was dann auf der anderen Seite 88 % Nichtbesuchern entspricht (Tab. 8.7).

Für die Interpretation nominalskalierter Daten bzw. Häufigkeitsdaten sind **absolute Häufigkeiten** allein oft-

Tabelle 8.7 Absolute und relative Häufigkeiten des evangelischen Gottesdienstbesuches. (Daten aus: EKD 2013, S. 14)

Gottesdienstbesuch (binäres Merkmal)	Sonntag	Heiligabend
Besuch des evangelischen Gottesdienstes	900 000 (1.3 %)	8.5 Mio. (12 %)
Kein Besuch des evangelischen Gottesdienstes	69.9 Mio. (98.7 %)	62.3 Mio. (88 %)
Summe	70.8 Mio. (100 %)	70.8 Mio. (100 %)

mals nicht so aussagekräftig wie **relative Häufigkeiten bzw. Prozentwerte**. Dabei kommt es stark darauf an, welche Basis man zur Berechnung der relativen Häufigkeiten wählt. So kann die absolute Zahl der Kirchgänger wie in Tab. 8.7 an der gesamten Bevölkerung relativiert werden, was zu sehr kleinen Anteilswerten führt. Je nach inhaltlichem Erkenntnisinteresse wäre auch eine Relativierung an der evangelischen Bevölkerung möglich.

Nominalskalierte Merkmale werden deskriptivstatistisch über absolute und relative Häufigkeiten bzw. Prozentwerte ausgewertet. Zur Hypothesenprüfung werden als inferenzstatistische Verfahren vor allem **Chi-Quadrat-Verfahren** eingesetzt. Zudem werden nominalskalierte Daten z. B. über die Konfigurationsfrequenzanalyse (KFA) sowie über loglineare Modelle und spezifische Korrelationstest wie z. B. die punkt-biseriale und die punkt-tetrachorische Korrelation ausgewertet (Bortz & Schuster, 2010; Leonhart 2009).

8.4.3 Ordinalskala

Ordinalskalierte Variablen sind z. B. die Position in einer Unternehmenshierarchie (z. B. Lower Management, Middle Management, Top Management), der Dienstgrad beim Militär (z. B. Mannschaftsgrad, Unteroffiziersgrad, Offiziersgrad), die Platzierung bei einem Sportturnier (erster, zweiter, dritter, vierter etc. Platz), die durch Kundenurteile gebildete Beliebtheitsrangreihe von Urlaubsorten (erste, zweite, dritte etc. Präferenz bei Reisezielen), der formale Schulabschluss (z. B. kein Schulabschluss, Hauptschulabschluss, Realschulabschluss, Abitur) oder die Intensität des Rauchens (Nichtraucher, Gelegenheitsraucher, regelmäßiger Raucher). Eine Ordinalskala erfordert ein empirisches Relativ, für dessen Objektmenge eine sog. **schwache Ordnungsrelation** gilt. Dies bedeutet, dass bei einem beliebigen Objektpaar a und b entscheidbar sein muss, welches Objekt über das andere bzgl. eines untersuchten Kriteriums dominiert oder ob beide Objekte hinsichtlich des betrachteten Merkmals äquiva-

lent sind. Ferner ist die bereits in ▶ Messtheorie erwähnte Transitivität gefordert, nach der bei Dominanz von a über b und bei Dominanz von b über c das Objekt a auch über c dominieren muss. Dominiert ein Objekt a über ein Objekt b, so erhält das Objekt a eine Zahl, die größer ist als die dem Objekt b zugeordnete Zahl. Sind Objekte äquivalent, erhalten sie eine identische Zahl. Einer Ordinalskala ist die Rangfolge der untersuchten Objekte bzgl. eines Dominanzkriteriums zu entnehmen (z. B. Beliebtheit von Schülern, gesellschaftliches Prestige von Berufen, Schwere von Strafdelikten). Eine Ordinalskala wird deshalb auch **Rangskala** genannt, wobei äquivalente Objekte sog. Verbundränge erforderlich machen.

8

Ordinalskala (Rangskala) – Eine Ordinalskala („ordinal scale") ordnet den Objekten eines empirischen Relativs Zahlen (Rangzahlen) zu, die so geartet sind, dass von jeweils zwei Objekten das dominierende Objekt die größere Zahl erhält. Bei Äquivalenz der Merkmalsausprägung sind die Zahlen identisch (Verbundränge).

Messungen auf einer Ordinalskala sind eindeutig bis auf hier zulässige **monotone Transformationen**, also Transformationen, durch die die Größer-Kleiner-Relationen der Objektmessungen nicht verändert werden (rangerhaltende Transformation). Dementsprechend sind diejenigen quantitativen Aussagen bedeutsam, die gegenüber monotonen bzw. rangerhaltenden Transformationen invariant sind. Die statistische Analyse von Ordinaldaten läuft also auf die Auswertung von Ranginformationen hinaus, über die z. B. bei Bortz, Lienert und Boehnke (2008, Kap. 6) oder bei Bortz und Lienert (2008, Kap. 3) berichtet wird. Ordinalskalierte Daten erfordern besondere statistische Auswertungsverfahren (mit ihnen lässt sich u. a. kein sinnvoll interpretierbarer Mittelwert berechnen) und werden deswegen in der Praxis – sofern möglich – vermieden, indem eine Messung auf Intervallskalenniveau favorisiert wird (zur Abgrenzung zwischen Ordinal- und Intervallskala ▶ Abschn. 8.4.4).

Während nominale Daten dadurch gesammelt werden, dass man die Anzahl der Objekte **zählt**, auf die jeweils die unterschiedlichen Ausprägungen der Variable zutreffen, ist für die Messung auf Ordinalskalenniveau nicht das Zählen, sondern das **Urteilen** besonders einschlägig. Die Urteile sollen es erlauben, die untersuchten Objekte, Ereignisse oder Verhaltensweisen hinsichtlich der Ausprägung des betrachteten Merkmals in eine Rangreihe zu bringen. Dabei ist die menschliche Urteilsfähigkeit entscheidend. Wir unterscheiden zwei Methoden, um **originäre subjektive Rangreihen** aufzustellen: direkte und indirekte Rangordnungen. Darüber hinaus lassen sich sog. **objektive Rangreihen** bilden, indem man kardinalskalierte Daten in Rangdaten überführt (▶ Abschn. 8.4.6).

Direkte Rangordnungen

Im einfachsten Fall entstehen ordinalskalierte Daten durch eine **direkte Rangordnung**. Man spricht auch von einer **originären Rangreihe**. Die Untersuchungsobjekte werden dabei gemäß ihren Ausprägungen auf dem interessierenden Merkmal in eine Rangreihe gebracht. Das Ordnen von Untersuchungseinheiten nach einem vorgegebenen Merkmal stellt eine auch im Alltag geläufige Form des Urteilens dar, die vor allem im Zusammenhang mit Präferenzen eine große Bedeutung hat. Wer im Vorfeld des Hauskaufs mehrere Objekte in Augenschein nimmt, im Rahmen des Online-Datings Kontakte mit mehreren Flirtwilligen aufnimmt oder mögliche Praktikumsplätze vergleicht, wird diese nicht selten in eine Rangreihe bringen. Rangdaten können sich nicht nur auf **Urteile über Objekte** beziehen (z. B. „Bringen Sie folgende Liste von Urlaubsorten in eine Rangreihe, indem Sie den Ort, an den Sie am liebsten fahren würden, mit 1, den zweitliebsten Urlaubsort mit 2 etc. beschriften"), sondern auch auf **Selbsteinschätzungen auf einer Rangskala** („Sind Sie Nichtraucher, Gelegenheitsraucher, regelmäßiger Raucher oder Kettenraucher"). Direkte Rangordnungen können sich auch aus **Reihenfolgen von Abläufen** ergeben (▶ Studienbeispiel „Direkte Rangordnungen als Abfolgen von Verhaltensweisen").

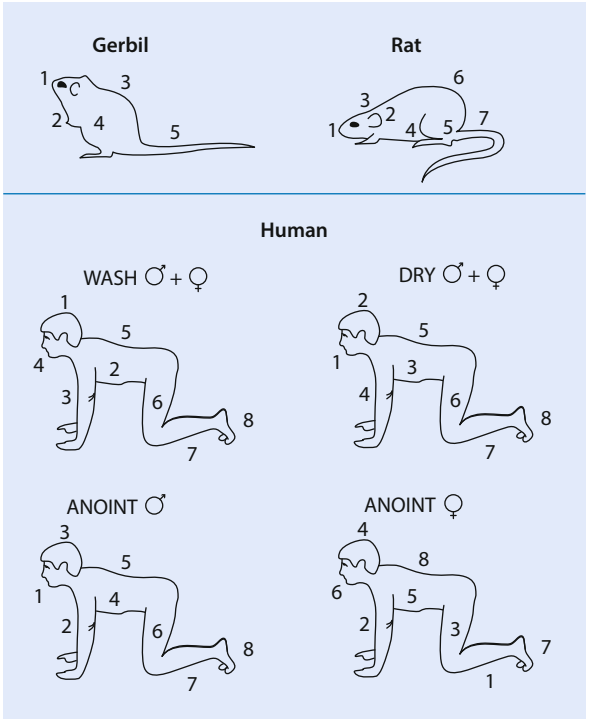
Die verbreitetste Form der direkten Rangordnung entsteht durch die bereits angesprochenen Urteile über Personen oder Objekte. Etwa wenn in der Notaufnahme Patienten gemäß Dringlichkeit der Behandlung oder Bewerber gemäß Eignung für eine berufliche Position in eine Rangreihe gebracht werden. Das Aufstellen einer Rangordnung geht von der Vorstellung aus, dass sich die untersuchten Objekte hinsichtlich der Ausprägung mindestens eines eindeutig definierten Merkmals unterscheiden. Der Urteiler weist demjenigen Objekt, bei dem das Merkmal am stärksten ausgeprägt ist, Rangplatz 1 zu, das Objekt mit der zweitstärksten Merkmalsausprägung erhält Rangplatz 2 und so fort bis hin zum letzten (dem n-ten) Objekt, das Rangplatz n erhält. Die so ermittelten Werte stellen eine Rangskala oder **Ordinalskala** dar. Objekte mit gleichen Merkmalsausprägungen erhalten sog. **Verbundränge** („ties"). Verbundränge sind immer erforderlich, wenn die Anzahl der Merkmalsabstufungen kleiner ist als die Anzahl der Objekte, die in eine Rangreihe gebracht werden sollen.

Die Grenzen der Urteilskapazität werden bei direkten Rangordnungen mit zunehmender **Anzahl der zu ordnenden Objekte** rasch erreicht. Wie viele Objekte noch sinnvoll in eine direkte Rangreihe gebracht werden können, hängt von der Komplexität des untersuchten Merkmals und der Kompetenz der Urteilenden ab. So dürfte die Anzahl verlässlich nach ihrem Gewicht zu ordnender Gegenstände (von sehr schwer bis feder-

Studienbeispiel

Direkte Rangordnungen als Abfolgen von Verhaltensweisen

Gegenstand der **komparativen Psychologie** („comparative psychology") ist der Vergleich von Verhaltensweisen zwischen verschiedenen Arten. Eine Studie zum Selbstpflegeverhalten ging der Frage nach, in welcher Reihenfolge Menschen ihre **Körperpflege** (Waschen, Abtrocknen, Eincremen) betreiben und ob diese Reihenfolge sich bei Tieren – etwa beim Putzverhalten von Wüstenmaus oder Laborratte – wiederfindet (Young & Thiessen, 1991). Die einzelnen menschlichen Körperteile wurden also gemäß der Reihenfolge beim Waschen, Abtrocknen und Eincremen in eine Rangreihe gebracht von 1, dem zuerst gepflegten Körperteil, bis 8, dem zuletzt gepflegten Körperteil. Diese direkten Rangdaten wurden von einer Stichprobe von Männern und Frauen mittels Befragung über drei Tage hinweg erfasst. Die Reihenfolge beim Putzverhalten der Tiere wurde beobachtet. Es zeigten sich tatsächlich über verschiedene Tage und Personen hinweg relativ stabile Rangreihen beim Menschen, die denen der untersuchten Nagetiere insofern ähnelten, dass die Pflegerichtung tendenziell vom Kopf zu den Füßen verläuft. Lediglich beim Eincremen zeigten sich Geschlechtsunterschiede in der Weise, dass die untersuchten Männer sich mehrheitlich zuerst das Gesicht, Frauen zuerst die Unterschenkel eincremten (◘ Abb. 8.2).



◘ **Abb. 8.2** Die Reihenfolgen des Pflegens einzelner Körperteile bei Mensch, Wüstenmaus und Laborratte stellen jeweils direkte Rangordnungen dar. (Aus Young & Thiessen, 1991, S. 341. Copyright © 2012 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA)

leicht) sicherlich größer sein als die maximale Anzahl von Politikerinnen und Politikern, die problemlos nach dem Merkmal „politischer Sachverstand" in eine Rangreihe gebracht werden können. Die Diskriminationsfähigkeit (Bewertungsgenauigkeit) der Urteilenden hängt dabei auch von ihren Sachkenntnissen ab; so wird es politisch gut Informierten leichter fallen, eine große Zahl von Politiker/innen rangzuordnen als weniger Informierten. Vorversuche oder Selbstversuche stellen geeignete Mittel dar, um bei einem konkreten Rangordnungsproblem die Höchstzahl sinnvoll zu ordnender Objekte festzustellen. In der Marktforschung geht man oft zweistufig vor und lässt Urteilsobjekte (z. B. Geschmacksrichtungen, Werbesujets, Produktideen) zunächst in Fokusgruppen diskutieren, um diejenigen Alternativen auszusortieren, die generell aus Sicht der Konsumenten „durchfallen". Nur die Urteilsobjekte, die prinzipiell positiv bewertet werden, werden dann in einer größer angelegten quantitativen Studie in eine Rangreihe gebracht.

Übersteigt die Anzahl der zu ordnenden Objekte die Diskriminationsfähigkeit der Urteilenden, so werden nicht alle Objekte in eine fortlaufende Rangreihe gebracht, sondern es wird jedes einzelne Objekt hinsichtlich seiner Merkmalsintensität auf einer **ordinalen Ratingskala** eingeordnet (z. B. schwach – mittel – stark). Werden hier nur drei oder vier Abstufungen vorgenommen, betrachtet man die Ratingskala meist als ordinale Messung, während bei fünf und mehr annähernd gleichabständigen Stufen auf der Ratingskala heute oft Intervallskalenniveau zugrunde gelegt wird (▶ Abschn. 8.4.4), was hinsichtlich der gewählten statistischen Analysemethoden und der Interpretierbarkeit der Befunde von Bedeutung ist. Gemäß der Methode der sukzessiven Intervalle und dem **Law of Categorial Judgement** können Rangplätze einer direkten Rangreihe in intervallskalierte Skalenwerte rechnerisch hochtransformiert werden (▶ Abschn. 8.4.6).

Indirekte Rangordnungen

Eine indirekte Rangordnung wird gebildet, indem alle Objekte paarweise miteinander verglichen werden. Indirekte Rangordnungen sind in der Regel **genauer als direkte Rangordnungen**, dafür aber auch **deutlich aufwändiger**. In der Psychometrie wird teilweise argumentiert, dass Paarvergleiche der Urteilsabgabe auf Ratingskalen grundsätzlich überlegen sind, weil die Entscheidung zwischen zwei Optionen eine evolutionär verankerte und gut beherrschte Form der Beurteilung darstellt, während die Abgabe abgestufter verbaler oder symbolischer Urteile auf einer Ratingskala meist kognitiv überfordernd sei. Gegen diesen Einwand spricht indessen die Tatsache, dass sich Ratingskalen in der Forschungspraxis gut

bewährt haben und viele Menschen damit inzwischen vertraut sind (zu Ratingskalen ▶ Abschn. 8.4.4).

Bei einem **Dominanzpaarvergleich** werden die Urteilenden aufgefordert anzugeben, bei welchem von zwei Objekten das untersuchte Merkmal stärker ausgeprägt ist bzw. welches Objekt bzgl. des Merkmals „dominiert" (Beispiele: Welche von zwei Aufgaben ist schwieriger, welcher von zwei Filmen ist interessanter, welche von zwei Krankheiten ist schmerzhafter etc.). Dominanzpaarvergleiche sind einfache Urteilsaufgaben, die allerdings sehr aufwändig werden, wenn viele Objekte zu beurteilen sind. Werden n Objekte untersucht, müssen für einen vollständigen Paarvergleich, bei dem jedes Objekt mit jedem anderen Objekt verglichen wird,

$$\binom{n}{2} = \frac{n \cdot (n-1)}{2} \tag{8.1}$$

Paarvergleichsurteile abgegeben werden (bei $n = 10$ sind damit 45 Paarvergleichsurteile erforderlich).

Ein vollständiger Paarvergleich von n Objekten führt zu Angaben darüber, wie häufig jedes Objekt den übrigen Objekten vorgezogen wurde. Ordnet man diesen Häufigkeiten nach ihrer Größe Rangzahlen zu, erhält man eine Rangordnung der untersuchten Objekte.

Ein kleines Beispiel soll dieses Verfahren erläutern. Nehmen wir an, es sollen 7 Urlaubsorte nach ihrer Attraktivität in eine Rangreihe gebracht werden. Der vollständige Paarvergleich dieser 7 Orte (nennen wir sie einfachheitshalber A, B, C, D, E, F und G) durch eine Urteilerin führte zu folgenden Präferenzhäufigkeiten (eine 1 in einer Zelle bedeutet, dass der Ort in der Spalte gegenüber dem Ort in der Zeile dominiert; ◘ Tab. 8.8).

Insgesamt wurden also von der Urteilerin

$$\binom{7}{2} = 21$$

Paarvergleichsurteile abgegeben. Ort D wurde in der Summe 6 Mal und somit am häufigsten gegenüber anderen Orten bevorzugt. Ort D erhält damit den Rangplatz 1. Die Orte A und F teilen sich die Rangplätze 2 und 3 (verbundener Rangplatz: 2.5), die Orte B und G teilen sich die Rangplätze 4 und 5 (verbundener Rangplatz 4.5), dem Ort C wird der Rangplatz 6 zugewiesen und Ort E erhält als der am wenigsten attraktive Ort Rangplatz 7.

Bei Dominanzpaarvergleichsurteilen erfährt man „nur" etwas über die relative Ausprägung von Objekten auf einer Urteilsskala und nichts über die absolute Ausprägung. Wenn im Beispiel Ort D als attraktiver beurteilt wird als Ort A, sagt dies nichts darüber aus, ob die beiden Orte überhaupt als besonders attraktiv erlebt wurden oder beide aus Sicht der Urteilerin eher uninteressante Ferienziele darstellen. Mit diesem Problem des Ursprungs bzw. der Verankerung von Paarvergleichsurteilen befasst sich eine Arbeit von Böckenholt (2004).

◘ **Tabelle 8.8** Bewertungsmatrix eines Dominanzpaarvergleichs für die Urlaubsorte A bis G

	A	B	C	D	E	F	G
A	–	0	0	1	0	1	0
B	1	–	0	1	0	0	1
C	1	1	–	1	0	1	1
D	0	0	0	–	0	0	0
E	1	1	1	1	–	1	1
F	0	1	0	1	0	–	0
G	1	0	0	1	0	1	–
Summe	4	3	1	6	0	4	3
Rangplätze	**2.5**	**4.5**	**6**	**1**	**7**	**2.5**	**4.5**

Mehrere Urteiler

Wird der Paarvergleich von mehreren Urteilern durchgeführt, resultiert deren gemeinsame Rangreihe durch Summation der individuellen Präferenzhäufigkeiten. Hierfür fertigt man sinnvollerweise eine Tabelle an, der zusätzlich entnommen werden kann, von wie vielen Urteilenden ein Objekt einem anderen vorgezogen wurde (Dominanzmatrix ▶ Dominanzpaarvergleiche).

Konsistenz und Konkordanz

Wie bei direkten Rangordnungen wird auch bei der indirekten Ermittlung einer Rangskala über Paarvergleiche vorausgesetzt, dass die Objekte bzgl. des untersuchten Merkmals ordinale Relationen aufweisen. Führen wiederholte Paarvergleiche derselben Objekte zu verschiedenen Rangreihen, ist diese Voraussetzung verletzt, es sei denn, man toleriert die Abweichungen als unsystematische Urteilsfehler.

Eine Verletzung der ordinalen Modellannahme liegt auch vor, wenn sog. **zirkuläre Triaden** (Kendall 1955) oder **intransitive** Urteile auftreten (auch Dunn-Rankin, Knezek, Wallace, & Zhang, 2004). Wird z. B. von zwei Gemälden (A, B) A als das schönere vorgezogen (A > B) und zudem Gemälde B einem dritten Bild C vorgezogen (B > C), müsste man folgern, dass A auch C vorgezogen wird (A > C). In der Praxis kommt es jedoch nicht selten zu dem scheinbar inkonsistenten Urteil C > A. Nachlässigkeit auf Seiten der Urteilenden und/oder nur gering-

Dominanzpaarvergleiche

Ermittlung der Beliebtheit von Schulfächern durch Dominanzpaarvergleiche

Dreißig Schüler wurden gebeten, in einem vollständigen Paarvergleich ihre Präferenzen für 5 Unterrichtsfächer anzugeben. Hierfür wurden für die Fächer Deutsch (De), Mathematik (Ma), Englisch (En), Sport (Sp) und Musik (Mu) alle 10 möglichen Paarkombinationen gebildet und jeder Schüler musste angeben, welches der jeweils 2 Fächer seiner Meinung nach das interessantere sei. Aus den Paarvergleichsurteilen resultierte folgende **Dominanzmatrix** (Begründung des Rechenganges im Text).

Die fett gesetzte Zahl in ◘ Tab. 8.9 gibt an, dass 24 Schüler Englisch interessanter fanden als Mathematik. Die Werte besagen, wie häufig die Fächer, die die Spalten bezeichnen, über die Fächer, die die Zeilen bezeichnen, „dominieren". Einander entsprechende Zellen ergänzen sich jeweils zu 30 (wenn von den 30 Schülern 24 Englisch interessanter als Mathematik beurteilten, dann fanden umgekehrt 6 Schüler offenbar Mathematik interessanter als Englisch).

Wollte man für alle Schüler eine gemeinsame Rangreihe bestimmen, wären die Spaltensummen nach ihrer Größe zu ordnen. **Es resultiert Sp > Mu > En > De > Ma.** Mit dem Law of Comparative Judgement können die Rangplätze in intervallskalierte Skalenwerte hochtransformiert werden (► Abschn. 8.4.6).

◘ **Tabelle 8.9** Dominanzmatrix, Dominanzpaarvergleich von Schulfächern

	De	Ma	En	Sp	Mu
De	–	10	12	24	22
Ma	20	–	**24**	26	23
En	18	6	–	19	20
Sp	6	4	11	–	14
Mu	8	7	10	16	–
Summe	52	27	57	85	79
Rangplätze	**4**	**5**	**3**	**1**	**2**

De Deutsch, *Ma* Mathematik, *En* Englisch, *Sp* Sport, *Mu* Musik

fügige Unterschiede in den Merkmalsausprägungen der Urteilsobjekte können für derartige „Urteilsfehler" verantwortlich sein.

Ein weiterer Grund für zirkuläre Triaden sind **mehrdimensionale Merkmale**, also Merkmale, die mehrere Aspekte oder Dimensionen aufweisen. So könnte die beim Gemäldevergleich aufgetretene zirkuläre Triade z. B. durch die Verwendung zweier Aspekte des Merkmals „Schönheit" zustande gekommen sein. Beim Vergleich der Bilder A und B wurde besonders auf die farbliche Gestaltung und beim Vergleich der Bilder B und C auf eine harmonische Raumaufteilung geachtet. Wird nun beim Vergleich der Bilder A und C erneut die farbliche Gestaltung (oder ein dritter Schönheitsaspekt) betont, kann es zu der oben aufgeführten intransitiven Urteilsweise kommen.

Über ein Verfahren, das die Zufälligkeit des Auftretens zirkulärer Triaden bzw. die Konsistenz der Paarvergleichsurteile überprüft, wird z. B. bei Bortz, Lienert und Boehncke (2008; Abschn. 9.5) berichtet (Dunn-Rankin, et al., 2004; Knezek, Wallace, & Dunn-Rankin, 1998). Übersteigt die Anzahl zirkulärer Triaden die unter Zufallsbedingungen zu erwartende Anzahl, muss man davon ausgehen, dass das untersuchte Merkmal mehrdimensional ist – es sei denn, die intransitiven Urteile sind auf Nachlässigkeit der Urteilenden (z. B. Zeitdruck, Motivationsmangel) zurückzuführen. Mit der statistischen Methode der Multidimensionalen Skalierung (MDS) lassen sich Ähnlichkeitsurteile über Objekte hinsichtlich der zugrunde liegenden Bewertungsdimensionen auswerten (zu explorativen multivariaten Auswertungstechniken ► Abschn. 12.3.1). Über einen Ansatz zur Überprüfung kognitiver Faktoren, die das Auftreten zirkulärer Triaden bedingen können, berichtet Böckenholt (2001).

Wird ein vollständiger Paarvergleich von mehreren Urteilern durchgeführt, informiert ein Verfahren von Kendall (1955) über die Güte der **Urteilerübereinstimmung** bzw. die **Urteilskonkordanz** (Bortz, et al., 2008, Abschn. 9.5.2). Eine Zusammenfassung individueller Paarvergleichsurteile setzt einen hohen Konkordanzwert voraus.

Stimmen die Paarvergleichsurteile der verschiedenen Urteilenden nicht überein, kann auch dies ein Hinweis auf Mehrdimensionalität des Merkmals sein, die in diesem Falle jedoch nicht intraindividuell, sondern interindividuell zum Tragen kommt. Bezogen auf den oben erwähnten Schönheitspaarvergleich von Gemälden hieße dies z. B., dass verschiedene Urteilende in ihren (möglicherweise konsistenten bzw. transitiven) Urteilen verschiedene Schönheitsaspekte beachtet haben. Auch in diesem Falle wäre dem eindimensionalen Paarvergleich ein mehrdimensionales Analysemodell vorzuziehen, das gleichzeitig individuelle Unterschiede in der Nutzung von Urteilsdimensionen berücksichtigt. Da Geschmäcker – nicht nur in Bezug auf die Schönheit von Gemälden – bekanntlich verschieden sind, wird man konkordante Urteile umso eher erzielen, je genauer man festlegt, hinsichtlich welcher Aspekte die Objekte im Einzelnen beurteilt werden sollen. Pauschale Bewertungen hinsichtlich „Schönheit" sind hier sicher nicht optimal.

Man beachte, dass eine hohe Konkordanz nicht an konsistente Individualurteile gebunden ist, denn eine hohe Konkordanz läge auch dann vor, wenn alle Urteilenden einheitlich inkonsistent urteilen. Weitere Informationen zu eindimensionalen Skalierungsverfahren findet man bei Borg, Müller und Staufenbiel (1990).

Konsistenz – Unter Konsistenz („intra-rater consistency", „intra-rater reliability") versteht man die Widerspruchsfreiheit der Paarvergleichsurteile, die *eine* Person über die Urteilsobjekte abgibt.

Konkordanz bzw. Urteilerübereinstimmung – Mit Konkordanz bzw. Urteilerübereinstimmung („inter-rater concordance", „inter-rater agreement", „inter-rater reliability") ist die Übereinstimmung der Paarvergleichsurteile von *zwei oder mehr* Urteilenden gemeint.

8

8.4.4 Intervallskala

Viele psychologische und soziale Merkmale von Personen wie etwa Stimmungen, Einstellungen, Vorurteile, Persönlichkeitseigenschaften, Vorlieben, Lernerfolge oder Motive werden in der empirischen Sozialforschung auf Intervallskalenniveau gemessen. Dabei geht man davon aus, dass die Ausprägungen dieser Merkmale genauer quantifizierbar sind als über die Größer-Kleiner-Relation, die von einer Ordinalskala erfasst wird. Eine Intervallskala erfordert ein empirisches Relativ, für das eine **schwache Ordnungsstruktur der Dominanzrelationen** aller Objektpaare gilt. Anders als bei einer Ordinalskala, bei der die Frage, wie stark ein Objekt über ein anderes dominiert, unerheblich ist, wird hier also gefordert, dass die paarweisen Dominanzrelationen nach ihrer Stärke in eine Rangordnung gebracht werden können. Interpretieren wir eine Dominanzrelation für a und b als Merkmalsunterschied zwischen den Objekten a und b, dann impliziert die Existenz einer schwachen Ordnungsrelation der Objektpaare, dass die Größe des Unterschiedes bei jedem Objektpaar bekannt ist.

Dieses empirische Relativ wird mit dem numerischen Relativ durch folgende Zuordnungsfunktion verknüpft: Wenn der Unterschied zwischen zwei Objekten a und b mindestens so groß ist wie der Unterschied zwischen zwei Objekten c und d, ist die Differenz der den Objekten a und b zugeordneten Zahlen $\phi(a) - \phi(b)$ mindestens so groß wie die Differenz der den Objekten c und d zugeordneten Zahlen $\phi(c) - \phi(d)$.

Intervallskala – Eine Intervallskala („interval scale") ordnet den Objekten eines empirischen Relativs Zahlen zu, die so geartet sind, dass die Rangordnung der Zahlendifferenzen zwischen je zwei Objekten der Rangordnung der Merkmalsunterschiede zwischen je zwei Objekten entspricht. Die Intervallskala zeichnet sich durch Äquidistanz bzw. Gleichabständigkeit der Messwerte aus.

Für eine Intervallskala gilt, dass gleich große Merkmalsunterschiede durch **äquidistante (gleichabständige) Zahlen** abgebildet werden, d. h., identische Messwertunterschiede zwischen Objektpaaren entsprechen identischen Merkmalsunterschieden. Hieraus folgt, dass Zahlenintervalle wie z. B. 1–2, 2–3, 3–4 etc. gleich große Merkmalsunterschiede abbilden.

Ein Beispiel für eine Intervallskala ist die Celsius-Skala. Der Temperaturunterschied zwischen 2 °C und 4 °C ist genauso groß wie z. B. der Temperaturunterschied zwischen 23 °C und 25 °C, und die Intervalle 1–2 °C, 2–3 °C, 3–4 °C etc. bilden gleich große Temperaturunterschiede ab. Man beachte, dass vergleichbare Aussagen für Ordinalskalen nicht gültig sind.

Eine Intervallskala ist eindeutig bis auf für sie zulässige **lineare Transformationen**:

$$\varphi' = \beta \cdot \varphi + \alpha \quad (\beta \neq 0). \tag{8.2}$$

Durch β und α werden die Einheit und der Ursprung der Intervallskala im numerischen Relativ festgelegt. Die Celsius-Skala z. B. wird durch folgende lineare Transformation in die Fahrenheit-Skala (F) überführt:

$$F = \frac{9}{5}C + 32. \tag{8.3}$$

Auch die Fahrenheit-Skala bildet identische Temperaturunterschiede durch äquidistante Zahlenintervalle ab.

Bei einer Intervallskala ist die Bedeutung einer numerischen Aussage gegenüber linearen Transformationen invariant. Dies gilt für Differenzen, Summen bzw. auch Mittelwerte von intervallskalierten Messwerten. Die am häufigsten eingesetzten statistischen Verfahren gehen für sinnvoll interpretierbare Ergebnisse von intervallskalierten Daten aus.

Zur Messung intervallskalierter psychologischer und sozialer Variablen greift man meist auf **Selbstauskünfte der Untersuchungsteilnehmenden** zurück, welche die jeweils interessierenden Erlebens- und Verhaltensweisen in ihrem Ausprägungsgrad auf **Ratingskalen** (engl. „rating" = Einschätzung) einstufen (z. B. „Ich bin sehr gläubig." – stimmt gar nicht – stimmt wenig – stimmt teils-teils – stimmt ziemlich – stimmt völlig). Aber auch **Fremdbeurteilungen** können über Ratingskalen erfasst werden, etwa wenn Führungskräfte die Mitarbeitenden oder Lehrkräfte die Lernenden bewerten (z. B. hinsichtlich ihrer Eignung oder Leistung auf einer Rating- bzw. Schulnotenskala). Die Ratingskala ist ein intervallskaliertes Messinstrument, wenn die vorgegebenen Antwortmöglichkeiten gleichabständig sind. Da Ratingskalen in der Forschungspraxis so eine große Bedeutung haben, müssen sie im Detail verstanden werden. Im Folgenden befassen wir uns mit

Tabelle 8.10 Empfehlungen zu verbalen Marken für Ratingskalen in englischer Sprache nach Rohrmann. (Mod. nach Rohrmann 2007)

Häufigkeit („frequency“)	never – seldom – sometimes – often – always
Intensität („intensity“)	not – a-little – moderately – quite a bit – very
Wahrscheinlichkeit („probability“)	certainly not – unlikely – about 50 : 50 – likely – for sure
Qualität („quality“)	bad – inadequate – fair – good – excellent
Zustimmung („agreement“)	fully disagree – mainly disagree – neutral – mainly agree – fully agree

- verschiedenen Varianten von Ratingskalen,
- messtheoretischen Problemen bei Ratingskalen,
- Urteilsfehlern beim Einsatz von Ratingskalen sowie
- der Anzahl der Urteilenden und der Urteilsobjekte beim Einsatz von Ratingskalen.

Varianten von Ratingskalen

Ratingskalen als Antwortformate für mündliche oder schriftliche Fragen bzw. Aussagen geben (durch Zahlen, verbale Beschreibungen, Beispiele etc.) **markierte Abschnitte eines Merkmalkontinuums** vor, die die Urteilenden als gleich groß bewerten sollen, d. h., man geht in der Regel davon aus, dass die **Stufen der Ratingskala** eine Intervallskala bilden. Die Urteilenden wählen diejenige Stufe der Ratingskala, die ihrem subjektiven Empfinden von der Merkmalsausprägung bei dem infrage stehenden Objekt entspricht. Einige methodische Varianten für Ratingskalen verdeutlicht ▶ Varianten von Ratingskalen.

Uni- und bipolare Ratingskalen

Unipolare Skalen werden am häufigsten eingesetzt, sofern es inhaltlich passt, können bipolare Skalen jedoch von Vorteil sein.

1. **Unipolare Ratingskalen:** Beispiel 1 und 2 in ▶ Varianten von Ratingskalen zeigen unipolare Ratingskalen, d. h., die Ratingskalen bilden graduell abgestuft die Intensität **eines** Merkmals ab. Dies ist besonders naheliegend für Merkmale, die keine negativen Werte bzw. keinen Gegenpol aufweisen (z. B. Ausmaß der Belästigung durch Lärm, Intensität von Kopfschmerzen, Häufigkeit von Verhaltensweisen).
2. **Bipolare Ratingskalen**: Demgegenüber präsentiert das Beispiel 3 bipolare Ratingskalen, bei denen jedes Skalenende ein anderes, gegensätzliches Merkmal repräsentiert (z. B. gespannt/gelöst; unzufrieden/zufrieden). Bipolare Skalen haben gegenüber unipolaren Skalen den Vorteil, dass sich die beiden gegensätzlichen Begriffe wechselseitig definieren, d. h., sie erhöhen die Präzision der Urteile (zur Eindeutigkeit bipolarer Skalen vgl. auch Kaplan 1972; Krebs & Hoffmeyer-Zlotnik, 2009; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991; Trommsdorff 1975). Bipolare Ratingskalen kommen z. B. im Semantischen Differenzial zum Einsatz (▶ Abschn. 8.6.4).

Darstellung der Skalenmarken

Die Abstufungen einer Ratingskala können durch **verbale Marken**, **numerische Marken**, **symbolische Marken** oder **ohne Marken** (in Form einer visuellen Analogskala) dargestellt werden. Nicht untypisch sind auch Kombinationen (z. B. verbale plus numerische Marken). Die Art der Darstellung sollte auf die Zielgruppe abgestimmt sein und ein zügiges Antworten ermöglichen. Gleichzeitig muss das gewünschte Skalenformat im Hinblick auf das Layout und die Länge eines Paper-Pencil-, Online- oder Mobilfragebogens praktisch umsetzbar sein.

a. Verbale Marken. Bei der verbalen Charakterisierung der numerischen Abstufungen von Ratingskalen (Beispiel 1 in ▶ Varianten von Ratingskalen; Tab. 8.10) ist darauf zu achten, dass die verwendeten Begriffe zumindest annähernd äquidistante Ausprägungen des Merkmalskontinuums markieren. Hierzu hat Rohrmann (1978) eine Untersuchung vorgelegt, die ergab, dass die Urteilenden bei 5-stufigen Skalen die folgenden sprachlichen Marken weitgehend als äquidistant auffassten. Hierbei ist die Balance zu finden, zwischen einer möglichst exakten Äquidistanz der Stufen einerseits und einer sprachlich eingängigen Formulierung andererseits.

- **Häufigkeit**
  Beispiel: Wie oft hat Ihr Kind Kopfschmerzen?
  nie – selten – gelegentlich – oft – immer
  sehr selten – selten – gelegentlich – oft – sehr oft
- **Intensität**
  Beispiel: Sind Sie mit Ihrem neuen Auto zufrieden?
  gar nicht – kaum – mittelmäßig – überwiegend – völlig
- **Wahrscheinlichkeit**
  Beispiel: Für wie wahrscheinlich halten Sie es, dass nach den nächsten Wahlen ein Regierungswechsel stattfindet?

Varianten von Ratingskalen

Mit den folgenden Beispielen soll die Vielfalt der Konstruktionsmöglichkeiten für Ratingskalen als Antwortskalen von standardisierten Fragebogenitems oder Interviewfragen verdeutlicht werden. Das Fragebogenitem besteht dabei jeweils aus einer Frage oder Aussage kombiniert mit der Ratingskala als Antwortformat. Teilweise wird dem Item auch eine kurze Instruktion (Ausfüllanleitung mit einer Erläuterung der Skala) beigefügt. Im Zuge der Online-Forschung vervielfältigen sich die Möglichkeiten, Skalen darzustellen. Originelle grafische und interaktive Formate (z. B. „Sniper-Skala" in Form einer Zielscheibe: Lütters 2008) sollen die Teilnahmemotivation erhöhen.

Beispiel 1: Unipolare verbale Ratingskala

Wie oft gehen Sie in die Kirche?

☐ nie
☐ sehr selten
☐ selten
☐ gelegentlich
☐ oft
☐ sehr oft

In der Mode kehrt alles wieder.

☐ stimmt gar nicht
☐ stimmt wenig
☐ stimmt teils-teils
☐ stimmt ziemlich
☐ stimmt völlig

Auszug aus der ASTS (Aktuelle Stimmungsskala von Dalbert 1992)

Instruktion. Nachfolgend finden Sie eine Liste mit Wörtern, die verschiedene Gefühle und Gefühlszustände beschreiben. Bitte lesen Sie sorgfältig jedes einzelne Wort und kreuzen Sie dann die Zahl an, die am besten Ihren Gefühlszustand **im Moment** beschreibt. Bitte machen Sie bei jeder Aussage ein Kreuz. 7 = sehr stark, 6 = stark, 5 = ziemlich, 4 = etwas, 3 = schwach, 2 = sehr schwach, 1 = überhaupt nicht.

1	zornig	7 6 5 4 3 2 1
2	abgeschlafft	7 6 5 4 3 2 1
3	unglücklich	7 6 5 4 3 2 1
4	traurig	7 6 5 4 3 2 1

Beispiel 2: Unipolare numerische Ratingskala

Bitte beurteilen Sie die Qualität dieser Lehrveranstaltung auf einer Schulnotenskala von Note 1 (sehr gut) bis Note 6 (ungenügend).
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

Beispiel 3: Bipolare Ratingskala

Instruktion. Im Folgenden zeige ich Ihnen Videoaufnahmen gruppentherapeutischer Sitzungen. Beurteilen Sie bitte die Gruppenatmosphäre bzgl. des Merkmals „gespannt gelöst". Hierfür steht Ihnen eine 5-stufige Skala zur Verfügung. Die einzelnen Skalenwerte haben folgende Bedeutung: 1 = Gruppenatmosphäre ist gespannt; 2 = Gruppenatmosphäre ist eher gespannt als gelöst; 3 = Gruppenatmosphäre ist weder gespannt noch gelöst; 4 = Gruppenatmosphäre ist eher gelöst als gespannt; 5 = Gruppenatmosphäre ist gelöst. Bitte urteilen Sie möglichst spontan! Uns interessiert Ihre persönliche Meinung, d. h., es gibt keine „richtigen" oder „falschen" Antworten.
Die Gruppenatmosphäre in der ersten Videoaufzeichnung empfinde ich als

gespannt	1	2	3	4	5	gelöst

Um die Polarisierung dieser bipolaren Skala besser zum Ausdruck zu bringen, können die Stufen auch in folgender Weise beziffert werden:

gespannt	-2	-1	0	+1	+2	gelöst

Das folgende Beispiel verzichtet auf eine Bezifferung der 5-stufigen Skala und arbeitet mit Minus- und Plus-Symbolen.

gespannt	--	-	·	+	++	gelöst

Beispiel 4: Symbolische Ratingskala

Instruktion. Auch in Ihrer Abteilung hat sich durch die Einführung von Gruppenarbeit in den letzten Monaten vieles verändert. Um diese Veränderungen zu erfassen, führen wir eine anonyme Mitarbeiterbefragung durch. Sie haben in dieser Umfrage die Möglichkeit, vollkommen anonym Ihre Meinung zu sagen. Das Ausfüllen des Fragebogens dauert etwa 10 Minuten und geht ganz leicht: Kreuzen Sie einfach die auf Sie zutreffenden Antworten an.
Wie zufrieden sind Sie mit der Beziehung zu Ihrem direkten Vorgesetzten?



Beispiel 5: Grafische Ratingskala bzw. Analogskala

Instruktion. Im Folgenden werden Ihnen verschiedene Berufspaare vorgelegt. Bitte beurteilen Sie bei jedem Berufspaar die Ähnlichkeit der beiden Berufe. Hierfür steht Ihnen eine Skala mit den Polen „extrem ähnlich" und „extrem unähnlich" zur Verfügung. Bitte markieren Sie durch ein Kreuz die von Ihnen eingeschätzte Ähnlichkeit.
Beispiel: Bäcker und Soziologe

extrem ähnlich ——————x—— extrem unähnlich

(Mit dieser Position des Kreuzes wird verdeutlicht, dass die im Beispiel zu vergleichenden Berufe von der urteilenden Person für sehr unähnlich gehalten werden.)

Beispiel 6: Ratingskala mit Ankerbeispielen

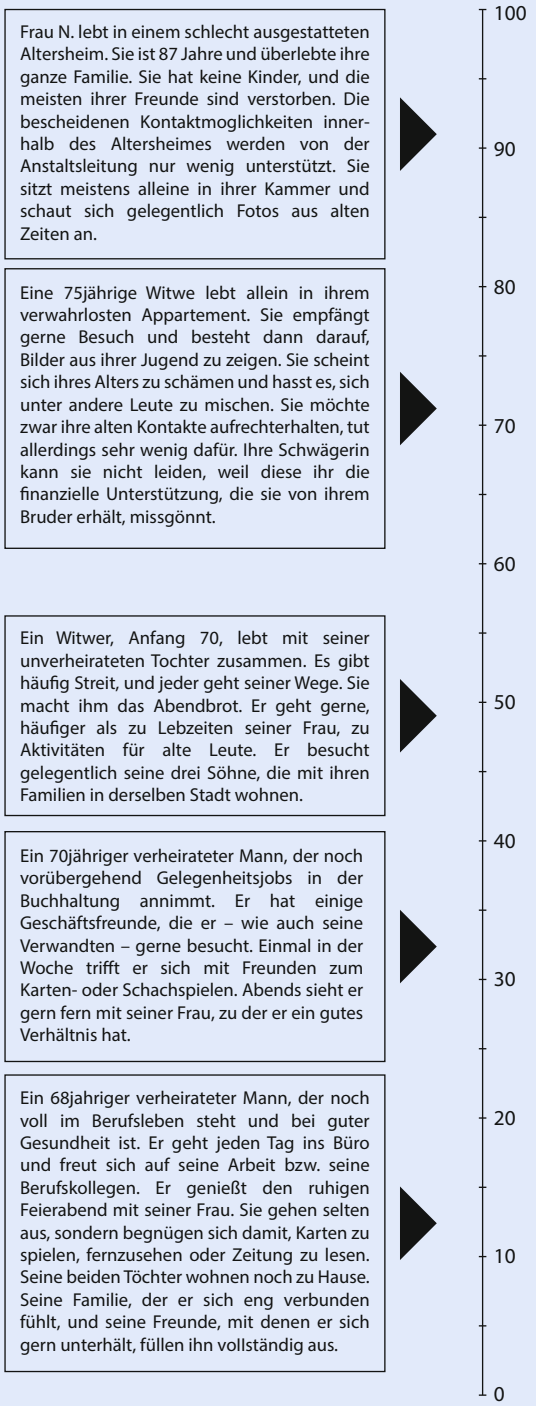
Instruktion. Im Folgenden geht es um die Beurteilung einiger Ihnen bekannter Strafgefangener. Bitte tragen Sie Ihren Eindruck von den zu beurteilenden Personen auf den folgenden Skalen ein. Verwenden Sie hierbei die Werte 3-2-1-0-1-2-3 als gleichmäßige Abstufungen des jeweils angesprochenen Merkmals.
Wie geht der Strafgefangene A mit Schwierigkeiten mit seinen Mitgefangenen um?

Er versucht, jeder Schwierigkeit aus dem Weg zu gehen.	3-2-1-0-1-2-3	Es reizt ihn, Schwierigkeiten zu überwinden.

(in Anlehnung an Waxweiler 1980).

Beispiel 7: Ratingskala mit Fallbeispielen

Instruktion. Im Folgenden geht es um die Einstufung der Hilfsbedürftigkeit Ihnen bekannter Personen. Hierfür steht Ihnen eine Skala mit 100 Punkten zur Verfügung. Je mehr Punkte Sie vergeben, desto hilfsbedürftiger ist Ihrer Ansicht nach die beurteilte Person. Um Ihnen die Arbeit mit der Skala zu erleichtern, wurden Personen mit unterschiedlicher Hilfsbedürftigkeit bereits einigen Punktwerten exemplarisch zugeordnet (nach Taylor et al., 1970; ◘ Abb. 8.3).



◘ **Abb. 8.3** Hilfsbedürftigkeit. (Mod. nach Taylor, Haefele, Thompson, & O'Donoghue, 1970. Mit freundlicher Genehmigung von Sage)

keinesfalls – wahrscheinlich nicht – vielleicht – ziemlich wahrscheinlich – ganz sicher

- **Bewertung**
  Beispiel: An den Universitäten sollte mehr geforscht werden!
  trifft gar nicht zu – trifft wenig zu – trifft teils, teils zu – trifft ziemlich zu – trifft völlig zu.

Da es verbal manchmal schwierig oder auch sehr platzgreifend ist, bei einer 7- oder 9-stufigen Skala alle Zwischenstufen äquidistant durch verbale Etiketten abzubilden, begnügt man sich auch oft damit, **nur die Skalenendpunkte** mit verbalen Marken zu versehen und die Abstufungen zu nummerieren oder mit Symbolen zu versehen (Beispiel 3 in ▶ Varianten von Ratingskalen). Anstelle unterschiedlicher Eigenschaften können an den Skalenenden z. B. auch „starke Ablehnung" („strongly disagree") vs. „starke Zustimmung" („strongly agree") stehen. Dieser Skalentyp wird häufig in Einstellungs- oder Persönlichkeitsfragebögen eingesetzt (▶ Abschn. 10.4.3). Eine vergleichende Analyse von Ratingskalen mit englischsprachigen Labels findet man bei Rohrmann (2007, ◘ Tab. 8.10). Bezogen auf ein Häufigkeitsrating weisen Newstead und Arnold (1989) auf die Vorzüge einer numerischen Prozentskala gegenüber verbalen Marken für Häufigkeiten hin (Beispiel: „An wie vielen Tagen hat Ihr Kind Kopfschmerzen?" Antwortmöglichkeiten: 0 % – 25 % – 50 % – 75 % – 100 %).

Neben Bewertungs-/Zustimmungsskalen sind Häufigkeitsratings am stärksten verbreitet. Sie kommen bei Selbstauskünften zu eigenen Verhaltensweisen (Mediennutzung, Sport, Ernährungs-, Arbeits-, Lernverhalten etc.) zum Einsatz. Eine **abstrakte Häufigkeitsratingskala** wie „nie – selten – gelegentlich – oft – immer" hat den Vorteil, dass sie universell einsetzbar ist. Sie hat jedoch mangels Konkretisierung den Nachteil, dass man nicht genau weiß, was genau die Befragten z. B. unter „gelegentlich" verstehen: einmal pro Woche, einmal pro Monat oder einmal im Jahr? Will man mit Häufigkeitsskalen arbeiten, die die **Zeiträume konkretisieren** (z. B. mehrmals täglich – täglich – mehrmals pro Woche – einmal pro Woche – mehrmals pro Monat – einmal pro Monat – seltener als einmal pro Monat – nie), so bereitet dies einerseits den Befragten höheren Bewertungsaufwand (sie müssen sich

genauer erinnern) und erfordert es zudem, dass vorab die realistischen Verhaltensfrequenzen bekannt sind oder in einer Vorstudie erhoben werden. Denn diese schwanken sehr stark in Abhängigkeit von der Art des Verhaltens sowie der jeweiligen Zielgruppe. Sie treten manche Verhaltensweisen bis zu mehrmals pro Tag auf, andere allenfalls einmal im Monat – die vorgegebenen Abstufungen einer zeitlich konkretisierten Ratingskala müssen jeweils auf die typischen Auftretenshäufigkeiten des gemessenen Merkmals zugeschnitten sein. Nicht zuletzt sind Häufigkeitsskalen mit konkreten Zeiträumen in der Regel nicht äquidistant und deswegen eher als Ordinalskalen zu behandeln, auch wenn in der Praxis hier eine liberale Auslegung als Intervallskala nicht unüblich ist.

b. Numerische Marken. Numerische Skalenbezeichnungen (Beispiel 2 in ► Varianten von Ratingskalen) sind knapp und eindeutig; ihre Verwendung ist jedoch nur sinnvoll, wenn die Untersuchungspersonen diese abstrakte Darstellungsform verstehen. Beispiel 3 in ► Varianten von Ratingskalen zeigt eine bipolare Ratingskala, deren Extreme durch zwei verschiedene gegensätzliche Begriffe markiert sind. Als Skalenwerte sind die Zahlen 1–5 vorgegeben, deren Bedeutung in der Instruktion erläutert wird. Um die Gegensätzlichkeit der Begriffe stärker zu betonen, werden gelegentlich positive und negative Zahlenwerte einschließlich einer neutralen Mitte (0) verwendet.

c. Symbolische Marken. Noch anschaulicher als verbale Marken sind symbolische Marken, die insbesondere bei Kindern gerne verwendet werden, aber auch Erwachsenen die Urteilsabgabe erleichtern. Die im Beispiel 4 in ► Varianten von Ratingskalen wiedergegebene Smiley-Skala wurde von Jäger (1998) entwickelt und auf Äquidistanz geprüft. Im Unterschied zu verbalen Marken, die erst gelesen werden müssen, kann man die Bedeutung der symbolischen Marken auf einen Blick erfassen. Durch die Visualisierung wirken symbolische Marken bei längeren Listen von Urteilsaufgaben auflockernd.

d. Keine vorgegebenen Marken bzw. grafisches Rating. Beispiel 5 in ► Varianten von Ratingskalen zeigt ein grafisches Rating, auch **visuelle Analogskala** genannt, das man häufig für die Schätzung von Ähnlichkeiten (Ähnlichkeitspaarvergleich) im Rahmen der Multidimensionalen Skalierung MDS verwendet (zu explorativen multivariaten Auswertungstechniken ► Abschn. 12.3.1). Die Ähnlichkeit (Unähnlichkeit) ergibt sich hierbei aus der Länge der Strecke zwischen einem Extrem der Skala und dem vom Urteiler gesetzten Kreuz. Hier wird also auf die Vorgabe von Merkmalsabstufungen gänzlich verzichtet.

Diese Skalenart bietet gute Voraussetzungen für intervallskalierte Ratings; sie erschwert jedoch die Auswertung erheblich, sofern die Datenerhebung nicht am Computer erfolgt. In der Onlineforschung wird ein solches Item mit einem Schieberegler dargestellt, der mit der Computermaus positioniert werden kann. Dies hat den Vorteil, dass der mit der Position verbundene Messwert automatisch zugeordnet werden kann. Problematisch an dieser Operationalisierung ist indessen, dass der Schieberegler in einer Ausgangsposition vorgegeben werden muss, wobei meist die Mittelposition gewählt wird. Dann bleibt bei der Auswertung aber offen, ob Befragte die Mittelposition aktiv gewählt oder die Frage gar nicht beantwortet haben.

Ausführliche Hinweise über Vor- und Nachteile grafischer Ratingskalen geben Ahearn (1997), Couper, Tourangeau, Conrad, & Singer (2006), Wewers und Lowe (1990) sowie McCormack, Horne und Sheather (1988). Mit der „**Stimulus Comparison Scale“ (SCS)** liegt eine auch im Face-to-Face-Interview verwendbare visuelle Analogskala vor, bei der die Befragungspersonen von vorne auf einer Art Lineal einen Schieber einstellen und nur der Interviewer von hinten den Merkmalswert ablesen kann (Frank & Schlund, 2000).

▪ Skalenverankerung durch Beispiele

Beispiel 6 in ► Varianten von Ratingskalen zeigt Ratingskalen, bei denen durch die Formulierung beispielhafter Extrempositionen bzw. **Ankerbeispiele** sehr gezielt Informationen erfragt werden können (Beaton & Allen, 1992; Kingstrom & Bass, 1981; „Example Anchored Scales“ nach Smith & Kendall, 1963). Derartige Skalen haben sich insbesondere in der klinischen Forschung bzw. der Persönlichkeitspsychologie bewährt. Gelegentlich erfolgt die Verankerung der Skalen auch durch typische Zeichnungen, Testreaktionen oder Fotografien.

Ratingskalen, deren Abstufungen durch konkrete Falldarstellungen bzw. **Fallbeispiele** verdeutlicht werden (Beispiel 7 in ► Varianten von Ratingskalen) finden nicht nur in der klinischen Psychologie, sondern auch in zahlreichen anderen Anwendungsgebieten wie z. B. bei der Beschreibung beruflicher Tätigkeiten, der Bewertung von Arbeitsleistungen oder im sozialen Bereich Verwendung (Smith & Kendall, 1963). Die Ermittlung der Skalenwerte für die Falldarstellungen von **„Behaviorally Anchored Rating Scales“** (BARS; DeCotiis 1978) basierte ursprünglich auf dem „Law of Categorical Judgement“ (► Abschn. 8.4.6) und wurde inzwischen erheblich verbessert (Bernardin & Smith, 1981; Campbell, Dunnette, Arvey, & Hellervik, 1973; Green, Sauser, Fagg, & Champion, 1981; Hauenstein, Brown, & Sinclair, 2010; Kinicki & Bannister, 1988; Kinicki, Bannister, Hom, & Denisi, 1985; Shapira & Shirom, 1980; Tziner, Joanis, & Murphy, 2000). Einen Literaturüberblick zu dieser Ratingtechnik findet man bei Schwab, Heneman und DeCotiis (1975)

und eine Analyse der psychometrischen Eigenschaften bei Kinicki et al. (1985; siehe auch Borman 1986; Hauenstein, et al., 2010; Kinicki & Bannister, 1988; Tziner et al., 2000).

Anzahl der Skalenstufen

Ein häufig diskutiertes Problem betrifft die Anzahl der Stufen einer Ratingskala. Dabei geht es um zwei Aspekte:

a. Gerade oder ungerade Stufenanzahl: Ungeradzahlige Ratingskalen enthalten eine **neutrale Mittelkategorie** und erleichtern damit bei unsicheren Urteilenden das Ausweichen auf diese Neutralkategorie. Geradzahlige Ratingskalen verzichten auf eine neutrale Kategorie und erzwingen damit von den Befragungspersonen ein zumindest tendenziell in eine Richtung weisendes Urteil. Diese Vorgehensweise empfiehlt sich, wenn man mit Verfälschungen der Urteile durch eine übermäßige **Tendenz zur Mitte** der Urteilenden rechnet.

Die Schwierigkeiten bei der Interpretation von Ratingskalen mit neutralen Antwortkategorien werden in der Literatur unter dem Stichwort **Ambivalenz-Indifferenz-Problem** diskutiert. Hierzu ein Beispiel: Eine Teamleiterin hat bei der Beurteilung einer Angestellten hinsichtlich ihres Kommunikationsstils in der Arbeitsgruppe auf der bipolaren Skala „bestimmt – zurückhaltend“ die neutrale Mittelkategorie gewählt. Dies kann bedeuten, dass die Teamleiterin bzgl. dieses Merkmals keine dezidierte Meinung vertritt, dass sie also **indifferent** ist bzw. Meinungslosigkeit vorliegt. Es kann aber auch bedeuten, dass sie das Auftreten ihrer Mitarbeiterin in bestimmten Situationen für bestimmt (z. B. bei Diskussionen zu Themen, in denen sie sich sehr gut auskennt), in anderen jedoch für zurückhaltend (z. B. wenn sie keine ausgewiesene Expertise in einem bestimmten Bereich hat) hält, dass ihre Meinung bzgl. dieses Merkmals also **ambivalent** ist. Weil bestimmte und zurückhaltende Seiten sich die Waage halten, wählt die Teamleiterin die neutrale Kategorie und bringt damit eine mittlere Merkmalsausprägung (und eben gerade nicht Meinungslosigkeit) zum Ausdruck. Welche methodischen Möglichkeiten es gibt, zwischen Ambivalenz bzw. mittlerer Merkmalsausprägung einerseits und Indifferenz bzw. Meinungslosigkeit andererseits zu unterscheiden, wird bei Kaplan (1972), Breckler (1994) und Bierhoff (1996, S. 65ff.) erörtert.

Eine pragmatische Lösung besteht z. B. darin, bei Items, bei denen man ein gehäuftes Auftreten von Meinungslosigkeit erwartet (z. B. weil die Beantwortung intensives Nachdenken erfordern würde oder weil das Thema für viele Personen einfach so unwichtig ist, dass sie sich bislang kaum darüber Gedanken gemacht haben), eine ausdrückliche **Sonderkategorie für Meinungslosigkeit** (z. B. „weiß nicht“, „keine Angabe“) neben der Ratingskala anzubieten. Auf diese Weise kann Meinungslosigkeit getrennt erhoben werden von ambivalenten bzw. mittleren Merkmalsausprägungen.

b. Optimale Stufenzahl: Eine Skala mit **sehr wenigen Stufen** (z. B. drei oder vier Stufen) liefert undifferenzierte Urteile, die maximal Ordinalskalenniveau besitzen (Matell & Jacoby, 1971). Eine Skala mit **sehr vielen Stufen** (z. B. zwanzig Abstufungen) wäre beim Ausfüllen sehr mühsam und anstrengend und würde die Differenzierungsfähigkeit der meisten Befragungspersonen übersteigen. Wählt man Ratingskalen mit extrem vielen (z. B. 100) Skalenstufen, ist zudem festzustellen, dass die Urteilenden überwiegend Stufen wählen, die durch 10 (bzw. durch 5) teilbar sind, was Henss (1989) auf die Prominenzstruktur des Dezimalsystems zurückführt. Interpretativ lässt sich dieser Befund so deuten, dass eine zu feine Differenzierung bei einer Ratingskala das Urteilsvermögen der Urteilenden überfordert mit der Folge, dass nur eine gröber segmentierte Teilmenge aller vorgegebenen Antwortkategorien verwendet wird.

Eine **Stufenzahl zwischen fünf und sieben –** zuweilen auch bis neun – hat sich in der Praxis bewährt und liefert psychometrisch auch die besten Validitäten und Reliabilitäten (Dawis 1987; Lissitz & Green, 1975; Lozano, García-Cueto, & Muñiz, 2008; Preston & Colman, 2000; B. Rohrmann 1978). Es ist zu beachten, dass die optimale Stufenzahl auch von der Datenerhebungsmethode abhängt: Sehr feine Abstufungen sind schriftlich im Fragebogen möglich, für mündliche oder telefonische Interviews aber ungeeignet.

Schätzurteile auf Ratingskalen überfordern die Urteilenden zuweilen, wenn sie zu sehr bemüht sind, durch sorgfältiges Nachdenken zu einem fundierten Urteil zu gelangen. Im Bemühen um eine rationale Begründung der Urteile können sie – vor allem bei überdifferenzierten Skalen – zu widersprüchlichen Eindrücken von der Ausprägung des untersuchten Merkmals kommen, die gelegentlich dazu führen, dass die Beurteilung gänzlich verweigert wird. Derartige Verweigerungen, auf die man z. B. im Rahmen eines Pretests stößt, sollten zum Anlass genommen werden, die Ratingskalen bzw. die Instruktion zu überarbeiten. Besteht jedoch der Verdacht, dass die Verweigerung auf übermäßige Skrupel zurückgeht, hilft oft ein ausdrücklicher Hinweis an die Urteilenden, dass ohne langes Nachdenken einfach ihr **spontanes Urteil**, d. h. der erste, subjektive Eindruck von der Merkmalsausprägung zum Ausdruck gebracht werden soll.

Gelegentlich steht man vor dem Problem, Urteile auf Ratingskalen mit unterschiedlichen Stufenanzahlen miteinander vergleichen oder ineinander überführen zu müssen. Hierfür geeignete Transformationsformeln findet man bei Aiken (1987) und Henss (1989).

Messtheoretische Probleme bei Ratingskalen

Ratingskalen sind zwar relativ einfach zu handhaben; sie werfen jedoch eine Reihe messtheoretischer Probleme auf, die im Folgenden kurz erörtert werden. Wir konzentrieren diese Diskussion auf die Frage nach dem Skalenniveau und nach der Verankerung von Ratingskalen (zur Bestimmung der in ► Abschn. 10.4.1 behandelten testtheoretischen Gütekriterien „Reliabilität“ und „Validität“ vgl. Aiken 1985b; Aiken 1996, 1997).

Zum Skalenniveau von Ratingskalen

Das gemeinsame Problem aller Ratingskalenarten betrifft ihr Skalenniveau. Garantieren eine detaillierte Instruktion und eine sorgfältige Skalenkonstruktion, dass die Befragungspersonen intervallskalierte Urteile abgeben?

Die Kontroverse zu diesem Thema hat eine lange Tradition und scheint bis heute noch kein Ende gefunden zu haben. Die messtheoretischen „Puristen“ behaupten, Ratingskalen seien nicht intervallskaliert; sie verbieten deshalb die statistische Analyse von Ratingskalen mittels parametrischer Verfahren, die – so wird häufig argumentiert – intervallskalierte Daten voraussetzen. Demgegenüber vertreten die „Pragmatiker“ den Standpunkt, die Verletzungen der Intervallskaleneigenschaften seien bei Ratingskalen nicht so gravierend, als dass man auf die Verwendung parametrischer interferenzstatistischer Verfahren gänzlich verzichten müsste.

Ein Missverständnis: In diesem Zusammenhang sei auf einen Irrtum aufmerksam gemacht, der seit der Einführung der vier wichtigsten Skalenarten durch Stevens (1946, 1951) anscheinend nur schwer auszuräumen ist. Die Behauptung, parametrische inferenzstatistische Verfahren wie z. B. der t-Test oder die Varianzanalyse setzten intervallskalierte Daten voraus, ist in dieser Formulierung nicht richtig. Die mathematischen Voraussetzungen der genannten Verfahren sagen nichts über die Skaleneigenschaften der zu verrechnenden Daten aus. (Die Varianzanalyse setzt z. B. normalverteilte, unabhängige und homogene Fehlerkomponenten voraus.) Vor diesem Hintergrund wäre z. B. gegen die Anwendung varianzanalytischer Verfahren auf Daten wie z. B. Telefonnummern oder Lottozahlen nichts einzuwenden, solange diese Zahlen die geforderten mathematischen Voraussetzungen erfüllen („The numbers do not know where they come from“; Lord 1953, S. 751).

Gaito (1980) diskutiert die Hartnäckigkeit dieses Missverständnisses anhand zahlreicher Literaturbeispiele und fordert nachdrücklich, bei der Begründung der Angemessenheit eines statistischen Verfahrens zwischen **messtheoretischen Interpretationsproblemen** und **mathematisch-statistischen Voraussetzungen** zu unterscheiden. Die Frage, ob verschiedene Zahlen tatsächlich unterschiedliche Ausprägungen des untersuchten Merkmales abbilden bzw. die Frage, ob – wie es die Intervallskala fordert – gleiche Zahlendifferenzen auch gleiche Merkmalsunterschiede repräsentieren, ist ein messtheoretisches und kein statistisches Problem. Der statistische Test „wehrt“ sich nicht gegen Zahlen minderer Skalenqualität, solange diese seine mathematischen Voraussetzungen erfüllen. Die Skalenqualität der Zahlen wird erst bedeutsam, wenn man die **Ergebnisse inhaltlich interpretieren** will. Es sind dann messtheoretische Erwägungen, die dazu veranlassen, die Ergebnisse einer Varianzanalyse über Nominaldaten für nichtssagend zu erklären, weil die Mittelwerte derartiger Zahlen, die in diesem Verfahren verglichen werden, keine inhaltliche Bedeutung haben (Hand 1996; King et al., 2010; Michell 1986; Stine 1989).

Für die Behauptung, parametrische inferenzstatistische Verfahren führten auch dann zu korrekten Entscheidungen, wenn das untersuchte Zahlenmaterial nicht exakt intervallskaliert ist, liefern Baker, Hardyck und Pertinovich (1966) einen überzeugenden Beleg (weitere Literatur z. B. Rasmussen 1989; Zumbo & Zimmerman, 1993). In einer aufwändigen Simulationsstudie wurde die Äquidistanz der Zahlen einer Intervallskala systematisch in einer Weise verzerrt, dass Verhältnisse resultieren, von denen behauptet wird, sie seien für Ratingskalen typisch. Die Autoren erzeugten

- Skalen mit zufällig variierten Intervallgrenzen,
- Skalen, deren Intervalle an den Extremen breiter waren als im mittleren Bereich (was z. B. von Intelligenzskalen behauptet wird),
- Skalen, die nur halbseitig intervallskaliert waren (was gelegentlich von einigen sozialen Einstellungsskalen behauptet wird).

Mit diesem Material wurden 4 000 t-Tests über Paare zufällig gezogener Stichproben ($n = 5$ bzw. $n = 15$) gerechnet. Die Autoren kommen zu dem Schluss, dass statistische Entscheidungen von der Skalenqualität des untersuchten Zahlenmaterials weitgehend unbeeinflusst bleiben.

Diese Unbedenklichkeit gilt allerdings nicht, wenn die in dieser Studie berechneten Mittelwerte inhaltlich interpretiert werden. Statistisch bedeutsame Mittelwertunterschiede sagen nichts aus, wenn das Merkmal mit einer Skala gemessen wurde, deren Intervallgrößen beliebig variieren.

In der empirischen Sozialforschung bleibt das Messen und insbesondere das Messen mit Ratingskalen damit – was die Skalenqualität der Messungen anbelangt – ein auf Hypothesen gegründetes Unterfangen. Die Hypothese der Intervallskalenqualität von Ratingskalen und die damit verbundene Interpretierbarkeit der Messungen werden in jeder konkreten Untersuchungssituation neu zu

begründen sein. Die Sozialwissenschaften wären allerdings schlecht beraten, wenn sie mangels Argumenten, die für den Intervallskalencharakter von Ratingskalen sprechen, gänzlich auf dieses wichtige Erhebungsinstrument verzichteten. Viele, vor allem junge Forschungsbereiche, in denen die inhaltliche Theorienbildung erst am Anfang steht, wären damit eines wichtigen, für die Untersuchungspersonen relativ einfach zu handhabenden Messinstrumentes beraubt. Solange die Forschung mit Ratingskalen zu inhaltlich sinnvollen Ergebnissen kommt, die sich in der sozialen Realität bewähren, besteht nur wenig Veranlassung, an der Richtigkeit der impliziten messtheoretischen Hypothesen zu zweifeln. Diese Position wird durch eine Untersuchung von Westermann (1985) gestützt, in der die Axiomatik einer Intervallskala in Bezug auf Ratingskalen empirisch erfolgreich geprüft werden konnte.

Ratingskalen („rating scales“) werden meist als _intervallskaliert_ aufgefasst. Somit können mit den Daten dann z. B. sinnvoll interpretierbare Mittelwerte gebildet und die für intervallskalierte Daten vorgesehenen statistischen Verfahren verwendet werden. Bei Zweifeln an der Gleichabständigkeit (Äquidistanz) der Stufen einer Ratingskala wird diese als _ordinalskaliert_ betrachtet. Dementsprechend sind dann Mittelwerte nicht sinnvoll interpretierbar, stattdessen ist mit Medianwerten sowie den auf ordinalskalierte Daten zugeschnittenen statistischen Verfahren zu operieren.

Ratingskalen mit gleichabständiger Etikettierung und mindestens fünf Stufen sowie auch Schulnoten als intervallskaliert aufzufassen, hat sich in der Forschungspraxis eingebürgert. Ansonsten dürfte man z. B. nicht mit Durchschnittsnoten operieren. Zuweilen wird aus messtheoretischer Sicht jedoch die Position vertreten, es handele sich hier nur um ordinalskalierte Daten. Ein Grundsatzstreit lohnt an dieser Stelle aus den oben genannten Gründen sowie aus forschungspragmatischer Sicht nicht. So können in einem strittigen Fall die Daten sowohl mit einem Verfahren für Ordinaldaten als auch einem Verfahren für Intervalldaten ausgewertet und die Ergebnisse verglichen werden. Sollen etwa zwei Gruppen hinsichtlich ihrer zentralen Tendenz kontrastiert werden, so kann unter der Annahme einer Intervallskala der t-Test für unabhängige Stichproben, unter der Annahme einer Ordinalskala der Mann-Whitney-U-Test zum Einsatz kommen (Bortz & Schuster, 2010, Abschn. 8.7.1). Ebenso wäre für intervallskalierte Daten als bivariates Zusammenhangsmaß die Produkt-Moment-Korrelation, bei ordinalen Daten Spearmans Rangkorrelation indiziert etc. In der Regel kommen beide Strategien zu nahezu identischen Resultaten, so dass eine großzügige Auslegung von Ratingskalen und Schulnoten als intervallskaliert keiner wissenschaftlich unzulässigen Begünstigung der eigenen Hypothese entspricht und somit vertretbar ist. Wie man Ratingskalen im Rahmen der probabilistischen Testtheorie mithilfe des Rasch-Modells (► Abschn. 10.4.4) hinsichtlich ihrer Skalierungseigenschaften analysiert, wird bei Rost (2004, Abschn. 3.3.2 und Abschn. 3.3.4) erklärt.

Einheit und Ursprung von Ratingskalen

Weitere Überlegungen zur Konstruktion intervallskalierter Ratingskalen betreffen die Einheit und die Verankerung bzw. den Ursprung der Skala. Untersuchungstechnische Hilfen sollten dazu beitragen, dass Einheit und Ursprung einer Ratingskala intra- und interindividuell konsistent verstanden werden. Wie stark scheinbar geringfügige Veränderungen in der Formulierung einer Frage bzw. im Skalenformat das Antwortverhalten beeinflussen, demonstrieren Kahnemann und Tversky (2000; vgl. hierzu auch Krosnick & Fabrigar, 2006).

Für ein einheitliches Verständnis des Ursprungs einer Skala ist es hilfreich, wenn die Urteilenden vor der eigentlichen Beurteilung sämtliche Untersuchungsobjekte (oder doch zumindest Objekte mit extremen Merkmalsausprägungen) kennenlernen. Nur so wird verhindert, dass Objekte mit extremen Merkmalsausprägungen nicht mehr korrekt eingestuft werden können, weil die Extremwerte zuvor bereits für Objekte mit weniger extremen Merkmalsausprägungen vergeben wurden. Durch dieses Vorgehen werden **Decken-Effekte** („ceiling effects“) sowie **Boden-Effekte** („floor effects“) vermieden. (Dies sind Effekte, die das „Zusammendrängen“ vieler Objekte mit starker, aber unterschiedlicher Merkmalsausprägung in der obersten Kategorie – der „Decke“ – oder mit schwacher, aber unterschiedlicher Merkmalsausprägung in der untersten Kategorie – dem „Boden“ – bezeichnen.) Die Urteilenden können sich so vom gesamten, durch die Objekte realisierten Merkmalskontinuum einen Eindruck verschaffen und dieses, evtl. unterstützt durch verbale Marken, in gleich große Intervalle aufteilen (McCarty & Shrum, 2000).

Zu beachten ist ferner die Verteilung der untersuchten Objekte über das Merkmalskontinuum. Werden viele positive, aber nur wenige negative Objekte auf einer Bewertungsskala eingestuft, ist damit zu rechnen, dass die positiven Objekte feiner differenziert werden als die negativen. Die Wahrscheinlichkeit intervallskalierter Ratingskalenurteile wird deshalb erhöht, wenn die Objekthäufigkeiten auf beiden Seiten der Skala symmetrisch sind bzw. wenn der mittlere Wert der Skala mit dem Medianwert der Häufigkeitsverteilung zusammenfällt (vgl. das „Range-Frequency-Model“ von Parducci 1963, 1965). Weitere theoretische Überlegungen über Urteilsprozesse, die für

die Konstruktion intervallskalierter Ratingskalen nutzbar gemacht werden können, findet man bei Eiser und Ströbe (1972), Upshaw (1962) bzw. Gescheider (1988; auch Igou, Bless, & Schwarz, 2002; Schwarz 2008; Schwarz & Oyserman, 2001; Schwarz, Wänke, Sedlmeier, & Betsch, 2002; Strack, Schwarz, Ash, & Sturm, 2007; Wänke & Fiedler, 2007). Wie man Ratingskalen mithilfe des Rasch-Modells (▶ Abschn. 10.4.4) analysiert, wird bei Rost (2004, Abschn. 3.3.2 und 3.3.4) beschrieben.

Urteilsfehler beim Einsatz von Ratingskalen

Die Brauchbarkeit von Daten, die mittels Ratingskalen gewonnen werden, ist zuweilen durch systematische Urteilsfehler bzw. Urteilsverzerrungen eingeschränkt. Damit ist gemeint, dass Untersuchungspersonen, wenn sie Merkmale ihres eigenen Erlebens und Verhaltens oder vorgegebener Objekte auf Ratingskalen einstufen, nicht die „wahren" Merkmalsausprägungen angeben, sondern die Merkmalsausprägungen unabsichtlich mehr oder minder stark über- oder unterschätzen. **Unsystematische Fehler**, die nur einzelne Untersuchungspersonen machen, gleichen sich bei den in der empirischen Sozialforschung typischen Gruppenstudien bei ausreichend großen Stichproben aus und sind somit unproblematisch. Mit den hier betrachteten Urteilsfehlern sind jedoch **systematische Fehler** gemeint, die bei vielen Menschen in ähnlicher Weise auftreten. Derartige Fehler werden auf **Besonderheiten der menschlichen Informationsverarbeitung** zurückgeführt. So können z. B. Gedächtniseffekte oder Urteilsheuristiken eine Rolle spielen, auch sind bei der Beurteilung von Personen wiederum bestimmte soziale Effekte auf die Urteilsbildung typisch. Manche Urteilsfehler sind auch mit **Persönlichkeitsmerkmalen** verknüpft. Statt von „Urteilsfehlern" wird – je nach Kontext und Perspektive der Betrachtung – auch von „kognitiven Verzerrungen", „Antwortfehlern", „Antworttendenzen", „sozialen Verzerrungen", „Beobachterfehlern" etc. gesprochen. Ein generelles Problem bei der Untersuchung von Urteilsfehlern betrifft die Trennung zwischen **„wahren" Merkmalsausprägungen** und **Fehleranteilen**. Da die „wahren" Merkmalsausprägungen meistens unbekannt sind, ist es nicht ohne weiteres möglich, Richtung und Stärke von Urteilsfehlern zu identifizieren.

Welche und wie viele Urteilsfehler im Zusammenhang mit Ratingskalen und anderen Operationalisierungsformen relevant sind, wie sie im Einzelnen zustande kommen und wie am besten mit ihnen umzugehen ist, lässt sich beim aktuellen Forschungsstand nicht umfassend beantworten (zum Überblick über Urteilsfehler und deren Kontrolle Saal, Downey, & Lahey 1980; Hoyt & Kerns, 1999; Hoyt 2000, 2002). Unsystematische sowie systematische Fehler und Verzerrungen spielen im Übrigen bei allen Datenerhebungsmethoden und zugehörigen Messinstrumenten eine Rolle: Bei der Beobachtung kann es zu Beobachtungsfehlern kommen (▶ Abschn. 10.1.1 „Beobachtungsfehler und Gegenmaßnahmen"), beim Interview zu Interviewfehlern (▶ Abschn. 10.2.1 „Interviewfehler und Gegenmaßnahmen"), bei der Fragebogenerhebung zu ähnlichen Antwortfehlern wie im Interview, bei psychologischen Tests zu Fehlern und Verfälschungen (▶ Abschn. 10.4.1 „Fehler beim Einsatz psychologischer Tests und Gegenmaßnahmen") und bei einer Dokumenten- bzw. Inhaltsanalyse sind diverse Kodierungsfehler möglich (▶ Abschn. 10.6.3 „Quantitative Inhaltsanalyse"). Lediglich physiologische Messungen, bei denen die Messwerte apparativ aufgezeichnet werden, bleiben von Verzerrungen der menschlichen Informationsverarbeitung unbeeinträchtigt, kämpfen dafür aber mit anderen Messproblemen (▶ Abschn. 10.5.1 „Methodische Grundlagen physiologischer Messungen").

Urteilsfehler bzw. Urteilsverzerrungen – Urteilsfehler bzw. Urteilsverzerrungen („rater bias") liegen vor, wenn Untersuchungspersonen Merkmale ihres eigenen Verhaltens und Erlebens bzw. Merkmale vorgegebener Urteilsobjekte auf Ratingskalen (oder anderen Messinstrumenten) nicht korrekt einstufen, sondern die „wahre" Merkmalsausprägung unwillentlich systematisch über- oder unterschätzen. Anzahl, Art und Ursachen aller möglichen Urteilsverzerrungen sind nicht vollständig bekannt. Typische Urteilsfehler sollten bei der Konstruktion von Ratingskalen (und anderen Messinstrumenten), bei der Gestaltung von Erhebungssituationen sowie bei der Interpretation von quantitativen Daten in reflektierter Weise berücksichtigt werden.

Teilweise lassen sich Urteilsfehler durch die spezifische Konstruktion von Messinstrumenten (z. B. Reihenfolge und Formulierung der Items einer psychometrischen Skala) und die Gestaltung von Erhebungssituationen (z. B. Aufklärung der Untersuchungspersonen über die Gefahr der betreffenden Verzerrung) verringern. Zudem sind mögliche Urteilsfehler bei der Daten- bzw. Ergebnisinterpretation zu berücksichtigen, was allerdings nicht darauf hinauslaufen darf, dass man hypothesenkonträre Befunde mit dem pauschalen Verweis auf Urteilsfehler einfach wegdiskutiert, während man bei hypothesenkonformen Ergebnissen die Unverzerrtheit der Urteile voraussetzt. Vielmehr muss im Sinne der Wissenschaftlichkeit jedes Ergebnis einer wissenschaftlichen Studie einer (selbst)kritischen Diskussion unterzogen werden. Einige bekannte Urteilsfehler, die im Zusammenhang mit Ratingskalen relevant sind, sollen im Folgenden kurz dargestellt werden.

▪ Halo-Effekt

Die Bezeichnung Halo-Effekt **(Ausstrahlungseffekt)** geht auf den amerikanischen Psychologen Edward Lee

Thorndike (1920) zurück und spielt metaphorisch auf den ausstrahlenden Effekt des Mondlichtes an, das um den Mond einen Hof („halo“) bildet. Der gleiche Urteilsfehler wurde von Newcomb (1931) als „logischer Fehler“ bezeichnet. Gemeint ist eine Tendenz, die Beurteilung mehrerer Merkmale eines Objektes von einem globalen Pauschalurteil abhängig zu machen (Borman 1975), die Unfähigkeit oder mangelnde Bereitschaft des Urteilers, auf unterschiedliche Ausprägungen verschiedener Merkmale zu achten (DeCotiis 1977) oder die Tendenz eines Urteilers, ein Objekt bzgl. vieler Merkmale gleich einzustufen (Bernardin 1977). Das Gemeinsame dieser Definitionen ist ein Versäumnis des Urteilers, konzeptuell unterschiedliche und potenziell unabhängige Merkmale im Urteil zu differenzieren (Cohen 1969, S. 41ff.; Hoyt 2000). Wird etwa eine Person als sehr sympathisch eingeschätzt, so kann dieser positive Eindruck auf die Beurteilung anderer Merkmale der Person (z. B. Gewissenhaftigkeit, Hilfsbereitschaft, fachliche Kompetenz) „ausstrahlen“.

Halo-Effekte treten verstärkt auf, wenn das einzuschätzende Merkmal ungewöhnlich, nur schwer zu beobachten oder schlecht definiert ist. Demzufolge können Halo-Effekte reduziert werden, wenn die Urteilenden vor der Beurteilung gründliche Informationen über die Bedeutung der einzustufenden Merkmale erhalten (Athey & McIntyre, 1987; Bernardin & Walter, 1977). Eine ähnliche Wirkung hat – wie Borman (1975) und Latham, Wexley, & Pursell (1975) zeigen – die Aufklärung der Urteilenden über mögliche, auf Halo-Effekte zurückgehende Urteilsfehler. Klauer und Schmeling (1990) kommen zu dem Schluss, dass vor allem schnell gefällte Urteile von Halo-Effekten durchsetzt sind.

Friedman und Cornelius (1976) weisen darauf hin, dass sich die Mitwirkung der Urteilenden an der Konstruktion der Ratingskalen günstig auf ihr Urteilsverhalten auswirkt. Eine geringe Verfälschung der Urteile durch Halo-Effekte wird nach Johnson und Vidulich (1956) auch erreicht, wenn bei der Einschätzung mehrerer Urteilsobjekte auf mehreren Urteilsskalen nicht objektweise, sondern skalenweise vorgegangen wird: Die Urteilenden beurteilen hierbei zunächst alle Objekte auf der ersten Skala, dann auf der zweiten Skala etc. Hinweise zur formalen Analyse von Halo-Effekten findet man bei Klauer (1989) bzw. Doll (1988; auch Li, Wang, & Zhang, 2002; Myford & Wolfe, 2003, 2004; Thomas, Palmer, & Feldman, 2009).

▪ Milde-Härte-Fehler (Leniency-Severity-Fehler)

Dieser Urteilsfehler, der – ähnlich wie auch der Halo-Effekt – vor allem bei Personenbeurteilungen auftreten kann, besagt, dass die zu beurteilenden Personen systematisch entweder zu positiv oder zu negativ eingestuft werden (Barr & Raju, 2003; Saal & Landy, 1977). So könnte es rollenbedingt zu einem Milde-Fehler kommen, wenn die Studierende andere Studierende und deren Referate beurteilen, jedoch zu einem Härte-Fehler, wenn sie Lehrende und deren Vorlesungen einschätzen.

Auch dieser Fehler kann weitgehend ausgeräumt werden, wenn die Urteilenden zuvor auf die Gefahr einer derartigen Urteilsverfälschung aufmerksam gemacht werden. Hilfreich sind zudem Diskussionen über die Wertigkeit der einzustufenden Merkmale bzw. über mögliche Konsequenzen, die mit den Einstufungen verbunden sind (Bernardin & Walter, 1977). Es gibt Hinweise darauf, dass die Tendenz, besonders hart oder besonders mild zu urteilen, auch ein Persönlichkeitsmerkmal darstellt und sich z. B. bei Vorgesetzten in verschiedenen Berufsgruppen zeigt, wenn sie Beschäftigte bewerten (Kane, Bernardin, Villanova, & Peyrefitte, 1995).

Methodische Varianten, um Milde- und Härte-Fehler nachzuweisen, diskutieren Saal, Downey und Lahey (1980) bzw. Bannister, Kinicki, Denisi, & Horn (1987), sowie Myford und Wolfe (2003, 2004; Barr & Raju, 2003; Hoyt 2000; Wolfe 2004). Die Frage, inwieweit Messungen des Milde-Härte-Fehlers mit Messungen des Halo-Effekts konfundiert sind, erörtern Alliger und Williams (1989).

▪ Zentrale Tendenz (Tendenz zur Mitte)

Dieser Urteilsfehler bezeichnet eine Tendenz, alle Urteilsobjekte im mittleren Bereich der Urteilsskala einzustufen, also **moderate Urteile** abzugeben und extreme Urteile zu vermeiden (Korman 1971, S. 180f.). Mit einer Tendenz zur Mitte ist vor allem zu rechnen, wenn die zu beurteilenden Objekte den Urteilenden nur wenig bekannt sind – eine Untersuchungssituation, die eigentlich generell zu vermeiden ist. Eine Massierung der Urteile im mittleren Skalenbereich tritt bevorzugt auch dann auf, wenn man es versäumt hat, die Skalen an Extrembeispielen zu verankern (► Abschn. 8.4.4 ► Varianten von Ratingskalen). Die Urteilenden „reservieren“ dann die Extremkategorien für evtl. noch auftauchende Objekte mit extremer Merkmalsausprägung. Bleiben diese aus, resultieren wenig differenzierende Urteile mit starker zentraler Tendenz.

Mangelnde Differenzierung muss jedoch nicht unbedingt eine Antworttendenz zur Mitte bedeuten. Sie tritt immer dann auf, wenn die Urteilenden nicht die gesamte Skalenbreite nutzen, sondern ihre Urteile in einem Bereich der Skala konzentrieren. In diesem Fall schafft eine Neukonstruktion der Ratingskala Abhilfe, die den Bereich, der für die meisten Urteilsobjekte typisch ist, feiner differenziert. Auch für den Nachweis dieser Urteilsfehler nennen Saal, Downey und Lahey (1980) verschiedene me-



Abb. 8.4 Ein weiterer Grund für eine Antworttendenz zur Mitte. (© Peter Roy / Search ID: pron128, Rights Available from CartoonStock.com)

8

thodische Varianten. Ein Test, mit dem die Vermeidung von zentraler Tendenz statistisch geprüft werden kann, wurde von Aiken (1985a) entwickelt (auch Wolfe 2004). Eine Antworttendenz zur Mitte kann zudem bedeuten, dass die Urteilenden nicht näher über die Items nachdenken wollen und einfach stereotyp die Skalenmitte ankreuzen. Dies wird vermieden, wenn man Skalen mit geradzahliger Stufenzahl nutzt, die keine Mittelkategorie aufweisen (Abb. 8.4). Nicht zuletzt kann die Tendenz zu moderaten Urteilen auch eine Persönlichkeitseigenschaft sein. Gleichzeitig existiert als komplementärer Urteilsfehler die **Tendenz zu den Extremen**, hier werden dann bevorzugt Urteile am oberen oder unteren Skalenende abgegeben und moderate Urteile vermieden. Auch dieser Urteilsfehler kann unterschiedliche situations- und/oder persönlichkeitsbezogene Ursachen haben.

Rater-Ratee-Interaktion

Bei Personenbeurteilungen können Urteilsverzerrungen in Abhängigkeit von der Position des Urteilers auf der zu beurteilenden Dimension entstehen. Dabei kann einerseits ein **Ähnlichkeitsfehler** auftreten, wenn Urteilende mit extremer Merkmalsausprägung die Merkmalsausprägungen anderer in Richtung ihrer eigenen Extremposition verschätzen, andere Menschen also für ihnen selbst ähnlicher halten als sie sind. Andererseits kann ein **Kontrastfehler** auftreten, bei dem Urteilende mit extremer Merkmalsausprägung die Merkmalsausprägung anderer in Richtung auf das gegensätzliche Extrem verschätzen, andere Menschen also für ihnen unähnlicher halten als sie sind (Sherif & Hovland, 1961). Einen Beitrag zur Klärung dieser Urteilsfehler liefert z. B. die Theorie der variablen Perspektive von Upshaw (1962). So zeigt sich z. B., dass Urteilende – nachdem sie selbst positive Leistungsergebnisse zurückgemeldet bekommen haben – die Leistungen anderer Personen besser bewerten, als wenn sie vorher selbst negative Beurteilungen bekommen haben (Shore & Tashchian, 2003). Dass es hinsichtlich soziodemografischer Merkmale wie Geschlecht oder Ethnizität zu Wechselwirkungen zwischen den Eigenschaften von Urteilenden und Beurteilten kommt, etwa in der Weise, dass Personen desselben Geschlechts oder derselben Hautfarbe generell positiver bewertet werden, hat sich in der Literatur bislang nicht pauschal bestätigt (Sackett & DuBois, 1991; Shore & Tashchian, 2003; Waldman & Avolio, 1991), obwohl im Detail Interaktionseffekte durchaus auftreten können (Mount, Sytsma, Hazucha, & Holt, 1997).

Primacy-Recency-Effekt

Dieser Effekt bezeichnet Urteilsverzerrungen, die mit der sequenziellen Position der zu beurteilenden Objekte einhergehen und auch als **Reihenfolgeneffekte** bezeichnet werden: Geht es um den besonderen **Einfluss von Anfangspositionen**, spricht man von einem Primacy- oder Primat-Effekt. Geht es um den besonderen **Einfluss von Endpositionen**, handelt es sich um einen Recency- bzw. Rezenz-Effekt. Werden z. B. Paarvergleiche durchgeführt, kann im Sinne eines Primacy-Effekts jeweils das zuerst präsentierte Objekt (z. B. eine Geschmacksprobe im Konsumententest) positiver bewertet werden, etwa weil beim zweitpräsentierten Objekt die Aufmerksamkeit verringert ist. Soll für eine Liste von präsentierten Objekten ein Präferenzurteil abgegeben werden, können im Sinne eines Recency-Effekts die letztgenannten Objekte bevorzugt werden, weil sie noch frisch im Gedächtnis sind.

Derartige Erklärungsansätze des Primacy-Recency-Urteilsfehlers diskutiert Scheuring (1991), und weitere Hinweise zur Bedeutung der Objektreihenfolge findet man bei Kane (1971) und Lohaus (1997). Eine verbreitete

Technik, um Reihenfolgeneffekte in Stichprobenuntersuchungen zu vermeiden, besteht darin, Urteilsreihenfolgen zwischen den Versuchspersonen bzw. den Urteilenden systematisch zu variieren, so dass sich dieser Verzerrungsfaktor im Gesamtergebnis „herausmittelt". Dazu müssen dann verschiedene Erhebungsinstrumente bzw. unterschiedliche Untersuchungsabläufe mit entsprechenden Variationen der Urteilsreihenfolge erzeugt werden.

Weitere Urteilsfehler

Weitere Urteilsfehler (Jäger & Petermann, 1992; Upmeyer 1985; Wessels 1994), die auch beim Einsatz von Ratingskalen auftreten können, sind

- der **Inter- und Intraklasseneffekt** (Merkmalsunterschiede zwischen Objekten werden vergrößert, wenn die Objekte zu unterschiedlichen Klassen oder Gruppen gehören, und sie werden verkleinert, wenn die Objekte zu einer Klasse gehören),
- der **fundamentale Attributionsfehler** (die Gründe und Ursachen für eigenes Fehlverhalten werden in der Situation gesucht, die Gründe für das Fehlverhalten anderer Menschen in deren Charakter),
- der **Self-Serving-Bias** (Selbstbeurteilungen werden mit dem Selbstkonzept in Einklang gebracht und fallen eher selbstwertstützend aus) und
- der **Baseline-Error** (die Auftretenswahrscheinlichkeit von Ereignissen wird falsch eingeschätzt, weil man sich nicht an der objektiven Häufigkeit, der sog. Baseline, orientiert, sondern irrtümlich besonders prägnante, im Gedächtnis gerade verfügbare oder typische Ereignisse irrtümlich für sehr wahrscheinlich hält).

Einige der genannten Fehler sind nur für bestimmte Arten von Urteilsaufgaben relevant (z. B. Selbsteinschätzungen, Wahrscheinlichkeitsratings). Bei Urteilsfehlern kommen die Urteilenden aufgrund von Besonderheiten der menschlichen Informationsverarbeitung irrtümlich und unbemerkt zu falschen Einschätzungen. Verzerrungen können aber auch durch Besonderheiten beim Antwortprozess entstehen, etwa durch **stereotypes Ankreuzen** oder durch **Akquieszenz** (d. h. die persönlichkeits- und kulturspezifische generelle Neigung zum Zustimmen bzw. Ja-Sagen, unabhängig vom konkreten Inhalt). Schließlich ist in Urteils-, Test- und Befragungssituationen auch mit willkürlichen, bewusst kalkulierten Antwortveränderungen zu rechnen. So kann z. B. eine Führungskraft einen geschätzten Mitarbeiter oder eine geschätzte Mitarbeiterin in einer Personalbewertung absichtlich schlechter einstufen, um z. B. deren Versetzung in einen anderen Bereich zu verhindern (North & Reinhardt, 2011). Patienten können ihre Beschwerden stärker einstufen, um eine Verlängerung der Behandlung oder eine Berentung zu erreichen. Entsprechende Probleme sind in der Anwendungspraxis deutlich gravierender als in der Grundlagenforschung, wo die Datenerhebung anonym erfolgt und für die Untersuchungsteilnehmenden meist keine unmittelbaren Konsequenzen hat (zu Verfälschungen und Fehlern beim Einsatz psychologischer Tests ▶ Abschn. 10.4.1 „Fehler beim Einsatz psychologischer Tests und Gegenmaßnahmen").

Anzahl der Urteilenden und Urteilsobjekte beim Einsatz von Ratingskalen

In der sozialwissenschaftlichen Forschung werden häufig Selbstauskunftsdaten mit Ratingskalen erfasst, d. h., die Befragungspersonen beurteilen ihre eigenen Gedanken, Gefühle, Einstellungen, Erfahrungen, Verhaltensweisen. Jedes abgegebene Ratingurteil bezieht sich dabei auf ein anderes, singuläres Urteilsobjekt, nämlich genau die jeweils eine Selbstbeurteilung abgebende Person. Die Anzahl der Urteilenden entspricht der **Anzahl der Urteilsobjekte**. Variabilität in den Daten spiegelt also interindividuelle Unterschiede hinsichtlich der Urteilsobjekte wider.

Eine andere Situation liegt vor, wenn mehrere Befragungspersonen alle **ein und dasselbe Urteilsobjekt** bewerten. Etwa wenn dreihundert Studierende die Qualität derselben, von ihnen besuchten Vorlesung auf verschiedenen Dimensionen (z. B. Aktualität des Stoffes, Kompetenz der Lehrkraft, Tempo des Vortrags etc.) auf Schulnotenskalen bewerten. Für die Charakterisierung der Vorlesung wird dann die durchschnittliche Beurteilung durch alle an der Lehrveranstaltungsevaluation teilnehmenden Studierenden als Maßzahl herangezogen. Durchschnittliche Urteile sind reliabler und valider als Individualurteile (Horowitz, Inouye, & Seigelmann, 1979; Strahan 1980). Die Zusammenfassung mehrerer Schätzurteile zu einem Gesamturteil setzt jedoch eine hinreichende Übereinstimmung der individuellen Urteile voraus. Methoden zur Überprüfung der Urteilerübereinstimmung (Konkordanz) werden z. B. Bortz und Lienert (2008, Kap. 6) und Wirtz und Caspar (2002) dargestellt und diskutiert (zur Überprüfung der Urteilerübereinstimmung ▶ Abschn. 10.1.3, darin ▶ Beobachterübereinstimmung und ▶ Abschn. 10.6.3 „Quantitative Inhaltsanalyse", ◘ Tab. 10.34 zur „Kodiererübereinstimmung".

Weichen die Urteile verschiedener Urteilender in ihren Mittelwerten und Streuungen so stark voneinander ab, dass eine Zusammenfassung nicht mehr zu rechtfertigen ist, kann Vergleichbarkeit durch eine sog. ***z*-Transformation** der individuellen Urteile hergestellt werden (▶ Begleitwebsite zu diesem Buch). Diese für alle Ur-

teilenden getrennt durchzuführenden Transformationen sorgen gewissermaßen im Nachhinein für eine Vergleichbarkeit der individuellen Urteile.

Im Übrigen sei auf Marcus und Schuler (2001) verwiesen, die unterschiedliche Varianten des Urteilertrainings erörtern, die das Ziel haben, Urteilsfehler zu vermeiden und Urteilsprozesse interindividuell zu vereinheitlichen.

Zudem kann es auch inhaltlich von Interesse sein, Personengruppen mit divergierenden Urteilen voneinander zu differenzieren. So mag dieselbe Vorlesung hinsichtlich des Vortragstempos von einem Teil der Studierenden als zu schnell, von einem anderen Teil dagegen als zu langsam eingestuft werden (zu diesen und weiteren Problemen bei der Auswertung von Lehrveranstaltungsevaluationen Döring 2005). Wenn im Zuge standardisierter Verhaltensbeobachtungen bestimmte Verhaltensweisen von den Beobachtern auf Ratingskalen eingestuft werden (z. B. Intensität des Flirtverhaltens), so werden meist auch pro Versuchsperson Urteile von mehreren Beobachtern eingeholt und dann zu einem Durchschnittswert zusammengefasst (▶ Abschn. 10.1.3 „Strukturierte Verhaltensbeobachtung“).

8.4.5 Verhältnisskala

Verhältnisskalierte Merkmale haben gleichabständige Ausprägungen und darüber hinaus im Unterschied zur Intervallskala auch einen absoluten Nullpunkt. Beispiele für solche Merkmale sind etwa die Fehlerzahl einer Person im Diktat oder die Anzahl ihrer Kinder sowie die Häufigkeit ihrer bisherigen Krankenhausaufenthalte, ihr Lebensalter, ihre Körpergröße und ihr Körpergewicht oder die Menge des wöchentlich konsumierten Alkohols. Im empirischen Relativ einer Verhältnis- bzw. Ratioskala sind typischerweise neben einer schwachen Ordnungsrelation der Objekte **Verknüpfungsoperationen** definiert wie z. B. das Aneinanderlegen zweier Bretter oder das Abwiegen von zwei Objekten in einer Waagschale. Dem Verknüpfungsoperator entspricht im numerischen Relativ die Addition.

Bei Merkmalen wie Länge oder Gewicht, auf die der Verknüpfungsoperator sinnvoll angewendet werden kann, sind Aussagen wie: „Durch das Zusammenlegen zweier Bretter a und b resultiert eine Brettlänge, die dem Brett c entspricht“ oder: „Zwei Objekte d und e haben gemeinsam das doppelte Gewicht von f“ möglich. Man beachte, dass derartige Aussagen bei intervallskalierten Merkmalen nicht zulässig sind, denn weder die Aussage: „An einem Tag mit einer Durchschnittstemperatur von 10 °C ist es doppelt so warm wie an einem Tag mit einer Durchschnittstemperatur von 5 °C“ noch die Aussage: „Durch das Zusammenfügen der Intelligenz zweier Personen a und b resultiert die Intelligenz einer Person c“ macht Sinn.

Ein empirisches Relativ mit den oben genannten Eigenschaften bezeichnet man als **extensive Messstruktur**. Man erhält eine Verhältnisskala, wenn ein empirisches Relativ mit einer extensiven Messstruktur wie folgt in ein numerisches Relativ abgebildet wird: Einem Objekt a, dessen Merkmalsausprägung mindestens so groß ist wie die eines Objektes b, wird eine Zahl $\phi(a)$ zugeordnet, die mindestens so groß ist wie $\phi(b)$. Die Zahl, die der Merkmalsausprägung zugeordnet wird, die sich durch die Verknüpfung von a und b ergibt, entspricht der Summe der Zahlen für a und b. Hieraus folgt (Helmholtz 1887, 1959; zitiert nach Steyer & Eid, 1993, Kap. 8.1):

Verhältnisskala bzw. Ratioskala – Eine Verhältnisskala bzw. Ratioskala („ratio scale“) ordnet den Objekten eines empirischen Relativs Zahlen zu, die so geartet sind, dass das Verhältnis zwischen je zwei Zahlen dem Verhältnis der Merkmalsausprägungen der jeweiligen Objekte entspricht. Die Verhältnisskala verfügt über einen absoluten Nullpunkt.

Messungen auf Verhältnisskalen sind eindeutig bis auf hier zulässige **Ähnlichkeitstransformationen** vom Typus $\phi' = \beta \cdot \phi$ ($\beta > 0$). Beispiele für diese Transformationen sind das Umrechnen von Meter in Zentimeter oder Inches, von Kilogramm in Gramm oder Unzen, von Euro in Dollar, von Minuten in Sekunden. Man beachte, dass die Ähnlichkeitstransformation – anders als die für Intervallskalen zulässige lineare Transformation – den Ursprung der Verhältnisskala, der typischerweise dem Nullpunkt des Merkmals entspricht, nicht verändert.

Die Bedeutung einer numerischen Aussage über verhältnisskalierte Messungen ist gegenüber Ähnlichkeitstransformationen invariant. Für die Aussage: „Ein Objekt a kostet doppelt so viel wie ein Objekt b“ ist es unerheblich, ob die Objektpreise z. B. in Euro oder US-Dollar angegeben sind.

Verhältnisskalierte Daten können in den Sozialwissenschaften mittels Befragung erhoben werden (z. B. Anzahl der eigenen Kinder, Anzahl der TV-Geräte im Haushalt, Anzahl der Erkältungen im letzten Jahr), es kann sich aber auch um Beobachtungsdaten (z. B. Reaktionszeit, Häufigkeit und Dauer des Blickkontakts; Anzahl der Klicks in der Protokolldatei eines Webservers) oder um Ergebnisse physiologischer Messungen (▶ Abschn. 10.5) handeln.

Verhältnisskalen kommen in der Forschung bei sozialwissenschaftlichen Merkmalen nur selten vor. Dementsprechend finden sie in der sozialwissenschaftlichen Statistik kaum gesonderte Beachtung. Da jedoch Verhältnisskalen genauere Messungen ermöglichen als Intervallskalen, sind alle mathematischen Operationen bzw. statistischen Verfahren für Intervallskalen auch für Verhältnisskalen gültig. Man verzichtet deshalb häufig auf

eine Unterscheidung der beiden Skalen und bezeichnet Intervall- und Verhältnisskalen zusammengenommen als **Kardinalskalen** oder auch **metrische Skalen**.

Kardinalskala bzw. metrische Skala – Mit der Kardinalskala bzw. metrischen Skala („metric scale") sind als Oberbegriff sowohl die Intervall- als auch die Verhältnisskala angesprochen. Denn in der sozialwissenschaftlichen Forschungsstatistik existieren keine wichtigen Verfahren, die speziell nur auf Verhältnisskalen zugeschnitten sind.

8.4.6 Skalentransformationen

Die vier Skalenniveaus unterscheiden sich in ihrem Informationsgehalt. Prinzipiell ist es immer möglich, Daten, die auf einem höheren Skalenniveau operationalisiert und erhoben wurden, auf ein niedrigeres Skalenniveau herunterzutransformieren. Dies ist jedoch mit einem Informationsverlust verbunden. Zudem gibt es einige wenige Methoden, die es erlauben, vorhandene Daten in ein höheres Skalenniveau zu überführen.

Daten auf ein niedrigeres Skalenniveau transformieren

Drei verschiedene Varianten des Heruntertransformierens sind zu unterscheiden: 1.) Kardinaldaten können in Ordinaldaten, 2.) Kardinaldaten können in Nominaldaten und 3.) Ordinaldaten können in Nominaldaten überführt werden. Dies kommt in der Praxis vor allem vor, wenn

- im Nachhinein Zweifel an der höheren Skalenqualität bestehen,
- die Merkmalsverteilung in der Stichprobe sehr asymmetrisch ist,
- Daten aus unterschiedlichen Studien, die auf verschiedenen Skalenniveaus gemessen wurden, verglichen werden sollen,
- Daten für eine Auswertung als binäre Indikatorvariablen kodiert werden sollen (▶ Abschn. 12.5.2),
- die Daten für die statistische Auswertung gruppiert werden sollen (z. B. Gruppenvergleich) oder
- die Daten für eine grafische Darstellung vereinfacht werden sollen (z. B. Histogramm).

Transformation einer Kardinalskala in eine Ordinalskala

Von einer Stichprobe von $n = 12$ Schülern (Jahrgang 1940) seien die Fehlerzahlen im Diktat (Kardinalskala) bekannt. Wie diese Kardinalskala in ordinalskalierte Daten bzw. Rangplätze zu überführen ist, zeigt Tab. 8.11. Hierbei treten auch sog. **Verbundränge** („ties") auf, wenn mehrere Schüler dieselbe Fehlerzahl zeigen.

In diesem Beispiel haben 4 Schüler in einem Diktat 0 Fehler erreicht. Ihr Rangplatz entspricht dem mittleren Rangplatz derjenigen Ränge, die zu vergeben wären, wenn die gleichen Schüler verschiedene, aufeinander folgende Rangplätze erhalten hätten. Dies sind die Rangplätze 1, 2, 3 und 4, d. h., diese 4 Schüler erhalten den Verbundrang $(1 + 2 + 3 + 4)/4 = 2.5$. Es folgen 2 Schüler mit jeweils einem Fehler, denen als Verbundrang der Durchschnitt der Rangplätze 5 und 6, also 5.5, zugeordnet wird. Die nächsthöhere Fehlerzahl (3 Fehler) kommt nur einmal vor, d. h., dieser Schüler erhält den Rangplatz 7. Die folgenden 3 Schüler mit jeweils 4 Fehlern teilen sich den Rangplatz $(8 + 9 + 10)/3 = 9$, der Schüler mit 5 Fehlern erhält den Rangplatz 11, und dem Schüler mit 7 Fehlern wird schließlich der Rangplatz 12 zugewiesen. Durch diese Umgangsweise mit Verbundrängen bleibt die Rangsumme – unabhängig von der Anzahl der Rangbindungen – konstant, was für die statistische Auswertung von Ordinaldaten wichtig ist. In anderen Kontexten (z. B. im Sport) wird mit Rangbindungen anders umgegangen, so dass z. B. zwei gleichberechtigte zweite Plätze existieren können.

Tabelle 8.11 Transformation einer Kardinalskala in eine Ordinalskala

Schüler	Fehlerzahl (Kardinalskala)	Rangplatz (Ordinalskala)
Kurt	0	2.5
Fritz	7	12
Alfred	4	9
Willi A.	5	11
Detlef	1	5.5
Dieter	1	5.5
Konrad	0	2.5
Heinz	3	7
Karl	4	9
Siegurt	0	2.5
Bodo	4	9
Willi R.	0	2.5

Ob die im vorliegenden Beispiel durchgeführte Transformation eines kardinalskalierten (bzw. durch den absoluten Nullpunkt sogar verhältnisskalierten) Merkmals „Fehlerzahl" in ein ordinalskaliertes Merkmal „Fehlerrangplatz" sinnvoll ist, hängt von den Zielsetzungen und Voraussetzungen der weiteren Datenanalyse ab und von der Genauigkeit der kardinalskalierten Daten. Im Beispiel könnte es fraglich sein, was als Fehler zu bewerten

ist, so dass „sicherheitshalber“ statt der kardinalen Informationen nur die ordinalen Informationen für weitere Analysen verwertet werden sollten. Durch die Transformation kardinaler Daten auf ordinales Datenniveau entsteht eine **objektive** Rangreihe (d. h. eine Rangreihe, die aus genaueren Daten objektiv abgeleitet wurde). Von einer **originären** Rangreihe dagegen sprechen wir, wenn mittels direkter oder indirekter Rangordnung unmittelbar Ordinaldaten erhoben werden.

■ Transformation einer Kardinalskala in eine Nominalskala

Kardinalskalierte Daten, insbesondere wenn diese sehr fein abgestuft sind, werden für bestimmte Auswertungen sowie für grafischen Darstellungen nicht selten **kategorisiert** bzw. in Gruppen eingeteilt, wobei die Merkmalsausprägungen oder die Merkmalsverteilungen zugrunde gelegt werden können. Bei den so gebildeten Kategorien handelt es sich um **geordnete Kategorien**, die über eine Nominalskala (unterschiedliche Gruppen) hinausgehend Ranginformation beinhalten. Es hat sich jedoch eingebürgert, kategorisierte Kardinaldaten (insbesondere bei geringer Zahl der Kategorien) als nominalskaliert zu bezeichnen.

Kategorienbildung gemäß Merkmalsausprägung Die Beschreibung von Untersuchungsobjekten durch quantitative Merkmale wie z. B. Körpergröße, Reaktionszeit, Testleistung, Pulsfrequenz etc. (kardinalskalierte Merkmale) beginnt mit der **Urliste**, d. h. mit einer Auflistung aller individuellen Merkmalsausprägungen, die sämtliche Informationen für weitere statistische Berechnungen enthält. Für die elektronische Datenerfassung und -verarbeitung besteht die Urliste in der Regel aus einer Rohdatendatei. Um sich ein Bild von der Verteilungsform des Merkmales zu verschaffen (z. B. um zu erkennen, ob Intelligenztestwerte bei Realschülern anders verteilt sind als bei Gymnasialschülern), ist es erforderlich, das Merkmal in Kategorien einzuteilen. Die Häufigkeiten in diesen Kategorien sind dann die Grundlage einer tabellarischen oder grafischen Darstellung des Datenmaterials.

Es ist darauf zu achten, dass die Kategorien weder zu eng noch zu breit sind, was letztlich auf die Festlegung der Anzahl der Kategorien hinausläuft. Zu breite Kategorien verdecken möglicherweise typische Eigenarten der Verteilungsform. Zu enge Kategorien führen dagegen zu einer überdifferenzierten Verteilungsform, in der zufällige Irregularitäten das Erkennen der eigentlichen Verteilungsform erschweren. Letzteres wird umso eher der Fall sein, je kleiner die Anzahl der Untersuchungsobjekte ist. Bei der Verwendung von Statistiksoftware ist darauf zu achten, dass die vom System vorgegebenen Kategorienbreiten keinesfalls immer die optimalen sind und nicht kritiklos übernommen werden sollten (ausführlichere Informationen zur Kategorisierung kardinaler Merkmale findet man z. B. bei Bortz, Lienert, & Boehnke, 2000, ► Abschn. 3.2).

In ► Kategorisierung des kardinalskalierten Merkmals Weitsprungleistung wird aus einer Urliste der Weitsprungleistungen von 500 Schülern eine Häufigkeitsverteilung erstellt.

Kategorisierung des kardinalskalierten Merkmals Weitsprungleistung

(Kategorienbildung gemäß Merkmalsausprägung)

Das folgende Beispiel zeigt, wie man aus einer Urliste Kategorien bildet bzw. ein kardinalskaliertes Merkmal in ein kategoriales Merkmal transformiert. Es handelt sich um Weitsprungleistungen (in Metern mit 2 Nachkommastellen) von $n = 500$ Schülern (auf die Wiedergabe der Urliste wird aus Platzgründen verzichtet). Die Einzelwerte werden in 11 gleichgroße Kategorien gruppiert, jede Kategorienbreite beträgt 25 cm.

Kategorienbildung des Merkmals Weitsprungleistung nach Merkmalsausprägungen:

Anzahl der Kategorien	11
Größter Wert	6.10 m
Kleinster Wert	3.40 m
Variationsbreite („range“)	6.10 m − 3.40 m = 2.70 m
Kategorienbreite	0.25 m
Kategoriengrenzen	3.40–3.64 m
	3.65–3.89 m
	3.90–4.14 m
	...

Berechnung der Kategorienmitten (veranschaulicht an der 1. Kategorie):

$$\frac{3.40 + 3.64}{2} = 3.52$$

Häufigkeitsverteilung anhand der neu gebildeten Kategorien: Häufigkeitsverteilung der gebildeten Kategorien zur Weitsprungleistung:

Kategoriengrenzen (in Metern)	Kategorienmitten (in Metern)	Absolute Häufigkeiten ($n = 500$)
3.40–3.64	3.52	1
3.65–3.89	3.77	9
3.90–4.14	4.02	18
4.15–4.39	4.27	33
4.40–4.64	4.52	51
4.65–4.89	4.77	108
4.90–5.14	5.02	192
5.15–5.39	5.27	61
5.40–5.64	5.52	19
5.65–5.89	5.77	6
5.90–6.14	6.02	2

Diese Häufigkeitsverteilung veranschaulicht ◘ Abb. 8.5 grafisch als Liniendiagramm.

Enthält eine Urliste Extremwerte, so dass mehrere aufeinander folgende Kategorien unbesetzt blieben, verwendet man einfachheitshalber **offene Randkategorien**, in die alle Werte gezählt werden, die größer sind als die Obergrenze der obersten Kategorie oder kleiner als die Untergrenze der untersten Kategorie. Für weitere mathematische Berechnungen sind derart gruppierte Daten allerdings unbrauchbar, es sei denn, die genauen Extremwerte sind bekannt.

Gruppenbildung gemäß Merkmalsverteilung (Perzentile)

Bei der oben dargestellten Kategorienbildung für die Weitsprungleistung gemäß Merkmalsausprägung (jede Kategorienbreite beträgt 25 cm Weitsprungleistung) resultieren 11 Kategorien bzw. Weitsprung-Leistungsgruppen mit sehr unterschiedlicher Kategorienbesetzung (von einer Person pro Kategorie bis 192 Personen pro Kategorie). Zuweilen möchte man Stichproben anhand eines interessierenden kardinalskalierten Merkmals jedoch in **gleichgroße Gruppen** bzw. in **Gruppen definierter Größe** einteilen. Hierfür können die **Perzentile** bzw. **Prozentränge** herangezogen werden. Sie teilen die Stichprobe in 1 %-Scheiben auf. Angenommen, man möchte eine Stichprobe von Arbeitnehmerinnen und Arbeitnehmern anhand der erhobenen Intelligenztestwerte in vier gleichgroße Gruppen einteilen: Dazu würde man den Minimal- und Maximalwert der Stichprobe sowie die 25., 50. und 75. Perzentile benötigen, um die vier gleichgroßen Gruppen zu bilden. Nach demselben Prinzip können auch drei oder fünf gleichgroße Gruppen gebildet werden. Im Unterschied zur Kategorienbildung gemäß Merkmalsausprägung entstehen bei der Gruppenbildung gemäß Merkmalsverteilung jetzt Kategoriengrenzen, die keinen gleichabständigen Merkmalsausprägungen entsprechen.

Ebenso lassen sich **Extremgruppen** herausgreifen, indem alle Messwerte einerseits vom Minimalwert bis zum 10. Perzentil sowie andererseits vom 90. Perzentil bis zum Maximalwert herausgegriffen werden. Die 10 % Arbeitnehmer mit dem geringsten und die 10 % mit dem höchsten IQ könnten dann z. B. hinsichtlich ihrer Arbeitsleistung, Arbeitszufriedenheit etc. verglichen werden. Derartige **Extremgruppenvergleiche** können explorativ interessant sein, führen jedoch tendenziell zu einer Überschätzung von Effekten. Besonders problematisch sind Extremgruppen, wenn mit Messwiederholungen gearbeitet wird: Der hierbei auftretende Methodeneffekt der „Regression zur Mitte“ (▶ Abschn. 12.5.4) darf dann nicht mit dem interessierenden inhaltlichen Effekt gleichgesetzt werden.

Zuweilen gibt es Anlass, eine Stichprobe anhand der Ausprägung des interessierenden Merkmals in **zwei gleichgroße Gruppen** einzuteilen. Hierfür wird an der 50. Perzentile bzw. am Medianwert die Gruppeneinteilung vorgenommen. Dieses Vorgehen wird auch als **Mediansplit** bezeichnet. Ein Heruntertransformieren von einem fein abgestuften kardinalskalierten Merkmal in eine polytome oder dichotome nominalskalierte Variable erfolgt z. B., wenn die Variable als unabhängige Variable fungiert und Gruppenvergleiche durchgeführt werden sollen. Ebenso wie die Extremgruppenbildung birgt auch das Mediansplit-Verfahren methodische Probleme bei der statistischen Auswertung und Interpretation der Ergebnisse. Zunächst einmal führt das Heruntertransformieren von verhältnis- bzw. intervallskalierten Daten auf eine dichotome Variable zu einem Informationsverlust: Mit der Reduktion der Varianz der Messwerte wird auch die Teststärke statistischer Signifikanztests verringert (zur Teststärke ▶ Abschn. 14.1). Bei mehrfaktoriellen Designs (z. B. wenn zwei durch Mediansplit gebildete unabhängige Variablen verwendet werden) können sich Scheineffekte ergeben (Maxwell & Delaney, 1993). Nicht zuletzt ist auch inhaltlich zu hinterfragen, welchen Stellenwert Gruppen haben (z. B. „Vielnutzer“ vs. „Wenignutzer“ eines Medienangebots), die nicht anhand theoretisch festgelegter Kriterien gebildet werden, sondern anhand eines – von Stichprobe zu Stichprobe schwankenden – Medianwertes. Bevor also ein Mediansplit vorgenommen wird, sollte eruiert werden, ob nicht a) eine theoretisch fundierte Gruppenbildung sinnvoller ist oder b) anstelle von Hypothesen über Gruppenunterschiede, die typischerweise mit *t*-Tests oder Varianzanalysen geprüft werden (▶ Abschn. 12.5.3), lieber Zusammenhangshypothesen zu formulieren und die kardinalskalierten Variablen ohne Transformation als Prädiktoren im Rahmen der bivariaten oder multivariaten Korrelations- und Regressionsanalyse einzusetzen sind (▶ Abschn. 12.5.2).

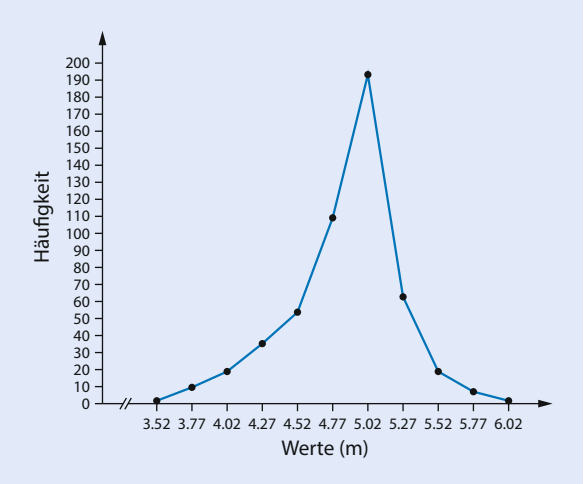


Abb. 8.5 Grafische Darstellung der Häufigkeitsverteilung der Kategorien zur Weitsprungleistung

- **Transformation einer Ordinalskala in eine Nominalskala**

Liegen ordinalskalierte Variablen mit sehr vielen Abstufungen vor, so kann bei Bedarf – analog wie bei kardinalskalierten Variablen – eine Kategorienbildung gemäß Merkmalsausprägung oder eine Gruppenbildung gemäß Merkmalsverteilung durchgeführt werden.

Daten auf ein höheres Skalenniveau transformieren

Normalerweise ist es nicht möglich, bereits erhobene Daten von einem informationsärmeren Skalenniveau auf ein informationsreicheres Niveau zu heben. Eine Ausnahme stellt – unter bestimmten Bedingungen – die Transformation ordinalskalierter Daten auf Intervallskalenniveau dar. Wurden Ordinaldaten durch direkte Rangordnung erhoben, so lassen sich diese nach dem „**Law of Categorial Judgement**“ in intervallskalierte Daten transformieren. Analog lassen sich Ordinaldaten, die mittels indirekter Rangordnung (Dominanzpaarvergleiche) generiert wurden, gemäß dem „**Law of Comparative Judgement**“ in Intervalldaten überführen. Eine weitere, mathematisch anspruchsvolle Methode zur Überführung von Ordinaldaten in Kardinaldaten stellt das hier nicht näher behandelte Markovketten-Monte-Carlo-Verfahren („markov chain monte carlo“; MCMC) dar (Granberg-Rademacker 2010).

- **Transformation einer direkten Rangordnung in eine Intervallskala mittels „Law of Categorial Jugdement“**

Die Grundidee des Skalierungsansatzes des „Law of Categorial Judgement“ geht auf den US-amerikanischen Psychologen Louis Leon Thurstone (1927) zurück. Nach Thurstones Terminologie basiert die Einschätzung der Merkmalsausprägungen von Objekten hinsichtlich psychologischer Variablen auf einem Diskriminationsprozess, der die Basis aller Identifikations- und Diskriminationsurteile darstellt. Jedem zu beurteilenden Objekt ist ein derartiger Diskriminationsprozess zugeordnet. Organismische Fluktuationen haben zur Konsequenz, dass Empfindungen, die ein Objekt bei wiederholter Darbietung auslöst, nicht identisch sind, sondern um einen „wahren“ Wert oszillieren. Es resultiert eine **Empfindungsstärkenverteilung**, von der angenommen wird, sie sei eine „glockenförmige“ Verteilung (Normalverteilung). Wird ein Objekt nicht wiederholt von einem Beurteiler, sondern einmal von vielen Beurteilern eingestuft, gilt die Annahme der Normalverteilung entsprechend auch für diese Urteile.

Für das „Law of Categorical Judgement“ – das gleiche Skalierungsprinzip wurde unter dem Namen „Method of Successive Categories“ von Guilford (1938), als „Method of Graded Dichotomies“ von Attneave (1949) und als „Method of Discriminability“ von Garner und Hake (1951) publiziert – resultieren hieraus die folgenden Annahmen (Torgerson 1958, Kap. 10):

1. Der Urteiler ist in der Lage, das Merkmalskontinuum in eine bestimmte Anzahl ordinaler Kategorien aufzuteilen.
2. Die Grenzen zwischen diesen Kategorien sind keine festen Punkte, sondern schwanken um bestimmte Mittelwerte.
3. Die Wahrscheinlichkeit für die Realisierung einer bestimmten Kategoriengrenze folgt einer Normalverteilung.
4. Die Beurteilung der Merkmalsausprägung eines bestimmten Objektes ist nicht konstant, sondern unterliegt zufälligen Schwankungen.
5. Die Wahrscheinlichkeit für die Realisierung eines bestimmten Urteils folgt ebenfalls einer Normalverteilung.
6. Ein Urteiler stuft ein Objekt unterhalb einer Kategoriengrenze ein, wenn die im Urteil realisierte Merkmalsausprägung des Objektes geringer ist als die durch die realisierte Kategoriengrenze repräsentierte Merkmalsausprägung.

Werden die Objekte wiederholt von einem Urteiler oder – was üblicher ist – einmal von mehreren Urteilern nach der Methode der sukzessiven Intervalle geordnet, erhalten wir für jede Rangkategorie Häufigkeiten, die angeben, wie oft ein bestimmtes Objekt in die einzelnen Rangkategorien eingeordnet wurde (► Emotionale Wärme in der Gesprächspsychotherapie).

Das Beispiel für das „Law of Categorical Judgement“ zeigt, wie nach Einführung einiger Modellannahmen aus einfachen ordinalen Informationen eine skalentheoretisch höherwertige Skala (Intervallskala) entwickelt werden kann. Dies setzt allerdings voraus, dass die Urteilsvorgänge in der von Thurstone beschriebenen Weise ablaufen. (Über Verfahren zur Überprüfung der Modellannahmen berichtet Torgerson 1958, S. 240f.).

Die Modellannahmen betreffen vor allem die Normalverteilung, die z. B. gefährdet ist, wenn Objekte mit extremen Merkmalsausprägungen zu beurteilen sind (Carbonell, Sendra, Bayarri, Izquierdo, & Tárrega, 2008; Hofacker 1984; Saito 1994). Extrem starke Merkmalsausprägungen werden eher unterschätzt (Attneave 1949), und extrem schwache Merkmalsausprägungen werden eher überschätzt, d. h., es werden rechtssteile bzw. linkssteile Urteilsverteilungen begünstigt (Lei & Lomax, 2005). Rozeboom und Jones (1956) konnten allerdings zeigen, dass die Ergebnisse, die nach dem „Law of Categorical Judgement“ erzielt werden, durch nichtnormale Empfindungsstärkenverteilungen wenig beeinflusst sind. Nach Jones (1959) sind sie zudem invariant gegenüber verschie-

Emotionale Wärme in der Gesprächspsychotherapie: ein Beispiel für das „Law of Categorical Judgement"

50 Studierende eines Einführungskurses in Gesprächspsychotherapie wurden gebeten, das Merkmal „Emotionale Wärme des Therapeuten" in 5 Therapieprotokollen einzustufen. Die Einstufung erfolgte anhand der folgenden 5 Rangkategorien: Therapeut zeigt sehr viel emotionale Wärme (= 1); Therapeut zeigt viel emotionale Wärme (= 2); Therapeut wirkt neutral (= 3); Therapeut wirkt emotional zurückhaltend (= 4); Therapeut wirkt emotional sehr zurückhaltend (= 5). Die 5 Therapieprotokolle wurden von den 50 Urteilern in der in Tab. 8.12 dargestellten Weise eingestuft.

Die fett markierte Zahl 12 in Tab. 8.12 besagt also, dass der Therapeut in Protokoll C nach Ansicht von 12 Studierenden viel emotionale Wärme zeigt (Kategorie 2). Die Zahlen addieren sich zeilenweise zu 50, weil jeweils 50 Studierende jedes Protokoll beurteilt haben.

Die absoluten Häufigkeiten lassen sich in relative Häufigkeiten überführen, indem sie durch die Gesamtzahl der Urteilenden dividiert werden. Die Aussage, dass $f_{abs} = 12$ Urteilende bei Protokoll C „viel emotionale Wärme" angegeben haben, ist gleichbedeutend mit der Aussage, dass $f_{rel} = 12/50 = 24\,\%$ der Urteilenden den Therapeuten so eingeschätzt haben (Tab. 8.13).

Diese relativen Häufigkeiten werden im nächsten Schritt zeilenweise über die einzelnen Urteilsstufen kumuliert (kumulierte relative Häufigkeiten). Dementsprechend ergibt sich, dass Therapeut C von $20\,\% + 24\,\% = 44\,\%$ der Urteilenden sehr viel oder viel emotionale Wärme zugeschrieben wurde (Tab. 8.14).

Der im ▶ Webangebot zu diesem Buch wiedergegebenen Standardnormalverteilungstabelle (Tab. 1) wird nun entnommen, wie die z-Werte (Abszissenwerte der Standardnormalverteilung) lauten, die die oben aufgeführten relativen Häufigkeiten (oder Flächenanteile) von der Standardnormalverteilung abschneiden (Tab. 8.15).

Der fett markierte Wert (−0.15) besagt also, dass sich in der Standardnormalverteilung zwischen $z = -\infty$ und $z = -0.15$ ein Flächenanteil von 44 % befindet. Die letzte Spalte (Urteilskategorie 5) bleibt unberücksichtigt, weil die kumulierten relativen Häufigkeiten in dieser Spalte alle $1.00 = 100\,\%$ (mit $z \to +\infty$) betragen.

Die Kategoriengrenzen entsprechen den Spaltenmittelwerten. Der Wert −0.91 markiert die Grenze zwischen den Kategorien „sehr viel emotionale Wärme" (1) und „viel emotionale Wärme" (2), der Wert −0.03 die Grenze zwischen „viel emotionale Wärme" (2) und „neutral" (3) etc.

Die Merkmalsausprägungen für die beurteilten Protokolle ergeben sich als Differenzen zwischen der durchschnittlichen Kategoriengrenze (0.33) und den Zeilenmittelwerten. Für Protokoll A resultiert also der Skalenwert $0.33 - (-0.61) = 0.94$. Insgesamt ergeben sich für die 5 Therapieprotokolle Ausprägungen in Bezug auf das Merkmal „emotionale Wärme", die in der letzten Spalte aufgeführt sind. Da es sich – wenn die Annahmen des „Law of Categorical Judgement" zutreffen – hierbei um Werte einer Intervallskala handelt, könnte zu allen Werten der Betrag des kleinsten Skalenwerts (−0.74) addiert werden; man erhält dadurch neue Werte auf einer Skala mit einem (künstlichen) Nullpunkt. Nach der hier gewählten Abfolge der Urteilskategorien wird im Protokoll E am meisten (trotz negativem Wert) und im Protokoll A am wenigsten emotionale Wärme gezeigt. Der Unterschied zwischen den Protokollen A und B ist größer als der zwischen B und C.

Tabelle 8.12 Absolute Häufigkeiten der Urteilsstufen

Urteilskategorien	1	2	3	4	5
Protokoll A	2	8	10	13	17
Protokoll B	5	10	15	18	2
Protokoll C	10	**12**	20	5	3
Protokoll D	15	20	10	3	2
Protokoll E	22	18	7	2	1

Tabelle 8.13 Relative Häufigkeiten der Urteilsstufen

Urteilskategorien	1	2	3	4	5
Protokoll A	.04	.16	.20	.26	.34
Protokoll B	.10	.20	.30	.36	.04
Protokoll C	.20	**.24**	.40	.10	.06
Protokoll D	.30	.40	.20	.06	.04
Protokoll E	.44	.36	.14	.04	.02

Tabelle 8.14 Kumulierte relative Häufigkeiten der Urteilsstufen

Urteilskategorien	1	2	3	4	5
Protokoll A	.04	.20	.40	.66	1.00
Protokoll B	.10	.30	.60	.96	1.00
Protokoll C	.20	**.44**	.84	.94	1.00
Protokoll D	.30	.70	.90	.96	1.00
Protokoll E	.44	.80	.94	.98	1.00

denen Urteilerstichproben, verschiedenen Kategorienbezeichnungen sowie der Anzahl der Kategorien.

Der in ▶ Emotionale Wärme in der Gesprächspsychotherapie wiedergegebene Rechengang geht davon aus, dass die Kovarianzen der Verteilungen von Kategoriengrenzen und Urteilsobjekten Null und die Varianzen der Verteilungen der Kategoriengrenzen konstant sind.

Tabelle 8.15 Flächenanteile der Standardnormalverteilung der relativen Häufigkeiten

Urteilskategorien	1	2	3	4	Zeilensummen	Zeilenmittel	Merkmals-ausprägung
Protokoll A	−1.75	−0.84	−0.25	0.41	−2.43	−0.61	0.94
Protokoll B	−1.28	−0.52	0.25	1.75	0.20	0.05	0.28
Protokoll C	−0.84	**−0.15**	0.99	1.55	1.55	0.39	−0.06
Protokoll D	−0.52	0.52	1.28	1.75	3.03	0.76	−0.43
Protokoll E	−0.15	0.84	1.55	2.05	4.29	1.07	−0.74
Spaltensummen	−4.54	−0.15	3.82	7.51	–	–	–
Kategoriengrenzen	−0.91	−0.03	0.75	1.50	–	0.33	–

8

Transformation einer indirekten Rangordnung in eine Intervallskala mittels „Law of Comparative Judgement“

Zur Veranschaulichung des „Law of Comparative Judgement“ (Thurstone 1927) greifen wir das bei der Darstellung von Dominanzpaarvergleichen und daraus resultierender indirekter Rangordnung bereits verwendete Beispiel zur Beurteilung von Schulfächern noch einmal auf (▶ Ist Sport beliebter als Englisch? Ein Beispiel für eine Paarvergleichsskalierung nach dem „Law of Comparative Judgement“; für ein aktuelles Studienbeispiel aus der Gesundheitsforschung Krabbe 2008).

Der Grundgedanke des **„Law of Comparative Judgement“** lässt sich vereinfacht in folgender Weise charakterisieren: Wie schon beim „Law of Categorical Judgement“ wird davon ausgegangen, dass wiederholte Beurteilungen einer Merkmalsausprägung nicht identisch sind, sondern – möglicherweise nur geringfügig – fluktuieren. Es resultiert eine (theoretische) Verteilung der Empfindungsstärken, von der angenommen wird, sie sei um einen „wahren“ Wert normalverteilt. Ein konkretes Urteil stellt dann die Realisierung dieser normalverteilten Zufallsvariablen dar.

Theorie Wegen der Bedeutung dieses Skalierungsansatzes sei der Rechengang im Folgenden ausführlicher begründet: Die Schätzung der Merkmalsausprägungen von zwei Objekten entspricht der Realisierung von zwei normalverteilten Zufallsvariablen. Die Differenz dieser beiden Schätzungen $(x_1 - x_2)$ stellt dann ihrerseits eine normalverteilte Zufallsvariable dar (Differenzen zweier normalverteilter Zufallsvariablen sind ebenfalls normalverteilt). Dividieren wir die Differenz durch die Streuung der Differenzenverteilung (über die im „Law of Comparative Judgement“ unterschiedliche Annahmen gemacht werden, ▶ unten), resultiert ein z-Wert der Standardnormalverteilung. Ein positiver z-Wert besagt, dass $x_1 > x_2$, ein negativer z-Wert, dass $x_1 < x_2$ ist, und $z = 0$ resultiert, wenn $x_1 = x_2$.

Der Wert $z = 0$ schneidet von der Fläche der Standardnormalverteilung 50 % ab (Abb. 8.6). Gleichzeitig gilt, dass für $z = 0$ (bzw. $x_1 = x_2$) im Paarvergleich die Präferenz für einen Reiz zufällig erfolgt, d. h., die Wahrscheinlichkeit, dass ein Reiz dem anderen vorgezogen wird, beträgt ebenfalls 50 %. Ist nun $x_1 > x_2$, resultiert ein positiver z-Wert, der mehr als 50 % der Fläche der Standardnormalverteilung abschneidet. Gleichzeitig ist auch die Wahrscheinlichkeit, dass Reiz 1 dem Reiz 2 vorgezogen wird, größer als 50 %. Auf dieser Korrespondenz basiert die Annahme, dass die Wahrscheinlichkeit, mit der ein Reiz einem anderen vorgezogen wird, dem durch die standardisierte Differenz $(x_1 - x_2)$ abgeschnittenen Flächenanteil der Standardnormalverteilung entspricht (Abb. 8.6).

Die Wahrscheinlichkeit, mit der ein Reiz einem anderen vorgezogen wird, wird aus den Paarvergleichsurteilen geschätzt (relative Häufigkeiten in ▶ Ist Sport beliebter als Englisch?). Gesucht werden nun diejenigen z-Werte, die von der Standardnormalverteilungsfläche genau diese Flächenanteile bzw. Prozentwerte abschneiden. Diese z-Werte repräsentieren die Differenzen zwischen je zwei Reizen auf einer Intervallskala (David 1963).

Der weitere Rechengang einer Paarvergleichsskalierung nach dem „Law of Comparative Judgement“ ist dann relativ problemlos. Wir berechnen die mittlere Abweichung eines jeden Objektes von allen übrigen Objekten und erhalten damit die Skalenwerte. (Die mittlere Abweichung eines Objektes von allen übrigen Objekten entspricht der Abweichung dieses Objektes vom Mittelwert aller übrigen Objekte.) Diese Skalenwerte haben einen Mittelwert von Null, d. h., es treten auch negative Skalenwerte auf. Sie werden vermieden, wenn in einer für

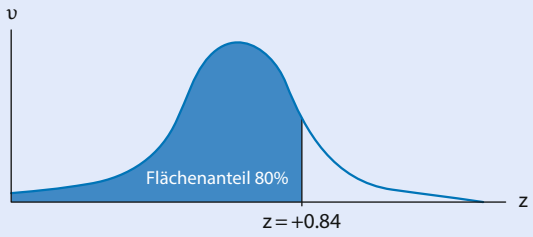


Abb. 8.6 Transformation der relativen Häufigkeiten in *z*-Werte der Standardnormalverteilung (Tab. 8.18)

Ist Sport beliebter als Englisch? Ein Beispiel für eine Paarvergleichsskalierung nach dem „Law of Comparative Judgement"

30 Schüler wurden gebeten, in einem vollständigen Paarvergleich ihre Präferenzen für 5 Unterrichtsfächer anzugeben. Hierfür wurden für die Fächer Deutsch (De), Mathematik (Ma), Englisch (En), Sport (Sp) und Musik (Mu) alle 10 möglichen Paarkombinationen gebildet und jeder Schüler musste angeben, welches der jeweils 2 Fächer seiner Meinung nach das interessantere sei. Aus den Paarvergleichsurteilen resultierte folgende **Dominanzmatrix** (Begründung des Rechenganges im Text; Tab. 8.16).

Die fett gesetzte Zahl gibt an, dass 24 Schüler Englisch interessanter fanden als Mathematik. Die Werte besagen, wie häufig die Fächer, die die Spalten bezeichnen, über die Fächer, die die Zeilen bezeichnen, „dominieren". Einander entsprechende (an der Diagonale gespiegelte) Zellen ergänzen sich jeweils zum Gesamtstichprobenumfang von 30 (6 Schüler finden Mathematik interessanter als Englisch).

Wollte man für alle Schüler eine gemeinsame Rangreihe bestimmen, wären die Spaltensummen nach ihrer Größe zu ordnen. **Es resultiert folgende Beliebtheitsrangreihe gemäß Spaltensummen: Sp > Mu > En > De > Ma.**

Für die weitere Auswertung nach dem **„Law of Comparative Judgement"** werden die oben genannten Präferenzhäufigkeiten in relative Häufigkeiten transformiert, indem sie durch die Anzahl der Schüler ($n = 30$) dividiert werden. Wenn also $f_{abs} = 24$ Schüler Englisch interessanter als Mathematik beurteilen, dann entspricht dies $f_{rel} = 24/30 = 80\,\%$ (Tab. 8.17).

Für die relativen Häufigkeiten entnimmt man der Standardnormalverteilungstabelle (Tab. 1 ▶ Webangebot zum Buch) die in Tab. 8.18 dargestellten *z*-Werte (die Werte in der Diagonale werden Null gesetzt).

Beispiel: $z = +0.84$ ergibt sich deshalb, weil sich zwischen $z = -\infty$ und $z = +0.84$ 80 % der Fläche der Standardnormalverteilung befinden (Flächenanteil: .80 = 80 % gemäß der Tabelle der relativen Häufigkeiten; Tab. 8.18; Abb. 8.6).

Man berechnet als nächstes die Spaltensummen und die Spaltenmittelwerte, deren Summe bis auf Rundungsungenauigkeiten Null ergibt. Addieren wir den Betrag des größten negativen Wertes (−0.63) zu allen Werten, resultieren die Skalenwerte. Offensichtlich ist Mathematik das am wenigsten interessante Fach. Englisch wird für geringfügig interessanter gehalten als Deutsch. Sport halten die befragten Schüler für das interessanteste Fach, dicht gefolgt von Musik.

Tabelle 8.16 Dominanzmatrix für Schulfächer

	De	Ma	En	Sp	Mu
De	–	10	12	24	22
Ma	20	–	**24**	26	23
En	18	**6**	–	19	20
Sp	6	4	11	–	14
Mu	8	7	10	16	–
Summe	52	27	57	85	79
Rangreihe	**4**	**5**	**3**	**1**	**2**

Tabelle 8.17 Transformation der Dominanzmatrix für Schulfächer in relative Häufigkeiten

	De	Ma	En	Sp	Mu
De	–	.33	.40	.80	.73
Ma	.67	–	**.80**	.87	.77
En	.60	.20	–	.63	.67
Sp	.20	.13	.37	–	.47
Mu	.27	.23	.33	.53	–

Tabelle 8.18 Transformation der relativen Häufigkeiten in *z*-Werte der Standardnormalverteilung (Abb. 8.6)

	De	Ma	En	Sp	Mu
De	0.00	−0.44	−0.25	+0.84	+0.61
Ma	+0.44	0.00	**+0.84**	+1.13	+0.74
En	+0.25	−0.84	0.00	+0.33	+0.44
Sp	−0.84	−1.13	−0.33	0.00	−0.07
Mu	−0.61	−0.74	−0.44	+0.07	0.00
Spaltensummen	−0.76	−3.15	−0.18	+2.37	+1.72
Spaltenmittel	−0.15	−0.63	−0.04	+0.47	+0.34
Skalenwerte	**0.48**	**0.00**	**0.59**	**+1.10**	**+0.97**

Intervallskalen zulässigen Lineartransformation zu allen Skalenwerten der Betrag des größten negativen Skalenwertes addiert wird. Dadurch verschiebt sich die gesamte Skala so, dass das Objekt mit der größten negativen Ausprägung den Nullpunkt der Skala repräsentiert. Mit diesen

Skalenwerten können sämtliche für Intervallskalen sinnvolle Operationen durchgeführt werden.

Der hier beschriebene Rechengang geht davon aus, dass alle Empfindungsstärkenverteilungen gleich streuen und dass die Korrelationen zwischen den Verteilungen konstant sind. Über den Rechengang, der sich für andere Annahmen bzgl. der Streuungen und Korrelationen ergibt, sowie über weitere Spezialprobleme (z. B. Wahrscheinlichkeitswerte von Null oder Eins, Tests zur Überprüfung der Güte der Skalierung, iterative Methoden für die Bestimmung der Skalenwerte etc.) berichten z. B. Sixtl (1967, Kap. 2c) und Torgerson (1958, Kap. 9). Intransitive Urteile in Paarvergleichsskalierungen behandeln Hull und Buhyoff (1981).

8

Unvollständige Paarvergleiche Paarvergleichsurteile geraten schnell zu einer mühevollen Aufgabe für die Urteilenden, wenn die Anzahl der zu skalierenden Objekte wächst. Resultiert für 10 Objekte die noch zumutbare Anzahl von 45 Paarvergleichen, sind bei 20 Objekten bereits 190 Paarvergleiche erforderlich – eine Aufgabe, die zumindest bei schwierigen Paarvergleichen das Konzentrations- und Durchhaltevermögen der Urteilenden übersteigen dürfte. In diesem Fall sollte statt des „Law of Comparative Judgement" das „Law of Categorical Judgement" (▶ Abschn. 8.4.6) eingesetzt werden, wenngleich Skalierungen nach dem „Law of Comparative Judgement" in der Regel zu stabileren Resultaten führen (Kelley, Hovland, Schwartz, & Abelson, 1955).

Es gibt jedoch auch Möglichkeiten, den Arbeitsaufwand für eine Paarvergleichsskalierung zu reduzieren. Sollen z. B. 20 Objekte skaliert werden, wählt man ca. 6 Objekte aus, die ein möglichst breites Spektrum des Merkmalskontinuums mit annähernd äquidistanten Abständen repräsentieren. Diese 6 Ankerobjekte werden untereinander und mit den verbleibenden 14 Objekten verglichen, so dass insgesamt statt der ursprünglich 190 nur noch

$$\binom{6}{2} + 14 \cdot 6 = 99$$

Paarvergleiche erforderlich sind. Die durchschnittlichen z-Werte basieren dann bei den Ankerobjekten jeweils auf 19 und bei den übrigen Objekten jeweils auf 6 relativen Häufigkeiten. Über weitere Möglichkeiten, den Aufwand bei Paarvergleichsskalierungen zu reduzieren, berichten Torgerson (1958, S. 191ff.), van der Ven (1980, Kap. 9.1) und Clark (1977).

Bei Chignell und Pattey (1987) findet man eine vergleichende Übersicht verschiedener Techniken, die es gestatten, mit einem reduzierten Paarvergleichsaufwand eindimensionale ordinale Skalen zu konstruieren.

Der am häufigsten gegen das „Law of Comparative Judgement" vorgebrachte Einwand betrifft die Annahme der normalverteilten Empfindungsstärken. Dieser Annahme folgend sind die Differenzen zwischen je zwei Objekten und die Wahlwahrscheinlichkeiten für Objektpräferenzen im Paarvergleich über die Verteilungsfunktion der Standardnormalverteilung miteinander verknüpft. Diese funktionale Verknüpfung wird von Bradley und Terry (1952) durch eine logistische Funktion ersetzt. Wie Sixtl (1967, S. 209ff.) jedoch zeigt, sind die Skalierungsergebnisse nach der von Bradley und Terry vorgeschlagenen Methode praktisch mit denen des „Law of Comparative Judgement" identisch, es sei denn, die relativen Häufigkeiten für die Objektpräferenzen basieren auf mehr als 2 000 Urteilen.

Ähnliches gilt für die von Luce (1959) vorgenommene Erweiterung des Modells von Bradley und Terry, bekannt als Bradley-Terry-Luce-(BTL-)Modell oder auch als Luce'sches Wahlaxiom. Nach Coombs, Dawes und Tversky (1970, S. 152) sind die nach diesem Ansatz erzielten Skalierungsergebnisse mit den Ergebnissen, die nach dem Thurstone-Modell ermittelt werden, praktisch identisch.

Subkoviak (1974) ging der Frage nach, wie sich Verletzungen der Modellannahmen des „Law of Comparative Judgement" auf das Skalierungsergebnis auswirken. Verletzungen der Normalverteilungsvoraussetzung vermochten die Skalierungsergebnisse nur unbedeutend zu beeinflussen. Ernsthafte Skalierungsfehler traten erst bei extrem heterogenen Verteilungsformen auf (Jones & Thurstone, 1955; Mosier 1941; Rambo 1963).

8.5 Messung mit Einzelindikatoren

Wir haben die Operationalisierung theoretischer Konzepte definiert als die Auswahl von beobachtbaren Indikatoren sowie die Festlegung der Messmethode, was gleichzeitig dann auch das Skalenniveau der erhobenen Daten bestimmt. Ebenso haben wir schon hervorgehoben, dass ein einfaches manifestes Merkmal meist über einen einzelnen Indikator operationalisiert wird, während für die Erfassung komplexer latenter Konstrukte oft multiple Indikatoren herangezogen und miteinander zu Skalen oder Indizes kombiniert werden.

Dieser Abschnitt widmet sich der Messung mit Einzelindikatoren, die für manifeste Variablen gängig und in Ausnahmefällen auch für latente Merkmale möglich ist.

8.5.1 Messung manifester Variablen mit Einzelindikatoren

Manifeste Variablen (► Abschn. 8.2) sind theoretisch in der Regel nicht weiter erklärungsbedürftig. Für sie werden selten ausdrückliche Nominaldefinitionen geliefert, auch erübrigen sich Konzeptspezifikationen mittels Bedeutungsanalyse oder dimensionaler Analyse. Stattdessen wird oft einfach nur die Operationalisierung durch einen Einzelindikator (z. B. ein Fragebogenitem) angegeben. Dies trifft z. B. auf die in jeder empirischen Untersuchung erhobenen **soziodemografischen Merkmale** (z. B. Alter, Wohnort, Geschlecht, Bildungsniveau, Familienstand, Nationalität etc.) zu. Auch andere einfache Personen-, Verhaltens- oder Umweltvariablen (z. B. Allergien, TV-Nutzungszeit, Wohnungsgröße) werden als manifeste Merkmale über nicht näher diskutierte Einzelindikatoren erfasst, bei denen **Augenschein-Validität** unterstellt wird. Ein ausführliches Beispiel für die Operationalisierung des manifesten Merkmals „Rauchverhalten" mittels Einzelindikatoren auf verschiedenen Skalenniveaus ist in ► Abschn. 8.4 zu finden.

Da die Operationalisierung soziodemografischer Merkmale in jeder empirischen Studie ansteht und dabei von Studierenden immer wieder ähnliche Fehler gemacht werden, möchten wir dazu noch einige praktische Hinweise geben (► Vier praktische Hinweise zur Messung soziodemografischer Merkmale mittels Einzelindikatoren). Soziodemografische Merkmale dienen vor allem der **Stichprobenbeschreibung**: Man möchte wissen, welche Merkmale die untersuchten Personen haben, auch um die eigene Stichprobe mit anderen Studien vergleichen zu können. Der gängige Satz an soziodemografischen Variablen wird meist um einige **Hintergrund-Variablen passend zum Studienthema** ergänzt. Diese können dann auch für die inhaltliche Auswertung von Interesse sein. Üblicherweise werden soziodemografische Daten am Ende (manchmal auch am Anfang) einer Untersuchung per Fragebogen erhoben, dies gilt auch für Studien, in denen andere Datenerhebungsverfahren (z. B. Interview, Beobachtung) im Zentrum stehen. Die Operationalisierung erfolgt also mittels **Fragebogen-Items**, wobei pro Variable in der Regel nur eine einzige Frage eingesetzt wird (Einzelindikator).

8.5.2 Messung latenter Variablen mit Einzelindikatoren

Für die Operationalisierung latenter Variablen bzw. theoretischer Konstrukte wird in der Regel ein sehr viel größerer Aufwand betrieben als bei manifesten Variablen. Hier liegen nach komplizierten Verfahren entwickelte und überprüfte psychometrische Skalen und Indizes vor. Die methodischen und ethischen Ansprüche, die an die Operationalisierung latenter Merkmale gestellt werden, sind umso größer, je wichtiger die jeweilige Variable für die betreffende Studie ist und je größer die Konsequenzen für die Betroffenen sind. So werden im Rahmen der Diagnostik grundsätzlich nur wissenschaftlich geprüfte psychologische Testverfahren eingesetzt, die die interessierenden Merkmale über multiple Indikatoren erfassen und die Ergebnisse mit Normwerten vergleichen (zu psychologischen Tests ► Abschn. 10.4).

Für empirische Studien, die auf Aggregatwerte (Gruppenwerte) hinauslaufen und bei denen manche Variablen eher am Rande des Interesses stehen, sind Operationalisierungen latenter Variablen mittels Einzelindikatoren vertretbar – wenn nicht sogar aus forschungsökonomischen Gründen notwendig, um die Untersuchungspersonen von überlangen und ermüdenden Erhebungsinstrumenten zu entlasten sowie die Forschungskosten (z. B. Zeitdauer bei standardisierten mündlichen Interviews) zu reduzieren (Gardner, Cummings, Dunham, & Pierce, 1998; Nagy 2002). In der praxisnahen Forschung sind Erhebungen mittels Einzelindikatoren sehr beliebt, da die Operationalisierung deutlich weniger aufwändig ist als die Entwicklung einer psychometrischen Skala oder eines Index (Bergkvist & Rossiter, 2007).

Die Methodenforschung hat gezeigt, dass eine Messung über Einzelitems („**Single-Item-Measures**") auch bei theoretisch anspruchsvollen latenten Variablen die üblichen Testgütekriterien erfüllen kann und somit brauchbar ist. Einige **psychometrisch überprüfte** Einzelitem-Messverfahren aus verschiedenen Inhaltsbereichen stellt exemplarisch ◘ Tab. 8.20 vor. Es wird auch argumentiert, dass die Testgüte von Einzelitems der von Mehr-Item-Testskalen sogar überlegen sein kann (Bergkvist & Rossiter, 2007; Nagy 2002). Für eine Methode zur Reliabilitätsschätzung von Einzelitems siehe Wanous und Hudy (2001) und Wanous, Reichers und Hudy (1997). Einzelindikatoren werden auch als **Screening-Fragen** („screener") verwendet, um bestimmte Personengruppen zu identifizieren, die dann für eine vertiefte Untersuchung angeworben werden.

Vier praktische Hinweise zur Messung soziodemografischer Merkmale mittels Einzelindikatoren

1. Konventionalität

Wir empfehlen, auf **etablierte Fragebogenitems** für die gängigen soziodemografischen Variablen zurückzugreifen und sich keine Items selbst auszudenken. Als Quellen können hierfür all jene wissenschaftliche Publikationen herangezogen werden, die in ihrem Anhang das vollständige Fragebogeninstrument inklusive sozialstatistischem Teil enthalten. Auch wissenschaftliche Archive, die Forschungsdaten und zugehörige Erhebungsinstrumente anbieten, können hier hilfreich sein (z. B. http://www.psychdata.zpid.de; http://www.gesis.org: Datenbestandskatalog; http://www.icpsr.umich.edu; http://www.cessda.org). Insbesondere Items, die Bildung oder Tätigkeit erfassen, sind recht komplex und sollten möglichst nicht „handgestrickt" werden, insbesondere weil sie dann meist nicht die an nominale Variablen gestellten Kriterien der Genauigkeit, Exklusivität und Exhaustivität erfüllen (► Abschn. 8.4.2). Bei Längsschnitt- und Panelstudien behält man einmal gewählte Operationalisierungen nach Möglichkeit bei, damit die Ergebnisse der einzelnen Erhebungswellen direkt vergleichbar sind. Es sei denn, die soziale Wirklichkeit verändert sich zwischenzeitlich (z. B. Abschaffung des Grundwehr- und somit auch des Zivildienstes in Deutschland im Jahr 2011, so dass diese Antwortalternative bei der Abfrage der aktuellen Tätigkeit zukünftig entfällt) ◘ Tab. 8.19.

2. Sparsamkeit

Generell sind Fragen zu soziodemografischen Merkmalen **sparsam einzusetzen**, nicht nur um die Untersuchungsteilnehmenden von überlangen Fragebögen zu entlasten, sondern auch weil es sich um personenbezogene und zum Teil sensible Daten handelt. In jedem Fall ist aus forschungsethischen Gründen sicherzustellen, dass durch die erhobenen Variablen – insbesondere bei einer kleinen Stichprobe – nicht ungewollt die **Anonymität** aufgehoben wird. Fragen nach Nationalität, Behinderung oder Geschlecht können sich aus Gründen der ethisch notwendigen strikten Anonymisierung verbieten, wenn sich in der untersuchten Gruppe nur sehr wenige Personen befinden, auf die eine bestimmte Merkmalsausprägung zutrifft. Wie heikel oder indiskret die Frage nach einem Personenmerkmal empfunden wird, hängt stark von der Zielgruppe ab. Während Schülerinnen und Schüler alters- oder einkommensbezogene Fragen kaum kritisch bewerten, sieht das bei Berufstätigen oft anders aus. Sofern auf entsprechende Fragen nicht ganz verzichtet werden kann, sind vorgegebene grobe Alters- oder Einkommenskategorien diskreter als offene Fragen, die präzise Angaben verlangen und zu Antwortverweigerung oder gar Untersuchungsabbruch führen können.

Ein weiteres Argument für Sparsamkeit ist der tatsächliche **Informationsnutzen der einzelnen Variablen** und ihrer mehr oder minder differenziert erfassten Ausprägungen. Häufig werden soziodemografische Angaben **übergenau** erfasst, die letztlich nie einer sinnvollen Auswertung zugeführt werden. So sieht die amtliche Statistik zur Erfassung des Bildungsstandes zunächst 6 Ausprägungen für den „allgemeinbildenden Schulabschluss" plus 12 Ausprägungen für den „höchsten Ausbildungs- oder (Fach)Hochschulabschluss" vor. Kann in der eigenen Studie wirklich eine sinnvolle Auswertung differenziert nach so vielen verschiedenen Bildungsgängen erfolgen oder reicht nicht vielleicht eine Grobunterteilung nach Abitur versus kein Abitur? Nicht zuletzt sind Fragen oder Variablenausprägungen obsolet, die in der untersuchten Gruppe keine oder so gut wie **keine Varianz** erzeugen. So ist es meist wenig ergiebig, Erstsemesterstudierende zu fragen, ob sie geschieden oder verwitwet sind oder welches ihr höchster allgemeinbildender Schulabschluss ist.

3. Diskriminierungsfreiheit

Die Abfrage soziodemografischer Merkmale ist durch ihre historische Tradition leider nicht immer diskriminierungsfrei. Nicht selten findet sich sowohl inhaltlich als auch formal in der Darstellung z. B. die Sichtweise, der Mann sei das „Familienoberhaupt" oder der Normalfall, der notorisch immer an erster Stelle genannt wird. Diskriminierung kann auch vorliegen, wenn Fragen so gestellt werden, dass bestimmte Merkmalsausprägungen bei allen Personen implizit vorausgesetzt werden, z. B. deutsche Nationalität oder Heterosexualität. So ist die Abfrage des Familienstandes mit den klassischen Ausprägungen „ledig/verheiratet/geschieden/verwitwet" auf die heterosexuelle Ehe gemünzt und muss eigentlich um die äquivalenten amtlichen Begriffe für gleichgeschlechtliche eingetragene Lebenspartnerschaften ergänzt werden: „ledig/verheiratet bzw. verpartnert/geschieden bzw. entpartnert/verwitwet bzw. partnerhinterblieben".

4. Theoretische Reflexion

Der **Familienstand** als juristisch relevantes Personenmerkmal ist in vielerlei Hinsicht im Alltag bedeutsam, jedoch unter theoretischen Gesichtspunkten oft kein guter Einzelindikator zur Kennzeichnung der in der sozialwissenschaftlichen Forschung tatsächlich interessierenden **Lebenssituation**. Da Ehe bzw. Verpartnerung meist relativ spät im Leben stattfinden, fallen fast alle Jüngeren undifferenziert in die Kategorie „ledig", obwohl einige in jahrelangen festen Paarbeziehungen leben. Ebenso können Ältere „verwitwet", aber gleichzeitig neue Paarbeziehungen eingegangen sein. Ergänzend zu oder anstelle des rechtlichen Familienstandes wird man deswegen in den meisten Fällen die Lebenssituation lieber direkt abfragen.

Welchen Familienstand haben Sie?

☐ Ehe/eingetragene Lebenspartnerschaft
☐ feste Paarbeziehung
☐ Single/alleinstehend
☐ anderes, und zwar: _____

Sind detaillierte Angaben zur Partner- und Familiensituation wichtig, so kann noch nach der Dauer der Paarbeziehung, nach gemeinsamer oder getrennter Haushaltsführung, nach Kindern im Haushalt etc. gefragt werden.

Dass sich eine theoretische Reflexion auch vermeintlich einfacher und konventionalisierter Einzelindikatoren lohnt, sei an einem letzten Beispiel illustriert. Wir alle kennen diese Form der Abfrage:

Geschlecht: ☐ männlich ☐ weiblich

Die Vorstellung, alle Menschen ließen sich trennscharf in „Männer“ und „Frauen“ einteilen, ist zwar im Alltag sehr verbreitet, wird jedoch in der Geschlechterforschung aus verschiedenen Gründen infrage gestellt und entspricht in Deutschland auch nicht mehr dem Personenstandsrecht. Die zunehmende gesellschaftliche Sichtbarkeit von Personen, die sich als intersexuell/intergeschlechtlich, transsexuell/transidentisch, androgyn, postgender, agender etc. definieren, kann es (insbesondere bei entsprechend sensibilisierten Zielgruppen) notwendig und wünschenswert erscheinen lassen, die herkömmliche Abfrage gemäß binärem Geschlechtermodell zu überwinden und mehr als zwei Ausprägungen zuzulassen. Zudem wäre im Sinne einer Konstruktspezifikation jeweils genauer zu definieren, welche Dimension von Geschlecht gemeint ist (biologisches Geschlecht, soziales Geschlecht etc.).
Die tradierte Geschlechtsabfrage (männlich/weiblich) genügt offenbar nicht den Kriterien der Genauigkeit (was ist mit „Geschlecht“ genau gemeint?), der Exklusivität (was ist mit Personen, die sich sowohl als weiblich als auch männlich definieren?) und der Exhaustivität (was ist mit Personen, die sich weder als weiblich noch als männlich verstehen?), die an nominalskalierte Variablen gestellt werden (► Abschn. 8.4.2). Die entsprechende Operationalisierung ist also theoretisch und methodisch unzureichend, wird aber in der empirischen Sozialforschung bislang so gut wie nie hinterfragt (zu alternativen Operationalisierungen von Geschlecht Döring 2013).
Das hier anhand der Fragebogenmethode erläuterte Operationalisierungsproblem der Variable „Geschlecht“ betrifft im Übrigen auch andere Datenerhebungsmethoden. Etwa wenn in einer Interview- oder Beobachtungsstudie das vermeintlich einfache manifeste Merkmal „Geschlecht“ anhand des Vornamens, der Stimme oder des äußeren Erscheinungsbildes dichotom durch die Forschenden zugewiesen wird.

Tabelle 8.19 Tätigkeit mit 9 Ausprägungen. (Aus Statistisches Bundesamt 2011)

Was trifft überwiegend auf Sie zu? Bitte kreuzen Sie nur *eine* Antwortmöglichkeit an.	
Ich bin erwerbs- bzw. berufstätig (inkl. Auszubildende, Personen in Elternzeit oder Altersteilzeit)	☐
Ich bin Grundwehr-/Zivildienstleistender	☐
Ich bin Schüler/-in	☐
Ich bin Student/-in	☐
Ich bin Rentner/-in, Pensionär/-in	☐
Ich lebe von Einkünften aus Kapitalvermögen, Vermietung oder Verpachtung	☐
Ich bin Hausfrau/-mann oder versorge Kinder und/oder pflegebedürftige Personen	☐
Ich bin arbeitslos	☐
Keine der genannten Auswahlmöglichkeiten (z. B. dauerhaft arbeitsunfähig)	☐

Tabelle 8.20 Beispiele für psychometrisch geprüfte Single-Item-Measures

Theoretisches Konstrukt	Psychometrisch geprüftes Einzelitem	Quelle
Depression	In the past 4 weeks have you often felt sad or depressed? “yes”/ “no”	B. McCormack, Boldy, Lewin, & G. R. McCormack (2011)
Depression	Do you think you suffer from depression? 1 = “not at all” bis 5 = “certainly yes”	Ayalon, Goldfracht, & Bech, (2010)
Emotionale Erschöpfung/ Burnout	I feel burned out from my work. “never” – “a few times a year or less” – “once a month or less” – “a few times a month” – “once a week” – “a few times a week” – “every day”	West, Dyrbye, Sloan, & Shanafelt (2009)
Zufriedenheit	Do you feel happy in general? 0-1-2-3-4-5-6-7-8-9-10	Abdel-Khalek (2006)
Selbstwertgefühl	I have high self-esteem. 1 = “not very true of me” bis 5 = “very true of me”	Robins, Hendin, & Trzesniewski (2001)

8.6 Messung mit psychometrischen Skalen und Skalenbildung

Unter einer **psychometrischen Skala**, Testskala oder kurz: Skala versteht man einen **Satz von Skalen-Items** (Selbstauskunftsfragen, Aussagen oder Aufgaben samt Antwortformat), die als **reflektive Indikatoren** dazu dienen, gemeinsam ein latentes Merkmal zu messen (zum reflektiven Messmodell ► Abschn. 8.3.1). Eine psychometrische Skala muss erwiesenermaßen messtheoretischen Skalierungseigenschaften und Gütekriterien (Objektivität, Reliabilität, Validität) genügen. Jede psychometrische Skala verfolgt das Ziel, über **multiple Indikatoren** (Skalen-Items) ein theoretisches Konzept präziser zu erfassen als dies mit einem Einzelindikator (Single-Item-Measure) möglich wäre.

Die **Hauptvorteile der psychometrischen Skala** gegenüber dem Einzelitem bestehen darin, dass a) ein komplexes Konstrukt in verschiedenen inhaltlichen Facetten und Nuancen erfassbar wird, was die Validität steigern sollte, b) Messfehler – z. B. ein Item wird missverstanden – sich im Kontext der anderen Items relativieren und somit die Reliabilität steigt, c) die psychometrischen Merkmale einer Skala (v. a. Dimensionalität und Messgenauigkeit bzw. Reliabilität) anhand der Daten der Hauptuntersuchung routinemäßig noch einmal statistisch überprüft werden können. **Hauptnachteile der Skala** gegenüber dem Einzelitem sind jedoch der hohe Zeitaufwand für die Befragten sowie deren zuweilen geringe Akzeptanz für langwierige und redundant wirkende Fragenbatterien einerseits und der Kostenaufwand für Entwicklung und Durchführung auf Seiten der Forschenden andererseits.

Skalen, die aus einer mehr oder weniger beliebigen Sammlung von Items bestehen und die als Skalenwert einer Untersuchungsperson schlicht die Summe der Punktwerte pro Item nutzen, ohne dass Skalierungseigenschaften und Gütekriterien geprüft wurden, sind als sog. **Ad-hoc-Skalen** für wissenschaftliche Studien nicht geeignet. Auch wenn eine eigenständig entwickelte Skala nur in einem begrenzten Rahmen Anwendung findet, sollten die folgenden Anforderungen bei der Skalenbildung beachtet werden (für Gütekriterien im Überblick ► Abschn. 10.4.1), die sich auf die psychometrische Testtheorie (► Abschn. 10.4.4) beziehen:

1. **Eigenschaften der Skalen-Items**: Die Items müssen alle das zu messende Merkmal inhaltlich angemessen und umfassend abbilden, zudem müssen sie klar und verständlich formuliert sein. Die Items sollten möglichst viele Ausprägungsgrade des zu messenden Merkmals – also geringe, mittlere und hohe Merkmalsausprägungen – repräsentieren (hohe Streuung der Schwierigkeitsindizes). Jedes Item sollte möglichst eindeutig Personen mit starker Merkmalsausprägung von Personen mit schwächerer Merkmalsausprägung trennen (hohe Trennschärfe – im Sinne einer exakten Unterscheidbarkeit – der Items). Anzahl und Reihenfolge der Items sollten so festgelegt werden, dass die Skala für die Untersuchungspersonen gut zu bearbeiten ist und gleichzeitig ihre psychometrischen Gütekriterien (v. a. Reliabilität und Validität) optimiert werden.
2. **Dimensionalität der Skala**: Es muss empirisch geprüft werden, ob die Skala eindimensional (berechnet wird ein Gesamtskalenwert) oder mehrdimensional ist. Bei Mehrdimensionalität muss bekannt sein, welche Items welche Subdimension des Merkmals erfassen und somit in die Berechnung der Subskalenwerte eingehen.
3. **Skalierbarkeit bzw. Skalenniveau der Skala**: Es sollte für die Berechnungsvorschrift des Gesamtskalenwertes bzw. der Subskalenwerte (z. B. ungewichtete oder gewichtete Summen- oder Durchschnittsbildung) geprüft werden, ob tatsächlich Messwerte auf dem in der Regel unterstellten Intervallskalenniveau resultieren (diese Überprüfung ist mithilfe der probabilistischen Testtheorie bzw. Item-Response-Theorie möglich; ► Abschn. 10.4.4).
4. **Objektivität der Skala**: Anwendungs- und Auswertungsvorschriften für die Skala müssen eindeutig bzw. objektiv festgelegt sein und dürfen nicht von der jeweiligen Person abhängen, welche die Skala zur Datenerhebung bei unterschiedlichen Untersuchungspersonen anwendet.
5. **Reliabilität der Skala**: Die Skalenwerte bzw. Subskalenwerte dürfen nur gering von Messfehlern belastet sein bzw. die Skalenwerte müssen hohe Messgenauigkeit aufweisen, was anhand mindestens eines geeigneten Reliabilitätskoeffizienten zu prüfen ist.
6. **Validität der Skala**: Die einzelnen Items sowie die Skalenwerte bzw. Subskalenwerte müssen inhaltlich tatsächlich das zu messende Merkmal umfassend repräsentieren und dürfen nicht versehentlich ein anderes (z. B. verwandtes) Konstrukt messen, was mit unterschiedlichen Techniken der Validitätsprüfung nachzuweisen ist.

Ein Itemsatz, der erwiesenermaßen diesen Bedingungen genügt, soll als „psychometrische Skala" bezeichnet werden. Für die Konstruktion einer psychometrischen Skala ist die Art des zu messenden Merkmals letztlich unerheblich. Es wird davon ausgegangen, dass z. B. für die Konstruktion einer Skala zur Messung eines Persönlichkeitsmerkmals (Aggressivität, Gedächtnisleistung, Belastbarkeit, räumliches Vorstellungsvermögen, emotionale Labilität etc.) die gleichen Regeln gelten wie für die Konstruktion von Einstellungsskalen (Einstellungen zu bestimmten Einstellungsobjekten wie Kirche, Demokratie, Kernkraft, Vegetarismus etc.).

Psychometrische Skala – Eine psychometrische Skala („psychometric scale") dient der Operationalisierung eines theoretischen Konstrukts im Rahmen eines **reflektiven Messmodells**. Sie besteht aus einem Satz von formal und inhaltlich ähnlichen bzw. reflektiven Indikatorvariablen in Form von Skalen-Items. **Skalen-Items** („scale items") sind Fragen, Aussagen oder Aufgaben plus standardisiertem Antwortformat (häufig werden als Antwortformat Ratingskalen verwendet). Anhand der Antworten auf die einzelnen Items wird der Skalenwert („score") gebildet (z. B. als ungewichteter oder gewichteter Summen- oder Durchschnittswert). Eine psychometrische Skala kann eindimensional (aus allen Items wird ein Gesamtskalenwert gebildet) oder mehrdimensional sein (aus jeweils einem Teil der Items werden Subskalenwerte gebildet, welche die Ausprägung von Unteraspekten des gemessenen Merkmals repräsentieren). Die Dimensionalität einer Skala muss geprüft und bekannt sein.

Ebenso müssen die psychometrischen Gütekriterien (v. a. Reliabilität und Validität) der Skala empirisch überprüft sein und Mindestanforderungen erfüllen. Die von psychometrischen Skalen erzeugten Messwerte werden üblicherweise als intervallskaliert aufgefasst. Mit Techniken der probabilistischen Testtheorie bzw. Item-Response-Theorie kann das Skalenniveau auch empirisch geprüft werden.

Wir möchten noch einmal darauf hinweisen, dass bevorzugt bereits **etablierte psychometrische Skalen** aus der Literatur für eine eigene Studie übernommen oder adaptiert werden sollten und die Eigenkonstruktion von Skalen nur dann sinnvoll ist, wenn sichergestellt wurde, dass noch keine überprüften Instrumente in der deutsch- und englischsprachigen Fachliteratur vorliegen. Dazu ist eine entsprechende Recherche in **Literaturdatenbanken** (► Abschn. 6.1.2) sowie **Testbibliotheken** (► Abschn. 10.4.1) vorzunehmen oder ein **Skalenhandbuch** zu konsultieren (z. B. Skalenhandbuch für die Kommunikationswissenschaft: Rössler 2011, für die Marktforschung: Bearden, Netemeyer, & Haws, 2011, für die klinisch-psychologische und psychiatrische Forschung und Praxis: Baer, & Blais, 2009).

Im Folgenden werden drei bekannte Skalierungsverfahren in der Reihenfolge ihrer praktischen Bedeutung dargestellt: 1. Die **Likert-Skala** („likert scale"; „likert-type scale", 1 815 bzw. 734 Treffer in der führenden Literaturdatenbank PsycINFO der American Psychological Association, Stand: Mai 2011), 2. die **Guttman-Skala** („guttman scale", 201 Treffer), 3. die **Thurstone-Skala** („thurstone scale", 61 Treffer). Die ebenfalls verbreitete **Rasch-Skala** („rasch scale", 23 Treffer) wird aufgrund ihrer Konstruktion auf Basis der probabilistischen Testtheorie bzw. Item-Response-Theorie (IRT) in ► Abschn. 10.4.4 behandelt. Für weitere sehr seltene Skalierungsformen wie z. B. die Coombs-Skala (1 Treffer) oder die Edwards-Kilpatrick-Skala (0 Treffer) sei auf das Webangebot zu diesem Buch sowie auf die weiterführende Literatur verwiesen (Edwards & Kilpatrick, 1948; Pepels 2007).

Ergänzend zu den drei Skalierungsmethoden nach Likert, Guttman und Thurstone wird zudem die Skalierung nach dem **Semantischen Differenzial** („semantic differential scale", 504 Treffer) dargestellt, mit dessen Hilfe weniger die rationalen Beurteilungen als vielmehr die emotionalen Reaktionen hinsichtlich eines Urteilsobjekts erhoben werden.

8.6.1 Likert-Skala

Die Likert-Skala ist die mit großem Abstand am häufigsten in den Sozialwissenschaften verwendete psychometrische Skala, die mit mehreren Indikatoren ein bestimmtes Konstrukt auf Intervallskalenniveau erfassen soll. Die Indikatoren sind dabei Items in der Form von Fragen („Wie oft haben Sie sich im letzten halben Jahr depressiv gefühlt?") oder Aussagen („Ich fühle mich oft niedergeschlagen und traurig."), die auf einer mindestens 5-stufigen Ratingskala zu beantworten sind. Entwickelt wurde dieses auch als „**Methode der summierten Ratings**" bezeichnete Skalierungsverfahren von dem US-amerikanischen Sozialforscher Rensis Likert (1932), dessen Name als /'lɪkert/ (nicht /'laɪkert/) auszusprechen ist.

Likert-Skala – Eine Likert-Skala („likert scale") ist eine psychometrische Skala, die aus mehreren (z. B. 10–20) Aussagen besteht, die alle dasselbe Merkmal messen und auf Ratingskalen hinsichtlich des Grades der Zustimmung einzuschätzen sind. Typischerweise werden 5-stufige Ratingskalen verwendet. Die Likert-Items repräsentieren unterschiedliche Intensitäten des gemessenen Merkmals. Der Skalenwert der Likert-Skala berechnet sich als Summen- oder Durchschnittsscore der einzelnen Ratings. Neben eindimensionalen Likert-Skalen existieren auch mehrdimensionale Likert-Skalen; dabei bilden dann Teilmengen der Items jeweils die Subskalen, die Unteraspekte des interessierenden Merkmals erfassen. Für jede der (z. B. 2–4) Subskalen wird ein eigener Skalenwert berechnet. Der Likert-Skala wird üblicherweise Intervallskalenniveau, manchmal auch nur Ordinalskalenniveau zugeschrieben.

Die Likert-Skalierung ist weniger aufwändig, aber nahezu ebenso zuverlässig wie das Vorgehen mittels Thurstones „Law of Categorial Judgement", bei dem eine originäre Rangreihe in Intervalldaten transformiert wird (Roskam 1996, S. 443; ► Abschn. 8.4.6).

Die Skalenkonstruktion erfolgt bei der Likert-Skala in folgenden vier Schritten:

1. **Item-Konstruktion**: Zu dem zu messenden Konstrukt wird ein Pool von rund 100 Items gesammelt, etwa auf der Basis von Theorien, Praxiserfahrungen und empirischen Befunden zum Konstrukt. Die Items sollten Augenschein-Validität aufweisen sowie einfach und eindeutig formuliert sein. Typisch ist die Kombination aus Statement und 5-stufiger Bewertungsskala von starker Ablehnung bis starker Zustimmung (z. B. stimmt gar nicht [1], stimmt wenig [2], stimmt teils, teils [3], stimmt ziemlich [4], stimmt völlig [5]). Um Antworttendenzen zu vermeiden, sollten sowohl **positive Items** (Statements, die für eine starke Ausprägung des Konstrukts sprechen) als auch **negative Items** (Statements, die gegen eine starke Ausprägung des Konstrukts sprechen) konstruiert werden. Im Rahmen der Konstruktion einer Depressionsskala wäre z. B. ein Statement wie „Ich bin oft traurig und niedergeschlagen" ein positives Item, „Ich bin meist fröhlich und gut gelaunt" dagegen ein negativ gepoltes Item. Für die Auswertung müssen die Messwerte der negativen Items umgepolt werden. Das heißt, wer ein negatives Item stark ablehnt, bekommt dafür nicht den Messwert 1, sondern den Messwert 5 (zur Umpolung von Items im Rahmen der Datenaufbereitung

► Abschn. 11.4.4). Der von den Forschenden entwickelte Itempool kann durch Beratung unter Kolleginnen und Kollegen in mehreren Korrekturdurchgängen überarbeitet und verfeinert werden.

2. **Skalen-Pretest:** Der korrigierte Item-Pool wird einer für die interessierende Population möglichst repräsentativen Skalenkonstruktionsstichprobe vorgelegt, die ihn im Sinne ehrlicher Selbstauskünfte beantworten soll. Anhand der in diesem Methoden-Pretest gewonnenen Daten wird pro Item die Itemschwierigkeit (Grad der Zustimmung zu einem Item) sowie die Trennschärfe (Korrelation eines einzelnen Items mit dem Gesamtmittelwert aller Items) berechnet. Zudem werden die Korrelationen der Items untereinander betrachtet, die positiv sein sollten, da alle Items dasselbe Konstrukt messen. (Für weitere Hinweise zu Fragebogen-Pretests ► Abschn. 10.3.3). Die Likert-Skalierung basiert auf der klassischen Testtheorie (KTT; ► Abschn. 10.4.4).
3. **Item-Selektion:** In die finale Skala gehen rund 10–20 Items ein, die alle über hohe Trennschärfen verfügen und deren Itemschwierigkeiten streuen sollen, d. h., die Skala soll sowohl leichte Items, die von vielen Untersuchungspersonen bejaht als auch schwere Items, die von vielen abgelehnt werden sowie Items mittlerer Schwierigkeit enthalten. Die Eindimensionalität der Skala sollte sich in einer 1-Faktor-Struktur gemäß exploratorischer oder konfirmatorischer Faktorenanalyse widerspiegeln. Der Gesamtwert der Skala (**Gesamtscore**), dessen Höhe die Ausprägung des gemessenen Konstrukts auf Intervallskalenniveau widerspiegeln soll, ergibt sich als **Summe oder Mittelwert der (korrekt gepolten) Ratings** über alle Items hinweg. Eine Depressionsskala mit 20 Items und jeweils 5-stufiger Ratingskala hat also einen Wertebereich von 20 bis 100 (Summenscore) bzw. von 1 bis 5 (Durchschnittsscore), wobei 20 bzw. 1 die Abwesenheit von Depression und 100 bzw. 5 eine stark ausgeprägte Depression anzeigt. Durchschnittsscores haben den Vorteil, dass fehlende Werte kompensiert werden, wenn man den Summenscore jeder Person durch die Zahl der von ihr beantworteten Items dividiert. Bei einer mehrdimensionalen Skala sind jeweils eindimensionale Subskalen aus Teilmengen der Items zu bilden. Die Dimensionalität ist über die Faktorstruktur zu prüfen (zu psychometrischen Gütekriterien im Überblick ► Abschn. 10.4.1).
4. **Finale psychometrische Likert-Skala**: Die fertige psychometrische Likert-Skala besteht aus allen selektierten Items, d. h., allen Statements plus Ratingskalen als Antwortformat in einer festen Reihenfolge sowie einer zusätzlichen Information darüber, welche Items negativ gerichtet und somit vor der Auswertung umzupolen sind. Auch eine standardisierte Instruktion (Ausfüllanweisung) gehört zur Skala. Ebenso sind die psychometrischen Kennwerte (Itemschwierigkeiten, Trennschärfen, Faktorladungen, interne Konsistenz) zu berichten, so dass die Eigenschaften der einzelnen Items, der Gesamtskala sowie – bei mehrdimensionalen Skalen – auch der Subskalen bekannt sind. Dieses Messinstrument kann dann kritisch bewertet sowie von anderen Forschenden übernommen werden (► Beispiel einer zweidimensionalen Likert-Skala zur Messung von Computer-Erfahrung).

Es ist zu beachten, dass nur ein Itemsatz, der den hier beschriebenen Skalierungsprozess durchlaufen hat, eine psychometrische Likert-Skala darstellt. Eine willkürlich zusammengestellte und psychometrisch ungeprüfte Menge von Statements wird in der Literatur zuweilen fälschlich auch als Likert-Skala bezeichnet, obwohl es sich um eine **Ad-hoc-Skala** handelt. Zuweilen wird fälschlich auch die Ratingskala als Antwortformat als Likert-Skala etikettiert und z. B. von einem dichotomen Antwortformat (ja/nein) abgegrenzt.

In der Forschungspraxis wird der **Skalenbildungsprozess oft abgekürzt**, d. h., die Bildung eines umfangreichen Itempools wird übersprungen und gleich mit dem Pretest einer kleinen Zahl von Items begonnen, in der Erwartung, dass diese ausreichende psychometrische Eigenschaften zeigen mögen, was offenbar nicht selten auch gelingt.

Befürworter der Messung mittels Einzelindikatoren weisen darauf hin, dass durch eine Messung desselben Konstrukts statt mit einem einzelnen Indikator mit 10–20 nahezu gleichlautenden Items die Untersuchungspersonen übermäßig strapaziert und demotiviert werden Einen Kompromiss zwischen einer 20 Items umfassenden Likert-Skala und einer Messung mit Einzelitem stellen psychometrische **Kurzskalen** dar, die mit 3–5 Items arbeiten. Hierfür werden einer längeren Likert-Skala die besten (trennschärfsten) Items entnommen.

Die eigentlich empfohlene Verwendung positiv und negativ gepolter Items erweist sich nicht selten insofern als nachteilig, als sich dadurch ungewollt eine **zweidimensionale Struktur der Skala** bilden kann, d. h., im Antwortmuster sind die positiv gepolten Items einander untereinander sehr ähnlich, und für die negativ gepolten Items gilt untereinander dasselbe, aber positive und negative Items korrelieren weniger miteinander. Deswegen finden sich in der Literatur auch viele Likert-Skalen, die nur mit positiv gepolten Items arbeiten.

Nicht selten stellt sich bei der Operationalisierung eines komplexen theoretischen Konstrukts mithilfe einer Likert-Skala heraus, dass die einzelnen Items nicht nur Nuancen desselben Konstrukts erfassen (eindimensiona-

Beispiel einer zweidimensionalen Likert-Skala zur Messung von Computer-Erfahrung

Computer haben heute in Ausbildung, Beruf und Privatleben einen festen Platz eingenommen. Dementsprechend ist **Computer-Erfahrung** ein latentes Personenmerkmal, das in immer mehr Bereichen bedeutsam ist: Unterschiede im Computer-Knowhow von Menschen zu messen, zu erklären und zu verändern kann z. B. in der Erziehungswissenschaft, der Psychologie oder auch der Wirtschaftswissenschaft von Bedeutung sein. Eine verbreitete und gut handhabbare Möglichkeit, um ein Konstrukt wie „Computer-Erfahrung" zu operationalisieren, ist die Erhebung von Selbstauskünften mittels einer Likert-Skala, wie z. B. der **„Computer Understanding and Experience (CUE) Scale"** (Potosky & Bobko, 1998).

Das Konstrukt „Computer-Erfahrung" wurde im Zuge der Entwicklung der CUE-Skala auf der Basis des Forschungsstandes definiert (**Nominaldefinition**, ▶ Abschn. 8.2.1) als Wissen und Erfahrung im Bereich Computernutzung. Bei den einzelnen Items (Indikatoren) der Skala handelt es sich teils um Verhaltensbeschreibungen („Ich lese oft Computerzeitschriften oder andere Informationsquellen über Computertechnologie."), teils um Bewertungen („Ich kann gut mit Computern umgehen."). Jede **Selbstauskunftsaussage** ist auf einer **5-stufigen Ratingskala** von starker Ablehnung bis starker Zustimmung zu beantworten, deren Stufen die Messwerte 1 („strongly disagree") bis 5 („strongly agree") zugewiesen werden.

Der idealtypische Entwicklungsprozess der Likert-Skala, der mit einem Pool von 100 Items beginnt, wurde pragmatisch verkürzt. Ausgangspunkt war ein Itempool von rund 15 Items, die im Zuge von zwei Pretests ($n = 20$ und $n = 5$) erprobt und umformuliert wurden. Es resultierte eine finale Likert-Skala **mit 12 Items**. Alle Items sind **positiv gepolt**, d. h., sprechen für vorhandene Computererfahrung und müssen somit vor einer Auswertung **nicht** mehr umgepolt werden. In ◘ Tab. 8.21 sind die **12 Items** der Likert-Skala zur Messung von Computer-Erfahrung im Wortlaut genannt.

Der **Summenscore dieser Likert-Skala** (bzw. der 12 Skalenitems) liegt zwischen 12 Punkten (geringste Computererfahrung) und 60 Punkten (höchste Computererfahrung). Die Skalenkonstruktionsstichprobe ($n = 272$ Studierende in den USA) erzielte im Durchschnitt einen Summenscore von $M = 34.79$ ($SD = 11.86$), was gemäß der Skala mittlerer Computer-Erfahrung entspricht.

Die Berechnung eines Gesamtscores ist nur gerechtfertigt, wenn die Skala tatsächlich eindimensional ist. Die Dimensionalität wird mittels Faktorenanalysen geprüft. Auch die interne Konsistenz ist ein Hinweis auf Eindimensionalität, vor allem aber ist sie ein Indikator für hohe Reliabilität. Die **interne Konsistenz** wird über den Koeffizienten **Cronbachs α** bestimmt, der inhaltlich die durchschnittliche Korrelation zwischen allen Einzelitems darstellt. Bei einer eindimensionalen und reliablen bzw. intern konsistenten Skala erwarten wir hohe positive Korrelationen der Items untereinander (schließlich sind alle Items Indikatoren desselben Konstrukts). Dies ist gleichbedeutend mit der Erwartung eines hohen Konsistenzkoeffizienten von $\alpha > .80$. Tatsächlich wurde für die CUE-Skala anhand der Konstruktionsstichprobe ein sehr zufriedenstellender Konsistenzkoeffizient von **Cronbachs $\alpha = .93$** ermittelt. Damit gilt nicht nur die **Eindimensionalität** der Skala, sondern auch ihre **Reliabilität** als gesichert. Denn hohe Item-Interkorrelationen (bzw. hohe interne Konsistenz) sind nur möglich, wenn die einzelnen Items wenig von unsystematischen Messfehlern verzerrt sind.

Gemäß **Klassischer Testtheorie (KTT;** ▶ Abschn. 10.4.4) wird für jedes Item die **Itemschwierigkeit** berichtet. Die Itemschwierigkeit entspricht dem **Mittelwert des jeweiligen Items** in der untersuchten Skalenkonstruktionsstichprobe und hat bei einer 5-stufigen Ratingskala den **Wertebereich 1–5** pro Item. Je höher der Itemmittelwert (d. h., umso stärker dem Item zugestimmt wird), umso „leichter" ist es auch im testtheoretischen Sinne. Im vorliegenden Datensatz war Item 8 („Ich weiß, was eine Datenbank ist.") mit einem Mittelwert von $M = 3.71$ das leichteste Item, Item 1 („Ich lese oft Computerzeitschriften oder andere Informationsquellen über Computertechnologie.") mit $M = 1.96$ dagegen das schwierigste bzw. unpopulärste Item. Generell ist es bei einer Likert-Skala erwünscht, Items unterschiedlicher Schwierigkeiten zu kombinieren, was bei der vorliegenden CUE-Skala realisiert ist.

Neben der Itemschwierigkeit wird als weiterer Itemkennwert die **Trennschärfe** jedes Items angegeben. Die Trennschärfe ist die **bivariate Korrelation des Items mit dem Gesamtscore** der Skala und hat einen **Wertebereich von −1 bis +1**. Erwartet werden generell hohe positive Korrelationen. Item 2 hatte mit $r_{it} = .75$ („Ich weiß, wie ich gelöschte oder verlorene Daten wiederherstellen kann.") die höchste Trennschärfe und Item 10 („Ich nutze den Computer regelmäßig für Textverarbeitung.") hatte mit $r_{it} = .58$ die geringste Trennschärfe und differenziert somit am schlechtesten zwischen erfahrenen und unerfahrenen Computernutzern. Will man eine Skala verkürzen und gleichzeitig ihre interne Konsistenz steigern, so können diejenigen Items mit den geringsten Trennschärfen eliminiert werden, sofern dies die Skala nicht zu sehr verkürzt und somit wieder ihre Messgenauigkeit reduziert. Im Rahmen einer Itemanalyse (z. B. mit dem Statistikprogrammpaket SPSS) kann man sich angeben lassen, wie sich Cronbachs α verändert, wenn wenig trennscharfe Items entfernt werden. Da im Falle der CUE mit 12 Items eine sehr hohe interne Konsistenz erreicht wurde, ist eine Eliminierung von Items nicht notwendig.

Schließlich kann bei Skalen, die eine größere Zahl an Items enthalten (etwa ab 10 Items aufwärts) noch geprüft werden, ob sich die eindimensionale Gesamtskala sinnvoll in **Subskalen** (Unteraspekte des theoretischen Konstrukts) differenzieren lässt. Hierfür kann im einfachsten Fall eine **explorative Faktorenanalyse** eingesetzt werden: Die Durchführung einer Hauptkomponentenanalyse (PCA) mit anschließender obliquer Rotation (Bortz & Schuster, 2010, Kap. 23) legte für die CUE-Skala eine zweifaktorielle Struktur nahe. Demnach bilden die Items 1–6 den ersten Faktor („technische Kompetenz", der fortgeschrittene Kenntnisse wie z. B. Programmierung anspricht) und die Items 7–12 den zweiten Faktor („generelle Kompetenz", der Grundkenntnisse wie die Bedeutung von „E-Mail" oder „Datenbank" adressiert). Wie gut welches Item zu welchem Faktor passt, wird über die **Faktorladungen** ausgedrückt, die die Korrelation des jeweiligen Items mit dem Faktor repräsentieren (Wertebereich −1 bis +1). Item 1 weist die höchste Faktorladung auf Faktor 1 auf, Item 7 die höchste Faktorladung auf Faktor 2. Beide Items sind somit „Markieritems" für den Faktor und helfen besonders bei der inhaltlichen Interpretation der Faktoren. Das Herausarbeiten der Zwei-Faktor-Struktur erlaubt es, ergänzend zum Gesamtscore der Skala auch die **Scores für die beiden Subskalen** zu berechnen und für differenzierte Analysen zu nutzen.

Eine noch genauere Überprüfung der Faktorstruktur der Skala als mit der exploratischen Faktorenanalyse kann mit der **konfirmatorischen Faktorenanalyse** bzw. mit **Strukturgleichungsmodellen** (▶ Kap. 17) vorgenommen werden. Entsprechende Analysen bestätigten die zweifaktorielle bzw. zweidimensionale Struktur der CUE-Skala (Schulenberg & Melton, 2007).

Tabelle 8.21 Items inklusive Itemstatistiken für eine zweidimensionale Likert-Skala zur Computer-Erfahrung („Computer Understanding and Experience [CUE] Scale"; Hauptkomponentenanalyse mit obliquer Rotation, Interfaktor-Korrelation = .60. Fehlende Faktorladungen wurden in der Originalpublikation nicht angegeben). (Mod. nach Potosky & Bobko, 1998. Copyright 1998, with permission from Elsevier)

Item-Nr.	Item	Itemschwierigkeit	Trennschärfe	Faktorladung auf Faktor 1 „Technical Competence"	Faktorladung auf Faktor 2 „General Competence"
1	I frequently read computer magazines or other sources of information that describe new computer technology.	1.96	.61	.94	–
2	I know how to recover deleted or "lost data" on a computer or PC.	2.53	.75	.83	–
3	I know what a LAN is.	2.34	.69	.73	–
4	I know what an operating system is.	2.93	.82	.66	–
5	I know how to write computer programs.	2.42	.64	.75	–
6	I know how to install software on a personal computer.	2.72	.70	.71	–
7	I know what e-mail is.	3.56	.59	–	.90
8	I know what a database is.	3.71	.68	–	.79
9	I am computer literate.	3.55	.74	–	.70
10	I regularly use a PC for word processing.	3.02	.58	–	.73
11	I often use a mainframe computer system.	2.85	.65	–	.67
12	I am good at using computers.	3.21	.74	–	.60

le Skala), sondern sich in zwei oder mehr inhaltliche Subdimensionen aufteilen lassen (mehrdimensionale Skala). Die **Dimensionalität einer Likert-Skala** wird typischerweise im Rahmen der Skalenkonstruktion mittels exploratorischer oder konfirmatorischer Faktorenanalyse bzw. Datenmodellierung mittels Strukturgleichungsmodellen (▶ Kap. 17) geprüft. Liegt eine stabile mehrfaktorielle Struktur vor, so wird die Skala nicht nur über einen Gesamtscore ausgewertet, sondern es können auch die Summen- oder Durchschnittsscores für die Subskalen berechnet werden, indem jeweils die Ratings der zu einer Subskala gehörenden Items zusammengefasst werden. Ein Beispiel für eine mehrdimensionale Likert-Skala wird in ▶ Beispiel einer zweidimensionalen Likert-Skala zur Messung von Computer-Erfahrung präsentiert.

8.6.2 Guttman-Skala

Die von dem amerikanisch-israelischen Sozialforscher und Methodiker **Louis (Eliyahu) Guttman** in den 1940er-Jahren entwickelte Skalierungsmethode (auch „Skalogrammanalyse" genannt; Guttman 1950) läuft auf einen Satz von Items hinaus, die hinsichtlich ihrer Itemschwierigkeiten strikt gestaffelt sind. Jedes weitere Item, das von den Untersuchungspersonen jeweils bejaht oder verneint werden kann, drückt eine stärkere Merkmalsausprägung aus. Je mehr Items eine Person also bejaht, umso stärker ist das Merkmal bei ihr ausgeprägt. Der **Gesamtscore** einer Guttman-Skala entspricht der **Summe der bejahten Items**.

Die Entwicklung einer Guttman-Skala ist deutlich aufwändiger als die einer Likert-Skala, weil Items zu konstruieren sind, die empirisch nachweisbar eine **schrittweise stärkere Merkmalsausprägung** widerspiegeln. Anders ausgedrückt: Es wird gefordert, dass eine Person mit höherer Merkmalsausprägung mindestens diejenigen Items bejaht (löst), die eine Person mit geringerer Merkmalsausprägung bejaht (löst).

Guttman-Skala – Eine Guttman-Skala („guttman scale") ist eine psychometrische Skala, die aus mehreren Aussagen besteht, die alle dasselbe Merkmal messen und jeweils mit Zustimmung oder Ablehnung (Ja-/Nein-Format) einzuschätzen sind. Die Guttman-Items einer Skala werden so konstruiert, dass sie graduell jeweils eine immer stärkere In-

tensität des gemessenen Merkmals repräsentieren. Eine Person, die dem vierten Item auf der Skala zustimmt, muss also alle vorherigen, leichteren Items ebenfalls bejaht haben, wenn das Skalierungsmodell zutrifft. Der Skalenwert der Guttman-Skala berechnet sich als Summe aller zugestimmten Items.

Folgender Satz von sieben Items zur Messung von permissiven Einstellungen zur vorehelichen Sexualität erfüllt die Merkmale einer Guttman-Skala: mit jedem Item steigt der Grad der sexuellen Freizügigkeit (**Premarital Sexual Permissiveness Scale**: Reiss 1964). Die Skalengesamtwerte einer Person variieren zwischen 0 (keinem Item wird zugestimmt, geringste sexuelle Freizügigkeit) und 7 (allen sieben Items wird zugestimmt, maximale sexuelle Freizügigkeit gemäß der Skala).

Das Beispiel illustriert nicht nur den Aufbau einer Guttman-Skala sondern gleichzeitig auch, wie stark theoretische Konstrukte und die zugehörigen Messinstrumente historischem Wandel unterliegen: „Voreheliche Sexualität“ ist heute in Deutschland der Normalfall, wird aber in unterschiedlichen Religionen und Kulturen durchaus noch kontrovers betrachtet und mit entsprechenden Permissivitätsinstrumenten gemessen.

a. Ich finde, dass Petting vor der Ehe erlaubt ist, wenn man verlobt ist. [ja/nein]
b. Ich finde, dass Petting vor der Ehe erlaubt ist, wenn man seine Partnerin (seinen Partner) liebt. [ja/nein]
c. Ich finde, dass Petting vor der Ehe erlaubt ist, wenn man für seine Partnerin (seinen Partner) starke Zuneigung empfindet. [ja/nein]
d. Ich finde, dass uneingeschränkte Sexualbeziehungen vor der Ehe erlaubt sind, wenn man verlobt ist. [ja/nein]
e. Ich finde, dass uneingeschränkte Sexualbeziehungen vor der Ehe erlaubt sind, wenn man seine Partnerin (seinen Partner) liebt. [ja/nein]
f. Ich finde, dass uneingeschränkte Sexualbeziehungen vor der Ehe erlaubt sind, wenn man für seine Partnerin (seinen Partner) starke Zuneigung empfindet. [ja/nein]
g. Ich finde, dass uneingeschränkte Sexualbeziehungen vor der Ehe erlaubt sind, auch wenn man keine besonders starke Zuneigung für seine Partnerin (seinen Partner) empfindet. [ja/nein]

Eine Person, die z. B. Item c ablehnt, müsste gemäß Guttman-Skalierung auch die Items d bis g ablehnen, die noch mehr sexuelle Freizügigkeit beinhalten als Item c. Wäre das Item b für diese Person akzeptierbar, müsste sie Item a ebenfalls akzeptieren.

Ein Beleg für Modellkonformität der gesamten Skala wäre die in Tab. 8.22a dargestellte Antwortmatrix.

Person 1 (oder eine Personengruppe mit diesem Antwortmuster) lehnt alle Items ab und bringt damit zum Ausdruck, dass sie gegen voreheliche Sexualität jeglicher Art ist. Person 5 hingegen befürwortet das relativ „liberale“ Item d und müsste damit bei einer modellkonformen Skala auch den Items a–c zustimmen, deren Bejahung für weniger sexuelle Freizügigkeit spricht als die Bejahung von Item d. Person 8 schließlich stimmt allen Items zu, wodurch die höchste, mit dieser Skala messbare, sexuelle Freizügigkeit zum Ausdruck gebracht wird.

In Tab. 8.22b haben zwei Personen nicht modellkonform reagiert: Person 2 befürwortet Item c, obwohl sie das schwächere Item b ablehnt. Und Person 6 dürfte bei einer modellkonformen Skala Item c nicht ablehnen, weil sie die stärkeren, d. h. für mehr sexuelle Freizügigkeit stehenden Items d und e bejaht. Bei diesen Personen ist also die Regel verletzt, dass aus dem stärksten bejahten Item das gesamte Reaktionsmuster rekonstruierbar sein muss.

Tabelle 8.22 Antwortmatrizen für Guttman-Skalen (7 Items – a bis g – der „Premarital Sexual Permissiveness Scale“, beantwortet durch 8 Personen) + Zustimmung, − Ablehnung

Items							
Person	**a**	**b**	**c**	**d**	**e**	**f**	**g**
a. Modellkonformes Antwortverhalten							
1	−	−	−	−	−	−	−
2	+	−	−	−	−	−	−
3	+	+	−	−	−	−	−
4	+	+	+	−	−	−	−
5	+	+	+	+	−	−	−
6	+	+	+	+	+	−	−
7	+	+	+	+	+	+	−
8	+	+	+	+	+	+	+
b. Nicht modellkonformes Antwortverhalten							
1	−	−	−	−	−	−	−
2	+	−	+	−	−	−	−
3	+	+	−	−	−	−	−
4	+	+	+	−	−	−	−
5	+	+	+	+	−	−	−
6	+	+	−	+	+	−	−
7	+	+	+	+	+	+	−
8	+	+	+	+	+	+	+

Mit einer perfekten Reproduktion aller bejahten Items aufgrund des Gesamttestwertes dürfte allerdings nur bei sehr präzise definierten, eindeutig operationalisierten eindimensionalen Merkmalen zu rechnen sein. Um die Anwendbarkeit dieses Skalentyps nicht allzu stark einzuengen, schlägt Guttman vor, sich mit einer 90 %igen Reproduzierbarkeit aller Itemantworten aufgrund des Gesamttestwertes zu begnügen (Schnell et al., 2008).

Das praktische Vorgehen zur Bestimmung der Reproduzierbarkeit lässt sich wie folgt beschreiben: Man bestimmt zunächst die Anzahl der Zustimmungen pro Item und die Anzahl der Zustimmungen pro Person. Als nächstes ordnet man die Items und die Personen nach der Anzahl der Zustimmungen. Dies ist in Tab. 8.22a geschehen. Der Skalenwert einer Person entspricht bei Modellkonformität der Anzahl der akzeptierten Items. Demnach wäre z. B. der Person 3 der Skalenwert 2 zuzuordnen.

In Tab. 8.22b hat Person 2 ebenfalls 2 Items akzeptiert, allerdings nicht modellkonform, denn das liberalere Item b wurde abgelehnt und das weniger liberale Item c akzeptiert. Gegenüber der modellkonformen Person 3 mit ebenfalls 2 Zustimmungen hat Person 2 auf 2 Items „fehlerhaft" reagiert, d. h., es werden 2 Fehler notiert. Zwei weitere „Fehler" hat Person 6 (mit dem Skalenwert 4) gemacht: Gegenüber der modellkonformen Person 5 wurde auf die Items c und e modellkonträr reagiert. Insgesamt ergibt die „Skalogrammanalyse" also 4 „Fehler" d. h., Modellverletzungen, die nach folgender Gleichung in einen **Reproduzierbarkeitskoeffizienten (REP)** überführt werden:

$$\text{REP} = 1 - \frac{\text{Anzahl der Fehler}}{\text{Anzahl der Befragten} \cdot \text{Anzahl der Items}}. \tag{8.4}$$

Für das Beispiel mit 8 Personen und 7 Items erhält man

$$\text{REP} = 1 - \frac{4}{8 \cdot 7} = .93.$$

Dieser Wert liegt über .90 und würde damit Modellkonformität der Skala signalisieren. Ein weiteres Maß zur Prüfung der Modellkonformität stellt Loevingers H-Koeffizient dar, der z. B. bei Roskam (1996, S. 439) beschrieben wird.

Die hier diskutierte Skala verdeutlicht, wie bereits angesprochen, den historischen und kulturellen Bezug theoretischer Konstrukte. So gehen alle Skalenitems ganz selbstverständlich davon aus, dass Menschen heiraten und Biografien in eine Phase „vor der Ehe" und eine Phase „in der Ehe" zerfallen. Wer diese Vorstellung nicht teilt, für den sind die Skalenitems sinnlos. Generell ist bei der Formulierung von Items darauf zu achten, dass sie keine impliziten Aussagen enthalten, die von den Untersuchungspersonen möglicherweise nicht geteilt werden und ihnen somit keine Möglichkeit zum adäquaten Antworten lassen. Ein Ausweg aus diesem Problem ist die Verwendung von vorgeschalteten Filterfragen, die unterschiedliche Personengruppen identifizieren, denen dann jeweils nur die zur aktuellen Lebenssituation oder zu den individuellen Lebenseinstellungen passenden Fragen vorgelegt werden (▶ Abschn. 10.3.3 „Ablauf einer vollstrukturierten Fragebogenstudie").

Das Spektrum der Konstrukte, für die in der Literatur bereits geprüfte Guttman-Skalen vorliegen, ist breit: Es liegen z. B. Guttman-Skalen zur Messung der Intensität häuslicher Gewalt (Keller & Wagner-Steh, 2005), der Kompetenz zur Kondomverwendung (Lindemann & Brigham, 2003) oder zur Karriereorientierung im Ehrenamt (Zakour 1994) vor.

8.6.3 Thurstone-Skala

Ebenso wie bei der Guttman-Skala werden auch bei der Thurstone-Skala (Thurstone & Chave, 1929) die Items nur mit ja/nein bzw. Zustimmung/Ablehnung und nicht wie bei der Likert-Skala mittels Ratings beantwortet. Während die Guttman-Skala durch die Itemformulierung selbst die Intensität des Merkmals abbildet, werden bei der Thurstone-Skala für jedes Item Expertenratings mitgeliefert, die den Grad der Merkmalsausprägung des Items repräsentieren sollen. Der Gesamtscore der Thurstone-Skala ergibt sich also aus der Summe der Expertenratings für alle von der Befragungsperson bejahten Items.

Thurstone-Skala – Eine Thurstone-Skala („thurstone scale") ist eine psychometrische Skala, die aus mehreren Aussagen besteht, die alle dasselbe Merkmal messen und jeweils mit Zustimmung oder Ablehnung (Ja-/Nein-Format) einzuschätzen sind. Für alle Thurstone-Items wurden im Zuge der Skalenkonstruktion Expertenurteile darüber eingeholt, welche Intensität des Merkmals das jeweilige Item ausdrückt. Diese Expertenurteile werden auf Ratingskalen abgegeben. Das durchschnittliche Experten-Rating pro Item fungiert als Gewichtungsfaktor, ist jedoch den Befragungspersonen nicht bekannt. Der Skalenwert der Thurstone-Skala berechnet sich als Summe der Gewichtungsfaktoren der zugestimmten Items.

Die „klassische" Thurstone-Skala verwendet als Items Behauptungen, die unterschiedliche Bewertungen des untersuchten Einstellungsgegenstandes enthalten. („Der Gottesdienst inspiriert mich und gibt mir Kraft für die ganze Woche"; oder „Ich meine, dass die Kirche nur für arme und alte Leute gut ist" – zwei Itembeispiele für eine Skala zur Messung von Einstellungen zur Kirche.) Als Skala zur Messung von Persönlichkeitsmerkmalen werden Behauptungen gesammelt, deren Bejahung auf

unterschiedliche Ausprägungen des untersuchten Merkmals schließen lässt (z. B. „Ich halte mich grundsätzlich an die Regel, Auge um Auge, Zahn um Zahn"' oder „Wenn mich jemand beschimpft, neige ich dazu, wortlos aus dem Felde zu gehen", als mögliche Behauptungen in einer psychometrischen Skala zur Messung von Aggressivität).

Diese gesammelten Items werden einer Gruppe von Expertinnen und Experten (z. B. mit einem fachlichen Hintergrund in Psychologie, Soziologie, Methodenlehre) mit der Bitte vorgelegt, die Merkmalsausprägung, die mit der Bejahung der einzelnen Items zum Ausdruck gebracht wird, auf einer 11-Punkte-Ratingskala einzustufen. In der Instruktion für dieses Rating ist besonders hervorzuheben, dass nicht das persönliche Zutreffen der Behauptungen interessiert, sondern die mit der Bejahung einer Behauptung verknüpfte Merkmalsausprägung (Goodstadt & Magid, 1977). Als Skalenwert für ein Item gilt die durchschnittliche Itemeinstufung durch alle Experten. Die Skalenwerte sollten möglichst das gesamte Merkmalskontinuum (von 1 bis 11) repräsentieren. Items mit hoher Streuung werden wegen mangelnder Urteilerübereinstimmung ausgeschieden und durch umformulierte oder neue Items ersetzt. Thurstone verwendete als Skalenwert den Median der Urteilsverteilung und als Streuung den Interquartilrange. Zumindest bei unimodalen symmetrischen Urteilsverteilungen können diese Kennwerte jedoch durch das arithmetische Mittel und die Standardabweichung ersetzt werden.

Eine Konstruktionsalternative stellt der in ▶ Abschn. 8.4.6 behandelte Dominanzpaarvergleich dar. Hierbei müssen die Urteiler (Experten) bei jedem Itempaar angeben, welcher Itempaarling hinsichtlich des untersuchten Einstellungsobjektes passender ist. Die Skalenwerte der Items werden auf der Basis der Paarvergleichsurteile nach dem „Law of Comparative Judgement" (▶ Abschn. 8.4.6) ermittelt.

Zur weiteren Überprüfung der Skalenqualität empfiehlt Thurstone, die vorerst als brauchbar erscheinenden Items einer Stichprobe von Testpersonen mit einer (von der Experteninstruktion abweichenden) Instruktion vorzulegen, nach der zu prüfen ist, ob die Items auf sie persönlich zutreffen oder nicht. Stellt sich hierbei heraus, dass einigen Items mit niedrigem Skalenwert (geringe Merkmalsausprägung) zugestimmt und anderen Items mit höherem Skalenwert (stärkere Merkmalsausprägung) nicht zugestimmt wird, sollten diese Items ebenfalls überprüft und ggf. herausgenommen werden.

Die so überarbeiteten Items stellen die endgültige Skala dar, die den Untersuchungspersonen mit der Bitte um Zustimmung oder Ablehnung (natürlich ohne Bekanntgabe der Skalenwerte) vorgelegt werden. Der Skalenwert einer Person ergibt sich als Summe der Skalenwerte der von ihr akzeptierten oder bejahten Behauptungen (zur Kritik dieser Skala, die vor allem die Festlegung der Skalenwerte durch eine mehr oder weniger willkürlich ausgewählte Expertengruppe betrifft, vgl. z. B. Schnell, Hill, & Esser, 1999, S. 180f.). Das Ergebnis einer Thurstone-Skalierung anhand eines Beispiels zeigt ▶ Menschliche Kontakte in Siedlungen.

Einen kritischen Vergleich von Likert- und Thurstone-Skala hinsichtlich ihrer Validität findet man bei Roberts, Laughlin, & Wedell (1999).

Während Thurstones „Law of Comparative Judgement" (▶ Abschn. 8.4.6) zur Transformation von Paarvergleichsdaten auf Intervallskalenniveau bis heute angewendet und zuweilen auch als Thurstone-Skalierung bezeichnet wird, spielt die hier beschriebene Thurstone-Skala in der Praxis kaum noch eine Rolle. Vorliegende ältere Thurstone-Skalen messen z. B. Einstellungen gegenüber Erwachsenenbildung (Blunt 1983) oder sexuelle Einstellungen in Indien (Chatterjee & Puhan, 1980).

Menschliche Kontakte in Siedlungen

Beispiel für eine Thurstone-Skala

Bongers und Rehm (1973) konstruierten eine Skala zur Kontaktsituation in Wohnsiedlungen. Experten (es handelte sich um Architekten, Psychologen und Stadtplaner) wurden gebeten, verschiedene Aussagen, die die Kontaktgestaltung in einer Siedlung betreffen, auf einer 11-Punkte-Skala von −5 bis +5 einzustufen. Die Skala war in folgender Weise „verankert":

−5	Nachbarschaftliche Kontakte sind extrem schlecht.
0	In Bezug auf nachbarschaftliche Kontakte neutral.
+5	Nachbarschaftliche Kontakte sind extrem gut.

Für jedes Item wurde ein mittleres Expertenrating berechnet. (Die entsprechenden Werte sind in Klammern aufgeführt.)

- Ich komme mir in dieser Siedlung oft vor wie ein Fremder. [ja/nein] (−2.00)
- Keinem Menschen in der Nachbarschaft würde es auffallen, wenn mir etwas zustieße. [ja/nein] (−3.05)
- Hier in der Siedlung haben die Menschen keine Geheimnisse voreinander. [ja/nein] (+3.30)
- Ich habe oft den Eindruck, dass sich die Menschen in meinem Wohnbezirk nur flüchtig kennen. [ja/nein] (−0.53)
- Ich kenne kaum jemanden in meinem Wohnbezirk, mit dem ich über private Dinge reden könnte. [ja/nein] (−0.33)
- In diesem Wohnbezirk ist es kaum möglich, sich auch nur für kurze Zeit von den anderen zurückzuziehen. [ja/nein] (+1.79)
- Ich kenne hier in der Nachbarschaft fast jeden mit Namen. [ja/nein] (+0.90)

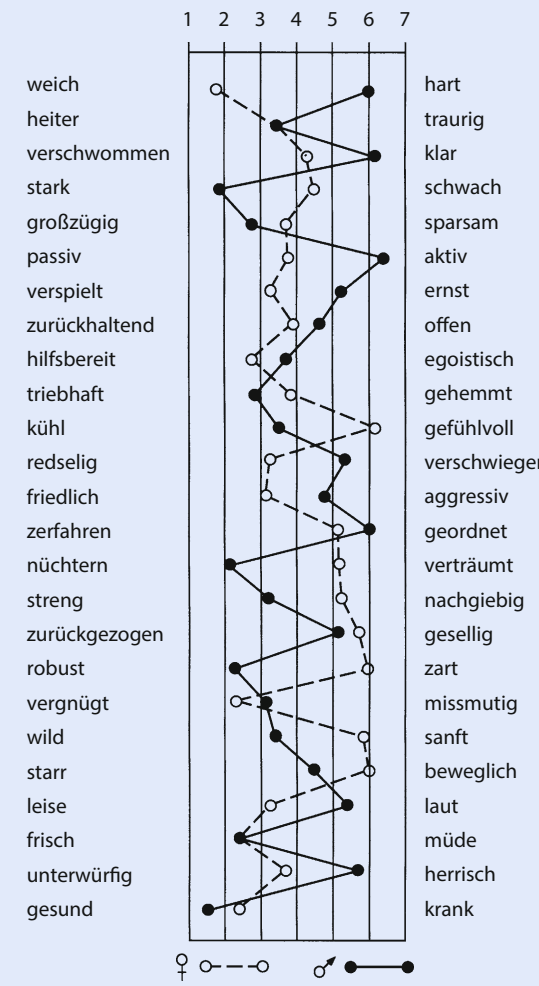
Der Wert einer Person ergibt sich als Summe der Skalenwerte der von ihr bejahten Items. (Ausgeschieden wurden Items mit einer Standardabweichung über 1.50.)

8.6.4 Semantisches Differenzial

Das Semantische Differenzial – oft auch noch geschrieben als „Differential“ – („semantic differential“ SD, „semantic differential scale“) wurde 1957 von dem US-amerikanischen Psychologen Charles Osgood und Kollegen entwickelt und hat seit seiner Einführung als Polaritätsprofil oder Eindrucksdifferenzial durch den österreichischen Psychologen Peter Robert Hofstätter (1957, 1977) auch im deutschsprachigen Raum weite Verbreitung gefunden. Es handelt sich um einen speziellen Typ der psychometrischen Skala, der zur Messung der konnotativen Bedeutung bzw. der affektiven Qualitäten beliebiger Objekte oder Begriffe dient („Schuhe, Schiffe und Siegellack, Kohl und Könige, ich, dein Vater, Fräulein Weber, mein Lehrer, die Schule, Algebra, ein Demokrat, dieses Buch, eine Büroklammer, die Vereinten Nationen, Eisenhower etc.“; nach Osgood, Suci, & Tannenbaum, 1957, S. 91).

Das Semantische Differenzial besteht aus 20–30 Adjektivpaaren (z. B. heiter/traurig, rund/eckig, laut/leise, beweglich/starr), hinsichtlich der das Urteilsobjekt auf siebenstufigen bipolaren Ratingskalen einzustufen ist. Das Semantische Differential der Begriffe „männlich“ und „weiblich“ ist in ◘ Abb. 8.7 dargestellt. Urteilsgrundlage ist die metaphorische Beziehung bzw. gefühlsmäßige Affinität des Urteilsgegenstandes zu den Adjektiven und weniger der sachliche oder denotative Zusammenhang, der häufig nicht gegeben ist. („Männlich“ bzw. „weiblich“ sind denotativ weder „laut“ noch „leise“ und wurden bei der Studie in den 1960er-Jahren trotzdem, wie ◘ Abb. 8.7 zeigt, unterschiedlich mit diesem Begriffspaar assoziiert.) Das Instrument eignet sich besonders für die Messung von sozialen Stereotypen.

Als Ergebnis einer Messung mit dem Semantischen Differenzial entsteht zunächst ein Profilverlauf für das Urteilsobjekt, wobei pro Item (d. h. pro Adjektivpaar) jeweils der Durchschnittswert der Untersuchungspersonen genutzt wird. Aussagekräftig werden Messungen mit dem Semantischen Differenzial vor allem dadurch, dass man die Beurteilung unterschiedlicher Urteilsobjekte vergleichend nebeneinander stellt (z. B. Beurteilung von „männlich“ und „weiblich“; von „alt“ und „jung“; von „Fernsehen“ und „Internet“ oder von „gesund“ und „krank“). Mithilfe der Korrelationsrechnung lässt sich die Ähnlichkeit der Profile verschiedener Urteilsgegenstände bestimmen. (Die Profile in ◘ Abb. 8.7 korrelieren zu −.07 miteinander.) Die Faktorenanalyse über derartige Korrelationen führt üblicherweise zu 2 bis 4 Dimensionen, die durch die Positionen der untersuchten Objekte auf den jeweiligen Dimensionen beschreibbar sind. Die Dimensionen des in ◘ Abb. 8.7 wiedergegebenen Begriffssystems lassen sich nach Hofstätter (1963) als kulturell und historisch geprägte Vorstellungen von Weiblichkeit (F1) und Männlichkeit (F2) interpretieren. Sprachvergleichende Untersuchungen von Osgood et al. (1957) über verschiedene Begriffe führten beim Semantischen Differenzial in der Regel zu einem dreidimensionalen System, einem „semantischen Raum“ mit den Dimensionen **Evaluation** (Bewertung, z. B. angenehm – unangenehm), **Potency** (Macht, z. B. stark – schwach) und **Activity** (Aktivität, z. B. erregend – beruhigend). Dieser semantische Raum wird vereinfachend auch als **EPA-Struktur** bezeichnet. Für die Dimensionen kann jeweils durch (ungewichtete oder gewichtete) Zusammenfassung der Ratings der zugehörigen Items ein Subskalenwert gebildet werden. Vergleiche zwischen Urteilsobjekten sind dann nicht nur anhand der Profil-



◘ Abb. 8.7 Semantisches Differenzial der Begriffe „weiblich“ und „männlich“. (Mod. nach Peter R. Hofstätter: Einführung in die Sozialpsychologie. Stuttgart: Kröner 1973, S. 259. Mit freundlicher Genehmigung)

verläufe (d. h. Mittelwerte der 20–30 Items), sondern komprimierter auch über die 2–4 Dimensions- bzw. Subskalenwerte möglich.

Semantisches Differenzial – Das semantische Differenzial („semantic differential" SD, „semantic differential scale") ist ein spezieller Typ einer psychometrischen Skala, mit dem vor allem die konnotative Bedeutung von Begriffen oder Objekten mit Hilfemithilfe eines Satzes von 20–30 bipolaren Adjektivpaaren erfasst wird. Das Ergebnis ist ein für das betreffende Urteilsobjekt charakteristischer Profilverlauf. Durch die Adjektivpaare werden verschiedene Dimensionen des Urteilsobjekts erfasst, beim klassischen Semantischen Differenzial sind das die drei Dimensionen Evaluation, Potency und Activity (EPA-Struktur). Es existieren Semantische Differenziale mit unterschiedlichen Sätzen von Adjektivpaaren und unterschiedlicher Anzahl von Dimensionen (typisch sind zwei bis vier Dimensionen). Für die Dimensionen werden Subskalenwerte gebildet.

Statt des von Osgood und Hofstätter vorgeschlagenen **universellen Semantischen Differenzials** ◘ Abb. 8.7 werden gelegentlich kontextspezifische, auf die Besonderheiten der Untersuchungsgegenstände zugeschnittene Polaritätsprofile mit anderen Adjektivpaaren eingesetzt (Peterson 1999). **Kontextspezifische Polaritätsprofile** erfassen erstrangig die denotativen, direkten Beziehungen der Urteilsobjekte zu den Urteilsskalen und führen deshalb zu anderen Resultaten (anderen „semantischen Räumen") als ein universelles Semantisches Differenzial (Flade 1978). Geht es um den Vergleich sehr unterschiedlicher Urteilsobjekte, ist ein universelles Semantisches Differenzial vorzuziehen. Die Reihenfolge, in der die Objekte beurteilt werden, sowie die Polung der Skalen (z. B. hart – weich oder weich – hart) sind nach Kane (1971) für die Ergebnisse unerheblich.

Mann, Phillips und Thompson (1979) weisen darauf hin, dass die Untersuchungsergebnisse nur unbedeutend beeinflusst werden, wenn statt bipolarer Ratingskalen unipolare Antwortformate verwendet werden (Cogliser & Schriesheim, 1994; Ofir, Reddy, & Bechtel, 1987). Probleme bereitet der in mehreren Arbeiten nachgewiesene Befund, dass dieselbe Ratingskala von unterschiedlichen Beurteilern zuweilen verschieden aufgefasst wird bzw. dass die Bedeutung einer Ratingskala aus Sicht der Urteilenden von der Art der zu beurteilenden Objekte abhängt (Rater-Concept-Scale-Interaction: Cronkhite 1976; Flynn 1993; Yu, Albaum, & Swenson, 2003). Skalentheoretische Probleme und vertiefte methodische Aspekte diskutieren z. B. Bintig (1980) oder Brandt (1978), als auch Evans (1980), H. H. Friedman, Gluck, & L. W. Friedman (1988), Garland (1990), Murakami und Kroonenberg (2003), und Zhikun und Fungfai (2008).

Erfahrungsgemäß stößt das Semantische Differenzial bei unvorbereiteten Untersuchungspersonen gelegentlich auf Akzeptanzprobleme, weil die geforderten Urteile sehr ungewohnt sind (ist „Algebra" eher „großzügig" oder „sparsam"?). Es ist daher empfehlenswert, bereits in der Instruktion „vorzuwarnen", etwa mit dem Hinweis: „Bei einigen Adjektiven wird es Ihnen vielleicht schwerfallen, ein Urteil abzugeben. Antworten Sie trotzdem einfach so, wie es Ihrem spontanen Gefühl am ehesten entspricht. Es gibt keine richtigen oder falschen Antworten! Wir interessieren uns für Ihren ganz persönlichen Eindruck."

Die Anwendungsvarianten des Semantischen Differenzials sind vielfältig. In der Originalarbeit von Osgood et al. (1957) werden bereits ca. 50 Anwendungsbeispiele genannt. Finstuen (1977) sammelte zwischen 1952 und 1976 insgesamt 751 psychologische Anwendungen (Himmelfarb 1993). Peterson (1999) hält fest, dass seit der Veröffentlichung der Originalarbeit über tausend wissenschaftliche Artikel und Bücher zum Semantischen Differenzial erschienen sind. Aktuelle Publikationen nutzen das Semantische Differenzial unter anderem, um so diverse Urteilsobjekte wie Musikstücke, Düfte und Geschmacksrichtungen, Supermärkte, technische Geräte oder Menschen, die stottern, beurteilen zu lassen. Gut etabliert im deutsch- und englischsprachigen Raum ist u. a. ein spezielles Semantisches Differenzial, mit dem Vorurteile gegenüber älteren Menschen erfasst werden (Aging Semantic Differential: Gonzales, Tan, & Morrow-Howell, 2010; Gluth, Ebner, & Schmiedek, 2010).

8.7 Messung mit Indizes und Indexbildung

Ein Index erfasst ein mehrdimensionales bzw. komplexes theoretisches Konstrukt mit Hilfe eines Satzes formativer Indikatorvariablen (zum formativen Messmodell des Index (▶ Abschn. 8.3.1). Da **formative Indikatorvariablen** inhaltlich und formal sehr heterogen beschaffen sein können, ist die psychometrische Testtheorie, die zur Bildung von Skalen mit ihren homogenen reflektiven Items herangezogen wird (▶ Abschn. 10.4.4), für die Indexbildung nicht nutzbar. Im Unterschied zur Skala, die üblicherweise Merkmale auf Individualebene erhebt, werden mit Indizes auch oft Merkmale auf Aggregatebene gemessen, dabei lassen sich als Indikatoren bei Bedarf Kennzahlen eines Landes (z. B. Bruttoinlandsprodukt, durchschnittliche Lebenserwartung) oder einer Organisation (z. B. Aktienkurs) heranziehen. Der Vorteil eines Index besteht darin, mehrdimensionale Sachverhalte auf einen einzigen greifbaren, vergleichbaren und gut kommunizierbaren Wert zu verdichten. Genau darin liegt gleichzeitig der Schwachpunkt eines Index, dem im Falle unreflektierter Konstruktionsregeln der Vorwurf zu machen wäre, dass hier „Äpfel und Birnen" zu einem fragwürdigen synthetischen Wert verrechnet werden.

Index – Ein Index („index“) dient der Operationalisierung komplexer bzw. mehrdimensionaler theoretischer Konstrukte auf individueller Ebene (z. B. Merkmale einer Person) sowie oft auch auf kollektiver Ebene (z. B. Merkmale von Märkten oder Gesellschaften) auf der Basis eines formativen Messmodells. Ein Index besteht aus einem Satz von formativen Indikatorvariablen, die formal und inhaltlich sehr heterogen sein können. Der Index schreibt vor, welche Indikatorvariablen in den Index eingehen und in welcher Weise sie zu standardisieren, zu gewichten und additiv oder multiplikativ zu einem Indexwert zu verrechnen sind. Die Aussagekraft von Indexwerten muss theoretisch begründet und möglichst auch empirisch geprüft werden.

Beispiel: Die Gesamtnote für einen Aufsatz berücksichtigt die Rechtschreibung, den Stil und den Inhalt des Textes. Jeder dieser drei Einzelindikatoren trägt zur Gesamtqualität des Aufsatzes bei. Entsprechend setzt sich die Gesamtnote aus den Werten für diese Teilaspekte zusammen. Werden Teilnoten für Rechtschreibung, Stil und Inhalt vergeben, so kann der Index „Gesamtnote“ als einfacher Mittelwert berechnet werden. Dabei würde jeder Indikator dasselbe Gewicht erhalten (ungewichteter additiver Index).

Man könnte nun aber argumentieren, dass letztlich der Inhalt das entscheidende Merkmal eines Aufsatzes sei und deswegen die Gesamtnote stärker beeinflussen sollte als z. B. Rechtschreibfehler. Dies würde nahe legen, dass man z. B. die Note für den Inhalt doppelt gewichtet. Daraus ergibt sich folgende Formel eines Indexes „Gesamtnote“ – intervallskalierte Noten vorausgesetzt (gewichteter additiver Index; Gl. 8.5):

$$\text{Note}_{\text{Gesamt}} = \frac{2 \times \text{Note}_{\text{Inhalt}} + \text{Note}_{\text{Stil}} + \text{Note}_{\text{Rechtschreibung}}}{4}. \tag{8.5}$$

Bei der Bildung eines Indexwertes sind theorie- und/oder empiriebasiert folgende Entscheidungen zu treffen:

1. Anzahl und Auswahl der im Index zu berücksichtigenden relevanten formativen Indikatorvariablen;
2. Art der Standardisierung der Indikatorvariablen;
3. Art der Verknüpfung der Indikatorvariablen (additiv oder multiplikativ);
4. Art der Gewichtung der Indikatorvariablen.

Indizes spielen vor allem in der **Ökonomie** und **Soziologie** eine große Rolle, wenn es z. B. darum geht, den Zustand eines Marktes oder einer Gesellschaft in einem einzelnen Messwert zu fassen und damit historisch und international vergleichbar zu machen. Wohlstand und Lebensqualität eines Landes werden z. B. über sog. **Wohlfahrtsindizes** erfasst, in die Dimensionen wie Einkommen, Gesundheit, Bildung, Umweltschutz etc. eingehen. Was genau unter „Wohlfahrt“ zu verstehen ist und welche Dimensionen sie hat, ist zunächst theoretisch im Rahmen der Konzeptspezifikation zu klären (► Abschn. 8.2), bevor die relevanten Indikatoren ausgewählt werden können. Gerade wenn sehr unterschiedliche Indikatoren einfließen (z. B. Alphabetisierungsrate in Prozent; durchschnittliche Lebenserwartung in Jahren, Bruttoinlandsprodukt pro Kopf in Euro), muss vor der Indexbildung eine Standardisierung der Variablen vorgenommen werden, damit die Einheiten vergleichbar sind. Über die Verrechnungsregel für die Indikatoren (z. B. additiv oder multiplikativ) sowie die Zuweisung von Gewichtungsfaktoren ist zu entscheiden. Es liegt auf der Hand, dass diese Entscheidungen gründlicher methodischer und theoretischer Reflexion bedürfen und im Ergebnis oft kritisch diskutiert werden (zu Funktionen, Ansätzen und Problemen der Indexbildung am Beispiel der Wohlfahrtsmessung Noll, 2002).

Indizes, die für die wissenschaftliche Forschung fruchtbar sein sollen, lassen sich ebenso wenig wie psychometrische Skalen „ad hoc“ erstellen, sondern erfordern einen systematischen, **theoretisch begründeten und empirisch abgesicherten Konstruktionsprozess**, da die Qualität des Instruments sichergestellt werden muss. Im Unterschied zur psychometrischen Skalenkonstruktion wird die genaue Vorgehensweise bei der Indexbildung in der Methodenliteratur bislang weniger ausführlich behandelt (Diamantopoulos & Winklhofer, 2001). Für die Indexbildung mit formativen Indikatoren wird u. a. das **C-OAR-SE-Modell** vorgeschlagen (Rossiter 2002, 2010), das aus den fünf Schritten 1.) Construct Definition, 2.) Object Classification, 3.) Attribute Classification (reflective, formative), 4.) Rater Identification, 5.) Scale Formation und 6.) Enumeration besteht. Das C-OAR-SE-Modell soll eine Alternative zur herkömmlichen psychometrischen Instrumententwicklung bieten und ist auch auf reflektive Indikatoren anwendbar (zur Kritik des Modells Diamantopoulos 2005; Finn & Kayande, 2005).

8.7.1 Auswahl und Art der Indikatoren

Die Qualität eines Index hängt wesentlich davon ab, ob alle relevanten Dimensionen bzw. Indikatoren ausgewählt und angemessen gewichtet wurden. Die Auswahl der Dimensionen erfolgt nach Maßgabe theoretischer Überlegungen (z. B. Bedeutungsanalyse oder dimensionale Analyse, ► Abschn. 8.2) sowie empirischer Vorkenntnisse und muss sich in der Praxis bewähren. Angenommen, man konstruiert auf der Basis von Expertenurteilen und klinischen Befunden einen **Index Operationstauglichkeit von Transplantationspatienten**, in den u. a. die Anzahl vorausgegangener Infektionen, das Lebensalter, die psychische Verfassung, die Stabilität des Herz-Kreislauf-Systems etc. eingehen. Diesen Index könnte man erproben, indem er zunächst ohne jeden Einfluss auf Operationsentscheidungen einfach bei allen Patienten berechnet wird. Zeigt sich ein substanzieller Zusammenhang zwischen späterem Operationserfolg und Indexwert, so

spricht dies für die praktische Tauglichkeit des berechneten Index.

Die für die Indexbildung ausgewählten Einzelindikatoren können dichotom (z. B. „vorhanden – nicht vorhanden“, „ja – nein“, „trifft zu – trifft nicht zu oder polytom (z. B. Einkommensgruppen, Schulabschlüsse) sowie intervallskaliert (z. B. Merkmale, die mit Ratingskalen erfasst werden) oder verhältnisskaliert sein (z. B. Körpergröße und Körpergewicht). In jedem Fall muss es sich im Hinblick auf das komplexe Zielmerkmal um geordnete Kategorien handeln (d. h. **mindestens Ordinalskalenniveau**, was bei den zuvor genannten dichotomen und polytomen Merkmalen zutrifft, denn es ist klar, in welcher Kategorie das Merkmal jeweils stärker oder schwächer ausgeprägt ist). Arbeitet man mit nominalen Indikatoren, stellt man bald fest, dass sich bei gleichzeitiger Berücksichtigung mehrerer Dimensionen eine Vielzahl von Merkmalskombinationen ergeben, die sich nicht ohne weiteres auf eine übersichtliche Zahl von Indexwerten reduzieren lassen (Schnell et al., 2008).

Die strukturierende Funktion der Indexbildung spielt auch bei einer der bekanntesten Anwendungen dieser Technik eine Rolle: dem Index für Wertorientierungen (▶ Der Inglehart-Index zur Messung der Wertorientierung).

8.7.2 Standardisierung der Indikatoren

Sollen Indikatoren zu einem Index verknüpft werden, die unterschiedlichen Maßeinheiten haben, so sind diese zuvor zu standardisieren, d. h., auf eine einheitliche Maßeinheit zu bringen.

Eine Möglichkeit besteht in der inhaltlichen Umrechnung aller Indikatoren auf **dieselbe Maßeinheit**. Beispiel: In einem zeitgemäßen **Wohlfahrtsindex** sollen neben dem Bruttoinlandsprodukt (gemessen in Euro) als Maßstab für den Wohlstand eines Landes u. a. auch Umweltbelastungen (z. B. Treibhausgase, Wasserverschmutzung) und gesundheitliche Einschränkungen (z. B. Alkoholabhängigkeit) berücksichtigt werden, da diese die Lebensqualität der Bevölkerung senken. Die Berücksichtigung entsprechender Phänomene erfolgt im Sinne der Standardisierung nicht über Indikatorvariablen unterschiedlicher Maßeinheiten (z. B. CO_2-Emmission in Tonnen pro Jahr; Verbreitung von alkoholassoziierten Erkrankungen pro 100 000 Einwohner), vielmehr werden sie jeweils in finanzielle Kosten umgerechnet, so dass in den Index letztlich nur Variablen in der Maßeinheit „Euro pro Jahr“ eingehen (Diefenbacher & Zieschank, 2008).

Häufig ist es allerdings nicht möglich, alle Indikatoren inhaltlich auf ein und dieselbe Maßeinheit zu bringen. In diesem Fall wird in der Praxis meist mit **Prozentuierungen** gearbeitet, d. h., die Indikatorwerte werden jeweils an ihrem Maximalwert relativiert. Beispiel: Ein **Genderindex**, der die Chancengleichheit von Frauen und Männern in unterschiedlichen Regionen der Bundesrepublik anhand von 19 Indikatoren erfasst, bestimmt für jeden Indikator einheitlich die relative Geschlechterdifferenz in Prozent von −100 % (absolute Benachteiligung der Frauen) bis +100 % (absolute Benachteiligung der Männer; http://www.gender-index.de). Wenn also in einer Region auf dem Indikator „Bürgermeister/innen“ 3 Bürgermeisterinnen und 8 Bürgermeister gezählt werden, erfolgt die Standardisierung des Indikators nach folgender Formel am Maximalwert von 11 Bürgermeisterpositionen insgesamt:

$$\begin{aligned}&\text{standardisierter Gender-Index Indikator}\\&= 100 \times \frac{\text{Indikator Frauen} - \text{Indikator Männer}}{\text{Indikator Frauen} + \text{Indikator Männer}}\\&= 100 \times \frac{3-8}{3+8} = -45\,\%\end{aligned} \tag{8.6}$$

Ein weiteres gebräuchliches Standardisierungsverfahren für Indikatoren ist die ***z*-Transformation**, bei der jeder Messwert eines Indikators vom Mittelwert subtrahiert und durch die Streuung geteilt wird. Die z-Werte haben dann einheitlich einen Mittelwert von 0 und eine Streuung von 1. Positive z-Werte zeigen überdurchschnittliche, negative z-Werte unterdurchschnittliche Merkmalsausprägungen an. Ganz unabhängig von der ursprünglichen Maßeinheit lassen sich zwei z-Werte direkt miteinander vergleichen.

Beispiel: Angenommen, man wollte in den in Gl. 8.5 vorgestellten **Index zur Aufsatzqualität** noch die Bearbeitungsgeschwindigkeit aufnehmen. Würde man die Werte unstandardisiert verrechnen, so würde eine um eine Minute kürzere Bearbeitungszeit im Index ebenso zu Buche schlagen, wie eine volle Schulnotendifferenz auf einem der anderen Indikatoren. Eine z-Standardisierung der Werte aller Indikatoren anhand des jeweiligen Stichprobenmittelwertes sowie der Stichprobenstreuung würde dagegen dazu führen, dass sich im Index über- und durchschnittliche Leistungen auf allen Indikatoren gleichermaßen abbilden. Allerdings mit dem Nachteil, dass der z-standardisierte Indexwert weniger anschaulich wirkt.

$$\begin{aligned}&z\text{-standardisierter Index Aufsatzqualität}\\&= \frac{(2 \times z\text{-Wert}_{\mathrm{NI}} + z\text{-Wert}_{\mathrm{NS}} + z\text{-Wert}_{\mathrm{NR}} + z\text{-Wert}_{\mathrm{B}})}{5}.\end{aligned} \tag{8.7}$$

wobei

NI	Note Inhalt
NS	Note Stil
NR	Note Rechtschreibung
B	Bearbeitungszeit

Der Inglehart-Index zur Messung der Wertorientierung

Datenerhebung

Welches der folgenden Ziele halten Sie persönlich für besonders wichtig?

A Aufrechterhaltung der nationalen Ordnung und Sicherheit.

B Verstärktes Mitspracherecht der Menschen bei wichtigen Regierungsentscheidungen.

C Kampf gegen steigende Preise.

D Schutz der freien Meinungsäußerung.

Welches dieser Ziele sehen Sie als das wichtigste an?

– Bitte tragen Sie den Buchstaben (A–D) in das Feld ein ☐

Welches dieser Ziele sehen Sie als das zweitwichtigste an?

– Bitte tragen Sie den Buchstaben (A–D) in das Feld ein ☐

Auswertung

Die obigen vier Aussagen lassen sich gemäß dem US-amerikanischen Politologen Ronald Inglehart (1977, 1997) als Indikatoren für die Wertorientierung nutzen. Die Aussagen A und C repräsentieren dabei **materialistische Werte** (physische Sicherheit und materielles Wohlergehen), während die Aussagen B und D **postmaterialistische Werte** zum Ausdruck bringen (Selbstverwirklichung und individuelle Freiheit). Indem aus den vier Aussagen eine Erst- und eine Zweitpräferenz gewählt werden, ergeben sich 12 mögliche Kombinationen.

Der Index sieht jedoch vor, dass die Erstpräferenz stärker gewichtet wird als die Zweitpräferenz, so dass durch die Erstwahl die Zuordnung zur Wertorientierung „Materialismus" (A oder C) bzw. „Postmaterialismus" (B oder D) festgelegt ist. Die Zweitpräferenz kann diese Wahl entweder verstärken, wenn die zweite Aussage aus derselben Wertorientierung stammt (reiner Materialismus bzw. reiner Postmaterialismus), oder abschwächen, indem eine Aussage aus der anderen Wertorientierung gewählt wird (eher materialistische Orientierung bzw. eher postmaterialistische Orientierung).

Nach diesem Schema lassen sich auf numerischer Ebene die 12 Kombinationen auf 4 Indexwerte reduzieren, indem man etwa für die Materialismus-Aussagen A oder C je 2 Punkte vergibt, wenn sie als Erstpräferenz gewählt werden und je 1 Punkt, wenn sie als Zweitpräferenz gewählt werden. Die Indexwerte würden dann zwischen 3 (reiner Materialismus = gar kein Postmaterialismus) und 0 (gar kein Materialismus = reiner Postmaterialismus) variieren (◘ Tab. 8.23).

Alternativen

Diese sehr einfache Form der Indexbildung aus wenigen gleichartigen Indikatoren (hier: politischen Aussagen) hat den Vorteil, dass eine ökonomische Datenerhebung mittels standardisierter mündlicher oder schriftlicher Befragung leicht möglich ist. So wird der Inglehart-Index auch im **ALLBUS** (regelmäßige allgemeine Bevölkerungsumfrage der Sozialwissenschaften des GESIS – Leibniz-Institut für Sozialwissenschaften) miterfasst, um einen möglichen gesellschaftlichen Wertewandel in der Bundesrepublik Deutschland zu messen. Materialistische und postmaterialistische Wertorientierung ließen sich freilich auch mit einem Index erfassen, in den eine Vielzahl anderer Indikatoren eingehen, die neben Meinungen konkrete Verhaltensweisen einschließen (z. B. politisches Engagement für Bürgerrechte, finanzielle Ausgaben und zeitlicher Aufwand für materielle vs. ideelle Güter etc.).

Aktuelle Forschung

In der Soziologie spielt die Erfassung von Wertvorstellungen eine große Rolle. Einerseits ist die historische Veränderung von Interesse, der sog. Wertewandel. Zudem sind generationen- und kulturvergleichende Studien besonders interessant. Das „**World Values Survey**" (http://www.worldvaluessurvey.org) ist ein globaler Forschungsverbund, der in mehreren Wellen die Wertvorstellungen der Menschen in 50 Ländern u. a. mit aktualisierten Fassungen des Inglehart-Index erfasst. Die Instrumente und Datensätze stehen über die Projekt-Website für Sekundäranalysen zur Verfügung.

8.7.3 Zusammenfassung der Indikatoren

Nach Art der rechnerischen Zusammenfassung der Einzelindikatoren werden verschiedene Arten von Indizes unterschieden:

- ungewichteter additiver Index
- multiplikativer Index und
- gewichteter additiver Index.

Ungewichteter additiver Index

Die einfachste Form der Indexbildung besteht darin, die Ausprägungen der Indikatorvariablen einfach zu addieren bzw. zu mitteln (z. B. Durchschnittswert der Noten für Rechtschreibung, Stil und Inhalt im oben genannten Beispiel). Bei dichotomen Antwortvorgaben führt dies zur Bildung der Summe aller positiv beantworteten Fragen. Dabei legt man zugrunde, dass alle Indikatoren das komplexe Merkmal mit derselben Präzision messen und theoretisch von gleicher Bedeutung sind. Diese Vorstellung ist genau zu begründen und in ihrem Vereinfachungsgrad nicht unproblematisch. Dennoch sind additive Indizes sehr verbreitet; auch additive Summenscores aus psychometrischen Skalen sind vom Verfahren her als ungewichtete additive Indexwerte zu kennzeichnen.

Inhaltlich ermöglicht ein additiver Index **Kompensationen**, d. h., ein geringer Wert auf einem Indikator kann durch einen höheren Wert auf einem anderen Indikator kompensiert werden. Dies ist etwa bei dem „klassischen" **Schichtindex** von Scheuch (1961) der Fall: Geringe Bildung kann durch hohes Einkommen kompensiert werden, d. h., eine Person mit hoher Bildung und geringem

Tabelle 8.23 Kombinationen der Präferenzen zur Wertorientierung und gebildete Indexwerte

Erstwahl	Zweitwahl	Kombination	Indexwert	
A	C	1	3	reiner Materialismus
C	A	2		
A	B	3	2	eher materialistische Orientierung
A	D	4		
C	B	5		
C	D	6		
B	A	7	1	eher postmaterialistische Orientierung
B	C	8		
D	A	9		
D	C	10		
B	D	11	0	reiner Postmaterialismus
D	B	12		

Einkommen kann denselben Indexwert für Schichtzugehörigkeit erhalten wie eine Person mit geringer Bildung und hohem Einkommen (zur Kritik vgl. z. B. Rohwer & Pötter, 2002, Kap. 6.3). Genauso ist es bei dem additiv zusammengesetzten **Index der Aufsatzqualität**: Schlechte Rechtschreibung kann durch guten Stil kompensiert werden und umgekehrt. Ein sehr bekanntes praktisches Beispiel für einen additiven Index ist der **Aktienindex Dow Jones** (Dow Jones Industrial Average DJIA), der als Durchschnittswert der Aktienkurse der 30 größten US-Unternehmen berechnet wird und ein Maß für die Entwicklung des US-amerikanischen Aktienmarktes darstellt. Der aus den Aktienkursen von 30 Unternehmen gebildete Index ist ein besserer Indikator für die Entwicklung des Aktienmarktes als der Aktienkurs eines einzelnen Unternehmens als Einzelindikator.

Ungewichteter multiplikativer Index

Wenn ein Index bestimmte Mindestausprägungen auf allen Indikatorvariablen voraussetzt, die sich wechselseitig **nicht kompensieren**, sollten die Teilindikatoren multiplikativ zu einem Gesamtindex verknüpft werden. Durch die multiplikative Verknüpfung erhält der zusammenfassende Index den Wert Null, wenn mindestens eine Indikatorvariable den Wert Null aufweist. Schnell et al. (2008, S. 172) nennen das folgende didaktisch vereinfachte Beispiel: Ein **Index zur Voraussage des Studienerfolgs** könnte sich multiplikativ aus den Indikatoren „Fleiß“ und „Begabung“ zusammensetzen. Sowohl völlig ohne Begabung als auch ohne jeden Fleiß scheint der Studienerfolg fraglich. Erhält nur einer der beiden Indikatoren den Wert Null, so ergibt sich auch für den Gesamtindex der Wert Null (kein Studienerfolg). Der **Human Development Index HDI** der Vereinten Nationen wird als ungewichteter multiplikativer Index aus drei Teilindizes für Lebenserwartung, Bildung und Bruttonationaleinkommen eines Lands berechnet und hat einen Wertebereich von 0 bis 1, wobei Deutschland im Jahr 2010 einen Indexwert von HDI = .88 erreichte – zum Vergleich: Norwegen: .93, China: .66, Simbabwe: .14.

Gewichteter additiver Index

Gewichtete additive Indizes ermöglichen eine differenzierte Behandlung der einzelnen Indikatoren. Über Techniken zur Bestimmung angemessener Gewichte informieren die folgenden Ausführungen. Ein Beispiel für einen gewichteten additiven Index ist der **Verbraucherpreisindex VPI**, der die Preise unterschiedlicher Warengruppen aus der Konsumwelt fortlaufend erfasst. Bei der Zusammenfassung des Index werden die einzelnen Güterarten im Warenkorb danach gewichtet, welcher Ausgabenanteil des durchschnittlichen Haushalts auf sie entfällt. Damit schlägt also z. B. eine Teuerung in einem selten gekauften Warenbereich weniger im Index zu Buche als eine Teuerung bei häufig gekauften Gütern oder Dienstleistungen (genauere Hinweise zu Berechnung des VPI finden sich auf der Website des statistischen Bundesamtes http://www.destatis.de).

Gewichteter multiplikativer Index

Ebenso wie ein additiver Index kann auch ein multiplikativer Index gewichtet werden, damit einzelne Indikatoren den Indexwert stärker oder schwächer als andere mitbestimmen. Der **chemische Gewässergüteindex** nach Erwin Bach (1980) wird z. B. als gewichteter multiplikativer Index aus acht Indikatoren der Wasserqualität gebildet (z. B. Temperatur, Sauerstoffsättigung, ph-Wert etc.). Die Gewichte entsprechen der ökologischen Bedeutung der Indikatoren. Wenn nur ein einziger Wert kritische Grenzwerte über- oder unterschreitet kann durch die multiplikative Verrechnung sichergestellt werden, dass die Wassergüte insgesamt einen geringen Indexwert erhält.

8.7.4 Gewichtung der Indikatoren

Entscheidet man sich für einen gewichteten Index, stellt sich die Frage, wie die Gewichtungsfaktoren zu bestimmen sind. Will man z. B. das Merkmal „Rechtschreibleistung“ operationalisieren, könnte sich herausstellen, dass die schlichte Addition von Schreibfehlern ein problematischer Indikator dieses Merkmals ist. Flüchtigkeitsfehler

z. B. könnten nachsichtiger behandelt werden, während Fehler, die grundlegende Rechtschreibregeln verletzen, härter zu „bestrafen“ wären. Wie jedoch soll ermittelt werden, wie gravierend verschiedene Rechtschreibfehlerarten sind, bzw. allgemein: Mit welchem Gewicht sollen die beobachteten Indikatorvariablen in die Indexberechnung eingehen? Man unterscheidet einerseits die Gewichtsbestimmung durch Expertenrating (normative Gewichtsbestimmung) und andererseits die empirisch-analytische Gewichtsbestimmung mittels Faktorenanalyse, multipler Regressionsrechnung sowie Datenmodellierung.

- **Gewichtsbestimmung durch Expertenrating.** Eine einfache Lösung des Gewichtungsproblems besteht darin, die Gewichtung der Indikatoren durch Experten vornehmen zu lassen (**normative Indexbildung**). Im Rechtschreibbeispiel wäre also das Wissen erfahrener Pädagoginnen und Pädagogen zu nutzen, um die relative Bedeutung verschiedener Rechtschreibfehler einzuschätzen (sog. Expertenrating). Zur Sicherung der Objektivität der Vorgehensweise ist es allerdings ratsam, die Gewichtung von mehreren unabhängig urteilenden Fachleuten vornehmen zu lassen. Erst wenn die Expertenurteile hinreichend gut übereinstimmen bilden die durchschnittlichen Bewertungen eine akzeptable Grundlage für eine gewichtete Indexbildung (zur Überprüfung der Urteilerübereinstimmung ▶ Abschn. 10.1.3 „Beobachterübereinstimmung“ und ▶ Abschn. 10.6.3 „Quantitative Inhaltsanalyse“, ◘ Tab. 10.34 zur Kodiererübereinstimmung).
- **Empirisch-analytische Gewichtsbestimmung.** Bei quantitativen Indikatorvariablen besteht die Möglichkeit, die relative Bedeutung der einzelnen Indikatoren empirisch mithilfe geeigneter statistischer Analysetechniken zu bestimmen (Perloff & Persons, 1988). Wenn z. B. das Merkmal „Schmerz“ durch unterschiedliche Ausprägungen von Indikatorvariablen wie „beißend“, „brennend“, „pochend“, „dumpf“ etc. charakterisiert wird, könnte die Frage interessieren, wie stark bzw. mit welchem Gewicht diese Empfindungsvarianten an typischen Schmerzbildern (Migräne, Muskelzerrung, Magenverstimmung etc.) beteiligt sind. Zur Beantwortung dieser Frage wäre die sog. Faktorenanalyse ein geeignetes Verfahren (Bortz & Schuster, 2010, Kap. 23). Dabei kann die **exploratorische Faktorenanalyse** zum Einsatz kommen. Mit ihrer Hilfe wird bei Eindimensionalität der Skala ein Ein-Faktor-Modell, bei Mehrdimensionalität ein Mehr-Faktor-Modell gebildet, wobei Enge und Richtung des Zusammenhangs der einzelnen Items mit den Faktoren durch **Faktorladungen im Wertebereich von –1 bis +1** quantifiziert werden. Diese Faktorladungen können als Gewichtungsfaktoren für die Indexbildung herangezogen werden (für ein Beispiel einer faktorenanalytischen Skalenauswertung mit entsprechenden Faktorladungen ◘ Tab. 8.21; zu explorativen multivariaten Auswertungstechniken inkl. explotorischer Fakorenanalyse ▶ Abschn. 12.3.1 „Explorative multivariate Auswertungstechniken“). Statistisch anspruchsvoller, aber genauer ist die Berechnung von Faktorladungen im Kontext der **konfirmatorischen Faktorenanalyse** und anderer Ansätze, um Messmodelle mit Strukturgleichungsmodellen zu prüfen (▶ Kap. 17).

  Eine weitere Technik zur empirisch-analytischen Gewichtsbestimmung für die Indexbildung stellt die **multiple Regressionsrechnung** dar (zur Datenanalyse bei multivariaten Zusammenhangshypothesen inkl. der multiplen Regressionsanalyse ▶ Abschn. 12.5.2 „Multivariate Zusammenhangshypothesen“). Hierbei wird ermittelt, welche Bedeutung verschiedene Indikatorvariablen für ein bestimmtes Kriterium haben (z. B.: Wie wichtig sind die letzte Mathematiknote, die Vorbereitungszeit, die Leistungsmotivation und das Konzentrationsvermögen für die Punktzahl in einer Statistikklausur?).

  Bei der empirisch-analytischen Gewichtsbestimmung mittels Faktorenanalyse oder multipler Regression sind – vor allem bei kleineren Stichproben – ungenaue bzw. instabile Gewichtsschätzungen zu erwarten. Große und möglichst repräsentative Stichproben, die eine sog. Kreuzvalidierung der Gewichte ermöglichen (Bortz & Schuster, 2010, S. 356), sind deshalb bei dieser Art der Gewichtsbestimmung für die Indexbildung von besonderem Vorteil.

8.7.5 Index als standardisierter Wert

Der Begriff „Index“ wird noch in einer zweiten Bedeutung verwendet, nämlich wenn es darum geht, quantitative Angaben zu standardisieren, etwa indem man sie zu einer festgelegten Größe in Beziehung setzt. Statt von „Index“ spricht man in dieser Bedeutung auch von „Rate“. So kann man mit der absoluten Zahl von Straftaten inhaltlich zunächst wenig anfangen, deswegen wird sie an Vergleichsgrößen (z. B. pro 1 000 Einwohner) relativiert. Die resultierenden **Kriminalitätsindizes** bzw. **Kriminalitätsraten** lassen sich dann z. B. zwischen verschiedenen Ländern vergleichen. Die Demografie arbeitet mit entsprechenden Verfahren, wenn es um Fruchtbarkeitsraten oder Scheidungsraten geht. Die Sicherheit von Verhütungsmitteln wird über den **Pearl-Index** angegeben, der die Zahl der trotz Verhütung entstandenen Schwangerschaften auf 100 sexuell aktive Frauen bezieht, die das

fragliche Kontrazeptivum ein Jahr nutzen. Der **Body-Mass-Index (BMI)** berechnet sich aus dem Körpergewicht dividiert durch das Quadrat der Körpergröße. Entsprechende „Raten" oder „Indizes" als standardisierte Werte haben Kardinalskalenniveau.

Index als standardisierter Wert – Neben dem Index als Messverfahren, das unterschiedliche formative Indikatoren zur Messung eines komplexen Merkmals verknüpft, existiert der Index („index") bzw. die „Rate" als ein standardisierter Kennwert für ein einfaches Merkmal. Er wird gebildet, indem man die interessierende Indikatorvariable (z. B. Häufigkeit von Straftaten, Ehescheidungen, Geburten, Körpergewicht, Schwangerschaften trotz Verhütung) mit einer festgelegten Größe in Beziehung setzt, wodurch sich die inhaltliche Aussagekraft und Interpretierbarkeit der Variable erhöht.

Lernquiz

1. Was versteht man unter Nominaldefinition und Nominalskala?
2. Was sind eine Ratingskala, eine Rangskala und eine psychometrische Skala?
3. Wie wird in der qualitativen und quantitativen Forschung jeweils zwischen theoretischen Konstrukten einerseits und empirischen Daten andererseits eine Verknüpfung hergestellt?
4. Was ist bei der Messung soziodemografischer Variablen zu beachten?
5. Charakterisieren Sie folgende Ratingskala: „stimmt gar nicht – stimmt wenig – stimmt teils-teils – stimmt ziemlich – stimmt völlig".
6. Was unterscheidet die Operationalisierung von „Depression" mittels Einzelindikator und eindimensionaler psychometrischer Skala?
7. Eine Likert-Skala zur Messung der Studienmotivation enthält 12 Items, die auf 5-stufigen Ratingskalen (1 = stimmt gar nicht bis 5 = stimmt völlig) zu beantworten sind. In welchem Wertebereich bewegen sich Summen- und Durchschnittsscore der Skala?
8. Eine Untersuchungsperson hat auf der Thurstone-Skala zur Kontaktsituation in Wohnsiedlungen (► Abschn. 8.6.3; ► Menschliche Kontakte in Siedlungen) die Items 1, 2 und 4 bejaht und die anderen Items verneint. Welchen Skalenwert erhält die Person und wie ist dieser inhaltlich zu interpretieren?
9. Eine Untersuchungsperson hat auf der Guttman-Skala zur vorehelichen Permissivität (► Abschn. 8.6.2) das Item a bejaht und die anderen Items verneint. Welchen Skalenwert erhält die Person und wie ist dieser inhaltlich zu interpretieren?
10. Auf welchem Skalenniveau sollten Daten erhoben werden? Was ist zu tun, wenn Daten in einem ungeeigneten Skalenniveau vorliegen?

Übungsaufgaben

1. Operationalisieren Sie das soziodemografische Merkmal „Bildungsstand"!
2. Führen Sie eine dimensionale Analyse des Konzeptes „Handykompetenz von Kindern" durch!
3. Suchen Sie ein psychometrisches Messinstrument für Motivation heraus, z. B. aus dem Elektronischen Testarchiv des ZPID (http://www.zpid.de).

Weiterführende Literatur

- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte Aufl.). München: Pearson Studium.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.
- Peterson, R. A. (1999). *Constructing Effective Questionnaires.* Thousand Oaks, California: SAGE Publications, Inc.

Literatur

Abdel-Khalek, A. M. (2006). Measuring happiness with single-item scale. *Social Behavior and Personality*, *34*(2), 139–150.

Ahearn, E. P. (1997). The use of visual analog scales in mood disorders: A critical review. *Journal of Psychiatric Research*, *31*(5), 569–579.

Aiken, L. R. (1985a). Evaluating ratings on bidirectional scales. *Educational and Psychological Measurement*, *45*, 195–202.

Aiken, L. R. (1985b). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*, 131–142.

Aiken, L. R. (1987). Formulars for equating ratings on different scales. *Educational and Psychological Measurement*, *47*, 51–54.

Aiken, L. R. (1996). *Rating scales and checklists: Evaluating behavior, personality, and attitudes.* Oxford: Wiley.

Aiken, L. R. (1997). *Psychological testing and assessment.* (9. Aufl.). Boston: Allyn & Bacon.

Alliger, G. M. & Williams, K. J. (1989). Confounding among measures of leniency and halo. *Educational and Psychological Measurement*, *49*, 1–10.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., et al. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151–173.

Athey, T. R. & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, *72*(4), 567–572.

Attneave, F. (1949). A method of graded dichotomies for the scaling of judgments. *Psychological Review*, *56*(6), 334–340.

Ayalon, L., Goldfracht, M., & Bech, P. (2010). „Do you think you suffer from depression?" Reevaluating the use of a single item question for the screening of depression in older primary care patients. *International Journal of Geriatric Psychiatry*, *25*(5), 497–502.

Bach, E. (1980). *Ein chemischer Index zur Überwachung der Wasserqualität von Fließgewässern.* (24. Aufl.). Frankfurt/Main: DGM.

Baer, L. & Blais, M. A. (Eds.). (2009). *Handbook of clinical rating scales and assessment in psychiatry and mental health (current clinical psychiatry)*. New York: Humana Press.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurement vs. strong statistics: An empirical critique of S. S. Stevens proscriptions of statistics. *Educational and Psychological Measurement*, *26*, 291–309.

Bannister, B. D., Kinicki, A. J., Denisi, A. S., & Horn, P. W. (1987). A new method for the statistical control of rating error in performance ratings. *Educational and Psychological Measurement*, *47*, 583–596.

Barr, M. A. & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods*, *6*(1), 15–43.

Bearden, W. O., Netemeyer, R. G., & Haws, K. L. (2011). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research* (vol. 3). Los Angeles: Sage.

Beaton, A. E. & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational and Behavioral Statistics*, *17*(2), 191–204.

Bergkvist, L. & Rossiter, J. R. (2007). The predictive validity of multiple-item vs. single-item measures of the same constructs. *Journal of Marketing Research*, *44*, 175–184.

Bernardin, H. J. (1977). Behavioral expectation scales vs. summated ratings: A fairer comparison. *Journal of Applied Psychology*, *62*, 422–427.

Bernardin, H. J. & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). *Journal of Applied Psychology*, *66*(4), 458–463.

Bernardin, H. J. & Walter, C. S. (1977). Effects of rater training and diary-helping on psychometric error in ratings. *Journal of Applied Psychology*, *62*, 64–69.

Bierhoff, H. W. (1996). Neue Erhebungsmethoden. In E. Erdfelder, R. Mausfeld & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 59–70). Weinheim: Beltz.

Bintig, A. (1980). The efficiency of various estimations of reliability of rating-scales. *Educational and Psychological Measurement*, *40*, 619–644.

Blunt, A. (1983). Development of a Thurstone scale for measuring attitudes toward adult education. *Adult Education Quarterly*, *34*(1), 16–28.

Böckenholt, U. (2001). Hierarchical modelling of paired comparison data. *Psychological Methods*, *6*, 49–66.

Böckenholt, U. (2004). Comparative judgements as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, *9*, 453–465.

Bongers, D. & Rehm, G. (1973). Kontaktwunsch und Kontaktwirklichkeit von Bewohnern einer Siedlung. Unveröffentlichte Diplomarbeit. Universität Bonn.

Borg, I., Müller, M., & Staufenbiel, T. (1990). Ein empirischer Vergleich von fünf Standard-Verfahren zur eindimensionalen Skalierung. *Archiv für Psychologie*, *142*, 25–33.

Borman, W. C. (1975). Effects of instructions to avoid error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, *60*, 556–560.

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (Ed.), *Performance Assessment: Methods and Applications.* (pp. 100–120). Baltimore: Johns Hopkins University Press.

Bortz, J. & Lienert, G. A. (2008). *Kurzgefaßte Statistik für die klinische Forschung.* (3. Aufl.). Berlin: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2000). *Verteilungsfreie Methoden in der Biostatistik.* (2. Aufl.). Heidelberg: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik.* (3. Aufl.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler (Lehrbuch mit Online-Materialien).* (7. Aufl.). Berlin: Springer.

Bradley, R. A. & Terry, M. E. (1952). The rank analysis of incomplete block designs. I: The method of paired comparison. *Biometrika*, *39*, 324–345.

Brandt, L. W. (1978). Measuring of a measurement: Empirical investigation of the semantic differential. *Probleme und Ergebnisse der Psychologie*, *66*, 71–74.

Breckler, S. J. (1994). A comparison of numerical indexes for measuring attitude ambivalence. *Educational and Psychological Measurement*, *54*(2), 350–365.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte Aufl.). München: Pearson Studium.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, *57*(1), 15–22.

Carbonell, L., Sendra, J. M., Bayarri, S., Izquierdo, L., & Tárrega, A. (2008). Thurstonian scales obtained by transformation of beta distributions. *Food Quality and Preference*, *19*(4), 407–411.

Chatterjee, B. B. & Puhan, B. N. (1980). A Thurstone scale for measuring attitude towards sex. *Indian Psychological Review*, *19*(3), 1–8.

Chignell, M. H. & Pattey, B. W. (1987). Unidimensional scaling with efficient ranking methods. *Psychological Bulletin*, *101*, 304–311.

Clark, J. A. (1977). A method of scaling with incomplete pair-comparison data. *Educational and Psychological Measurement*, *37*, 603–311.

Cogliser, C. C. & Schriesheim, C. A. (1994). Development and application of a new approach to testing the bipolarity of semantic differential. *Educational and Psychological Measurement*, *54*(3), 594.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.

Conrad, E. & Maul, T. (1981). *Introduction to experimental psychology*. New York: Wiley.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology*. Englewood Cliffs: Prentice Hall.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales. A web experiment. *Social Science Computer Review*, *24*(2), 227–245.

Crawshaw, L. (2009). Workplace bullying? Mobbing? Harassment? Distraction by a thousand definitions. *Consulting Psychology Journal: Practice and Research*, *61*(3), 263–267.

Cronkhite, G. (1976). Effects of rater-concept-scale interactions and use of different factoring procedures upon evaluative factor structure. *Human Communication Research*, *2*, 316–329.

Dalbert, C. (1992). Subjektives Wohlbefinden junger Erwachsener: Theoretische und empirische Analysen der Struktur und Stabilität. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *13*, 207–220.

David, H. A. (1963). *The method of paried comparison*. London: Griffin.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34*(4), 481–489.

De Cotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance*, *19*, 247–266.

De Cotiis, T. A. (1978). A critique and suggested revision of behaviorally anchored rating scales developmental procedures. *Educational and Psychological Measurement*, *38*, 681–690.

Diamantopoulos, A. (2005). The C-OAR-SE procedure for scale development in marketing: A comment. *International Journal of Research in Marketing*, *22*, 1–9.

Diamantopoulos, A. & Winklhofer, H. M. (2001). Index construction with formative indicators: an alternative to scale development. *Journal of Marketing Research*, *38*(269–277).

Diefenbacher, H. & Zieschank, R. (2008). Wohlfahrtsmessung in Deutschland. Ein Vorschlag für einen neuen Wohlfahrtsindex. Statusbericht zum Forschungsprojekt FKZ 3707 11 101/01. Zeitreihenrechnung zu Wohlfahrtsindikatoren. Abgerufen 22. Februar, 2012, unter http://www.beyond-gdp.eu/download/BMU_UBA_Endbericht_v20_endg.pdf.

Doll, J. (1988). Kognition und Präferenz: Die Bedeutung des Halo-Effektes für multiattributive Einstellungsmodelle. *Zeitschrift für Sozialpsychologie*, *19*, 41–52.

Döring, N. (2005). Für Evaluation und gegen Evaluitis. Warum und wie Lehrevaluation an deutschen Hochschulen verbessert werden sollte. In B. Berendt, H.-P. Voss, & J. Wildt (Hrsg.), *Neues Handbuch Hochschullehre* (S. 1–22). Berlin: Raabe.

Döring, N. (2013). Zur Operationalisierung von Geschlecht im Fragebogen: Probleme und Lösungsansätze aus Sicht von Mess-, Umfrage-, Gender- und Queer-Theorie. *Gender*, *2*, 94–113.

Dunn-Rankin, P., Knezek, G. A., Wallace, S., & Zhang, S. (2004). *Scaling Methods*. Mahwah: Erlbaum.

Edwards, A. L. & Kilpatrick, F. P. (1948). A technique for the construction of attitude scales. *Journal of Applied Psychology*, *32*, 374–384.

Eiser, J. R. & Ströbe, W. (1972). *Categorisation and social judgement*. New York: Academic Press.

EKD. (2013). Evangelische Kirche in Deutschland. Zahlen und Fakten zum kirchlichen Leben. Abgerufen 21. August, 2013, unter http://www.ekd.de/download/zahlen_und_fakten_2013.pdf.

Evans, R. H. (1980). The upgraded semantic differential: a further test. *Journal of the Market Research Society*, *22*(2), 143–147.

Ferguson, C. J. & Rueda, S. M. (2010). The Hitmann study: Violent video game exposure effects on aggressive behavior, hostile feelings, and depression. *European Psychologist*, *15*(2), 99–108.

Finn, A. & Kayande, U. (2005). How fine is C-OAR-SE? A generalizability theory perspective on Rossiter's procedure. *International Journal of Research in Marketing*, *22*, 11–21.

Finstuen, K. (1977). Use of Osgood's semantic differential. *Psychological Reports*, *41*, 1219–1222.

Flade, A. (1978). Die Beurteilung umweltpsychologischer Konzepte mit einem konzeptspezifischen und einem universellen semantischen Differential. *Zeitschrift für experimentelle und angewandte Psychologie*, *25*, 367–378.

Flynn, L. R. (1993). Do standard scales work in older samples? *Marketing Letters*, *4*(2), 127–137.

Frank, D. & Schlund, W. (2000). Eine neue Lösung des alten Skalenproblems. *Planung und Analyse*, *6*, 56 ff.

Friedman, B. A. & Cornelius III, E. T. (1976). Effect of rater participation on scale construction on the psychometric characteristics of two ratingscale formats. *Journal of Applied Psychology*, *61*, 210–216.

Friedman, H. H., Friedman, L. W., & Gluck, B. (1988). The effects of scale-checking styles on responses to a semantic differential scale. *Journal of the Market Research Society*, *30*(4), 477–481.

Gaito, J. (1980). Measurement scales and statistics. Resurgence of an old misconception. *Psychological Bulletin*, *87*, 564–567.

Galovski, T. E., Malta, L. S., & Blanchard, E. B. (2006). *Road rage: Assessment and treatment of the angry, aggressive driver*. Washington: American Psychological Association.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item vs. multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement*, *58*, 898–915.

Garland, R. (1990). A comparison of three forms of the semantic differential. *Marketing Bulletin*, *1*, 19.

Garner, W. R. & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review*, *58*(6), 446–459.

Gescheider, G. A. (1988). Psychophysical scaling. *Annual Revue of Psychology*, *33*, 169–200.

Glaser, B. G. (2002). Conceptualization: On theory and theorizing using grounded theory. *International Journal of Qualitative Methods*, *1*(2), 3rd Article. Retrieved August 29, 2011, from http://www.ualberta.ca/~iiqm/backissues/1_2Final/pdf/glaser.pdf.

Gluth, S., Ebner, N. C., & Schmiedek, F. (2010). Attitudes toward younger and older adults: The German aging semantic differential. *International Journal of Behavioral Development*, *34*(2), 147–158.

Gonzales, E., Tan, J., & Morrow-Howell, N. (2010). Assessment of the refined Aging Semantic Differential: Recommendations for enhancing validity. *Journal of Gerontological Social Work*, *53*(4), 304–318.

Goodstadt, M. S. & Magid, S. (1977). When Thurstone and Likert agree: A confounding of methodologies. *Educational and Psychological Measurement*, *37*(4), 811–818.

Granberg-Rademacker, J. S. (2010). An algorithm for converting ordinal scale measurement data to interval/ratio scale. *Educational and Psychological Measurement*, *70*(1), 74–90.

Green, S. B., Sauser, W. I., Fagg, J. N., & Champion, C. H. (1981). Shortcut methods for deriving behaviorally anchored rating scales. *Educational and Psychological Measurement*, *41*(3), 761–775.

Greenberg, J. (1990). Organizational justice: Yesterday, today, and tomorrow. *Journal of Management*, *16*, 399–432.

Guilford, J. P. (1938). The computation of psychological values from judgements in absolute categories. *Journal of Experimental Psychology*, *22*(1), 32–42.

Guttman, L. (1950). The basis of scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction. Studies in social psychology in World War II* (vol. 4, pp. 60–90). Princeton: Princeton University Press.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*(3), 445–492.

Hauenstein, N. M. A., Brown, R. D., & Sinclair, A. L. (2010). BARS and those mysterious, missing middle anchors. *Journal of Business and Psychology*, *25*(4), 663–672.

Helmholtz, H. (1887). *Zur Geschichte des Princips der kleinsten Action*. Berlin: Reichsdruckerei.

Helmholtz, H. (1959). *Die Tatsachen in der Wahrnehmung. Zählen und Messen erkenntnistheoretisch betrachtet*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Henss, R. (1989). Zur Vergleichbarkeit von Ratingskalen unterschiedlicher Kategorienzahl. *Psychologische Beiträge*, *31*, 264–284.

Himmelfarb, S. (1993). The measurement of attitudes. In A. H. Eagly & S. Chaiken (Eds.), *Psychology of attitudes* (pp. 23–88). Belmont: Thomson/Wadsworth.

Hofacker, C. F. (1984). Categorical judgment scaling with ordinal assumptions. *Multivariate Behavioral Research*, *19*(1), 91–106.

Hofstätter, P. R. (1957). *Psychologie*. Frankfurt/Main: Fischer.

Hofstätter, P. R. (1963). *Einführung in die Sozialpsychologie*. Stuttgart: Kröner.

Hofstätter, P. R. (1977). *Persönlichkeitsforschung*. Stuttgart: Kröner.

Horowitz, L. M., Inouye, D., & Seigelmann, E. Y. (1979). On avaraging judges' rating to increase their correlation with an external criterion. *Journal of Consulting and Clinical Psychology*, *47*, 453–458.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*(1), 64–86.

Hoyt, W. T. (2002). Bias in participant ratings of psychotherapy process: An initial generalizability study. *Journal of Counseling Psychology*, *49*(1), 35–46.

Hoyt, W. T. & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424.

Hull, R. B. & Buhyoff, G. J. (1981). On the „Law of Comparative Judgement": Scaling with intransitive observers and multidimensional stimuli. *Educational and Psychological Measurement*, *41*, 1083–1089.

Igou, E. R., Bless, H., & Schwarz, N. (2002). Making sense of standardized survey questions: The influence of reference periods and their repetition. *Communication Monographs*, *69*(2), 179–187.

Inglehart, R. (1977). *The silent revolution: Changing values and political styles among western publics*. Princeton: Princeton University Press.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic and political change in 43 societies*. Princeton: Princeton University Press.

Jäger, R. (1998). Konstruktion einer Ratingskala mit Smilies als symbolische Marken. Institut für Psychologie, Technische Universität Berlin.

Jäger, R. S. & Petermann, F. (1992). *Psychologische Diagnostik.* (2. Aufl.). Weinheim: Psychologie Verlags Union.

Johnson, D.–M. & Vidulich, R. N. (1956). Experimental Manipulation of the Halo-Effect. *Journal of Applied Psychology*, *40*, 130–134.

Jones, L. V. (1959). Some Invariant Findings under the Method of Successive Intervalls. *American Journal of Psychology*, *72*, 210–220.

Jones, L. V. & Thurstone, L. L. (1955). The psychophysics of semantics: An empirical investigation. *Journal of Applied Psychology*, *39*, 31–36.

Kahneman, D. & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

Kane, R. B. (1971). Minimizing order effects in the semantic differential. *Educational and Psychological Measurement*, *31*(137–144).

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, *1995*, 1036–1051.

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, *77*(5), 361–372.

Keller, J. & Wagner-Steh, K. (2005). A Guttman scale for empirical prediction of level of domestic violence. *Journal of Forensic Psychology Practice*, *5*(4), 37–48.

Kelley, H. H., Hovland, C. J., Schwartz, M., & Abelson, R. P. (1955). The influence of judges attitudes in three modes of attitude scaling. *Journal of Social Psychology*, *42*, 147–158.

Kendall, M. G. (1955). Further contributions to the theory of paired comparison. *Biometrics*, *11*, 43–62.

Kessler, J. (2009). Der Mythos vom globalen Dorf. Zur räumlichen Differenzierung von Globalisierungsprozessen. In J. Kessler & C. Steiner (Hrsg.), *Facetten der Globalisierung: Zwischen Ökonomie, Politik und Kultur* (S. 28–79). Wiesbaden: VS Verlag.

King, B. M., Rosopa, P. J., & Minium, E. W. (2010). *Statistical Reasoning in the Behavioral Sciences* (6. Aufl.). Hoboken: John Wiley & Sons.

Kingstrom, P. O. & Bass, A. R. (1981). A Critical Analysis of Studies Comparing Behaviorally Anchored Rating Scales (BARS) and Other Rating Formats. *Personnel Psychology*, *34*(2), 263–289.

Kinicki, A. J. & Bannister, B. D. (1988). A test of the measurement assumptions underlying behaviorally anchored rating scales. *Educational and Psychological Measurement*, *48*(1), 17–27.

Kinicki, A. J., Bannister, B. D., Hom, P. W., & Denisi, A. S. (1985). Behaviorally anchored rating scales vs. summated rating scales: Psychometric properties and susceptibility to rating bias. *Educational & Psychological Measurement*, *45*(3), 535–549.

Klauer, K. C. (1989). Untersuchungen zur Robustheit von Zuschreibungs-mal-Bewertungsmodellen: Die Bedeutung von Halo-Effekten und Dominanz. *Zeitschrift für Sozialpsychologie*, *20*, 14–26.

Klauer, K. C. & Schmeling, A. (1990). Sind Halo-Fehler Flüchtigkeitsfehler? *Zeitschrift für experimentelle und angewandte Psychologie*, *37*, 594–607.

Knezek, G., Wallace, S., & Dunn–Rankin, P. (1998). Accuracy of Kendall's chi-square. Approximation to circular triad distributions. *Psychometrica*, *63*, 23–34.

Korman, A. K. (1971). *Industrial and organizational psychology*. Englewood Cliffs: Prentice Hall.

Krabbe, P. F. M. (2008). Thurstone scaling as a measurement method to quantify subjective health outcomes. *Medical Care*, *46*(4), 357–365.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (2006a). *Foundations of measurement volume II: Geometrical, threshold, and probabilistic representations*. Mineola: Dover Publications.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (2006b). *Foundations of measurements volume I: Additive and polynomial representations*. Mineola: Dover Publications.

Krebs, D. & Hoffmeyer–Zlotnik, J. H. P. (2009). *Bipolar vs. unipolar scale format in fully vs. endpoint verbalized scale*. Paper presented at the Cognition in Survey Research, 3rd Conference of the European Survey Research Association. Warschau, 29th June – 3rd July, 2009.

Kromrey, H. (2000a). *Empirische Sozialforschung: Modelle und Methoden der standardisierten Datenerhebung und Datenausweitung: Modelle und Methoden der Datenerhebung und Datenauswertung* (12. Aufl.). Stuttgart: UTB.

Kromrey, H. (2000b). Qualität und Evaluation im System Hochschule. In R. Stockmann (Hrsg.), *Evaluationsforschung* (S. 233–258). Opladen: Leske & Budrich.

Krosnick, J. A. & Fabrigar, L. R. (2006). *Designing great questionnaires: Insights from psychology*. New York: Oxford University Press.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating error in the observation of behavior. *Journal of Applied Psychology*, *60*, 550–555.

Lei, M. & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*(1), 1–27.

Leonhart, R. (2009). *Lehrbuch Statistik. Einstieg und Vertiefung* (2. Aufl.). Bern: Huber.

Li, F., Wang, E., & Zhang, F. (2002). The multitrait-multirater approach to analyzing rating biases. *Acta Psychologica Sinica*, *34*(1), 89–96.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Lindemann, D. F. & Brigham, T. A. (2003). A Guttman scale for assessing condom use skills among college students. *AIDS and Behavior*, *7*(1), 23–27.

Lissitz, R. W. & Green, S. B. (1975). Effect of number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, *60*, 10–13.

Lohaus, D. (1997). Reihenfolgeeffekte in der Eindrucksbildung. Eine differenzierte Untersuchung verschiedener Meßzeiträume. *Zeitschrift für Sozialpsychologie*, *28*, 298–308.

Lord, F. M. (1953). On the statistical treatmen of football numbers. *American Psychologist*, *8*, 750–751.

Lozano, L. M., García–Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Science*, *4*(2), 73–79.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Lütters, H. (2008). Serious fun in market research: The sniper scale. *Marketing Review St. Gallen*, *25*(6), 17–22.

Maier, J., Maier, M., Maurer, M., Reinemann, C., & Meyer, V. (Eds.). (2009). *Real-time response measurement in the social sciences: Methodological perspectives and applications*. Frankfurt/Main: Lang.

Maier, J., Maurer, M., Reinemann, C., & Faas, T. (2006). Reliability and validity of real-time response measurement: A comparison of two studies of a televised debate in Germany. *International Journal of Public Opinion Research*, *19*(1), 53–73.

Mann, I. T., Phillips, J. L., & Thompson, E. G. (1979). An examination of methodological issues relevant to the use and interpretation of the semantic differential. *Applied Psychological Measurement*, *3*(2), 213–229.

Marcus, B. & Schuler, H. (2001). Leistungsbeurteilung. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 397–433). Stuttgart: Schäffer-Poeschel.

Mari, L. (2005). The problem of foundations of measurement. *Measurement*, *38*(4), 259–266.

Matell, M. S. & Jacoby, J. (1971). Is there an optimal number for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657–674.

Maxwell, S. E. & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*(1), 181–190.

McCarty, J. A. & Shrum, L. J. (2000). The measurement of personal values in survey research. A test of alternative rating procedures. *Public Opinion Quarterly*, *64*, 271–298.

McCormack, B., Boldy, D., Lewin, G., & McCormack, G. R. (2011). Screening for depression among older adults referred to home care services: A single-item depression screener vs. the geriatric depression scale. *Home Health Care Management and Practice*, *23*(1), 13–19.

McCormack, H. M., Horne, D. J., & Sheather, S. (1988). Clinical applications of visual analogue scales: a critical review. *Psychological Medicine*, *18*, 1007–1019.

Michell, J. (1986). Measurement scales and statistics. A clash of paradigms. *Psychological Bulletin, 100.* 398–407.

Michell, J. (2005). The logic of measurement: A realistic overview. *Measurement*, *38*(4), 285–294.

Mosier, C. J. (1941). A psychometric study of meaning. *Journal of Social Psychology*, *13*, 123–140.

Mount, M. K., Sytsma, M. R., Hazucha, J. F., & Holt, K. E. (1997). Rater-ratee race effects in developmental performance rating of managers. *Personnel Psychology*, *50*(1), 51–69.

Murakami, T. & Kroonenberg, P. M. (2003). Three-mode models and individual differences in semantic differential data. *Multivariate Behavioral Research*, *38*(2), 247–283.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*, 77–86.

Neumann, W. L. (2003). *Social research methods. Qualitative and quantitative approaches* (5th edn.). Bosten: Pearson.

Newcomb, T. (1931). An experimant designed to test the validity of a rating technique. *Journal of Educational Psychology*, *22*(4), 279–289.

Newstead, S. E. & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational and Psychological Measurement*, *49*(1), 33–43.

Niederée, R. & Mausfeld, R. (1996a). Das Bedeutsamkeitsproblem in der Statistik. In E. Erdfelder, R. Mausfeld, & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 399–410). Weinheim: Psychologie Verlags Union.

Niederée, R. & Mausfeld, R. (1996b). Skalenniveau, Invarianz und „Bedeutsamkeit". In E. Erdfelder, R. Mausfeld, & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 385–398). Weinheim: Psychologie Verlags Union.

Noll, H.-H. (2002). Globale Wohlfahrtsmaße als Instrumente der Wohlfahrtsmessung und Sozialberichterstattung: Funktionen, Ansätze und Probleme. In W. Glatzer, R. Habich, & K. U. Mayer (Hrsg.), *Sozialer Wandel und Gesellschaftliche Dauerbeobachtung. Festschrift für Wolfgang Zapf* (S. 317–336). Opladen: Leske & Budrich.

North, K. & Reinhardt, K. (2011). *Kompetenzmanagement in der Praxis: Mitarbeiterkompetenzen systematisch identifizieren, nutzen und entwickeln* (2. Aufl.). Wiesbaden: Gabler.

Ofir, C., Reddy, S. K., & Bechtel, G. G. (1987). Are semantic response scales equivalent? *Multivariate Behavioral Research*, *22*(1), 21.

Orpinas, P. & Horne, A. M. (2006). Bullies and victims: A challenge for schools. In J. R. Lutzker (Ed.), *Preventing violence: Research and evidence-based intervention strategies* (pp. 147–165). Washington: American Psychological Association.

Orth, B. (1983). Grundlagen des Messens. In H. Feger & J. Bredenkamp (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich B, Serie I Forschungsmethoden der Psychologie, Bd. 3: Messen und Testen* (S. 136–180). Göttingen: Hogrefe.

Osgood, C. E., Suci, G. J., & Tannenbaum, D. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Parducci, A. (1963). Range-frequency compromise in judgement. *Psychological Monographs*, *77*(2), 1–29.

Parducci, A. (1965). Category-judgement: a range-frequency model. *Psychological Review*, *72*, 407–418.

Pepels, W. (2007). *Market Intelligence: Moderne Marktforschung für Praktiker: Auswahlverfahren, Datenerhebung, Datenauswertung, Praxisanwendung, Marktprognose.* Düsseldorf: Publics Publishing.

Perloff, J. M. & Persons, J. B. (1988). Biases resulting from the use of indexes: An application to attributional style and depression. *Psychological Bulletin*, *103*(1), 95–104.

Peterson, R. A. (1999). *Constructing effective questionnaires.* Thousand Oaks: Sage.

Potosky, D. & Bobko, P. (1998). The Computer Understanding and Experience Scale: A Self-report measure of computer experience. *Computers in Human Behavior*, *14*(2), 337–348.

Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15.

Rambo, W. W. (1963). The distribution of successive interval judgements of attitude statements: A note. *Journal of Social Psychology*, *60*, 251–254.

Ramírez, J. M. & Andreu, J. M. (2009). The main sympthoms of the AHA-syndrome: Relationships between anger, hostility and agression on a normal population. In S. Bhave & S. Saini (Eds.), *The AHA-syndrome and cardiovascular diseases 2009* (pp. 16–29). New Delhi: Anamaya.

Rasmussen, J. L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin*, *105*(1), 167–170.

Reinemann, C., Maier, J., Faas, T., & Maurer, M. (2005). Reliabilität und Validität von RTR-Messungen. Ein Vergleich zweier Studien zur zweiten Fernsehdebatte im Bundestagswahlkampf 2002. *Publizistik*, *20*, 56–73.

Reiss, I. L. (1964). The scaling of premarital sexual permissiveness. *Marriage Family*, *26*, 188–198.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational & Psychological Measurement*, *59*, 211–233.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*(2), 151–161.

Rohrmann, B. (1978). Empirische Studie zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, *9*, 222–245.

Rohrmann, B. (2007). Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Project Report. Retrieved 22.02.2012, from http://www.rohrmannresearch.net/pdfs/rohrmann-vqs-report.pdf

Rohwer, G. & Pötter, U. (2002). *Methoden sozialwissenschaftlicher Datenkonstruktion*. Weinheim: Juventa.

Roskam, E. E. (1996). Latent-trait Modelle. In E. Erdfelder, R. Mausfeld, & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 431–458). Weinheim: Psychologie Verlags Union.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*, 305–335.

Rossiter, J. R. (2010). Measurement for the social sciences: The C-OAR-SE method and why it must replace psychometrics. New York: Springer.

Rössler, P. (2011). *Skalenhandbuch Kommunikationswissenschaft.* Wiesbaden: VS Verlag.

Rost, J. (2004). *Lehrbuch Testtheorie Testkonstruktion.* (2. Aufl.). Bern: Huber.

Rozeboom, W. W. & Jones, L. V. (1956). The validity of the successive intervals method of psychometric scaling. *Psychometrika*, *21*, 165–183.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Saal, F. E. & Landy, F. J. (1977). The Mixed Standard Rating Scale: An evaluation. *Organizational Behavior and Human Performance*, *18*, 19–35.

Sackett, P. R. & DuBois, C. L. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology*, *76*(6), 873–877.

Saito, T. (1994). Psychological scaling of the asymmetry observed in comparative judgement. *British Journal of Mathematical and Statistical Psychology*, *47*(1), 41–62.

Scheuch, E. K. (1961). Sozialprestige und soziale Schichtung. In D. W. Glass & R. König (Hrsg.), *Soziale Schichtung und soziale Mobilität. Sonderheft 5 der „Kölner Zeitschrift für Soziologie und Sozialpsychologie"* (S. 65–103). Opladen: Westdeutscher Verlag.

Scheuring, B. (1991). Primacy-Effekte, ein Ermündungseffekt? Neue Aspekte eines alten Phänomens. *Zeitschrift für Sozialpsychologie*, *22*, 270–274.

Schmeisser, D. R., Bente, G., & Isenbart, J. (2004). Am Puls des Geschehens. Die integrierte Rezeptionsprozessanalyse. Zum Mehrwert rezeptionsbegleitender Untersuchungsmethoden in der Werbewirkungsforschung. *Planung und Analyse*, *2004*(1), 28–34.

Schneider, F. M., Erben, J., Altzschner, R.-S., Kockler, T., Petzold, S., & Satzl, I. (2011). Die Übungssequenz macht den Meister.. . . Eine experimentelle Studie zu Kontext-Effekten von Übungsstimuli bei Real-Time Response Messungen. In M. Suckfüll, H. Schramm, & C. Wünsch (Hrsg.), *Rezeption und Wirkung in zeitlicher Perspektive* (S. 253–270). Baden-Baden: Nomos.

Schnell, R., Hill, P. B., & Esser, E. (1999). *Methoden der empirischen Sozialforschung*. München: Oldenbourg.

Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der empirischen Sozialforschung* (8. Aufl.). München: Oldenbourg.

Schulenberg, S. E. & Melton, A. M. (2007). Confirmatory factor analysis of the Computer Understanding and Experience Scale. *Psychological Reports*, *100*(3), 1263–1269.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales. A review of the literature. *Personnel Psychology*, *28*(4), 549–562.

Schwarz, N. (2008). Self-Reports: How the questions shape the answers. In R. H. Fazio & R. E. Petty (Eds.), *Attitudes: Their structure, function, and consequences* (pp. 49–67). New York: Psychology Press.

Schwarz, N., Knäuper, B., Hippler, H.-P., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaing of scale labels. *Public Opinion Quarterly*, *55*, 570–582.

Schwarz, N. & Oyserman, D. (2001). Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction. *American Journal of Evaluation*, *22*(2), 127–160.

Schwarz, N., Wänke, M., Sedlmeier, P., & Betsch, T. (2002). Experiential and contextual heuristics in frequency judgement: Ease of recall and response scales. In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency processing and cognition* (pp. 89–108). New York: Oxford University Press.

Shapira, Z. & Shirom, A. (1980). New Issues in the use of behaviorally anchored rating scales: Level of analysis, the effects of incident frequency, and external validation. *Journal of Applied Psychology*, *65*(5), 517–523.

Sherif, M. & Hovland, C. I. (1961). *Social judgement. Assimilation and contrast effects in communication and attitude change.* New Haven: Yale University Press.

Shore, T. H. & Tashchian, A. (2003). Effects of sex on raters' accountability. *Psychological Reports*, *92*(2), 693–702.

Sixtl, F. (1967). *Meßmethoden der Psychologie*. Weinheim: Beltz.

Smith, P. C. & Kendall, L. M. (1963). Retranslation of expectations: An approach to unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*, 149–155.

Statistisches Bundesamt. (2011). Haushaltsbefragung auf Stichprobenbasis zum Zensus 2011. Abgerufen 17. Juni, 2013, unter https://cdn.zensus2011.de/live/uploads/tx_templavoila/Fragebogen_Haushaltebefragung_20101007a.pdf.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 1–49). New York: Wiley.

Steyer, R. & Eid, M. (1993). *Messen und Testen*. Heidelberg: Springer.

Stine, W. W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin*, *105*(1), 147–155.

Strack, F., Schwarz, N., Ash, M. G., & Sturm, T. (2007). Asking questions: Measurement in the social sciences. In M. G. Ash & T.Sturm. (Eds.) *Psychology's territories: Historical and contemporary perspectives from different disciplines* (pp. 225–250). Mahwah: Lawrence Erlbaum Associates.

Strahan, R. F. (1980). More on averaging judges' ratings: Determining the most reliable composite. *Journal of Consulting and Clinical Psychology*, *48*, 587–589.

Subkoviak, M. J. (1974). Remarks on the method of paired comparisons: The effect on non-normality in Thurstone's Comparative Judgement Model. *Educational & Psychological Measurement*, *34*, 829–834.

Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (2006). *Foundations of measurement volume III: Representation, axiomatization, and invariance*. Mineola: Dover Publications.

Taylor, J. B., Haefele, E., Thompson, P., & O'Donoghue, C. (1970). Rating scales as measures of clinical judgement II: The reliability of example-anchored scales und conditions of rater heterogeneity and divergent behavior sampling. *Educational and Psychological Measurement*, *30*(2), 301–310.

Thomas, A., Palmer, J. K., & Feldman, J. M. (2009). Examination and measurement of halo via curvilinear regression: A new approach to halo. *Journal of Applied Social Psychology*, *39*(2), 350–358.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 469–477.

Thurstone, L. L. (1927). A „Law of Comparative Judgement". *Psychological Review*, *34*, 273–286.

Thurstone, L. L. & Chave, E. J. (1929). *The measurement of attitudes*. Chicago: University of Chicago Press.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Trommsdorff, V. (1975). *Die Messung von Produktimages für das Marketing. Grundlagen und Operationalisierung*. Köln: Heymanns.

Tziner, A., Joanis, C., & Murphy, K. R. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perception, and ratee satisfaction. *Group and Organization Management*, *25*(2), 175–190.

Upmeyer, A. (1985). *Soziale Urteilsbildung.* Stuttgart: Kohlhammer.

Upshaw, H. S. (1962). Own attitude as an anchor in equal appearing intervals. *Journal of Abnormal and Social Psychology*, *64*, 85–96.

Van der Ven, A. (1980). *Einführung in die Skalierung.* Bern: Huber.

Wade Savage, C. & Ehrlich, P. (Eds.). (1991). *Philosophical and foundational issues in measurement theory*. Hillsdale: Erlbaum.

Waldman, D. A. & Avolio, B. J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology*, *76*(6), 897–901.

Wänke, M. & Fiedler, K. (2007). What is said and what is meant: Conversational implicatures in natural conversations, research settings, media, and advertising. *Social Communication* (pp. 223–255). New York: Psychology Press.

Wanous, J. P. & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods*, *4*(4), 361–375.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252.

Waxweiler, R. (1980). *Psychotherapie im Strafvollzug. Eine empirische Erfolgsuntersuchung am Beispiel der sozialtherapeutischen Abteilung in einer Justizvollzugsanstalt.* Basel: Beltz.

Wessels, M. G. (1994). *Kognitive Psychologie* (3. Aufl.). München: Reinhardt.

West, C. P., Dyrbye, L. N., Sloan, J. A., & Shanafelt, T. D. (2009). Single item measures of emotional exhaustion and depersonalization aure useful for assessing burnout in medical professionals. *Journal of General Internal Medicine*, *24*(12), 1318–1321.

Westermann, R. (1985). Empirical tests of scale type for individual ratings. *Applied Psychological Measurement*, *9*, 265–274.

Wewers, M. E. & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*, *13*(4), 227–236.

Wirtz, M. A. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen.* Göttingen: Hogrefe.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35–51.

Young, R. K. & Thiessen, D. D. (1991). Washing, drying, and anointing in adult humans (Homo sapiens): Commonalities with grooming sequences in rodents. *Journal of Comparative Psychology*, *105*(4), 340–344.

Yu, J. H., Albaum, G., & Swenson, M. (2003). Is a central tendency error inherent in the use of semantic differential scales in different cultures? *International Journal of Market Research*, *45*(2), 213–228.

Zakour, M. J. (1994). Measuring career-development volunteerism: Guttman scale analysis using red cross volunteers. *Journal of Social Service Research*, *19*(3–4), 103–120.

Zhikun, D. & Fungfai, N. (2008). A new way of developing semantic differential scales with personal construct theory. *Construction Management & Economics*, *26*(11), 1213–1226.

Zumbo, B. D. & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology/Psychologie Canadienne*, *34*(4), 390–400.

Stichprobenziehung

Nicola Döring, Jürgen Bortz

9.1 Population und Stichprobe – 292

9.2 Stichproben für qualitative Studien – 302

9.3 Stichproben für quantitative Studien – 305

Literatur – 318

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_9, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Wissen, wie Population und Stichprobe definiert sind.
- Das Konzept der Repräsentativität erläutern und problematisieren können.
- Die wichtigsten qualitativen Stichprobenarten kennen.
- Die wichtigsten quantitativen Stichprobenarten kennen.
- Die Vor- und Nachteile probabilistischer und nichtprobabilistischer Stichproben abwägen können.

Die meisten qualitativen und quantitativen sozialwissenschaftlichen Studien untersuchen weder sämtliche interessierenden Personen, Objekten oder Verhaltensweisen noch einen singulären Einzelfall, sondern arbeiten mit Stichproben („samples"), d. h. mit einer Auswahl von Fällen. In der qualitativen Forschung spricht man im Zusammenhang mit der Bildung von Stichproben meist von **Fallauswahl**. In der quantitativen Forschung hat sich die Bezeichnung **Stichprobenziehung** eingebürgert. Im Folgenden wird zunächst das vor allem für quantitative Stichproben wichtige Verhältnis zwischen Population und Stichprobe behandelt, mit dem das Qualitätskriterium der Repräsentativität verbunden ist (► Abschn. 9.1). Anschließend stellen wir die wichtigsten Verfahren der Stichprobenbildung für qualitative Untersuchungen (► Abschn. 9.2) sowie für quantitative Studien (► Abschn. 9.3) vor. Während die „Theoretische Stichprobe" den Prototyp einer qualitativen Fallauswahl und die „probabilistische Stichprobe" den Prototyp eines quantitativen Samples darstellt, existieren eine Reihe von Verfahren der Fallauswahl bzw. Stichprobenziehung („sampling"), die sowohl in qualitativen als auch in quantitativen Studien zum Einsatz kommen (z. B. Gelegenheitsstichprobe, Schneeball-Stichprobe).

9.1 Population und Stichprobe

Die Gesamtheit aller Fälle, über die in einer Studie wissenschaftlich etwas ausgesagt werden soll, heißt in der quantitativen Forschung **Population** (auch: Grundgesamtheit, Universum; „population"). In der qualitativen Forschung wird eher vom (Untersuchungs-)**Feld** und den dort vorfindbaren **Fällen** gesprochen, um die Verortung der Fälle in ihrem lebensweltlichen Kontext zu betonen (zu wissenschaftstheoretischen Grundlagen der qualitativen Forschung ► Abschn. 2.3). In der empirischen Sozialforschung haben wir prinzipiell die Möglichkeit, alle interessierenden Fälle zu untersuchen (Vollerhebung) oder eine Stichprobe von Fällen auszuwählen (Teilerhebung). Stichprobenuntersuchungen sind der Regelfall, wobei die Qualität von Stichproben hinsichtlich ihrer Aussagekraft über die Gesamtheit der interessierenden Fälle sehr unterschiedlich zu bewerten ist. Methodologische Überlegungen zum Verhältnis zwischen Population und Stichprobe und die Differenzierung verschiedener Stichprobentypen spielen in der quantitativen Forschung eine wichtige Rolle, während methodischen Fragen der Fallauswahl in der qualitativen Forschung deutlich weniger Aufmerksamkeit geschenkt wird (Glaser & Strauss, 1999, S. 238).

9.1.1 Vollerhebung der Population

Die **Zielpopulation** (auch: angestrebte Grundgesamtheit; „target population") ergibt sich aus dem jeweiligen **Forschungsinteresse einer Studie**. Zielpopulationen können unterschiedlich definiert werden: die erwachsene Bevölkerung eines Landes, die Gesamtheit aller Besucher einer Veranstaltung, die Fünftklässler in Spanien, 70- bis 80-jährige Diabetespatienten in Deutschland etc. Neben einzelnen Personen können auch soziale Gruppen, Maßnahmen, Objekte oder Verhaltensweisen die Untersuchungseinheiten sein, so dass Stichproben dann aus entsprechenden Populationen zu ziehen sind (z. B. Population aller mittelständischen Unternehmen in Rumänien; Population aller Großfamilien einer Stadt; Population aller Zwischenrufe in Bundestagsdebatten im letzten Jahr; Population aller bisherigen Anti-Rauch-Kampagnen der Bundeszentrale für gesundheitliche Aufklärung [BZgA] etc.).

Population bzw. Zielpopulation – Unter Population bzw. Zielpopulation („population"; „target population") versteht man die Gesamtmenge aller Untersuchungseinheiten (z. B. Personen, Verhaltensweisen, Objekte), über die im Rahmen einer quantitativen empirischen Studie wissenschaftliche Aussagen getroffen werden sollen.

Neben dem Erkenntnisinteresse müssen in der empirischen Forschung auch **pragmatische Überlegungen** hinsichtlich der Durchführbarkeit einer Untersuchung bei der Festlegung der Zielpopulation beachtet werden. Will man Belästigungen durch Fluglärm untersuchen, ist hierfür die Grundgesamtheit der Personen, die höchstens einen Kilometer vom Flughafen entfernt wohnen (dieses populationsdefinierende Kriterium ist problemlos messbar), einer Populationsdefinition vorzuziehen, die sich auf alle Personen bezieht, die sich durch Fluglärm beeinträchtigt fühlen. Bei der **Definition einer Zielpopulation** sind einfache manifeste Merkmale zu verwenden, die leicht operationalisierbar sind. Auch wird man Zielpopulationen meist zeitlich, örtlich und kulturell einschränken und nicht z. B. Aussagen über alle Menschen aller Altersgruppen zu

allen historischen Zeiten in allen Ländern oder Kulturen anstreben.

Vollerhebung – Unter Vollerhebung (Totalerhebung, „total population survey“) versteht man die Untersuchung aller Objekte einer definierten Population.

Vollerhebungen sind **möglich**, wenn es sich um überschaubare und gut erreichbare Populationen handelt. Sie sind **notwendig**, wenn eine Population sehr klein und heterogen ist, so dass Stichproben das Gesamtbild verfehlen. Interessiert man sich z. B. für Arbeitsmotivation, Zufriedenheit und Verbesserungsvorschläge aller Mitarbeiterinnen und Mitarbeiter eines Unternehmens ($N = 22$, Population), so sollte eine Vollerhebung durchgeführt werden. Vollerhebungen sind nicht nur bei Populationen von Personen, sondern auch bei anderen Untersuchungsobjekten möglich und sinnvoll: So können in einer kommunikationswissenschaftlichen Analyse z. B. alle Folgen einer TV-Serie ausgewertet werden oder in der Verkehrsforschung alle Unfälle betrachtet werden, die sich während eines Jahres an einer bestimmten Straßenkreuzung ereignet haben.

Ein Sonderfall der Vollerhebung ist die **Volkszählung** (Zensus, „census“), bei der mit großem Aufwand für die **amtliche Statistik** Daten ganzer nationaler Bevölkerungen erhoben werden. Die Teilnahme an Volkszählungen ist in der Regel gesetzlich verpflichtend (s. z. B. für Deutschland das Zensusgesetz 2011; Abb. 9.1).

Oftmals ist es jedoch **nicht** möglich, eine Vollerhebung durchzuführen. Dies ist immer dann der Fall, wenn

- die Population nicht endlich (finit), sondern **unendlich** (infinit) groß ist (Beispiel: Verbreitung nationaler Stereotype in allen Ausgaben der in Deutschland – täglich neu – erscheinenden Tageszeitungen),
- die Population **nur teilweise bekannt** ist (Beispiel: Erfassung des Gesundheitszustandes aller medikamentenabhängigen Männer in der Schweiz),
- die Art der Untersuchung die Population zu stark **beeinträchtigt oder gar zerstört** (Beispiel: „Crashtests“ zur Qualitätskontrolle der gesamten Jahresproduktion eines Automobilherstellers) oder
- die Untersuchung der gesamten Population **zu aufwändig** wäre (Beispiel: Untersuchung zum Musikgeschmack bei allen europäischen Jugendlichen zwischen 14 und 16 Jahren).

Insbesondere in Beobachtungsstudien, in denen es um die direkte Erfassung von Verhaltensweisen geht (z. B. Flirtverhalten, Streitverhalten), sind Vollerhebungen des gesamten Verhaltensstroms zu allen Zeitpunkten und an allen Orten kaum möglich. Stattdessen müssen **Zeit- oder Ereignisstichproben des interessierenden Verhaltens** gezogen werden (▶ Abschn. 10.1.1). Relativ verbreitet sind Vollerhebungen z. B. in der Soziologie und Politikwissenschaft, wenn mit **Makrodaten** gearbeitet wird (z. B. Wahlbeteiligung in allen EU-Ländern; Bruttoinlandsprodukt in allen OECD-Ländern; Telefonanschlüsse in allen Ländern weltweit etc.).

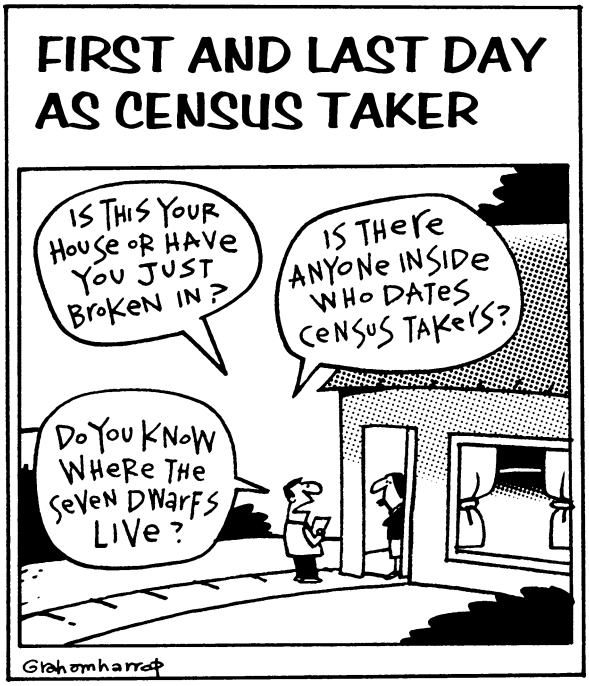


Abb. 9.1 Die Teilnahme an Volkszählungen ist verpflichtend, die Interviewenden müssen dementsprechend sorgfältig ausgewählt und geschult werden. (© Graham Harrop/Search ID: ghrn508, Rights Available from CartoonStock.com)

Findet im Rahmen einer quantitativen Studie eine Vollerhebung statt, so sind bei der statistischen Auswertung **deskriptiv-statistische Angaben** meist ausreichend: Wenn vollständige Daten über die Population vorliegen, ist **kein** Inferenzschluss mithilfe eines statistischen Signifikanztests (▶ Abschn. 12.5.1) notwendig, um zu prüfen, ob die Stichprobendaten auf die Population zu verallgemeinern sind, denn es liegen ja keine Stichproben-, sondern direkt die Populationsdaten vor. Es wird jedoch diskutiert, ob unter bestimmten theoretischen Annahmen nicht doch die **Anwendung von statistischen Signifikanztests auf Vollerhebungen** sinnvoll ist (Behnke, 2005; Broscheid & Geschwend, 2005). Ein anderer kontroverser Punkt im Zusammenhang mit Populationen ist die Abgrenzung zwischen finiten (abzählbaren, endlichen) und **infiniten Populationen** und den daraus resultierenden Konsequenzen für Vollerhebungen, Teilerhebungen und statistische Analysen (Kozak, 2005).

9.1.2 Teilerhebung der Population: Stichprobenverfahren

Teilerhebungen bzw. Stichprobenuntersuchungen sind weniger aufwändig als Vollerhebungen. Sie lassen sich schneller durchführen und auswerten und sind deshalb aus forschungsökonomischen Gründen meist der Vollerhebung vorzuziehen. Für Stichprobenuntersuchungen spricht zudem die Möglichkeit, wegen der im Vergleich zur Vollerhebung deutlich reduzierten Anzahl von Untersuchungsfällen eine größere Anzahl von Merkmalen sorgfältiger und kontrollierter erfassen bzw. individuelle Sichtweisen differenzierter rekonstruieren zu können. Dies ist ein Umstand, der gelegentlich zu der Behauptung veranlasst hat, Vollerhebungen seien weniger aussagekräftig als gut geplante und sorgfältig durchgeführte Stichprobenuntersuchungen (Scheuch, 1974, S. 5).

Stichprobenerhebung – Eine Stichprobenerhebung (Teilerhebung, „sample") liegt vor, wenn nur eine Auswahl von Fällen aus der Zielpopulation untersucht wird.

Jede Stichprobe zeichnet sich durch zwei zentrale Merkmale aus:

1. **Stichprobenart** („sampling method"): Nach welchem Auswahlverfahren Elemente aus der Population in die Stichprobe gelangen, hat wesentlichen Einfluss auf die Aussagekraft einer Studie. Grundsätzlich sind zufallsgesteuerte (**probabilistische**) Auswahlmethoden, bei denen die Stichprobenauswahl nach einem statistischen Zufallsverfahren erfolgt, von nicht-zufälligen (**nicht-probabilistischen**) Methoden zu unterscheiden. Probabilistische Auswahlverfahren sind nur bei großen Stichproben im Rahmen der quantitativen Forschung sinnvoll einsetzbar. Nicht-probabilistische Stichproben sind für die qualitative Forschung einschlägig, sie werden jedoch oft auch in der quantitativen Forschung eingesetzt. Bei nicht-zufälligen Stichproben differenziert man zwei Vorgehensweisen: Entweder es werden ganz gezielt bestimmte Elemente aus der Population herausgegriffen (**bewusste bzw. absichtsvolle Auswahl**; z. B. Quotenstichprobe: Befragung einer definierten Anzahl von Personen aus der Population, die gemäß Quotenplan bestimmte Merkmale oder Merkmalskombinationen aufweisen sollen) oder es werden relativ ungezielt vor allem gut erreichbare Elemente ausgewählt (**willkürliche Auswahl**; z. B. Selbstselektionsstichprobe: Befragung aller Personen, die sich auf eine veröffentlichte Umfrageeinladung melden).
2. **Stichprobenumfang** („sample size"): Der **Stichprobenumfang** ist definiert als Anzahl aller Objekte oder Personen in der Stichprobe. In der quantitativen Forschung symbolisiert n oder N den Stichprobenumfang, N die Populationsgröße. Die Fallzahl eines qualitativen Samples wird ohne „n" berichtet. Der **Populationsumfang** wird mit N symbolisiert. Mit wachsendem Stichprobenumfang steigt die Aussagekraft einer Studie in der qualitativen wie in der quantitativen Forschung. Bei der Theoretischen Stichprobenbildung werden in der qualitativen Forschung im Zuge der Datenerhebung und Datenauswertung schrittweise zusätzliche Fälle hinzugezogen, bis **theoretische Sättigung** eintritt, also weitere Fälle keine neuen Erkenntnisse liefern (▶ Abschn. 9.2.1). Bei quantitativen Studien wird idealerweise im Vorfeld festgelegt, welche Effektgrößen im statistischen Signifikanztest noch nachweisbar sein sollen (hypothesenprüfende Studien) bzw. wie präzise Parameterschätzungen ausfallen sollen (populationsbeschreibende Studien) und auf dieser Basis ein **optimaler Stichprobenumfang** (▶ Abschn. 14.3) bestimmt. Zu kleine Stichproben würden in quantitativen Studien aufgrund mangelnder statistischer Teststärke zu uneindeutigen Resultaten führen und zu große Stichproben würden unnötig Ressourcen verbrauchen. Um die – mit dem Stichprobenumfang variierende – Genauigkeit oder Ungenauigkeit von statistischen Parameterschätzungen auszudrücken, spricht man in der Umfrageforschung vom **Stichprobenfehler** („sampling error"), der als statistisches Maß über den Standardfehler („margin of error") quantifiziert wird (zu verschiedenen Arten von **Umfragefehlern gemäß dem Total Survey Error-Modell** ▶ Abschn. 10.2.3).

Betrachtet man das Verhältnis von quantitativer Stichprobe und Grundgesamtheit, so sind **drei verschiedene Populationskonzepte** zu differenzieren: Ziel-, Auswahl- und Inferenzpopulation. Idealerweise liegt eine vollständige Liste aller Elemente der (als finit definierten) **Zielpopulation** vor (**Auswahlgrundlage**, „sample frame", „sampling frame"), aus der eine Auswahl von Elementen für die Stichprobe gezogen werden kann. In diesem Fall haben prinzipiell alle Elemente der Zielpopulation eine positive Auswahlwahrscheinlichkeit, somit sind Zielpopulation und **Auswahlpopulation** („frame population") identisch. Beispiel: Ein Unternehmen möchte die Zufriedenheit seiner Kundschaft erfassen und nutzt zur Auswahl der Stichprobe das vollständige Kundenverzeichnis. Da jedoch nicht alle angeschriebenen Kunden an der Umfrage teilnehmen, können die Umfrageergebnisse nicht auf den gesamten Kundenstamm (Zielpopulation), sondern nur auf die auskunftswilligen und erreichbaren Kunden bezogen werden (**Inferenzpopulation**).

Auswahl- und Inferenzpopulation – Im Zusammenhang mit Stichprobenerhebungen müssen **drei Populationskonzepte** differenziert werden: Während die **Zielpopulation** („target population“) die Gesamtheit aller Objekte meint, über die Aussagen getroffen werden sollen, umfasst die **Auswahlpopulation** („frame population“) die Gesamtheit aller Objekte, die prinzipiell eine Chance haben, in die Stichprobe zu gelangen. Angesichts von Stichprobenausfällen erlaubt eine konkrete Stichprobe meist keinen direkten Rückschluss auf die Auswahlpopulation, sondern nur auf eine sog. **Inferenzpopulation** („inference population“), also die Gesamtheit aller Objekte, aus der die Stichprobe tatsächlich stammt.

Eine vollständige Zielpopulationsliste liegt leider nur in seltenen Fällen vor und kann bei infiniten Populationen definitionsgemäß gar nicht erstellt werden. Selbst wenn eine Liste vorliegt, ist sie z. B. aus Datenschutzgründen oft nicht für Forschungszwecke verfügbar. Meist muss deswegen mit **unvollständigen** bzw. **verzerrten Auswahlgrundlagen** gearbeitet werden. In einem solchen Fall weicht die Auswahlpopulation als Gesamtheit aller Objekte, die prinzipiell eine Chance haben, in die Stichprobe zu gelangen, von der Zielpopulation ab. Die Abweichung zwischen Zielpopulation und Auswahlpopulation bezeichnet man als **Abdeckungsfehler** („coverage error“). Man spricht von **Überabdeckung** („overcoverage“), wenn Elemente, die nicht zur Population gehören, im Auswahlrahmen erscheinen oder Populationselemente mehrfach im Auswahlrahmen auftauchen. Umgekehrt liegt ein Abdeckungsfehler im Sinne von **Unterabdeckung** („undercoverage“) vor, wenn Elemente im Auswahlrahmen fehlen, die zur Population gehören (► Studienbeispiel „Eine sehr große und sehr verzerrte Stichprobe“). Definiert man z. B. die Bevölkerung einer Stadt als Zielpopulation einer Befragungsstudie und verwendet für die Stichprobenziehung als Auswahlgrundlage das örtliche **Telefonbuch**, so entsteht ein beträchtlicher Unterabdeckungsfehler dadurch, dass alle Personen bzw. Haushalte, deren Nummern nicht im Telefonbuch veröffentlicht sind, überhaupt nicht in die Stichprobe gelangen können. Die entsprechende Studie kann somit keine Aussagen über die **Zielpopulation** (lokale Einwohner), sondern nur über die **Auswahlpopulation** (lokale Einwohner, deren Telefonnummern im Telefonbuch veröffentlicht sind) treffen (zu speziellen **Stichprobenverfahren für Telefonumfragen**, bei denen z. B. Telefonnummern nach dem Zufallsprinzip erzeugt werden ► Abschn. 10.2.3 „Vollstrukturiertes Telefoninterview“). Das Beispiel der Telefonumfrage illustriert, dass die Methode der Stichprobenziehung (z. B. mittels Telefonbuch oder künstlich generierten Telefonnummern) von der Methode der Datenerhebung (hier: Telefoninterview) abhängt. Bei der Datenerhebung mittels persönlichem Interview können zur Ziehung einer Stichprobe der lokalen Bevölkerung wiederum andere Stichprobentechniken zum Einsatz kommen: Aus dem **Melderegister des Einwohnermeldeamtes** als Auswahlrahmen lässt sich eine Stichprobe von Wohnadressen für persönliche Interviews ziehen. Alternativ lässt sich eine Zufallsauswahl von Wohnadressen ziehen, indem man durch die Stadt geht und per vordefiniertem blindem Zufallsprinzip jeweils an verschiedenen Türen klingelt (sog. **Random-Route-Verfahren** der Stichprobenziehung), um die Interviews durchzuführen (zu **Stichprobentechniken für mündliche Interviews**

Studienbeispiel

Eine sehr große und sehr verzerrte Stichprobe

Legendär ist eine misslungene Stichprobenuntersuchung der **Zeitschrift *Literary Digest***, anhand derer das Wahlverhalten der amerikanischen Bevölkerung im Jahr 1936 prognostiziert werden sollte: 10 Mio. Wahlberechtigte in den USA, deren Adressen man über Telefonbücher, Mitgliedskarteien von Clubs und Vereinen etc. ermittelt hatte (willkürliche Auswahl), erhielten den Untersuchungsfragebogen postalisch zugeschickt. Eine **Stichprobe von rund $n = 2.40$ Mio. Wahlberechtigten** sendete ausgefüllte Fragebögen zurück. Diese Stichprobe sollte Auskunft geben über das politische Stimmungsbild in der **Population aller Wahlberechtigten in den USA**, also etwa $N = 100$ **Mio. Wahlberechtigte**. Die erhobenen Fragebogendaten legten den Schluss nahe, dass die Demokraten mit ihrem Spitzenkandidaten Franklin Roosevelt dem republikanischen Kandidaten Alfred Landon unterliegen würden und nur 43 % der Stimmen auf sich vereinigen könnten. Tatsächlich erreichte Roosevelt jedoch eine Stimmenmehrheit von 62 %. Eine der größten Stichproben in der Geschichte der Meinungsforschung führte so zu einer Fehlschätzung von knapp 20 % (Freedman, Pisani, & Purves, 1978, S. 302ff.).

Zwei Fehlerquellen waren im Spiel: Zunächst wurden durch die willkürliche (nicht-probabilistische) Anwerbung über Mitgliedskarteien und Telefonbücher (im Jahr 1936!) Angehörige der Mittel- und Oberschicht unverhältnismäßig häufig angesprochen im Sinne eines Überabdeckungsfehlers („oversampling“, „overcoverage“), während Angehörige der unteren Schichten eine sehr viel geringere Auswahlwahrscheinlichkeit hatten im Sinne eines Unterabdeckungsfehlers („undersampling“, „undercoverage“). Zum **Abdeckungsfehler** kam noch die für postalische Umfragen charakteristische **hohe Ausfallrate** („non-response“) hinzu, die wiederum die Angehörigen der unteren Schichten benachteiligte (sie antworteten seltener). Da aber gerade die unterprivilegierten Schichten in der Bevölkerung Roosevelts Politik befürworteten, konnte die stark in Richtung Oberschicht verzerrte Stichprobe das Wahlergebnis schlecht vorhersagen. Daran hätte sich auch nichts geändert, wenn man über dieselben Rekrutierungswege zusätzliche Personen befragt und die Stichprobengröße z. B. verdoppelt hätte. Merke: Bei einer systematisch verzerrten Auswahl hilft auch ein großer Stichprobenumfang nicht, den Fehler zu beheben. Er wiederholt sich nur in großem Stil.

Mithilfe einer sehr viel kleineren **Stichprobe von $n = 2\,000$ Wahlberechtigten**, die nach wichtigen soziodemografischen Merkmalen quotiert zusammengesetzt war (Quotenstichprobe) und somit höhere Repräsentativität erreichte, konnte das **Gallup-Institut** den Wahlerfolg von Roosevelt dagegen recht präzise vorhersagen (Diekmann, 2005, S. 325).

► Abschn. 10.2.3 „Vollstrukturiertes persönliches Interview“).

Werden aus einem Auswahlrahmen zufällig Elemente in die Stichprobe gezogen, so können nur dann Aussagen über die Auswahlpopulation getroffen werden, wenn tatsächlich alle für die Stichprobe vorgesehenen Personen oder Objekte tatsächlich untersucht werden. Dies ist jedoch oft nicht der Fall: Manche Personen sind z. B. für Befragungen nicht erreichbar oder verweigern die Teilnahme. Auch kann es durch technische Probleme oder Fehler bei der Datenerhebung oder Datenauswertung zu Informationsverlusten kommen. Die tatsächlich realisierte Stichprobe erlaubt also streng genommen keinen direkten Rückschluss auf die Auswahlpopulation, sondern nur auf eine sog. **Inferenzpopulation** als Gesamtheit aller Objekte, aus der die Stichprobe tatsächlich stammt (also **Auswahlpopulation minus Ausfälle**). Die Inferenzpopulation ist in der Regel nicht exakt bekannt. Sie kann nur erschlossen werden auf der Basis von Informationen oder Vermutungen über Art und Umfang der Stichprobenausfälle durch Nichterreichbarkeit, Teilnahmeverweigerung etc.

In der Umfrageforschung spricht man vom **Non-Response-Fehler** („non-response error“), um die Abweichung zwischen der laut Auswahlrahmen geplanten Stichprobe und der tatsächlich realisierten Stichprobe zu kennzeichnen (zu Umfragefehlern ► Abschn. 10.2.3). Wurde als Auswahlrahmen das Telefonbuch verwendet und daraus eine Zufallsauswahl von Anschlüssen ausgewählt, so wird man im Zuge der Datenerhebung feststellen, dass eine Reihe der Nummern nicht erreichbar sind oder dass die angerufenen Personen die Teilnahme an der Umfrage ablehnen. Die entsprechende Stichprobe trifft am Ende nur Aussagen über die Population der lokalen Einwohner, deren Telefonnummern im Telefonbuch veröffentlicht sind und die im Untersuchungszeitraum unter ihrer Telefonnummer erreichbar und zur Teilnahme an der Umfrage bereit waren. Inhaltlich ist dies so zu interpretieren, dass in der Inferenzpopulation z. B. alle Einwohner fehlen, die temporär nicht telefonisch erreichbar sind (z. B. wegen Krankenhaus-, Gefängnis- oder Auslandsaufenthalt) oder die aus verschiedenen Gründen keine Teilnahmebereitschaft haben (z. B. wenig Zeit, mangelndes Vertrauen in die Forschung, kein Interesse am Befragungsthema etc.; zu weiteren Details von Telefonumfragen ► Abschn. 10.2.3). Bei der Rekrutierung von Stichproben ist es ein wichtiges Anliegen, den Non-Response-Fehler zu verringern, etwa dadurch, dass man vorgesehene Stichprobenmitglieder **mehrfach kontaktiert**, wenn sie beim ersten Versuch nicht erreichbar sind. Zudem können die **Formulierung der Anfrage** und das **Auftreten der Forschenden beim Erstkontakt** darüber entscheiden, ob ein vorgesehenes Stichprobenmitglied sich letztlich für oder gegen die Teilnahme an der Studie entscheidet und somit in der geplanten Stichprobe verbleibt oder aus der Stichprobe herausfällt.

Das Ausmaß der Ausfälle wird über die **Ausschöpfungsrate** oder **Rücklaufquote** quantifiziert. Generell ist eine hohe Ausschöpfungsrate (bzw. ein geringer Non-Response) wünschenswert. Abgesehen vom quantitativen Ausmaß der Ausfälle ist es für die Repräsentativität der Stichprobe aber noch wichtiger, ob es sich um unsystematische Ausfälle handelt oder um **systematische Ausfälle**, die zu Stichprobenverzerrungen führen. Eine systematische Verzerrung würde z. B. vorliegen, wenn der Non-Response bei Berufstätigen in einer Telefonumfrage deutlich höher ist als z. B. bei Rentnern, weil die Anrufzeiten innerhalb üblicher Arbeitszeiten liegen.

Unterschiedliche **Befragungsmodi** (z. B. mündliches Interview, telefonisches Interview, Online-Befragung, mobile Befragung etc.) gehen bei der Stichprobenziehung mit unterschiedlichen Verzerrungen einher (spezifische Über- oder Unterabdeckungsfehler, unterschiedliche Anteile und Merkmale der Antwortverweigerer etc.). Eine Möglichkeit, die Verzerrungen eines bestimmten Befragungsmodus zumindest teilweise auszugleichen ist die **Multi-Mode-Befragung** („multi mode survey“), bei der mindestens zwei Befragungsmodi kombiniert werden, um eine weniger einseitig verzerrte Stichprobe zu erlangen (z. B. bei einer kombinierten Online- und Telefonbefragung gelangen sowohl Personen in die Stichprobe, die Online gut erreichbar sind, als auch diejenigen, die telefonisch gut erreichbar sind; zu unterschiedlichen Befragungsmodi in der quantitativen Forschung ► Abschn. 10.2.3 und ► Abschn. 10.3.3).

Bei der Festlegung von Auswahlmethode und Umfang der Stichprobe geht es einerseits um die Aussagekraft der geplanten Untersuchung, aber auch um forschungsökonomische Aspekte, etwa den Zeit- und Kostenaufwand. Im Zusammenhang mit Stichprobenziehung („sampling“) sind drei Phasen zu unterscheiden:

1. **Stichprobenkonstruktion**: Entscheidung für ein bestimmtes Verfahren der Stichprobenauswahl, Festlegung des Stichprobenumfangs und des Stichprobenplans (z. B. Entscheidung für eine bewusste Auswahl nach dem Quotenverfahren und Festlegung des Quotenplans).
2. **Rekrutierung**: Kontaktierung und Anwerbung von Untersuchungspersonen entsprechend der vorangegangenen Stichprobenkonstruktion (z. B. Versendung von personalisierten Einladungs-E-Mails; Kontaktanrufe durch das Forschungsteam; Aufsuchen der Befragungspersonen an ihren Wohnadressen).
3. **Stichprobenerhebung**: Tatsächliche Datenerhebung mit allen rekrutierten Personen bzw. Objekten (z. B. Austeilen und Einsammeln der Fragebögen unter den ausgewählten Befragungspersonen). Die Anzahl der vollständig und plausibel ausgefüllten Fragebögen

(zur Datenaufbereitung ▶ Kap. 11) entspricht dann der endgültigen Stichprobe.

Während bei der Stichprobenkonstruktion methodische Fragen hinsichtlich der Aussagekraft der zu ziehenden Stichprobe im Mittelpunkt stehen, sind im Zusammenhang mit der **Rekrutierung** vor allem **praktische** und **organisatorische Fragen** zu lösen, allen voran die Frage: Wie können möglichst schnell und kostengünstig genügend Untersuchungsteilnehmende aus der Zielpopulation gewonnen werden? Das Spektrum der Vorgehensweisen reicht von einer Anwerbung über den **eigenen Bekanntenkreis** bis zum Rückgriff auf große Datenbanken von Personen, die sich freiwillig zur regelmäßigen Teilnahme an Studien bereit erklärt haben: Derartige **Access-Panels** werden vor allem von Forschungsinstituten aufgebaut und verwaltet. Die Panel-Mitglieder werden durch Incentives (z. B. Gutscheine, Verlosungen) zur langfristigen Teilnahme motiviert. Für alle Mitglieder eines Access-Panels liegen Kontaktdaten und soziodemografische Basisdaten vor. Bei Bedarf können aus diesem Freiwilligen-Pool jederzeit unterschiedliche Stichproben gezogen und z. B. für Fragebogenerhebungen genutzt werden. Bei Stichproben aus Access-Panels handelt es sich meist um nichtprobabilistische Quotenstichproben. Für probabilistische Stichproben greift man auf entsprechende **Auswahlrahmen** zurück, die möglichst viele Populationsmitglieder enthalten (z. B. Einwohnermeldeamtsregister, Kundenkartei, Mitgliederverzeichnis).

Manche Populationen sind sehr schwer erreichbar, etwa weil sie klein, geografisch verstreut oder ihre Mitglieder nicht ohne weiteres erkennbar sind bzw. sich nicht öffentlich zu erkennen geben wollen. Typische Beispiele für solche **speziellen Populationen** sind Drogenkonsumenten, Prostituierte, Obdachlose, aber auch Anhänger eines seltenen Hobbys oder Experten mit seltenen Fachgebieten. Vollständige Populationslisten als Auswahlgrundlagen für probabilistische Stichproben wird man hier kaum erhalten. Statistische Angaben zur Zusammensetzung der jeweiligen Populationen fehlen meist ebenfalls, so dass Quotenstichproben nicht möglich sind. Es bleiben zunächst nur **Gelegenheitsstichproben** in der Variante **Selbstselektionsstichprobe**, etwa indem man öffentliche Teilnahme-Aufrufe startet (▶ Abschn. 9.3.1). So lassen sich viele seltene Populationen heute über entsprechend spezialisierte Online-Foren ansprechen, in denen man Studieneinladungen veröffentlicht. Eine andere Möglichkeit besteht im **Screening**: Hierbei wird eine sehr große und heterogene Stichprobe gezogen und zunächst mit einer Screeningfrage bzw. einem Screening-Kurzfragebogen erfasst, wer zur gesuchten speziellen Population gehört, die untersucht werden soll. Es kann in diesem Fall auch auf spezielle **Stichprobenverfahren für schwer erreichbare Populationen** zurückgegriffen werden, etwa die **Schneeball-Stichprobe** (▶ Abschn. 9.3.1).

9.1.3 Repräsentativität von Stichproben

Welche Stichprobenmerkmale anzustreben sind, hängt von der Zielsetzung einer Studie ab: Für qualitative wie quantitative Erkundungsstudien **(explorative Studien)** sind kleine, nicht-zufällige Stichproben ausreichend. Auch für theorieprüfende oder theoriebildende Untersuchungen (explanative Studien) greift man in der quantitativen sowie qualitativen Forschung in der Regel auf kleine bis mittlere, nicht-zufällige Samples zurück. Große probabilistische Stichproben sind für **populationsbeschreibende quantitative Studien** (deskriptive Studien), insbesondere für die quantitative **Umfrageforschung** („survey research") nötig: Will man politische Einstellungen oder Merkmale des Gesundheitszustandes der breiten Bevölkerung möglichst genau erfassen, so sollte die Stichprobe möglichst unverzerrt die Populationsverhältnisse widerspiegeln.

Populationsbeschreibende quantitative Studien zielen darauf ab, die Ausprägung bestimmter Merkmale in der Population **(Populationsparameter)** möglichst präzise anhand der entsprechenden Merkmalsausprägungen in der Stichprobe **(Stichprobenkennwerte)** zu schätzen, entweder durch Berechnung eines einzelnen Schätzwertes **(Punktschätzung)** oder – das ist wesentlich aussagekräftiger – durch Angabe eines Schätzbereiches, in dem der wahre Populationsparameter mit hoher Wahrscheinlichkeit liegt **(Intervallschätzung)**. Die Intervallschätzung gibt durch die Breite des **Konfidenzintervalls** für den geschätzten Populationsparameter an, wie genau oder ungenau die Schätzung ausfällt (zur Datenanalyse mittels Parameterschätzung im Rahmen deskriptiver Studien ▶ Abschn. 12.4). Grundsätzlich können alle eine Population beschreibenden uni-, bi- oder multivariaten Parameter mittels Stichprobendaten geschätzt werden (z. B. arithmetischer Mittelwert, Medianwert, Modalwert, absolute Häufigkeit, relative Häufigkeit, Standardabweichung, Spannweite, Schiefe, Exzess, Regressions- und Korrelationskoeffizient, Kovarianz etc.; zur Bedeutung dieser Maße s. Bortz & Schuster, 2010; ◘ Tab. 9.1). Die konkrete Berechnung dieser statistischen Schätzungen sowie deren Genauigkeit hängen davon ab, welche Variante einer probabilistischen Stichprobe eingesetzt wird. So kann z. B. ein Populationsmittelwert in der Regel genauer geschätzt werden, wenn die Daten aus einer geschichteten Zufallsstichprobe anstelle einer einfachen Zufallsstichprobe (▶ Abschn. 9.3.2) stammen. Denn die geschichtete Zufallsstichprobe bildet die Population hinsichtlich aus-

Tabelle 9.1 Populationsbeschreibende quantitative Stichproben-Studien: Unbekannte Populationsparameter werden durch die empirischen Stichprobenkennwerte geschätzt (drei Beispiele)

	Populationsparameter (griechische Buchstaben)	Stichprobenkennwerte (lateinische Buchstaben)
Arithmetischer Mittelwert	μ (my) z. B. durchschnittliche TV-Nutzungszeit in der Population	$M = \bar{x}$ z. B. durchschnittliche TV-Nutzungszeit in der Stichprobe
Relative Häufigkeit	π (pi) z. B. Anteil der Personen mit Depression in der Population	p z. B. Anteil der Personen mit Depression in der Stichprobe
Bivariates lineares Zusammenhangsmaß	ρ (rho) z. B. linearer Zusammenhang (bivariater Korrelationskoeffizient) zwischen TV-Nutzungszeit und Internet-Nutzungszeit in der Population	r z. B. linearer Zusammenhang (bivariater Korrelationskoeffizient) zwischen TV-Nutzungszeit und Internet-Nutzungszeit in der Stichprobe

gewählter Schichtungsmerkmale (z. B. Einkommen) noch präziser ab als eine einfache Zufallsstichprobe.

Hat man sich anstelle einer Vollerhebung für eine **Teilerhebung** bzw. **Stichprobenerhebung** entschieden, so stellt sich – insbesondere bei populationsbeschreibenden Studien – die Frage, wie die Stichprobe zusammengesetzt werden soll, damit diese möglichst aussagekräftige Informationen über die Population liefert. Das Verhältnis zwischen Stichprobe und Population wird mit dem Qualitätsmerkmal der **Repräsentativität** („sample representativeness") beschrieben, wobei es sich jedoch streng genommen **nicht** um einen statistischen Fachbegriff handelt (Schnell, Hill, & Esser, 2008, S. 305f.). Idealerweise stellt eine repräsentative Stichprobe ein exaktes **Miniaturabbild der Population** dar: Die Merkmale und Effekte in der Population sollten sich im Kleinen 1:1 in der Stichprobe wiederfinden.

Repräsentativität einer Stichprobe – Die Repräsentativität einer Stichprobe („sample representativeness") gibt an, wie gut bzw. unverzerrt die Merkmalszusammensetzung in der Stichprobe die Merkmalszusammensetzung in der Population widerspiegelt. Zu unterscheiden sind dabei merkmalsspezifisch-repräsentative sowie global-repräsentative Stichproben.

Merkmalsspezifisch-repräsentative Stichprobe

Angenommen, eine Hochschule möchte ihre Lehrangebote im fächerübergreifenden **Studium Generale** stärker den aktuellen Studierendeninteressen anpassen und zu diesem Zweck eine schriftliche Befragung durchführen, die mit möglichst geringem Zeit- und Kostenaufwand umgesetzt werden soll. Zielpopulation ist hierbei die **Gesamtheit aller Studierenden**, die zum aktuellen Zeitpunkt an der betreffenden Hochschule eingeschrieben sind (z. B. $N = 6\,294$ Studierende an der TU Ilmenau im WS 2010/2011). Die Hochschulleitung geht in diesem fiktiven Beispiel davon aus, dass die Themeninteressen der Studierenden sich stark in Abhängigkeit vom Studienfach unterscheiden und strebt somit eine Stichprobe an, die hinsichtlich der Verteilung der Studienfächer der Population entspricht (merkmalspezifisch-repräsentative Stichprobe im Hinblick auf das **Merkmal Studienfach**). Zunächst wird also der Hochschulstatistik die Fächerverteilung in der Population entnommen (z. B. von den $N = 6\,294$ Studierenden in der Population sind 62 % in den Ingenieurwissenschaften, 13 % in Mathematik und Naturwissenschaften, 13 % in Rechts-, Wirtschafts- und Sozialwissenschaften und 12 % in Sprach- und Kulturwissenschaften eingeschrieben). Dieselbe Verteilung wird nun im Sinne eines Quotenverfahrens für die Stichprobe zugrunde gelegt, wobei der Stichprobenumfang aus forschungsökonomischen Gründen auf $n = 400$ Studierende begrenzt wird.

Merkmalsspezifisch-repräsentative Stichprobe – Eine merkmalsspezifisch-repräsentative Stichprobe liegt vor, wenn die Zusammensetzung der Stichprobe hinsichtlich einiger relevanter Merkmale der Populationszusammensetzung entspricht. Dies wird typischerweise durch eine nicht-probabilistische Quotenstichprobe erreicht.

Der **Quotenplan** für die Stichprobe von $n = 400$ sieht somit vor, 248 bzw. rund 250 Studierende aus den Ingenieurwissenschaften (ca. 62 %) und jeweils rund 50 Studierende (ca. 12–13 %) aus den anderen drei Fächergruppen zu befragen. Die entsprechende Anzahl an Paper-Pencil-Fragebögen wird in verschiedenen fachspezifischen Lehrveranstaltungen verteilt und nach dem Ausfüllen direkt wieder eingesammelt. Die merkmalsspezifisch-repräsentative Stichprobe basiert also auf einem vergleichsweise unaufwändigen nicht-probabilistischen Quotenverfahren. Dieses schützt leider nicht vor deutlichen **Verzerrungen der Stichprobe**. So kann im beschriebenen Beispiel das Studierenden-Sample in verschiedener Hinsicht von der Population abweichen: Beispielsweise könnte die Stich-



Abb. 9.2 Bei Verhaltensstichproben ist oft unklar, wie repräsentativ sie eigentlich sind. (Mod. nach Goscinny & Sempé, 1976. Mit freundlicher Genehmigung von Diogenes)

probe unbeabsichtigt einen viel geringeren Anteil an politisch interessierten Studierenden enthalten als die Population. Die Umfrage würde dann die Akzeptanz von und Nachfrage nach politikwissenschaftlichen Lehrangeboten im Studium Generale unterschätzen. Ob dies der Fall ist, weiß man nicht. Wird mit merkmalsspezifisch-repräsentativen Stichproben gearbeitet, so ist es wichtig, die für die Untersuchungsfrage wichtigen Populationsmerkmale zu identifizieren, anhand derer dann die Quotierung vorgenommen werden soll. Zudem muss die Tatsache, dass eine merkmalsspezifisch-repräsentative Stichprobe von zahlreichen Verzerrungen hinsichtlich nicht erfasster Merkmale betroffen sein kann, in die Ergebnisinterpretation einbezogen werden.

Global-repräsentative Stichprobe

Leider kann man nur selten ausschließen, dass neben den bekannten, mit dem untersuchten Merkmal (hier: Themeninteressen der Studierenden bzgl. Lehrangebot im Studium Generale) kovariierenden Merkmalen (im Beispiel hier: Studienfach der Studierenden) auch noch andere Variablen das untersuchte Merkmal beeinflussen. Besonders gravierend ist dieses Problem bei Untersuchungen, die ein neuartiges Produkt, eine technische Innovation oder bisher unerprobte Vorschriften und Richtlinien evaluieren oder bei Studien, in denen eine Population gleichzeitig bzgl. vieler, sehr unterschiedlicher Merkmale beschrieben werden soll (sog. **Omnibus-Untersuchungen** = Mehr-Themen-Untersuchungen). Da man nicht weiß, welche weiteren Variablen mit dem untersuchten Merkmal zusammenhängen oder da man – wie bei Omnibus-Untersuchungen – davon ausgehen muss, dass sich mangelnde Repräsentativität der Stichprobe auf die vielen untersuchten Merkmale in unterschiedlicher Weise auswirkt, wird man eine Stichprobe bevorzugen, die der Population in möglichst allen Merkmalen entspricht: eine Stichprobe, die für die Population **global-repräsentativ** ist.

Global-repräsentative Stichprobe – Eine global-repräsentative Stichprobe liegt vor, wenn die Zusammensetzung der Stichprobe in allen Merkmalen und Merkmalskombinationen der Populationszusammensetzung entspricht. Dies kann nur durch probabilistische Stichprobenverfahren sichergestellt werden, sofern gleichzeitig ein Mindeststichprobenumfang eingehalten wird.

Globale Repräsentativität kann nur durch eine probabilistische Stichprobenziehung sichergestellt werden, bei der die Stichprobe nach einem blinden Zufallsverfahren zusammengestellt wird. Zudem muss die Stichprobe eine Mindestgröße haben, da nur bei entsprechend großen Stichproben die Zufallsauswahl greift (Gesetz der großen Zahl) bzw. eine präzise statistische Schätzung von Populationsparametern ebenfalls vom Stichprobenumfang abhängt. Betrachtet man den Spezialfall von bevölkerungsrepräsentativen Stichproben (Zielpopulation: Wohnbevölkerung in Deutschland), so haben global-repräsentative Stichproben, die auf probabilistischen Auswahlverfahren basieren, typischerweise einen Mindestumfang von rund 2 000 Befragten.

Vielen Laien ist unklar, was „Repräsentativität“ im stichprobentheoretischen Sinne bedeutet und sie glauben oft, dass große Stichproben per se Repräsentativität garantieren (Abb. 9.2). Immer wieder wird das Etikett „repräsentativ“ auch pauschal zum Gütemerkmal von Wissenschaft erklärt. Eine „repräsentative Studie“, so wird suggeriert, sei besonders „seriös“ und „aussagekräftig“. Dabei bezieht sich Repräsentativität lediglich auf das Verhältnis von Population und Stichprobe, ist somit ein **Qualitätsmerkmal allein der Stichprobe**. Auch eine Studie, die mit einer in diesem Sinne perfekt global-repräsentativen Stichprobe operiert, kann wissenschaftlich wertlos sein, etwa wenn die interessierenden Merkmale unzureichend operationalisiert wurden, so dass die

Messwerte fehlerhaft sind. Umgekehrt stellt eine nicht-repräsentative Stichprobe keineswegs pauschal den Wert einer empirischen Studie infrage. Explorative und explanative Studien – darunter fast die gesamte Experimentalforschung (zum Experiment ▶ Abschn. 7.6.1) – sind von vorne herein gar nicht darauf angelegt, Merkmalsausprägungen in der gesamten Bevölkerung oder in anderen großen Populationen möglichst genau zu schätzen und arbeiten so gut wie nie mit repräsentativen bzw. probabilistischen Stichproben. Einer empirischen Studie darf also nur dann gerechtfertigt die „mangelnde Repräsentativität" ihrer Stichprobe(n) als Qualitätsmangel vorgeworfen werden, wenn sie als deskriptive Studie tatsächlich das primäre **Erkenntnisziel der Populationsbeschreibung** verfolgt (▶ Abschn. 7.5) und/oder wenn ihre Ergebnisse unzulässig auf die Allgemeinheit bezogen werden.

Die *Repräsentativität* ist ein *Qualitätsmerkmal*, das sich lediglich auf das Verhältnis zwischen Stichprobe und Population bezieht, nicht auf die wissenschaftliche Qualität der gesamten Studie. Global-repräsentative bzw. probabilistische Stichproben sind bei *populationsbeschreibenden Studien* notwendig, um die Ausprägung oder Verbreitung von Merkmalen in der Population präzise anhand der Stichprobendaten schätzen zu können.

Die meisten empirischen Studien in der quantitativen empirischen Sozialforschung, die nicht auf Populationsbeschreibung, sondern z. B. auf Theoriebildung oder Hypothesenprüfung angelegt sind, arbeiten mit nicht-global-repräsentativen bzw. nicht-probabilistischen Stichproben und haben dennoch einen hohen wissenschaftlichen Wert.

Wann immer Stichproben für „repräsentativ" erklärt werden, sollte man detailliert nach der Definition der Zielpopulation und dem Verfahren der Stichprobenauswahl fragen, um dann Möglichkeiten und Grenzen der Generalisierbarkeit der Stichprobenergebnisse auf die Populationsverhältnisse abzuwägen. Für derartige Beurteilungen ist in der Regel Sachkenntnis über den Untersuchungsgegenstand notwendig. Wenn z. B. die Lebenssituation von jungen Müttern anhand einer Stichprobe von Teilnehmerinnen von Mutter-Kind-Kursen untersucht wird, sind Aussagen über die Generalisierbarkeit der Befunde v. a. dann möglich, wenn man aus anderen Quellen zuverlässig weiß, inwiefern sich die Teilnehmerinnen von Mutter-Kind-Kursen von Müttern unterscheiden, die nicht an solchen Kursen teilnehmen (z. B. hinsichtlich Bildung, Familienstand, sozialer Integration oder Erwerbstätigkeit).

Dass eine Zufallssteuerung der Stichprobenauswahl aus der Population ab einem Mindeststichprobenumfang eher zu einer unverzerrten bzw. global-repräsentativen Stichprobe führt als eine nicht-zufallsgesteuerte Auswahl, ist unstrittig. Unklarheit besteht aber oft dahingehend, ob und inwiefern echte **Zufallsstichproben eine notwendige Voraussetzung für statistische Signifikanztests** sind. In Statistik-Lehrbüchern wird zu den gängigen statistischen Signifikanztests in der Regel die „Zufallsstichprobe" als eine der notwendigen Voraussetzungen zur sachgerechten Durchführung des Tests angegeben (Bortz & Schuster, 2010; Eid, Gollwitzer & Schmitt, 2013). Echte Zufallsstichproben sind jedoch, wie bereits angesprochen, in der akademischen Sozialforschung eher der Ausnahmefall. Ganze Forschungsfelder – wie etwa die psychologische Experimentalforschung – arbeiten nahezu ausschließlich mit nicht-probabilistischen Stichproben und werten ihre Daten dennoch mit parametrischen statistischen Signifikanztests (z. B. t-Test, Varianzanalyse, Regressionsanalyse) aus (Hager, 2013, S. 258).

Zu dieser Situation gibt es v. a. drei unterschiedliche Positionen:

1. **Pragmatische Position**: Bislang in der Forschungspraxis am häufigsten anzutreffen ist die – meist implizite (also nicht ausdrücklich diskutierte) – Annahme, die verwendeten nicht-probabilistischen Stichproben könnten bei der inferenzstatistischen Auswertung einfach **wie Zufallsstichproben behandelt** und somit mit herkömmlichen **parametrischen (verteilungsgebundenen) Signifikanztests** ausgewertet werden (Westermann, 2000). Diese Annahme ist natürlich problematisch. Zu ihrer Begründung wird zuweilen argumentiert, dass auch für eine Gelegenheitsstichprobe im Nachhinein eine **(fiktive) Inferenzpopulation** konstruiert werden kann, aus der die vorhandenen Fälle eine Zufallsauswahl bilden. Wird z. B. ein Experiment mit freiwilligen studentischen Versuchspersonen durchgeführt, die sich auf einen Teilnahmeaufruf gemeldet haben, so kann diese Gelegenheits- bzw. Selbstselektionsstichprobe (nicht-zufallsgesteuerte Auswahl aus der Population sämtlicher Studierenden) auch als Zufallsstichprobe aus der Inferenzpopulation speziell derjenigen Studierenden verstanden werden, die prinzipiell ebenfalls an genau diesem Experiment freiwillig teilnehmen würden. Ergebnisse dürften dann auch nur im Hinblick auf diese konstruierte Inferenzpopulation interpretiert werden.
2. **Mathematisch-statistische Position**: Aus mathematisch-statistischer Sicht wird dagegen betont, dass parametrische Signifikanztests nur dann sinnvoll sind, wenn tatsächlich eine zufallsgesteuerte Stichprobenauswahl erfolgt und die Daten somit als Ausprägungen von Zufallsvariablen gelten können. In Studien, in denen keine probabilistischen Stichproben vorliegen, dürfen dementsprechend parametrische Signifikanztests nicht eingesetzt werden, sondern statt-

dessen müssen **nicht-parametrische (verteilungsfreie) Signifikanztests** zum Einsatz kommen, etwa **Resampling-Verfahren** wie die **Randomisierungstests** (Bortz, Lienert, & Boehnke, 2008; Edgington & Onghena, 2007). Diese arbeiten nicht mit theoretischen Verteilungsannahmen über die Population, sondern bilden empirische Prüfverteilungen aus den vorliegenden Stichprobenergebnissen. Sie erlauben somit auch keinen Rückschluss auf eine Population, sondern nur eine kritische Bewertung des Stichprobenergebnisses. Manche Autoren fordern, dass die gängige statistische Auswertungspraxis angesichts überwiegend fehlender Zufallsstichproben weitgehend von parametrischen auf non-parametrische statistische Verfahren umgestellt wird.

3. **Erkenntnistheoretische Position**: Eine dritte, hier als „erkenntnistheoretisch" etikettierte Position plädiert dafür, die Frage nach der Angemessenheit statistischer Auswertungsverfahren und der Interpretierbarkeit statistischer Befunde nicht zu stark auf mathematisch-statistische Aspekte und die Frage der Zufälligkeit von Stichproben zu verengen. Die Verwendung parametrischer Signifikanztests bei nicht-zufallsgesteuerten Stichproben lässt sich nämlich insofern gut rechtfertigen, als **parametrische Signifikanztests** als **Approximationen der nicht-parametrischen Tests** (Randomisierungstests) gelten können (Hager, 2013, S. 258). Das eigentliche Problem liegt aus Sicht der erkenntnistheoretischen Position nicht bei der Frage, ob Daten aus nicht-probabilistischen Stichproben nun parametrisch oder lieber non-parametrisch ausgewertet werden sollten (was eben in der Regel zu vergleichbaren Resultaten führt), sondern bei der viel grundlegenderen Frage, inwiefern empirisch-statistische Ergebnisse (die sich auf **statistische Hypothesen** beziehen) begründet und nachvollziehbar der eigentlichen Zielsetzung hypothesenprüfender Studien dienen können. Diese Zielsetzung besteht gemäß der Wissenschaftstheorie des Kritischen Rationalismus darin, Aussagen über den **Bewährungsgrad von inhaltlichen bzw. substanziellen Hypothesen** und somit von **Theorien** zu treffen, die wiederum v. a. Aussagen über die Existenz von Ursache-Wirkungs-Relationen treffen (► Abschn. 2.2.4). Wenn also in einer psychologischen Experimentalstudie hinsichtlich der interessierenden abhängigen Variablen in hypothesenkonformer Richtung ein Mittelwertunterschied zwischen Experimental- und Kontrollgruppe gefunden wird, der sich im parametrischen t-Test (verstanden als Approximation an einen nicht-parametrischen Test) als statistisch signifikant erweist, so heißt dies nicht, dass dieser Effekt die **Höhe des Populationseffektes** schätzt (auf die genauen Populationsverhältnisse ist kein Rückschluss möglich, wenn es sich um nicht-probabilistische Stichproben handelt). Stattdessen ist die statistisch signifikante Effektgröße als grober Hinweis auf den **Bewährungsgrad der substanziellen Hypothese bzw. Theorie** zu interpretieren, die den gefundenen Gruppenunterschied kausal vorhergesagt hatte (Hager, 2013, S. 261). Wenn sich der Effekt in weiteren Studien replizieren lässt, steigt der Bewährungsgrad der Theorie. Fragen der Stichprobenziehung und statistische Auswertung sowie deren Voraussetzungen sind also stärker in eine **erkenntnistheoretische Gesamtreflexion des Forschungsprozesses einzubetten** (ausführlicher dazu Hager, 2013).

Zur Notwendigkeit global-repräsentativer bzw. probabilistischer Stichproben in der quantitiven Sozialforschung bleibt festzuhalten: **Populationsbeschreibende Untersuchungen** erfüllen nur dann ihren Zweck, wenn man sich vor der Stichprobenziehung darüber Klarheit verschafft, über welche Population Aussagen formuliert werden sollen. Erst nachdem die Merkmale der Zielpopulation präzise festgelegt sind, erfolgt die Entwicklung eines Stichprobenplans, der die zufallsgesteuerte Ziehung einer **global-repräsentativen Stichprobe** gewährleistet, auf deren Basis die Ausprägungen von Populationsparametern geschätzt werden. In der populationsbeschreibenden Umfrageforschung wird jedoch auch verstärkt über die Einsatzmöglichkeiten nicht-probabilistischer Stichproben diskutiert (Baker et al., 2013), insbesondere vor dem Hintergrund der hohen Kosten und oft mangelnden Realisierbarkeit (z. B. wegen Nichterreichbarkeit oder Teilnahmeverweigerung) von echten Zufallsstichproben. Anstelle globaler Repräsentativität von Stichproben im Hinblick auf Populationen wird als angemesseneres Kriterium z. B. in der Markt- und Meinungsforschung u. a. das **Kriterium der Zweckmäßigkeit** („fit for purpose") vorgeschlagen, d. h. es muss vorab definiert und dann im Forschungsprozess sichergestellt werden, dass Stichproben in der Weise gezogen (und ggf. noch gewichtet) werden, dass sie die Population in – hinsichtlich des Forschungsproblems – zweckmäßiger Weise abbilden (dabei kann es sich dann z. B. um lediglich merkmalsspezifisch-repräsentative bzw. nicht-probabilistische Stichproben wie Quotenstichproben handeln).

Bei **gegenstandserkundenden** und **hypothesenprüfenden Studien** kann sinnvoll mit nicht-probabilistischen Stichproben gearbeitet werden, denn hier sollen nicht Populationsparameter möglichst präzise geschätzt, sondern Theorien über Ursache-Wirkungs-Verhältnisse gebildet und hinsichtlich ihres Bewährungsgrades geprüft werden. Effektgrößen sind hier als Indikatoren des Bewährungsgrades zu deuten. Die wichtigsten probabilistischen sowie nicht-probabilistischen Stichprobenverfahren für quanti-

tative Studien werden in ▶ Abschn. 9.3 noch genauer vorgestellt.

In der qualitativen Forschung wird ein grundlegend anderes Stichprobenverständnis zugrunde gelegt, bei dem zufallsgesteuerte Auswahl sowie Repräsentativität im Hinblick auf eine exakt definierte Zielpopulation keine Rolle spielen, sondern es darauf ankommt, durch die gezielte Auswahl von einigen wenigen Fällen die Bandbreite der Erscheinungsformen des untersuchten sozialen Sachverhaltes möglichst erschöpfend abzubilden (▶ Abschn. 9.2).

9.2 Stichproben für qualitative Studien

Qualitative Studien arbeiten meist mit relativ **kleinen Stichproben** im ein- bis zwei-, selten im dreistelligen Bereich. So liegt die untersuchte Fallzahl bei qualitativen Doktorarbeiten in den Sozialwissenschaften oft zwischen 20 und 30 (Mason, 2010; für eine Übersicht der Stichprobengestaltung der in der Zeitschrift *Health Education and Behavior* publizierten qualitativen gesundheitswissenschaftlichen Studien s. Safman & Sobal, 2004). Da die differenzierte interpretative Rekonstruktion der einzelnen Fälle im Kontext ihrer Lebenswelt angestrebt wird und damit ein hoher Arbeitsaufwand für jeden einzelnen Fall verbunden ist, wären größere Stichprobenumfänge bei qualitativen Beobachtungs-, Befragungs- oder Dokumentenanalyse-Studien aus forschungsökonomischen Gründen gar nicht zu bewältigen. Während bei großen Stichproben eine Auswahl nach dem statistischen Zufallsprinzip der beste Garant für globale Repräsentativität des Samples ist, greift dieser Mechanismus bei kleinen Stichproben nicht. Hier würde eine blinde Zufallsauswahl zu verzerrten und wenig aussagekräftigen Stichproben führen. Nicht zuletzt deswegen hat sich im qualitativen Forschungsansatz anstelle der Zufallsauswahl die nicht-zufällige Auswahl – und zwar speziell die **bewusste bzw. absichtsvolle Auswahl von Fällen** („purposive/purposeful sampling") – als Stichprobenverfahren eingebürgert (Glaser & Strauss, 1999; Marshall, 1996). Damit ist gemeint, dass auf der Basis theoretischer und empirischer Vorkenntnisse gezielt solche Fälle in das Sample aufgenommen werden, die besonders aussagekräftig für die Fragestellung sind.

Wird keine Vollerhebung, sondern eine Teilerhebung durchgeführt, so sind im qualitativen Ansatz drei absichtsvolle Sampling-Strategien zu unterscheiden: Die Theoretische Stichprobe, die Fallauswahl gemäß einem qualitativen Stichprobenplan sowie die gezielte Auswahl bestimmter Arten von Fällen (Flick, 2010, S. 155ff.; Glaser & Strauss, 1999, S. 244ff.).

9.2.1 Theoretische Stichprobe

Das Verfahren der Theoretischen Stichprobenbildung („theoretical sampling"), bei dem am Ende eine sog. Theoretische Stichprobe („theoretical sample") vorliegt, wurde von den Soziologen Barney Glaser und Anselm Strauss (1967, 1999) im Rahmen ihrer **Grounded-Theory-Methodologie** (GTM) entwickelt, einem umfassenden qualitativen Forschungsansatz, der auf eine gegenstandsverankerte Theoriebildung hinausläuft (zur Theoriebildung gemäß GTM ▶ Abschn. 6.3.5; zur Datenauswertung mittels Kodierung von vorgefundenen Dokumenten oder Interviewtranskripten gemäß GTM ▶ Abschn. 10.6.2 „Kodierung gemäß Grounded-Theory-Methodologie"). Die Besonderheit der Theoretischen Stichprobe besteht darin, dass die Fallauswahl ganz bewusst so gesteuert wird, dass ein **maximaler theoretischer Erkenntniswert** resultiert. Dabei erfolgt die Fallauswahl schrittweise im Zuge der Datenerhebung und Datenauswertung: Anhand der bisherigen Ergebnisse der Studie wird jeweils entschieden, welche *weiteren* Fälle (Personen, Gruppen, Institutionen, Ereignisse) *als nächstes* in das Sample aufzunehmen sind. Stichprobenauswahl, Datenerhebung und Datenanalyse werden nicht linear, sondern **mehrfach zirkulär abgearbeitet**, was der Logik des qualitativen Forschungsprozesses entspricht. Bei der Theoretischen Stichprobenbildung können verschiedene und wechselnde Kriterien zur bewussten Auswahl von Fällen herangezogen werden. Die Stichprobenziehung wird im Idealfall erst dann abgeschlossen, wenn auf der Basis der bereits erarbeiteten Zwischenergebnisse der Studie der Eindruck entstanden ist, dass weitere Fälle keinen neuen Informationsgehalt für die Theoriebildung versprechen **(theoretische Sättigung)**. Aus forschungsökonomischen Gründen kann jedoch auch ein früheres Abschließen der Studie geboten sein, etwa wenn sich im Laufe der Erhebung herausstellt, dass viele weitere Kriterien für das untersuchte Phänomen wichtig sind, die notwendigen Ressourcen für eine entsprechend große Stichprobe jedoch nicht zur Verfügung stehen (Glaser & Strauss, 1999, S. 245). Damit ist die theoriebildende Aussagekraft der betreffenden Studie dann jedoch deutlich reduziert.

Die konkreten **Auswahlkriterien**, die bei der Theoretischen Stichprobenziehung zur Anwendung kommen, werden für jede einzelne Studie gemäß interessierendem (Untersuchungs-)Feld und Forschungsproblem individuell festgelegt. Allgemein lässt sich das Vorgehen so beschreiben:

» Auswahlstrategien beschreiben Wege, um ein Feld zu erschließen. Dies kann – ausgehend von **Extremfällen, negativen, kritischen oder abweichenden Fällen** –

> von seinen Rändern aus geschehen. Ein Feld kann auch von innen heraus erschlossen werden – ausgehend von **besonders typischen oder besonders entwickelten Fällen**. Es lässt sich von seiner vermuteten Struktur ausgehend erschließen – indem etwa möglichst **unterschiedliche Fälle** in ihrer Variationsbreite einbezogen werden. (Flick, 2010, S. 169f.; Hervorhebungen hinzugefügt; für qualitative Auswahlstrategien s. auch Patton, 1990)

Haben die Forschenden entschieden, welche Merkmale die als nächstes in das Sample aufzunehmenden Fälle (z. B. zwei weitere Fälle) haben sollen, so müssen die dafür passenden **Rekrutierungsstrategien** gewählt werden (z. B. Aushang in einer Institution, Vermittlung durch Kontaktpersonen im Feld, Aufruf in einem Internetforum, Anfrage im Bekanntenkreis etc.). Im Sinne der Wissenschaftlichkeit muss bei einer Theoretischen Stichprobe im Rahmen der jeweiligen Studie und ihres Forschungsproblems nachvollziehbar begründet werden, warum bei einer bestimmten Fallzahl schließlich von „theoretischer Sättigung" ausgegangen wird.

In der Forschungspraxis wird das Prinzip der theoretischen Sättigung bislang offenbar **nicht immer konsequent umgesetzt**. So finden sich überzufällig viele qualitative Studien (z. B. unter Doktorarbeiten), die angeben, mit Theoretischen Stichproben zu arbeiten, deren Stichprobenumfänge aber **exakt bei 10, 20 oder 30 Fällen** liegen (Mason, 2010). Diese „glatten" Zahlen deuten darauf hin, dass ein bestimmter, **überzeugend wirkender Stichprobenumfang** vorab festgelegt wurde und nicht das Prinzip der theoretischen Sättigung entscheidend war. Dieses würde nämlich häufiger auch zu ungeraden Stichprobenumfängen – etwa von 9, 27 oder 31 – führen. Für eine wissenschaftlich überzeugende Theoretische Stichprobenbildung ist es deswegen notwendig, die im Laufe des Forschungsprozesses erarbeiteten Auswahlkriterien für die jeweils nächsten Fälle inhaltlich zu begründen und insbesondere das Erreichen theoretischer Sättigung nachvollziehbar zu erläutern. Andernfalls droht die Gefahr, dass auch eine unreflektierte Auswahl beliebiger Fälle (also eine **Gelegenheitsstichprobe**; ▶ Abschn. 9.3.1) als „Theoretische Stichprobe" ausgegeben wird.

Studienbeispiel

Theoretische Stichprobe

Eine qualitative Studie aus dem Bereich der **Pflegeforschung** ging der Frage nach, inwiefern **Angehörige von Demenzpatienten** gegenüber deren Pflegeeinrichtungen eine aktive Fürsprecherrolle einnehmen (Legault & Ducharme, 2009).

Da es vor allem die Töchter sind, die ihre alten, demenzkranken Eltern pflegen, konzentrierte sich die Studie auf weibliche Befragungspersonen. Die Studie beschränkte sich zudem auf ein konkretes kanadisches Pflegeheim, das zur Kooperation bereit war.

Im ersten Schritt sollten für die Stichprobe **3 Töchter** rekrutiert werden, deren demenzkrankes Elternteil seit mehr als sechs Monaten und weniger als sechs Jahren in der Einrichtung lebte. Die Töchter wurden zunächst über das Pflegepersonal von der Studie informiert. Die teilnahmebereiten Töchter wurden dann von der Studienleitung telefonisch kontaktiert und zu einem Interviewtermin eingeladen.

Nach der Auswertung der ersten 3 Interviews wurden schrittweise weitere Untersuchungspersonen rekrutiert, dabei wurden dann – auf der Basis der sich entwickelnden Theorie – ganz gezielt Töchter mit unterschiedlichem **sozioökonomischem Status**, mit unterschiedlicher **Besuchshäufigkeit**, mit unterschiedlichem **Problemverhalten auf Seiten des demenzkranken Elternteils** und schließlich auch mit unterschiedlicher **Zufriedenheit mit der Pflegeeinrichtung** hinzugezogen – denn diese Merkmale wurden im Verlauf des Forschungsprozesses aus theoretischen Überlegungen im Hinblick auf die Fallauswahl für relevant erachtet. Dabei erwies es sich als besonders wichtig – und gleichzeitig besonders schwierig – Töchter in das Sample aufzunehmen, die mit dem Pflegeheim unzufrieden waren und nur selten zu Besuch kamen.

Theoretische Sättigung wurde aus Sicht der Forschenden erreicht, nachdem **14** Töchter interviewt worden waren.

9.2.2 Stichprobe gemäß qualitativem Stichprobenplan

Während bei der Theoretischen Stichprobenbildung die Stichprobe ergebnisoffen erst nach und nach auf der Basis theoretischer Überlegungen vervollständigt wird, arbeiten andere qualitative Stichproben mit einem vorab festgelegten **qualitativen Stichprobenplan** („qualitative sampling plan"). Dieses Vorgehen ähnelt der Quotenstichprobe im quantitativen Forschungsansatz. Doch während bei der Quotenstichprobe (▶ Abschn. 9.3.1) mittels Quotenplan die quantitative Zusammensetzung der Population nachgebildet wird, soll gemäß qualitativem Stichprobenplan die Fallauswahl am Informationsgehalt der Fälle ausgerichtet sein. Der qualitative Stichprobenplan wird bewusst so zusammengestellt, dass möglichst alle für den untersuchten Sachverhalt besonders wichtigen Merkmale und Merkmalskombinationen im Sample vorkommen. Bei Personenstichproben wird der Stichprobenplan meist an **soziodemografischen Kriterien** orientiert (z. B. Alter, Geschlecht, Beruf, Nationalität, Studiengang, Krankheitsbild etc.). Üblicherweise werden in einem qualitativen Stichprobenplan ca. **3 Merkmale** berücksichtigt. Je nach Anzahl von deren Ausprägungen ergibt sich daraus bereits eine Kreuztabelle mit recht vielen Zellen; ◘ Tab. 9.2). Pro Zelle werden meist **1–3 Fälle** untersucht. Es ist auch möglich, einzelne Zellen unbesetzt zu lassen, etwa wenn die entsprechende Merkmalskombination in der Praxis un-

Studienbeispiel

Stichprobe gemäß qualitativem Stichprobenplan

In einer qualitativen **Befragungsstudie zur Wahrnehmung technischen Wandels** wurde ein **Stichprobenplan mit 3 Merkmalen** der Befragungspersonen zugrunde gelegt: 1.) Herkunftsland (Ost- und Westdeutschland sowie Frankreich), 2.) Geschlecht (weiblich, männlich) sowie 3.) Beruf (Informatiker, Sozialwissenschaftler, Lehrer; Flick, 1996; Tab. 9.2). Somit ergibt sich anhand der Ausprägungen der drei nominalskalierten Merkmale im Stichprobenplan eine **Kreuztabelle mit $3 \times 2 \times 3 = 18$ Zellen**. Jede Zelle wurde mit 2 Fällen besetzt, so dass der **Gesamtstichprobenumfang bei 36 Befragungspersonen** liegt. Zur Umsetzung dieses Stichprobenplans müssen den Forschenden Rekrutierungsstrategien zur Verfügung stehen, um z. B. französische Informatikerinnen und Informatiker zu erreichen. Zudem müssen bei einer solchen internationalen Studie mögliche Sprachbarrieren überwunden werden (z. B. Einsatz französischsprachiger Interviewer), was normalerweise mit zusätzlichem Zeit- und Kostenaufwand verbunden ist (Tab. 9.2).

wichtig ist. Der **Gesamtstichprobenumfang** muss mit den vorhandenen Ressourcen (Zeit, Geld, Personal) zu bewältigen sein. Zudem müssen für die laut Stichprobenplan vorgesehenen Fälle passende **Rekrutierungsstrategien** existieren.

Im Rahmen einer **qualitativen Studie zur Krebsprävention** wurden 64 heterosexuelle nordamerikanische Männer in qualitativen Interviews zu ihrem Wissen und ihren Einstellungen gegenüber Prostatakrebs- und Darmkrebs-Vorsorgeuntersuchungen befragt (Winterich et al., 2009). Die Stichprobe wurde bewusst hinsichtlich Bildungsstand (gering, mittel, hoch), Hautfarbe (weiß, schwarz) sowie Wohnort (Land, Stadt) quotiert ($3 \times 2 \times 2 = 12$ Zellen), so dass Männer all dieser Gruppen im Sample etwa zu gleichen Teilen vertreten waren (ca. 5 Befragungspersonen pro Zelle). Die Rekrutierung vor Ort erfolgte jeweils über Kirchen, soziale Dienste, Männergruppen etc.

Der **Nachteil** der Vorabfestlegung der Samplestruktur besteht darin, dass auf diese Weise nur Unterschiede oder Gemeinsamkeiten zwischen den vorausgewählten Gruppen sichtbar gemacht werden können. Die Einschränkung der Vergleichsgruppen in der Stichprobe kann dazu führen, dass für die Theoriebildung ggf. wichtige Merkmalskombinationen, die in der Population existieren, übersehen werden. So könnte in der Beispielsstudie zur Technikwahrnehmung anstelle des Geschlechts möglicherweise das Lebensalter ein viel ausschlaggebenderer Faktor sein. Wenn verschiedene Altersgruppen jedoch bei der Vorabfestlegung der Samplestruktur nicht systematisch berücksichtigt wurden, können diese Effekte nach der Erhebung gegebenenfalls nicht mehr untersucht werden (z. B. weil dann zu wenige ältere Personen im Sample vorhanden sind). Das Arbeiten mit qualitativen Stichprobenplänen ist vor allem dann indiziert, wenn über den Untersuchungsgegenstand schon **genügend Vorwissen** existiert, so dass die Auswahl der Merkmale für den Stichprobenplan gut fundiert erfolgen kann. Entsprechende Vorabauswahlen gemäß einem qualitativen Stichprobenplan sind nicht nur bei Personenstichproben zu treffen, sondern z. B. auch bei Feldbeobachtungen (Auswahl von Beobachtungsorten und -zeiten) oder bei Stichproben von vorgefundenen Dokumenten (z. B. Auswahl von Zeitschriften-Covern, Briefen, Akten oder Sitzungsprotokollen).

Bei qualitativen Studien mit größeren Stichproben kommen anstelle von qualitativen Stichprobenplänen teilweise auch **Quotenpläne** zum Einsatz (zur Quotenstichprobe ▶ Abschn. 9.3.1).

9.2.3 Gezielte Auswahl bestimmter Arten von Fällen

Das Theoretical Sampling sowie die Stichprobenziehung auf der Basis eines qualitativen Stichprobenplans laufen beide auf eine bewusst heterogene Stichprobenstruktur hinaus: Eine Vielfalt an Fällen soll berücksichtigt werden, um das interessierende Untersuchungsfeld breit abzubilden. Demgegenüber werden auch oft Fragestellungen untersucht, die sich auf eine ganz **spezielle Zielgruppe** beziehen. In diesen Studien wird dann meist eine **gezielte Auswahl bestimmter Arten von Fällen** eingesetzt. Dabei sind zwei Varianten gezielter Stichproben zu unterscheiden:

1. **Homogene gezielte Stichproben**: Die Untersuchungseinheiten aus der interessierenden Zielgruppe werden über einen einzigen oder über wenige **Rekrutierungswege** angesprochen und es wird ein **relativ kleines Sample** zusammengestellt. Beispiele: Eine qualitative Studie ging der offenen Forschungsfrage nach, welche psychologischen Folgen eine schwere Kopfverletzung für jugendliche Patienten und ihre Familien hat. Dazu wurde zu einem einzigen Zeitpunkt in einer einzelnen Klinik in den USA ein Sample von 12 entsprechenden Patienten rekrutiert (Berland & Thomas, 1991). Eine andere Studie beschäftige sich mit der offenen Forschungsfrage, wie Überlebende des Atombombenabwurfs über Hiroshima und Nagasaki im Jahr 1945 die Ereignisse verarbeiteten. Dazu wurden im Jahr 2010 – also 65 Jahre später – insgesamt 8 Überlebende befragt, die mittlerweile in den USA ansässig waren (Knowles, 2011).
2. **Heterogene gezielte Stichproben**: Die Fälle aus dem interessierenden Feld werden breit gestreut über

Tabelle 9.2 Beispiel einer Vorabfestlegung der Samplestruktur für eine qualitative Stichprobe. (Mod. nach Flick, 2010, S. 157. Aus: Uwe Flick, „Qualitative Sozialforschung. Eine Einführung" Copyright © 1995, 2002 Rowohlt Taschenbuch Verlag GmbH, Reinbek bei Hamburg)

	Westdeutschland		Ostdeutschland		Frankreich		Summe
	weiblich	männlich	weiblich	männlich	weiblich	männlich	
Informatiker/in	2	2	2	2	2	2	12
Sozialwissenschaftler/in	2	2	2	2	2	2	12
Lehrer/in	2	2	2	2	2	2	12
Summe	6	6	6	6	6	6	**36**

unterschiedliche Rekrutierungswege angesprochen und es wird ein **relativ großes Sample** zusammengestellt. Auf diese Weise erhält man eine Auswahl von Fällen, die sich zwar alle hinsichtlich des Zielmerkmals ähneln, deren sonstige **interne Heterogenität** man jedoch im Vorfeld nicht genau kennt und auch nicht bewusst steuert, im Nachhinein dann jedoch im Zuge der Datenanalyse näher untersuchen kann (z. B. durch Typenbildung). Beispiel: Welche Erfahrungen lesbische Frauen mit dem Gesundheitssystem sammeln, wurde anhand einer qualitativen schriftlichen Befragung von 128 Respondentinnen erkundet, die sich selbst als lesbisch identifizierten und über unterschiedliche Onlinekanäle rekrutiert wurden (Bjorkman & Malterud, 2009). Die Fallzahl bei qualitativen Stichproben wird meist nicht mit „n =" angegeben.

9.3 Stichproben für quantitative Studien

Stichprobenumfänge für quantitative Studien sind im Allgemeinen deutlich größer als die für qualitative Studien und liegen meist im drei- bis vierstelligen Bereich oder noch höher. In der Forschungspraxis werden häufig nicht-probabilistische Stichproben mit eingeschränkter Repräsentativität genutzt (▶ Abschn. 9.3.1). Sofern inhaltlich sinnvoll und forschungsökonomisch realisierbar, kommen natürlich auch probabilistische Stichproben zum Einsatz, die der beste Garant für globale Repräsentativität eines Samples sind (▶ Abschn. 9.3.2).

9.3.1 Nicht-probabilistische Stichproben

Drei Typen nicht-probabilistischer Stichproben spielen in der quantitativen Forschung eine wichtige Rolle:

1. die **Gelegenheitsstichprobe**, die auf einer willkürlichen Auswahl von Fällen basiert, die gerade verfügbar sind;
2. die **Quotenstichprobe**, die auf einer bewussten Auswahl gemäß einem auf die Populationszusammensetzung abgestimmten Quotenplan basiert, sowie
3. die **Stichprobenziehung mittels spezieller Verfahren**, die vor allem für schwer erreichbare Populationen entwickelt wurden.

Nicht-probabilistische (nicht zufallsgesteuerte) Stichprobe – Bei einer **nicht-probabilistischen (nicht zufallsgesteuerten) Stichprobe** erfolgt die Auswahl der Untersuchungsobjekte **willkürlich** und/oder **bewusst/absichtsvoll** durch die Forschenden. Bei nicht-probabilistischen Stichproben ist unbekannt, welche Auswahlwahrscheinlichkeit die einzelnen Elemente der Population haben bzw. in welcher Weise und in welchem Ausmaß die Stichprobenzusammensetzung gegenüber einer Zielpopulation verzerrt ist.

Gelegenheitsstichprobe

Die **willkürliche Auswahl** bzw. **Gelegenheitsstichprobe** („convenience sample") ist ein in der quantitativen (und qualitativen) akademischen Sozialforschung sehr häufig anzutreffender Stichprobentyp. Denn Gelegenheitsstichproben sind definitionsgemäß mit dem geringsten Aufwand verbunden: Man bezieht einfach diejenigen Personen oder Objekte in die Studie ein, bei denen gerade eine **günstige Gelegenheit** besteht, sie ohne großen Aufwand zu untersuchen (Higginbottom, 2004). In der akademischen Forschungspraxis wird deswegen bekanntlich oft auf Studierende zurückgegriffen: Einen Fragebogen in der Vorlesung zu verteilen oder per Aushang im Institutsflur studentische Versuchspersonen für ein Experiment anzuwerben, ist die beste Gelegenheit, um schnell an Untersuchungspersonen zu kommen. Im Internet veröffentlichte Umfragen, die sich an beliebige Internetnutzende richten, oder Zeitungsanzeigen, mit denen Teilnehmende angeworben werden, haben ebenfalls den Charakter von Gelegenheitsstichproben. Auch qualitative Studien, bei denen willkürlich und ohne theoretische Reflexion einzelne Fälle untersucht werden, auf die man im Rahmen der Forschungstätigkeit im Feld gerade stößt, haben den Charakter einer Gelegenheitsstichprobe und stellen keine Theoretische Stichprobe dar (▶ Abschn. 9.2.1).

Gelegenheitsstichprobe oder Ad-hoc-Stichprobe – Bei einer Gelegenheitsstichprobe oder Ad-hoc-Stichprobe („convenience sample", „ad hoc sample", „accidental sample", „haphazard sample", „opportunistic sample"), manchmal auch: **anfallende/angefallene Stichprobe**, werden willkürlich Personen oder Objekte für die Stichprobe einer Untersuchung ausgewählt, die gerade zur Verfügung stehen oder leicht zugänglich sind (z. B. öffentliche Online-Befragung, Passantenbefragung).

Zuweilen werden willkürliche Auswahlen, bei denen nicht gezielt einzelne Personen angesprochen werden (z. B. durch personalisierte E-Mails), sondern ein **allgemeiner Teilnahmeaufruf** veröffentlicht wird (z. B. per Aushang, Presseanzeige, Newsletter oder Online-Banner) als **Selbstselektions-Stichproben** bezeichnet. Dahinter steht die Überlegung, dass bei nicht-personalisierten Einladungen zur Studienteilnahme überwiegend besonders motivierte Personen aus eigenem Antrieb teilnehmen (z. B. aufgrund ihres besonderen Interesses am Thema), während die persönliche Einladung durch ihren direkteren Aufforderungscharakter auch weniger motivierte Personen erreicht. Skeptische Personen können bei persönlicher Einladung ggf. zur Teilnahme überzeugt werden. Es ist jedoch zu beachten, dass eine gewisse Selbstselektion bei allen empirischen Studien vorliegt, da die Teilnahme aus forschungsethischen Gründen grundsätzlich freiwillig ist und es somit bei einem Teil der zu rekrutierenden Personen wegen fehlender Bereitschaft von vorne herein zu einer Teilnahmeverweigerung oder währenddessen zu einem Untersuchungsabbruch kommt (zur Bedeutung der Freiwilligkeit in der Forschungsethik ► Abschn. 4.1.1).

Eine Sonderform der **Selbstselektions-Stichprobe** stellt das **Access-Panel** dar. Ein Access-Panel besteht aus einer Menge an Personen, deren soziodemografische Merkmale und Kontaktdaten bekannt sind, und die sich aufgrund eines Teilnahmeaufrufs bereit erklärt haben, regelmäßig an wissenschaftlichen Studien teilzunehmen (Selbstselektion). Prinzipiell lassen sich durch aktive Rekrutierung auch probabilistische Panels zusammenstellen, in der Praxis sind aus ökonomischen Gründen aber oft nicht-probabilistische bzw. Selbstselektions-Panels anzutreffen. Markt- und Meinungsforschungsinstitute verfügen in der Regel über große, teilweise internationale Access-Panels, aus denen sie für aktuelle Studien jeweils Stichproben ziehen können. Access-Panels lassen sich im Rahmen der akademischen Forschung auch selbst aufbauen (z. B. Access-Panel befragungsbereiter Studierender). Access-Panels werden heute meist über Online-Plattformen verwaltet und müssen sorgfältig gepflegt werden (z. B. regelmäßige Kontaktierung und Incentivierung der Panel-Mitglieder = Panelisten), um Ausfälle (Panelmortalität) gering zu halten. Zudem müssen für ausgefallene Mitglieder regelmäßig neue Panelisten nachrekrutiert werden. Zum Umgang mit Access-Panels in der Markt-, Meinungs- und Sozialforschung existiert die ÖNORM ISO 26362:2010. Es sind bereits einige **Access-Panel-Anbieter** in Österreich und Deutschland nach dieser Norm zertifiziert (u. a. GfK Austria, Respondi, marketagent). Aus einem Access-Panel lassen sich nicht nur willkürliche Auswahlen ziehen, sondern durch entsprechende Steuerung der Einladungen auch Quotenstichproben zusammenstellen oder sehr spezielle Zielgruppen rekrutieren (sofern das Panel groß genug ist und das betreffende Merkmal in der Datenbank vorliegt, z. B. Stichprobe von Diabetespatienten).

Die Aussagekraft von Gelegenheitsstichproben im Rahmen der **quantitativen Forschung** ist naturgemäß eng begrenzt. Ihnen liegen in der Regel **keine** klare Definition der Zielpopulation und auch kein definierter Auswahlrahmen zugrunde. Stattdessen kann nur **indirekt auf die Inferenzpopulation zurückgeschlossen** werden: Greift man etwa für Interviews in der Fußgängerzone die ersten $n = 50$ Passanten heraus, die Teilnahmebereitschaft zeigen, so repräsentieren diese die Grundgesamtheit aller Personen, die an diesem Tag an diesem Ort mittags durch die Fußgängerzone schlendern und zu einem Interview bereit sind. Keineswegs handelt es sich bei dieser willkürlichen Auswahl um eine Zufallsauswahl der lokalen Einwohner. Denn für eine Zufallsauswahl müssten alle Bürgerinnen und Bürger des Ortes die gleiche (oder zumindest eine bekannte) Chance haben, in das Sample zu gelangen. Das willkürliche – **im Laienverständnis vermeintlich „zufällige"** – Herausgreifen von Passanten schließt aber systematisch viele Einwohnerinnen und Einwohner aus (z. B. alle, die tagsüber zu Hause oder im Betrieb arbeiten, in der Schule oder am Ausbildungsplatz sind, es sehr eilig haben etc.). Leider werden Gelegenheitsstichproben oft fälschlich als „Zufallsstichproben" bezeichnet. Wer für eine Studie, die nur auf einer Gelegenheitsstichprobe basiert, in Anspruch nimmt, eine „Zufallsstichprobe" erhoben zu haben, begeht einen schwerwiegenden „Etikettenschwindel". Denn die Gelegenheitsstichprobe weist die **niedrigste**, die Zufallsstichprobe dagegen die **höchste** Stichprobenqualität im Sinne globaler Repräsentativität auf.

Methodisch bestehen die Einschränkungen der Gelegenheitsstichprobe in der willkürlichen Gestaltung des „Auswahlrahmens" (der „Auswahlrahmen" liegt hier nicht als vollständige Liste der Elemente der Auswahlpopulation vor, sondern wird indirekt durch die Wahl der Untersuchungsumstände festgelegt) sowie in der willkürlichen Gestaltung des Auswahlprozesses. Beide Aspekte können durch entsprechende Vorüberlegungen verbessert werden, indem nicht nur die „beste" und damit die forschungsökonomisch sparsamste Gelegenheit ins Auge gefasst wird, sondern auch eine Erhöhung der Aussagekraft der Stichprobe im Sinne einer Annäherung an die relevante Zielpopulation berücksichtigt wird. Handelt es

sich bei der Zielpopulation um die Einwohnerschaft eines Ortes, so ist eben zu überlegen, ob die Fußgängerzone zur Mittagszeit der beste „Auswahlrahmen" ist, oder ob es vielleicht andere/ergänzende Rahmenbedingungen gibt, unter denen ein besserer Querschnitt der Bevölkerung des Ortes zu erreichen ist (z. B. Volksfest, Supermarkt). Zudem ist beim Herausgreifen der einzelnen Befragten sicherzustellen, dass keine grundlegende Verzerrung vorliegt (z. B. indem nur besonders freundlich und aufgeschlossen wirkende Personen, die alleine unterwegs sind, angesprochen werden). Hier kann die Festlegung eines „blinden" Auswahlschemas (z. B. angesprochen wird jede 10. Person, die den Supermarkt verlässt, oder jede 5. Person, die im Restaurant sitzt) entsprechende Verzerrungen aufgrund des persönlichen Eindrucks vermeiden.

Da man umgangssprachlich sagt, man habe „nach dem Zufallsprinzip" einige Personen (z. B. Kursteilnehmer, Passanten etc.) befragt, werden *Ad-hoc-Stichproben* leider allzu oft fälschlich als „Zufallsstichproben" bezeichnet. Bei Gelegenheitsstichproben handelt es sich jedoch um nicht-probabilistische Stichproben, es wird kein „blindes" statistisches Zufallsverfahren auf eine vollständige Populationsliste (oder einen anderen geeigneten Auswahlrahmen) angewendet, sondern die Forschenden wählen aus forschungsökonomischen Gründen willkürlich bestimmte Erhebungsbedingungen aus und greifen unter diesen Bedingungen willkürlich Untersuchungseinheiten heraus, was mit diversen (und überwiegend unbekannten) Verzerrungen im Hinblick auf die Populationsverhältnisse verbunden ist.

In der Forschungspraxis fehlen oft die finanziellen, personellen und zeitlichen Ressourcen für aufwändige probabilistische Stichproben. Anstatt in dieser Situation gänzlich auf empirische Forschung zu verzichten und somit das Themengebiet allein der Spekulation zu überlassen, wird pragmatisch mit Gelegenheitsstichproben gearbeitet. Dahinter steht die Überzeugung, dass eine – wenn auch schwache – Datenbasis immer noch besser ist als keinerlei empirische Basis. Damit das Arbeiten mit Gelegenheitsstichproben wissenschaftlichen Kriterien entspricht, müssen die engen Grenzen ihrer Aussagekraft deutlich gemacht und Überinterpretationen strikt vermieden werden. Insbesondere dürfen Daten aus Gelegenheitsstichproben nicht zur Populationsbeschreibung herangezogen werden. Für Theoriebildung und Hypothesenprüfung können sie indessen nützlich sein. Manche Autoren lehnen jedoch alle Gelegenheitsstichproben als unwissenschaftlich und für die Sozialforschung untauglich ab (z. B. Schnell, Hill, & Esser, 2008, S. 298).

Quotenstichprobe

Bei der Quotenstichprobe („quota sample") greift man ganz bewusst bzw. systematisch in den Auswahlprozess ein, so dass sich eine bestimmte Zusammensetzung der Stichprobe im Sinne **merkmalsspezifischer Repräsentativität** ergibt (► Abschn. 9.1.3). Typischerweise werden als Quotierungsmerkmale einige für das Untersuchungsthema besonders wichtige **soziodemografische Merkmale** herangezogen (z. B. Alter, Beruf, Bildungsstand, Geschlecht etc.). Indem man mittels Quotierungsverfahren z. B. eine Personenstichprobe so zusammenstellt, dass ihre Alters- und Bildungsverteilung derjenigen der Zielpopulation entspricht, erhofft man sich, dass auch die Stichprobenergebnisse hinsichtlich des Untersuchungsthemas (z. B. Mediennutzungsverhalten, Drogenkonsum) den Populationsverhältnissen besser entsprechen als z. B. bei einer Gelegenheitsstichprobe. Quotenstichproben sind nicht nur für Personenstichproben, sondern auch für andere Objektbereiche geeignet (z. B. Quotenstichprobe von Spielfilmen unterschiedlicher Filmgattungen und unterschiedlicher nationaler Herkunft).

Quotenverfahren – Die Stichprobenziehung nach dem **Quotenverfahren** („quota sampling") setzt voraus, dass man die Populationszusammensetzung hinsichtlich einiger, für die Untersuchungsfrage wichtiger soziodemografischer Merkmale kennt. Diese Merkmalsverteilung wird anhand des **Quotenplans („quota plan")** dann durch bewusste bzw. systematische Auswahl von **Personen oder Objekten** mit bestimmten Merkmalen (oder Merkmalskombinationen) in der Stichprobe nachgebildet. Die resultierende Quotenstichprobe („quota sample") kann als nicht-probabilistische Stichprobe zwar keine globale Repräsentativität, durch den Quotenplan aber merkmalsspezifische Repräsentativität (hinsichtlich der Quotierungsmerkmale) beanspruchen.

Voraussetzung für Quotenstichproben sind Kenntnisse darüber, a) welche soziodemografischen oder sonstigen einfachen manifesten Merkmale für das Untersuchungsthema besonders wichtig sind und b) wie die Ausprägungen dieser Merkmale in der Population verteilt sind. Wie man aus der Population aller Studierenden einer Hochschule eine Quotenstichprobe ziehen kann, wurde in ► Abschn. 9.1.3 „Merkmalsspezifisch-repräsentative Stichprobe" bereits anhand eines einzelnen Quotierungsmerkmals, nämlich der Fächergruppe, dargestellt. Quotenstichproben beschränken sich meist nicht auf ein Merkmal, sondern nutzen mehrere Quotierungsmerkmale gleichzeitig. In dem genannten Beispiel, in dem es um das Interesse der Studierenden an unterschiedlichen Themen im Studium Generale geht (► Abschn. 9.1.3 „Merkmalsspezifisch-repräsentative Stichprobe"), könnte neben der Fächergruppe auch die Herkunft der Studierenden für die Stichprobenzusammensetzung wichtig sein. Innerhalb der einzelnen Fächergruppen sind deswegen noch anteilsmäßig internationale Studierende zu berücksichti-

gen. Man spricht von **kombinierten Quoten**, wenn die Quotierungsvorgaben Merkmalskombinationen betreffen (z. B. genau 7 % der Stichprobenmitglieder sollen internationale Studierende sowie gleichzeitig Studierende der Ingenieurwissenschaften sein). Dementsprechend würden **unabhängige Quoten** (Randquoten) jeweils separate Vorgaben für die einzelnen Quotierungsmerkmale machen (z. B. insgesamt 62 % der Stichprobenmitglieder sollen Studierende der Ingenieurwissenschaften und insgesamt 10 % internationale Studierende sein – für die Merkmalskombination wird keine exakte Vorgabe gemacht). Kombinierte Quotenpläne bilden die Populationsverhältnisse besser ab, erhöhen aber den Rekrutierungsaufwand. Betrachtet man den kombinierten Quotenplan in Tab. 9.3, so fällt auf, dass bei einem Stichprobenumfang von $n = 400$ Studierenden manche Zellen eine sehr geringe Besetzung aufweisen. Auf der Basis von nur 4 internationalen Studierenden der Fächergruppe Mathematik und Naturwissenschaften wird man kaum tragfähige Schlussfolgerungen über die spezifischen Themenwünsche dieser Gruppe ziehen können. Eine Vergrößerung der Zellbesetzung ist auf zwei Wegen möglich:

1. Man erhöht den **Stichprobenumfang**: Werden statt 400 insgesamt 1 200 Studierende befragt, steigt bei einer Quote von 1 % die Zellenbesetzung von 4 auf 12. Dies ist jedoch mit großem Zusatzaufwand verbunden.
2. Man arbeitet anstelle der in Tab. 9.3 verwendeten kombinierten **proportionalen Quotenstichprobe** mit einer **nicht-proportionalen Quotenstichprobe**. Das bedeutet, dass man den prozentualen Anteil der internationalen Studierenden in der Stichprobe höher ansetzt als in der Population (z. B. 5 % statt 1 % in bestimmten Fächergruppen), um – bei konstantem Stichprobenumfang von $n = 400$ – genügend Befragte aus dieser Gruppe im Sample zu haben. Vor der quantitativen Datenanalyse kann eine nicht-proportionale Quotenstichprobe im Zuge der Datenaufbereitung durch Gewichtungsverfahren wieder den Populationsverhältnissen angeglichen werden (▶ Abschn. 11.4.4).

In der **Online-Forschung** werden Quotenstichproben neuerdings als sog. **River Samples** erstellt (DMS Research, 2008): Dazu werden Umfrageeinladungen möglichst breit im Internet gestreut. Aus dem ständigen **Zustrom an neuen Internet-Nutzenden**, die auf die Studieneinladungen stoßen (deswegen River Sampling, River = Fluss, Strom), entscheiden sich einige für die Studienteilnahme. Würde man alle Freiwilligen nun im Rahmen einer einzigen Studie untersuchen, wäre das eine einfache Gelegenheits- oder Selbstselektions-Stichprobe. Stattdessen werden beim River-Sampling die Teilnehmenden aber zunächst nach ihren soziodemografischen Angaben gefragt und auf dieser Basis vom Umfrage-Server gemäß entsprechende Quoten-Plänen verschiedenen Studien zugeordnet, die dann mit Quoten-Stichproben arbeiten. Im Unterschied zur Nutzung von **Quoten-Stichproben aus Online-Access-Panels** (die aus selbst-selektierten dauerhaften Panelisten bestehen), sollen River-Panels durch die ständig „frischen" Befragten aussagekräftiger sein. Systematische Methodenstudien, die die verschiedenen Stichprobentechniken vergleichen, stehen jedoch noch aus.

Stichprobenverfahren für spezielle Populationen

Um im Rahmen quantitativer (oder auch qualitativer) Studien ganz spezielle Populationen zu untersuchen, bei denen bereits das Auffinden einzelner Populationsmitglieder schwierig sein kann, wird zur Rekrutierung und Zusammenstellung der Stichprobe am häufigsten das sog. **Schneeballverfahren** eingesetzt, bei dem die sozialen Netzwerke der Populationsmitglieder genutzt werden („snowball sampling"; Goodman, 1961). Eine Weiterentwicklung des Schneeballverfahrens in Richtung erhöhter Repräsentativität stellt das **„Respondent Driven Sampling"** (RDS) dar. Für schwer erreichbare Populationen, deren Mitglieder untereinander nicht besonders eng vernetzt sind, bietet sich die Methode des **„Targeted Samplings"** an, bei der man bevorzugte Aufenthalts- oder Wohnorte der Mitglieder der Zielpopulation identifiziert und dann systematisch vor Ort rekrutiert.

Schneeballverfahren – Die Stichprobenziehung nach dem **Schneeballverfahren** („snowball sampling", „chain sampling", „chain-referral sampling", „referral sampling") ist geeignet für Populationen, die für die Forschenden schwer erreichbar, deren Mitglieder untereinander jedoch gut vernetzt sind. Hierbei werden einzelne Mitglieder aus der Population gebeten, über ihre persönlichen sozialen Netzwerke weitere Untersuchungspersonen zu rekrutieren. Die **Schneeballstichprobe** („snowball sample") kann als nicht-probabilistische Stichprobe keine Repräsentativität beanspruchen.

Beim Schneeballverfahren wird auf die **persönlichen sozialen Netzwerke** einzelner Populationsmitglieder zurückgegriffen, die von den Forschenden gebeten werden, weitere Untersuchungspersonen zu benennen bzw. zu rekrutieren. Je nach Datenerhebungsmethode werden im Schneeballverfahren entweder zunächst nur Kontaktdaten gesammelt (z. B. für spätere Interviews) oder es wird gleich ein Fragebogeninstrument weitergegeben. Ein Schneeballverfahren kann einstufig erfolgen, indem eine Gruppe von Populationsmitgliedern gebeten wird, im eigenen Bekanntenkreis einmalig eine bestimmte Anzahl an Untersuchungspersonen zu rekrutieren. Es ist aber auch möglich, dass die über die Kontaktpersonen angeworbenen Untersuchungspersonen ihrerseits gebeten werden, weitere Populationsmitglie-

Tabelle 9.3 Zusammensetzung der Population der Studierenden der TU Ilmenau im WS 2010/2011 laut Hochschulstatistik sowie Quotenplan für eine Stichprobe im Umfang von $n = 400$ nach den Quotierungsmerkmalen Fächergruppe und Herkunft der Studierenden

Zusammensetzung der Population ($N = 6439$)				Kombinierter proportionaler Quotenplan ($n = 400$)	
Fächergruppe	Anteil	Herkunft	Anteil	Anzahl	Anzahl
	[%]		[%]		
Ingenieurwissenschaften	62	DE	55	248	220
		INT	7		28
Mathematik und Naturwissenschaften	13	DE	12	52	48
		INT	1		4
Rechts-, Wirtschafts- und Sozialwissenschaften	13	DE	12	52	48
		INT	1		4
Sprach- und Kulturwissenschaften	12	DE	11	48	44
		INT	1		4
Gesamt	100		100	400	400

DE Deutschland; *INT* International

der anzuwerben. Auf diese Weise wächst die Zahl der Untersuchungsteilnehmenden wie ein **rollender Schneeball**.

Beim Schneeballverfahren kann es in Abhängigkeit von der Auswahl der ersten Kontaktpersonen und ihrer Netzwerkwerkzusammensetzung zu zahlreichen Verzerrungen kommen. Die rekrutierenden Populationsmitglieder müssen möglichst gut instruiert werden, welche und wie viele Personen sie auswählen (hier können bei Bedarf auch Quoten vorgegeben werden), was sie über das Forschungsprojekt mitteilen und wie sie den Datenschutz sichern sollen. Wenn die Schneeballrekrutierung in einer kleinen Population unprofessionell abläuft, kann dies zu Misstrauen, Widerstand oder gar öffentlichem Protest (z. B. kritische Internet- und Pressebeiträge über die Studie) führen.

Studien, die auf Schneeballstichproben basieren, untersuchten z. B.

- britische Bürger/innen, die gegen Mobilfunkmasten politisch protestieren (Drake, 2011),
- israelische Computerhacker/innen (Turgeman-Goldschmidt, 2005),
- afroamerikanische Crack-Konsumentinnen mit Kindern (Lam et al., 2007) oder
- kenianische Prostituierte, die unter Alkoholabhängigkeit leiden (Chersich et al., 2007).

Bei der Schneeballstichprobe handelt es sich um eine willkürliche, nicht-probabilistische Stichprobe, deren Repräsentativität für die Zielpopulation unbekannt ist.

Es gibt jedoch eine Weiterentwicklung des Verfahrens, mit deren Hilfe die Repräsentativität deutlich gesteigert werden kann: Der als **Respondent Driven Sampling** (RDS) eingeführte Ansatz arbeitet mit einem bestimmten Incentivierungsverfahren, dokumentiert den Rekrutierungsprozess detailliert und nutzt ein mathematisches Modell zur Gewichtung der rekrutierten Personen (Heckathorn, 1997; http://www.respondentdrivensampling.org). Bestandteil des RDS-Ansatzes ist es, dass die rekrutierten Personen jeweils immer nur eine begrenzte Zahl weiterer Personen rekrutieren dürfen und für jede tatsächlich rekrutierte Person incentiviert werden. Zudem wird erfragt, wie viele Personen aus der Zielpopulation ein Stichprobenmitglied jeweils kennt, um Auswahlwahrscheinlichkeiten zu schätzen. Vorliegende Studien mit Stichprobenziehung gemäß RDS beziehen sich z. B. auf spezielle Populationen wie Drogenkonsumenten oder Jazzmusiker. Der Ansatz lässt sich auch im Rahmen der Online-Forschung einsetzen (Wejnert & Heckathorn, 2008).

Respondent Driven Sampling – Die Stichprobenziehung nach der Methode des Respondent Driven Sampling (RDS) stellt eine Weiterent-

wicklung der Schneeballstichprobe dar und ist wie diese geeignet für Populationen, die für die Forschenden schwer erreichbar, deren Mitglieder untereinander jedoch gut vernetzt sind. Das Respondent Driven Sampling ist wesentlich aufwändiger als das Schneeballverfahren, garantiert jedoch durch eine bessere Kontrolle der Rekrutierungswege eine höhere Repräsentativität des Samples.

Der **Targeted-Sampling-Ansatz** (Watters & Biernacki, 1989) ist indiziert, wenn Stichproben aus schwer erreichbaren Populationen gezogen werden sollen, deren Mitglieder nicht stark untereinander vernetzt sind. Um z. B. eine Stichprobe von Drogenkonsumenten zu ziehen, werden gemäß Target Sampling in einer oder mehreren Städten zunächst typische Aufenthaltsorte der entsprechenden Drogenszene identifiziert und aufgelistet (z. B. durch eigene Beobachtung, Hinweise von Szenemitgliedern, Auskünfte der Polizei etc.). Dann wird ein Stichprobenplan entworfen, dem zu entnehmen ist, wie viele Personen pro Treffpunkt zu rekrutieren sind (Peterson et al., 2008). Um eine Stichprobe von Paaren zu ziehen, in denen es zu häuslicher Gewalt kommt, wurden gemäß dem Ansatz des „Targeted Neighborhood Sampling" zunächst Stadtviertel identifiziert, aus denen besonders häufig Notrufe wegen häuslicher Gewalt bei der Polizei eingehen. In den betreffenden Vierteln wurden dann Flyer in den Briefkästen verteilt, um Betroffene zu rekrutieren (Christopher, Pflieger, Canary, Guerrero, & Holtzworth-Munroe, 2008).

Targeted Sampling – Die Stichprobenziehung nach der Methode des Targeted Sampling ist besonders geeignet für schwer erreichbare Populationen, deren Mitglieder untereinander nicht besonders eng vernetzt sind. Sie stellt eine ortsbasierte Alternative zu netzwerkbasierten Verfahren wie dem Schneeballverfahren und dem Respondent Driven Sampling dar.

9.3.2 Probabilistische Stichproben

Probabilistische Stichproben basieren auf einem **Auswahlrahmen** („sampling frame") und einem darauf angewendeten statistischen Zufallsverfahren. Sie sind aufwändig und teuer, aber insbesondere für populationsbeschreibende Untersuchungen unerlässlich, in denen es auf globale Repräsentativität der Stichprobe ankommt. Zufallsgesteuerte Stichproben werden am häufigsten im Rahmen der Meinungs- und Wahlforschung, aber auch in der Gesundheitsforschung sowie der Politikwissenschaft und Soziologie eingesetzt, wo es eben um die präzise Feststellung von Merkmalsausprägungen in großen Grundgesamtheiten wie etwa der Wohnbevölkerung ganzer Länder geht. Man unterscheidet vier Arten der probabilistischen Stichprobe: die **einfache Zufallsstichprobe**, die **geschichtete Zufallsstichprobe**, die **Klumpenstichprobe** und die **mehrstufige Stichprobe**.

Der **Prozess der probabilistischen Stichprobenziehung** muss zwingend mit einer präzisen Definition der Zielpopulation („target population") sowie der Spezifizierung des Auswahlrahmens beginnen (▶ Abschn. 9.1.2). Dann folgt die Stichprobenkonstruktion im Sinne der Wahl der Auswahlmethode („sampling method"), der Festsetzung des Stichprobenumfangs („sample size") und Festlegung des Stichprobenplans („sampling plan"). Dabei sind Informationen über die Zusammensetzung der Zielpopulation zu berücksichtigen, aber auch forschungsökonomische Aspekte im Blick zu behalten. Nach oder gleichzeitig mit der Rekrutierung der Untersuchungspersonen oder Untersuchungsobjekte gemäß Stichprobenplan erfolgt die Datenerhebung und damit die eigentliche praktische Stichprobenziehung.

Einfache Zufallsstichprobe

Die Ziehung einer **einfachen Zufallsstichprobe** („simple random sample") setzt voraus, dass jedes zur finiten Population gehörende Untersuchungsobjekt einzeln identifizierbar ist bzw. dass als Auswahlrahmen eine vollständige Populationsliste vorliegt. So ist das Einwohnermelderegister ein vollständiger Auswahlrahmen für die lokale Wohnbevölkerung. Ebenso stellt das Immatrikulationsverzeichnis einen vollständigen Auswahlrahmen der Zielpopulation aller Studierenden einer Hochschule dar. Für die Qualität der Stichprobe ist es von Bedeutung, dass die Entscheidung darüber, welche Untersuchungsobjekte zur Stichprobe gehören und welche nicht, ausschließlich vom Zufall abhängt. Besteht eine finite Population aus N Objekten und sollen sich in der Stichprobe n Objekte befinden, können gemäß dem Binomialkoeffizienten insgesamt (Gl. 9.1)

$$C = \binom{N}{n} = \frac{N!}{n! \cdot (N-n)!} \tag{9.1}$$

verschiedene Stichproben bzw. Kombinationen (C) gezogen werden (der mittlere Ausdruck wird gelesen als „N über n"; Erläuterung dieser Formel durch ▶ Chancengleichheit für Skatspieler – Ziehung einer einfachen Zufallsstichprobe). Die Menge aller zu einer Population gehörenden Stichproben nennt man **Stichprobenraum** („sample space"). Setzen wir voraus, dass die Wahrscheinlichkeit, in die Stichprobe aufgenommen zu werden, für jedes Untersuchungsobjekt gleich ist („equal probability selection method", EPSEM), hat jede der C verschiedenen Stichproben die gleiche Auswahlwahrscheinlichkeit. Derartige Stichproben werden Zufallsstichproben genannt.

Idealerweise geht man bei der Entnahme einer einfachen Zufallsstichprobe wie folgt vor: Die gesamte Populationsliste bzw. der vollständige Auswahlrahmen wird von 1 bis N durchnummeriert. Mithilfe von **Zufallszahlen**

Chancengleichheit für Skatspieler – Ziehung einer einfachen Zufallsstichprobe

$N = 5$ Skatspieler (wir nennen sie einfachheitshalber A, B, C, D und E) treffen sich in einem Lokal und wollen Skat spielen. Da für eine Skatrunde jedoch nur $n = 3$ Spieler benötigt werden, müssen zwei Personen zusehen. Man einigt sich darauf, die drei aktiven Spieler auszulosen. Alle fünf Anwesenden schreiben ihren Namen auf einen Zettel und werfen den zusammengefalteten Zettel in ein leeres Bierglas (in der Statistik wird hierfür – zumindest symbolisch – eine Urne verwendet). Nach gründlichem Durchmischen werden nacheinander die Zettel B, E und D gezogen. Die Skatrunde steht fest.
Zunächst einmal ist es einleuchtend, dass sich an dieser Runde nichts geändert hätte, wenn die gleichen Zettel in einer anderen Reihenfolge, etwa E, D und B gezogen worden wären. Alle möglichen $3! = 3 \cdot 2 \cdot 1 = 6$ verschiedenen Reihenfolgen (BDE, BED, DBE, DEB, EBD und EDB) bilden die gleiche Stichprobe. Insgesamt hätten sich zehn verschiedene Skatrunden bilden können.

$$\binom{N}{n} = \frac{N!}{n! \cdot (N-n)!} = \frac{5!}{3! \cdot (5-3)!}$$
$$= \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1) + (2 \cdot 1)} = 10$$

Ein Spieler fragt sich nun, ob dieses Verfahren gerecht sei, ob jeder der fünf anwesenden Skatspieler tatsächlich die gleiche Chance hatte, in die Runde aufgenommen zu werden. Er argumentiert in folgender Weise: Bei der ersten Zettelentnahme besteht für jeden der fünf Spieler eine Auswahlwahrscheinlichkeit von $1/5$. Steht aber der 1. Spieler fest, erhöht sich für die verbleibenden Spieler bei der 2. Zettelentnahme die Auswahlwahrscheinlichkeit auf $1/4$ und für die 3. Zettelentnahme auf $1/3$. Also hat offensichtlich nicht jeder Spieler die gleiche Chance, in die Skatrunde aufgenommen zu werden.
Diese Argumentation ist unvollständig. Es wurde übersehen, dass es sich bei den Auswahlwahrscheinlichkeiten der 2. und 3. Ziehung um sog. **bedingte Wahrscheinlichkeiten** handelt. Die Wahrscheinlichkeit, bei der 2. Ziehung ausgewählt zu werden, beträgt $1/4$, vorausgesetzt, man wurde in der 1. Ziehung nicht berücksichtigt. Diese Wahrscheinlichkeit hat den Wert $4/5$. Die Wahrscheinlichkeit, dass jemand in der 1. Ziehung nicht ausgewählt und in der 2. Ziehung ausgewählt wird, lautet nach dem Multiplikationstheorem der Wahrscheinlichkeiten

$$\frac{N-1}{N} \cdot \frac{1}{N-1} = \frac{1}{N} = \frac{1}{5}. \qquad (9.2)$$

Für die 3. Ziehung sind die Wahrscheinlichkeiten, sowohl bei der 1. als auch bei der 2. Ziehung nicht ausgewählt worden zu sein, zu beachten. Für diese Wahrscheinlichkeiten erhält man die Werte $4/5$ und $3/4$. Damit ergibt sich zusammengenommen für die 3. Ziehung eine Trefferwahrscheinlichkeit von

$$\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{1}{N-2} = \frac{1}{N} = \frac{1}{5}$$

Alle fünf anwesenden Skatspieler hatten damit beim **Ziehen ohne Zurücklegen** die gleiche Chance, in die Skatrunde aufgenommen zu werden.

Tabelle 9.4 Zufallszahlen. (Aus The Rand Corporation, 1955)

88473	86062	26357
00677	42981	84552
25227	51260	14800
15386	68200	21492
42021	40308	91104
63058	06498	49339
32548	69104	89073
03521	52177	24816
39975	90626	35889
58252	56687	60412

(diese sind Tabellenwerken oder Zufallszahlengeneratoren wie http://www.random.org zu entnehmen) werden aus dieser Liste n Nummern bzw. die dazugehörenden Untersuchungsobjekte ausgewählt (zur Bestimmung von Zufallszahlen Billeter, 1970, S. 15ff.). Soll z. B. aus einer Population von $N = 8\,000$ Untersuchungsobjekten eine Stichprobe des Umfangs $n = 100$ gezogen werden, würde man mit den in Tab. 9.4 aufgeführten Zufallszahlen die Untersuchungsobjekte auswählen, wobei von den fünfstelligen Zufallszahlen dann nur jeweils maximal 4 Ziffern verwendet werden, da der Stichprobenumfang ebenfalls vierstellig ist.

Unter Berücksichtigung der **ersten 4 Ziffern** der Zufallszahlen in der ersten Spalte müsste als erstes Untersuchungsobjekt die Nummer 8 847 ausgewählt werden. Da die Population jedoch nur 8 000 Elemente enthält, wird diese Nummer ausgelassen. Das erste Untersuchungsobjekt hat dann die Nummer 67, das zweite die Nummer 2 522 und so fort. Eine gleichwertige Zufallsstichprobe würde resultieren, wenn man die Auswahl z. B. anhand der letzten 4 Ziffern der zweiten Zahlenkolonne oder anderer Viererkombinationen von Einzelzahlen zusammengestellt hätte. Jede beliebige Auswahl von Zufallszahlen garantiert eine Zufallsstichprobe, vorausgesetzt, eine bereits ausgewählte Zufallszahl wird nicht wieder verwendet. Man nennt dies eine Stichprobenentnahme **ohne Zurücklegen**.

Einfache Zufallsstichprobe – Man zieht eine einfache Zufallsstichprobe („simple random sample“), indem man aus einer vollständigen Liste aller Objekte der finiten Zielpopulation (Auswahlrahmen, „sample frame“) nach einem „blinden“ statistischen Zufallsprinzip (z. B. mittels Zufallszahlen) eine Anzahl von Objekten auswählt, wobei die Auswahlwahrscheinlichkeiten aller Objekte gleich groß sein müssen.

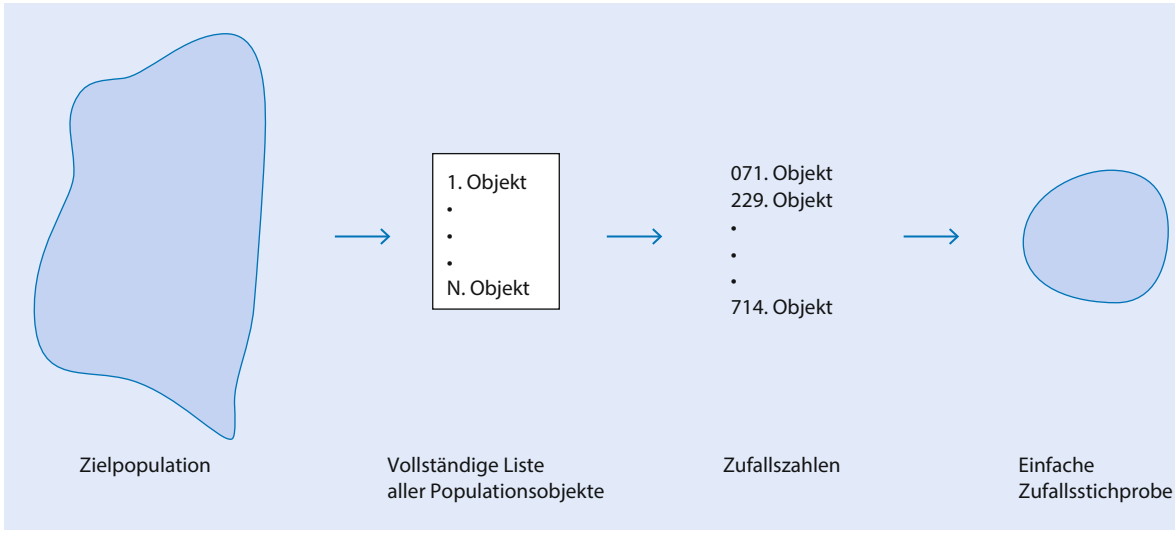


Abb. 9.3 Ziehung einer einfachen Zufallsstichprobe

An dieser Stelle könnte man vermuten, dass sich die Auswahlwahrscheinlichkeiten mit sukzessiver Entnahme von Untersuchungsobjekten ändern, dass sie also nicht – wie gefordert – für alle Untersuchungsobjekte mit einem „**Auswahlsatz**“ von n/N konstant sind. Wächst die Wahrscheinlichkeit, dass ein bestimmtes Untersuchungsobjekt ausgewählt wird, nicht mit fortschreitender Stichprobenentnahme? Dass dem nicht so ist, erläutert ▶ Chancengleichheit für Skatspieler – Ziehung einer einfachen Zufallsstichprobe (Abb. 9.3).

Die Ziehung einer einfachen Zufallsstichprobe setzt voraus, dass jedes Untersuchungsobjekt einer finiten Population erfasst ist und nach dem **Zufallszahlenprinzip** – oder einem anderen Auswahlverfahren, das ebenfalls eine zufällige Auswahl garantiert – ausgewählt werden kann. Eine praktisch häufig eingesetzte Alternative zum Zufallszahlenprinzip ist eine **systematische Auswahl**, bei der wiederum „blind“ (d. h. ohne Ansehen der konkreten Untersuchungseinheit) **jede *n*-te Einheit** ausgewählt wird. Dieses Verfahren hat den Vorteil, dass es auch auf infinite Populationen anwendbar ist. Beispiel: Soll eine Zufallsstichprobe aus der (ständig wachsenden) Population aller Besucher einer konkreten Website online befragt werden, so kann man über mehrere Tage hinweg z. B. jedem 10. Besucher der Website die Befragungseinladung präsentieren. Die infinite Zielpopulation (alle aktuellen und zukünftigen Website-Besucher) wird bei diesem Verfahren eingeschränkt auf eine Auswahlpopulation all jener Website-Besucher, die die Website im Erhebungszeitraum anklicken. Die **systematische Stichprobe** („systematic sample“) wird als gute Annäherung an die einfache Zufallsstichprobe gemäß Zufallszahlenprinzip aufgefasst.

Systematische Stichprobe – Man zieht eine systematische Stichprobe („systematic sample“), indem man ab einem definierten Anfangspunkt jedes n-te Element der Population zieht. Die in diesem Sinne systematische Stichprobenauswahl stellt eine pragmatische Alternative zur einfachen Zufallsstichprobe dar und ist auch auf infinite Populationen anwendbar.

Neben der systematischen Stichprobe sind **weitere quasi-zufallsgesteuerte Auswahlverfahren aus infiniten Populationen** möglich. So kann z. B. eine Zufallsauswahl von Kunden eines Fast-Food-Restaurants gezogen werden, indem man immer genau die Person herausgreift (z. B. für eine Kundenbefragung), die nach einem Kunden an der Reihe ist, der einen Coupon eingelöst hat. Da das Einlösen des Coupons durch einen Kunden völlig unabhängig davon geschieht, welcher Kunde als nächstes an der Reihe ist, kann das geschilderte Vorgehen als Annäherung an eine Zufallsauswahl für die infinite Population der Kunden verstanden werden. Eine solche quasi-zufallsgesteuerte Auswahl nach einer festen Regel ist auf jeden Fall unverzerrter als eine willkürliche Auswahl, bei der dann unbewusst z. B. besonders sympathische Kunden bevorzugt befragt würden.

Geschichtete Zufallsstichprobe

Bei der geschichteten Zufallsstichprobe erfolgt die Zufallsauswahl der Untersuchungseinheiten nicht direkt aus der Gesamtpopulation, sondern jeweils aus den Schichten der Population, die sich ergeben, wenn man die Population gemäß eines für das Forschungsproblem relevanten soziodemografischen Merkmals strukturiert: Anhand des Merkmals Einkommen mit den drei Ausprägungen niedriges, mittleres und hohes Einkommen lässt sich eine Population in drei Schichten bzw. Teilpopulationen gruppieren. Wird aus jeder dieser drei Schichten ei-

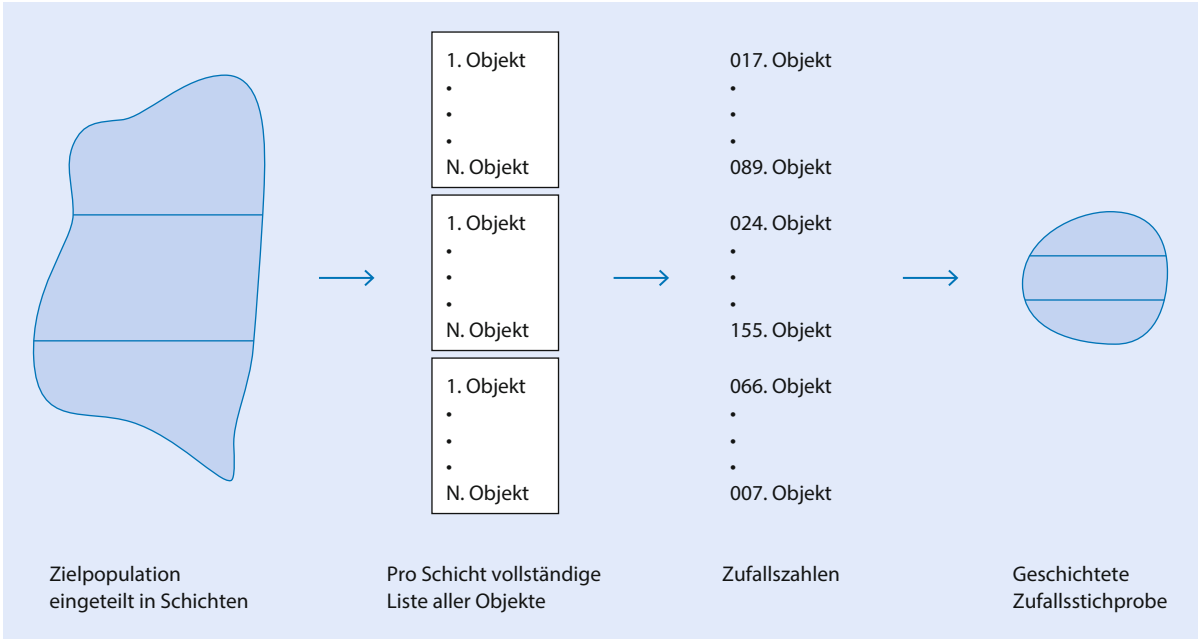


Abb. 9.4 Ziehung einer geschichteten Zufallsstichprobe

ne separate Zufallsstichprobe gezogen, so bilden diese drei einfachen Zufallsstichproben der Schichten zusammen eine geschichtete Stichprobe der Gesamtpopulation (Abb. 9.4).

Im Vergleich zu einfachen Zufallsstichproben wird in populationsbeschreibenden Studien die Parameterschätzung präziser, wenn es gelingt, die Population nach einem Merkmal zu schichten, von dem bekannt ist, dass es mit dem untersuchten Merkmal hoch korreliert (für entsprechende Berechnungsformeln ▶ Webangebot zu diesem Buch). Typische Schichtungsmerkmale sind neben dem Einkommen auch der sozioökonomische Status oder der Wohnort. Der Vorteil präziserer Parameterschätzungen geschichteter Stichproben gegenüber einfachen Zufallsstichproben kommt erst dann voll zum Tragen, wenn zusätzlich zum Schichtungsmerkmal die Größen der Teilpopulationen sowie deren Streuungen bekannt sind.

Schichtungsmerkmale sollten nicht nur mit dem untersuchten Merkmal hoch korrelieren, sondern zugleich einfach erhebbar sein. Intelligenz- und Einstellungsvariablen sind z. B. Merkmale, die zwar mit vielen sozialwissenschaftlich interessierenden Sachverhalten zusammenhängen; der Aufwand, der zu ihrer Erfassung erforderlich ist (Einsatz psychologischer Skalen und Tests; ▶ Abschn. 10.4), macht sie jedoch als Schichtungsmerkmale praktisch unbrauchbar. Für die meisten human- und sozialwissenschaftlichen Studien, die Personen als Merkmalsträger untersuchen, erweisen sich **soziodemografische Merkmale** (Alter, Geschlecht, soziale Schicht, Bildung etc.) als günstige Schichtungsmerkmale. Hat ein Schichtungsmerkmal z. B. k Ausprägungen, wird für jede Ausprägung eine Zufallsstichprobe des Umfanges n_j benötigt ($j = 1, 2, \ldots, k$). Bei einem gegebenen Schichtungsmerkmal entscheidet allein die Aufteilung der Gesamtstichprobe auf die einzelnen Schichten, also die Größe n_j der Teilstichproben, über die Präzision der Parameterschätzung. Bei günstiger Aufteilung erreicht man durch eine geschichtete Stichprobe **genauere Parameterschätzungen** als durch eine einfache Zufallsstichprobe.

Geschichtete bzw. stratifizierte Stichprobe – Man zieht ein geschichtete bzw. stratifizierte Zufallsstichprobe („stratified random sample"), indem man die Zielpopulation auf der Basis eines Merkmals oder mehrerer Merkmale in Teilpopulationen (Schichten) einteilt – pro Merkmalsausprägung bzw. Merkmalskombination entsteht eine Teilpopulation – und aus jeder dieser Schichten eine einfache Zufallsstichprobe entnimmt. Ziel der geschichteten zufälligen Stichprobenauswahl ist es, gegenüber der einfachen Zufallsstichprobe im Rahmen populationsbeschreibender Studien die Genauigkeit von Parameterschätzungen zu erhöhen.

Bisher gingen wir davon aus, dass die Schichtung nur in Bezug auf **ein** Merkmal vorgenommen wird. Diese Beschränkung ist jedoch keineswegs erforderlich und bei Untersuchungen, die gleichzeitig mehrere Sachverhalte untersuchen (Omnibus-Untersuchungen), auch nicht sehr sinnvoll. Hängen nämlich die einzelnen zu untersuchenden Sachverhalte mit jeweils verschiedenen Schichtungsmerkmalen zusammen, können bei einer nur nach

einem Merkmal geschichteten Stichprobe natürlich nur für diejenigen Sachverhalte genauere Parameterschätzungen berechnet werden, die mit dem Schichtungsmerkmal korrelieren. Die Parameterschätzungen zu den übrigen Sachverhalten fallen dann genauso exakt aus wie bei einer einfachen Zufallsstichprobe. Wenn eine Studie also mehrere Sachverhalte untersucht oder wenn der interessierende Sachverhalt mit **mehreren Schichtungsmerkmalen** zusammenhängt, empfiehlt es sich, die Schichtung gleichzeitig nach mehreren Merkmalen vorzunehmen (analog dem kombinierten Quotenplan bei einer Quotenstichprobe, ▶ Abschn. 9.3.1). Wären z. B. sowohl der Landesteil (Ost-/Westdeutschland) als auch das Einkommen (geringes/mittleres/hohes Einkommen) der Untersuchungspersonen wichtige Schichtungsmerkmale, hätte man Stichproben aus 6 Teilgesamtheiten bzw. Schichten, die als Kombinationen dieser beiden Merkmale resultieren, zu untersuchen. Dieser Aufwand zur genaueren Parameterschätzung lohnt sich allerdings nur, wenn wiederum auch die Umfänge und Streuungen dieser Teilpopulationen bekannt sind.

Grundsätzlich lassen sich zwei Varianten der geschichteten Stichprobe anhand des Auswahlsatzes (d. h. der Anzahl der Stichprobenelemente in der jeweiligen Schicht im Verhältnis zur Anzahl der Populationselemente in der Schicht: n_j / N_j) unterscheiden:

- **Proportional geschichtete Stichprobe**: Wenn die Stichprobenumfänge n_j zu ihren jeweiligen Teilpopulationen N_j proportional sind, sprechen wir von einer proportional geschichteten Stichprobe. In diesem Falle haben alle Objekte der Gesamtpopulation – wie bei der einfachen Zufallsstichprobe – dieselbe Auswahlwahrscheinlichkeit $n_j / N_j = \text{konst.}$
- **Disproportional geschichtete Stichprobe**: Bei disproportional geschichteten Stichproben sind die Auswahlwahrscheinlichkeiten für die einzelnen Teilpopulationen unterschiedlich. So können dann z. B. aus einer eher kleinen Populationsschicht (z. B. gesellschaftliche Elite) überproportional viele Personen in die Stichprobe gezogen werden, wenn insbesondere über diese Teilpopulation genaue Aussagen getroffen werden sollen. In ihrer Gesamtzusammensetzung weist die disproportional geschichtete Stichprobe systematische – aber aktiv hergestellte und somit bekannte – Abweichungen von der Zielpopulation auf. Werden diese bei der statistischen Parameterschätzung berücksichtigt, resultieren präzisere Ergebnisse als bei der proportional geschichteten Stichprobe.

Die geschichtete Stichprobe ist als aufwändiges probabilistisches Verfahren nicht mit der Quotenstichprobe, ihrem nicht-probabilistischen Pendant, zu verwechseln. Beide Verfahren verfolgen die Zielsetzung, eine Stichprobe entlang zentraler soziodemografischer Merkmale (Schichtungs- bzw. Quotierungsmerkmale) auf die Zielpopulation hin zu optimieren. Bei der geschichteten Stichprobe werden dabei jedoch aus den vordefinierten Populationsschichten jeweils echte Zufallsstichproben gezogen. Bei der Quotenstichprobe dagegen wird für jede Ausprägung der Quotierungsmerkmale eine bestimmte Zahl von Personen willkürlich ausgewählt.

Klumpenstichprobe

Wann immer eine Population aus vielen Gruppen oder natürlich zusammenhängenden Teilkollektiven (Klumpen) besteht, bietet sich die Ziehung einer Klumpenstichprobe („cluster sample“) an. Bei der Klumpenstichprobe wird eine **vollständige Liste der Klumpen** als Auswahlrahmen zugrunde gelegt. Die aus dieser Liste zufällig ausgewählten Klumpen werden dann jeweils vollständig untersucht (◘ Abb. 9.5).

Klumpenstichprobe – Man zieht eine Klumpenstichprobe („cluster sample“), indem man aus einer in natürliche Gruppen (Klumpen) gegliederten Population nach dem Zufallsprinzip eine ausreichende Anzahl von Klumpen auswählt und diese Klumpen dann vollständig untersucht.

Bei einer Untersuchung von Schulkindern würde man z. B. die Klumpenstichprobe aus mehreren zufällig ausgewählten Schulklassen (= Klumpen) zusammensetzen und bei einer Untersuchung von Betriebsangehörigen einzelne Betriebe oder Abteilungen per Zufall aus der Gesamtliste auswählen und die Klumpen dann vollständig erheben. Untersucht man Krankenhauspatienten, könnte man z. B. alle Patienten einiger zufällig ausgewählter Krankenhäuser zu einer Klumpenstichprobe zusammenfassen. Man beachte: Werden willkürlich einige Schulklassen untersucht, so handelt es sich **nicht** um eine Klumpenstichprobe, sondern um eine Gelegenheitsstichprobe (▶ Abschn. 9.3.1). Die Klumpenstichprobe als probabilistisches Verfahren basiert auf einer echten Zufallsauswahl einer ausreichenden Mindestzahl von Klumpen aus einer **Liste aller Klumpen**, aus denen sich die geklumpte Population zusammensetzt.

Die Klumpenstichprobe erfordert **weniger organisatorischen Aufwand als die einfache Zufallsstichprobe**. Während eine einfache Zufallsstichprobe strenggenommen voraussetzt, dass alle Untersuchungsobjekte der Population einzeln im Auswahlrahmen erfasst sind und einzeln rekrutiert werden, benötigt die Klumpenstichprobe lediglich eine vollständige Liste aller in der Population enthaltenen Klumpen (z. B. aller Krankenhäuser, aller Betriebe, aller Wahlkreise etc.), die dann auch oft als Gruppe für die Datenerhebung rekrutiert werden können. Eine

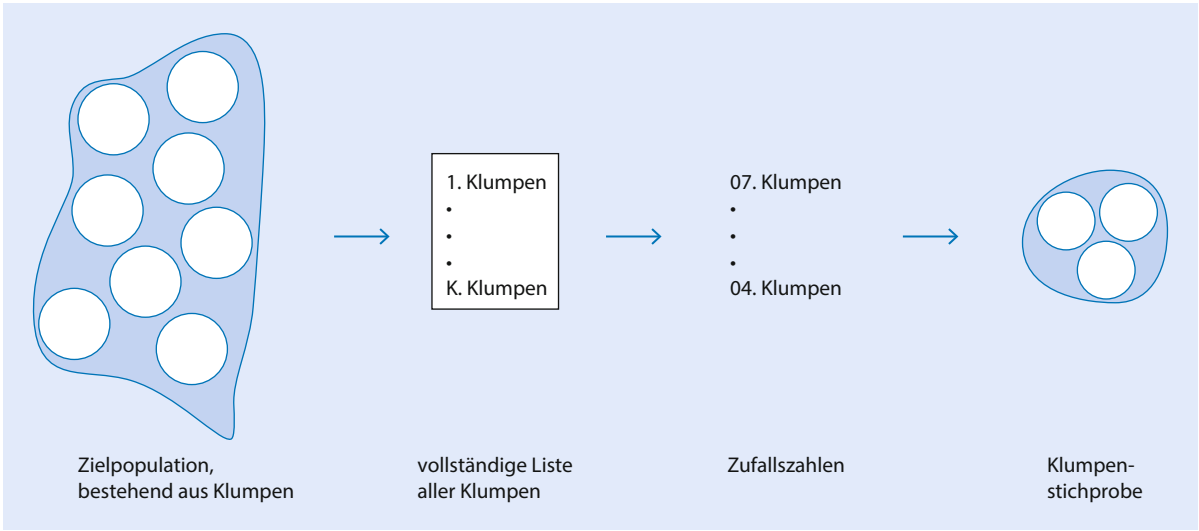


Abb. 9.5 Ziehung einer Klumpenstichprobe

vollständige Liste aller Klumpen einer geklumpten Population ist in der Regel einfacher anzufertigen als eine vollständige Liste aller einzelnen Populationsmitglieder.

Der Auswahlvorgang bezieht sich bei der Klumpenstichprobe nicht, wie bei der einfachen Zufallsstichprobe, auf die einzelnen Untersuchungsobjekte, sondern auf die Klumpen, wobei sämtliche ausgewählten Klumpen vollständig, d. h., mit allen Untersuchungsobjekten erfasst werden ($n_j = N_j$). Die Auswahlwahrscheinlichkeit ist für jeden Klumpen gleich. Es ist darauf zu achten, dass jedes Untersuchungsobjekt nur einem Klumpen angehört, dass sich also die Klumpen nicht wechselseitig überschneiden (Abb. 9.5).

Die Schätzung von Populationsparametern (z. B. Populationsmittelwert μ, Populationsanteilswert π) auf der Basis von Stichprobenkennwerten ist bei Klumpenstichproben präziser als bei einfachen Zufallsstichproben, sofern **Gleichheit aller Klumpen** und **Verschiedenartigkeit der Untersuchungsobjekte innerhalb der Klumpen** vorausgesetzt werden können (für entsprechende Berechnungsformeln ▶ Webangebot zu diesem Buch). Wird eine Klumpenstichprobe von z. B. zehn Arbeitsteams eines Großbetriebes gezogen, so bedeutet Gleichheit der Klumpen, dass sich die Teams untereinander ähneln (z. B. hinsichtlich Mitgliederzahl, Arbeitsaufgaben, Qualifikationsprofil der Mitglieder). In sich sind die Klumpen (Arbeitsteams) jedoch heterogen, etwa hinsichtlich Persönlichkeitsmerkmalen, Einstellungen, Hobbys der einzelnen Teammitglieder. Sind diese Voraussetzungen verletzt, kommt es zu einem unerwünschten **Klumpungseffekt** bzw. Klumpeneffekt („cluster effect"), der die Repräsentativität der Stichprobe reduziert: Wird z. B. in einer inhaltsanalytischen Studie von TV-Nachrichtensendungen mit einer Klumpenstichprobe gemäß der „künstlichen Woche" gearbeitet (d. h. alle Nachrichtensendungen eines zufällig ausgewählten Montags, eines zufällig ausgewählten Dienstags, eines zufällig ausgewählten Mittwochs etc. werden erhoben), so entsteht ein starker Klumpungseffekt. Denn die Nachrichtensendungen eines Tages (= Klumpens) ähneln sich untereinander hinsichtlich ihrer Themen sehr stark, gleichzeitig weichen sie von jedem anderen Klumpen deutlich ab (Jandura, Jandura, & Kuhlmann, 2005). Bei diesem Beispiel ist aufgrund der Problematik des Klumpungseffekts lieber auf eine einfache Zufallsauswahl zurückzugreifen (d. h. direkte Auswahl einzelner Nachrichtensendungen über das gesamte Jahr hinweg anstelle der Auswahl ganzer Tage als Klumpen).

Mehrstufige Zufallsstichprobe

Eine Klumpenstichprobe – so zeigte der vorige Abschnitt – setzt sich aus mehreren per Zufallsprinzip aus der Grundgesamtheit gezogenen Klumpen zusammen, die alle vollständig untersucht werden. In der Praxis kommt es häufig vor, dass die angetroffenen Klumpen zu groß sind, um sie vollständig zu erheben. In dieser Situation zieht man statt einer einfachen Klumpenstichprobe eine **zwei- oder mehrstufige Zufallsstichprobe** („multistage sampling"). Die erste Stufe betrifft die Zufallsauswahl der Klumpen aus dem Auswahlrahmen und die zweite Ziehungsstufe die Zufallsauswahl der Untersuchungsobjekte aus den einzelnen Klumpen. Damit erfasst eine zweistufige Stichprobe im Unterschied zur einstufigen Klumpenstichprobe die einzelnen Klumpen nicht vollständig, sondern nur in zufälligen Ausschnitten. Soll z. B. eine probabilistische Stichprobe von Schulkindern

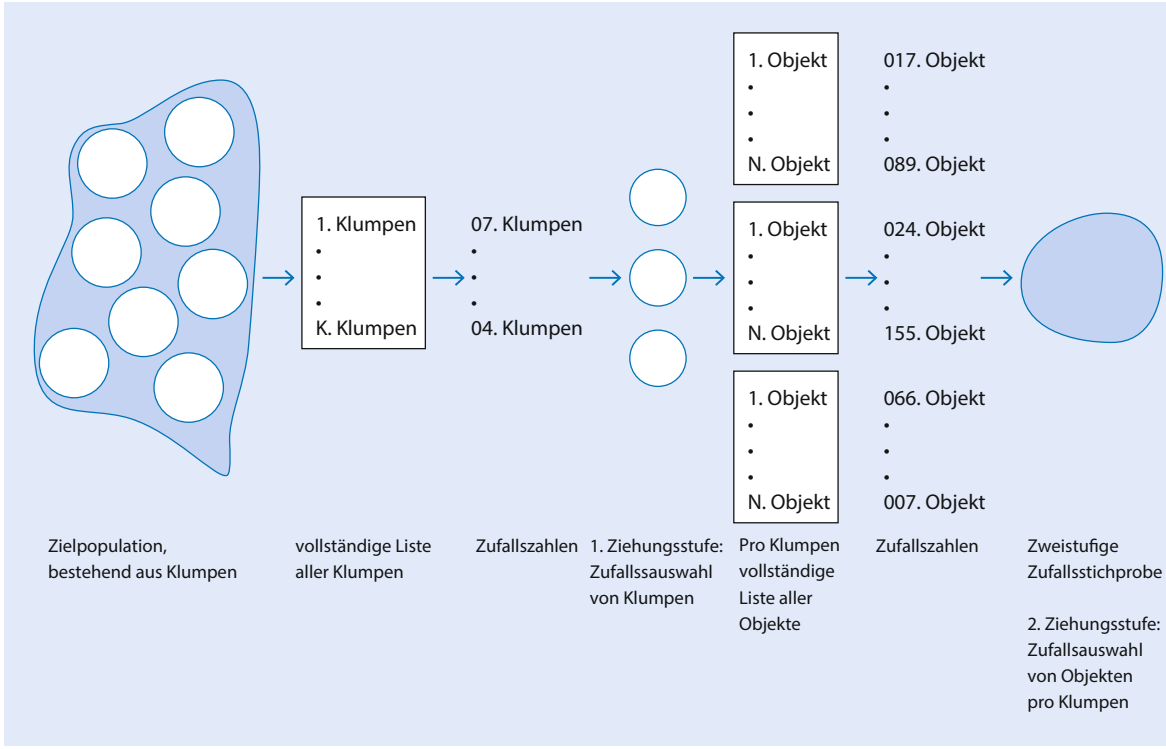


Abb. 9.6 Ziehung einer zweisufigen Zufallsstichprobe

einer Großstadt gezogen werden, so kann man in der ersten Ziehungsstufe eine Zufallsauswahl von Schulen (Klumpen) ziehen, in der zweiten Ziehungsstufe werden aus allen ausgewählten Schulen jeweils einfache Zufallsstichproben von Schüler/innen ausgewählt. Im Ergebnis liegt eine **zweistufige Klumpen- bzw. Zufallsstichprobe** vor (Abb. 9.6). Mit wachsender Klumpenanzahl (bzw. abnehmender Klumpengröße) verbessern in sich heterogene, aber untereinander homogene Klumpen die Parameterschätzung bei einer zweistufigen Stichprobe.

Eine **dreistufige Zufallsstichprobe** von Grundschulkindern erhält man z. B., wenn aus der Liste aller zur Population gehörenden Grundschulen eine Zufallsauswahl gezogen wird (1. Stufe), aus diesen Schulen zufällig Schulklassen (2. Stufe) und aus den Klassen wiederum zufällig eine festgelegte Anzahl von Schulkindern ausgewählt werden (3. Stufe). Ein populäres Beispiel für ein dreistufiges Stichprobensystem, das Repräsentativität für die Gesamtbevölkerung anstrebt, ist das Stichprobensystem des Arbeitskreises Deutscher Marktforschungsinstitute (**ADM-Mastersample**). Dieses Stichprobensystem kommt u. a. beim **ALLBUS** (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften; http://www.gesis.org/allbus/) zum Einsatz, mit dem seit 1980 in zweijährigem Abstand Mehrthemen-Umfragen (Omnibus-Umfragen) zu Bereichen wie Arbeit, Soziales, Umwelt, Politik etc. durchgeführt werden (Lipsmeier, 1999, S. 102f.). Bei bevölkerungsrepräsentativen mehrstufigen Stichproben sind die Klumpen in der ersten Ziehungsstufe meist geografische Flächen (Wahlkreise). In der zweiten Ziehungsstufe werden aus den Wahlkreisen jeweils Haushalte (wiederum Klumpen) ausgewählt. Auf der dritten Ziehungsstufe schließlich werden per Zufallsprinzip aus den Haushalten (z. B. mittels des sog. Schweden-Schlüssels oder der Next-Birthday-Methode; ▶ Abschn. 10.2.3 „Vollstrukturiertes persönliches Interview“) die Befragungspersonen ausgewählt.

Mehrstufige Stichprobe – Man zieht eine mehrstufige Zufallsstichprobe („multi-stage random sample“), indem man zunächst eine Klumpenstichprobe mit großen Klumpen zieht (1. Ziehungsstufe). Diese Klumpen werden nicht vollständig untersucht, sondern nur in zufälligen Ausschnitten. Werden aus den Klumpen mittels einfacher Zufallsstichprobe die Untersuchungsobjekte gezogen (2. Ziehungsstufe), so liegt eine zweistufige Klumpen- bzw. Zufallsstichprobe vor („two stage cluster/random sample“). Zieht man auf der zweiten Stufe wieder eine Klumpenstichprobe, ergibt sich durch Ziehung einer Zufallsstichprobe aus diesen Klumpen eine 3. Ziehungsstufe und somit eine dreistufige Klumpen- bzw. Zufallsstichprobe („three stage cluster/random sample“).

Diverse **kombinierte Stichprobenpläne** mit mehreren Ziehungsstufen und unterschiedlichen Auswahlverfahren sind möglich (z. B. auch stratifizierte mehrstufige Zufallsstichproben). Allerdings steigt mit der Komplexität der

Auswahlstufen auch die Komplexität der Berechnungsmodelle für Parameterschätzungen. Genauere Hinweise und Formeln zur Schätzung von Populationsparametern auf der Basis unterschiedlicher probabilistischer Stichproben sind dem ▶ Webangebot zu diesem Buch zu entnehmen.

Lernquiz

1. Was versteht man unter einer Vollerhebung und wann ist sie der Stichprobenziehung vorzuziehen?
2. Erklären Sie die Unterschiede zwischen Ziel-, Auswahl- und Interferenzpopulation!
3. In der Presse ist oft von „repräsentativen Studien" die Rede. Warum ist diese Bezeichnung irreführend?
4. Was versteht man unter einer „Theoretischen Stichprobe" und wann wird sie eingesetzt?
5. Wie ist die „einfache Zufallsstichprobe" definiert?
6. Ist die Gelegenheitsstichprobe eine Zufallsstichprobe? Begründen Sie Ihre Antwort!
7. Was versteht man unter „probabilistischen Stichproben", wozu werden sie eingesetzt und welche Arten von probabilistischen Stichproben sind zu unterscheiden?
8. Worin unterscheidet sich die Klumpenstichprobe von der geschichteten Stichprobe?
9. In der Zeitung lesen Sie unter der Überschrift „Haschisch macht müde und faul" folgende Meldung: „Wie eine amerikanische Repräsentativstudie zeigt, haben 70 % aller Haschischkonsumenten unterdurchschnittliche Schulleistungen. Gleichzeitig schlafen sie überdurchschnittlich lange. Diese Befunde belegen eindrücklich, wie gefährlich eine liberale Drogenpolitik ist." In dieser Nachricht sind 5 Fehler versteckt. Welche?
10. Nennen Sie Vor- und Nachteile probabilistischer Stichprobenverfahren.

Übungsaufgaben

In einem privaten Verkehrsbetrieb sind 20 Busfahrer und Busfahrerinnen beschäftigt (Population), deren Unfallzahlen in der Tab. 9.5 abgedruckt sind (Geschlecht: 1 = Frau, 2 = Mann).

Ziehen Sie aus dieser Population verschiedene Arten von Stichproben (aus didaktischen Gründen wird hier mit einem sehr kleinen Datensatz gearbeitet):

- eine **einfache Zufallsstichprobe** von $n = 10$ (verwenden Sie dazu die Zufallszahlen aus Tab. 9.4 in ▶ Abschn. 9.3.2 „Einfache Zufallsstichprobe", und zwar die 2. Kolonne von rechts, beginnend mit „85734"; streichen Sie von den 5-stelligen Zahlen die hinteren 3 Ziffern weg und suchen Sie von oben nach unten 10 Zufallszahlen im Wertebereich 1–20 aus, doppelte Zahlen werden übersprungen);
- eine **systematische Stichprobe** von $n = 4$ (ziehen Sie jedes 5. Objekt der Population, so dass Sie $n = 4$ Objekte erhalten),
- eine **Quotenstichprobe** von $n = 10$ mit der Vorgabe 80 : 20 für das Verhältnis von Frauen und Männern (greifen Sie bewusst die ersten 8 Frauen und die ersten 2 Männer heraus).
- Da sich in der Population prozentual gleichviele Männer und Frauen befinden (50 : 50), ist eine Stichprobe mit einem Geschlechterverhältnis 80 : 20 unglücklich gewählt. Durch Höhergewichtung der unterrepräsentierten Männer und Heruntergewichtung der überrepräsentierten Frauen in der obigen Quotenstichprobe lässt sich im

Tabelle 9.5 Unfallzahlen von allen Busfahrern und Busfahrerinnen eines Verkehrsbetriebs ($N = 20$)

Person	Geschlecht	Unfallzahl
1	1	0
2	1	3
3	2	2
4	2	0
5	1	0
6	1	1
7	2	0
8	1	1
9	2	3
10	2	2
11	2	0
12	2	4
13	1	0
14	1	1
15	1	1
16	1	2
17	1	0
18	2	3
19	2	2
20	2	0

Nachhinein ein Verhältnis von 50 : 50 erzeugen (**gewichtete Quotenstichprobe** von $n = 10$). Berechnen Sie die notwendigen Gewichtungsfaktoren für alle Stichprobenmitglieder als einfache Soll/Ist-Gewichte (Soll: angestrebte Anzahl der Fälle mit bestimmter Merkmalsausprägung dividiert durch Ist: tatsächliche Anzahl von Fällen in der ungewichteten Stichprobe; Kontrolle: die Summe der Soll/Ist-Gewichte über alle Stichprobenelemente hinweg muss 10 ergeben).

- Berechnen Sie für die Population und für alle 4 Stichproben jeweils die durchschnittliche Unfallzahl und vergleichen Sie die Ergebnisse.

9

Weiterführende Literatur

- Kauermann, G. & Küchenhoff, H. (2010). *Stichproben: Methoden und praktische Umsetzung mit R*. Berlin: Springer.
- Schreier, M. (2011). Qualitative Stichprobenkonzepte. In: G. Naderer & E. Balzer (Hrsg.). *Qualitative Marktforschung in Theorie und Praxis. Grundlagen, Methoden und Anwendungen* (2. Aufl., S. 241–256). Wiesbaden: Gabler
- Tashakkori, A. & Teddlie, C. (2009). Sampling strategies for mixed methods research. In: Teddlie, C. & Tashakkori, A. (Eds.). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences* (pp. 168–196). Thousand Oaks: Sage.

Literatur

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., et al. (2013). *Report of the AAPOR Task Force on non-probability sampling*: American Association for Public Opinion Research AAPOR. Retrieved October 4, 2013, from http://www.aapor.org/AM/Template.cfm?Section=Reports1&Template=/CM/ContentDisplay.cfm&ContentID=5963.

Behnke, J. (2005). Lassen sich Signifikanztests auf Vollerhebungen anwenden? Einige essayistische Anmerkungen. *Politische Vierteljahresschrift, 46*(1), 2–15.

Berland, M. A. & Thomas, K. R. (1991). Psychosocial issues following severe head injury in adolescence: Individual and family perceptions. *Rehabilitation Counseling Bulletin, 35*(1), 5–22.

Billeter, E. P. (1970). *Grundlagen der repräsentativen Statistik*. Wien: Springer.

Bjorkman, M. & Malterud, K. (2009). Lesbian women's experiences with health care: A qualitative study. *Scandinavian Journal of Primary Health Care, 27*(4), 238–243.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3rd ed.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7th. ed.). Berlin, Heidelberg: Springer.

Broscheid, A. & Geschwend, T. (2005). Zur statistischen Analyse von Vollerhebungen. *Politische Vierteljahresschrift, 46*(1), 16–26.

Chersich, M. F., Luchters, S. M. F., Malonza, I. M., Mwarogo, P., King'ola, N., & Temmerman, M. (2007). Heavy episodic drinking among Kenyan female sex workers is associated with unsafe sex, sexual violence and sexually transmitted infections. *International Journal of STD and AIDS, 18*(11), 764–769.

Christopher, F. S., Pflieger, J. C., Canary, D. J., Guerrero, L. K., & Holtzworth-Munroe, A. (2008). Targeted neighborhood sampling: A new approach for recruiting abusive couples. *Journal of Family Violence, 23*(2), 89–100.

Diekmann, A. (2005). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. Reinbeck: Rowohlt.

DMS Research. (2008). A Deep Dive into the River – Comparing Multiple Streams of River Respondents. Retrieved September 9, 2013, from http://www.usamp.com/objects/pdf/pr/DMS/DMS_River_Deep_Dive_White_Paper.pdf.

Drake, F. (2011). Protesting mobile phone masts: Risk, neoliberalism, and governmentality. *Science, Technology and Human Values, 36*(4), 522–548.

Edgington, E. & Onghena, P. (2007). *Randomization Tests* (4th ed.). Boca Raton, FL: Taylor & Francis.

Eid, M., Gollwitzer, M., & Schmitt, M. (2013). *Statistik und Forschungsmethoden* (3rd ed.). Weinheim: Beltz.

Flick, U. (1996). *Psychologie des technisierten Alltags – Soziale Konstruktion und Repräsentation technischen Wandels in verschiedenen kulturellen Kontexten*. Opladen: Westdeutscher Verlag.

Flick, U. (2010). *Qualitative Sozialforschung. Eine Einführung* (3rd ed.). Reinbeck: Rowohlt.

Freedman, D., Pisani, R., & Purves, R. (1978). *Statisics*. New York: Norton.

Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.

Glaser, B. G. & Strauss, A. L. (1999). *The discovery of grounded theory* (8th ed.). Chicago: Aldine.

Goodman, L. A. (1961). Snowballsampling. *Annals of Mathematical Statistics, 32*(1), 148–170.

Goscinny, R. & Sempé, J.-J. (1976). *Der kleine Nick und die Mädchen*. Zürich: Diogenes.

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: A methodological perspective. *Theory and Psychology, 23*(2), 251–270.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44*(2), 174–199.

Higginbottom, G. M. A. (2004). Sampling issues in qualitative research. *Nurse Researcher, 12*(1), 7–19.

Jandura, G., Jandura, O., & Kuhlmann, C. (2005). Stichprobenziehung in der Inhaltsanalyse. Gegen den Mythos der künstlichen Woche. In V. Gehrau, B. Fretwurst, B. Krause & G. Daschmann (Hrsg.), *Auswahlverfahren in der Kommunikationswissenschaft* (S. 71–116). Köln: Halem.

Knowles, A. (2011). Resilience among japanese atomic bomb survivors. *International Nursing Review, 58*(1), 54–60.

Kozak, M. (2005). Finite and infinite populations in biological statistics: Should we distinguish them? *The Journal of American Science 4*(1), 59–62.

Lam, W. K. K., Cance, J. D., Eke, A. N., Fishbein, D. H., Hawkins, S. R., & Williams, J. C. (2007). Children of African-American mothers who use crack cocaine: Parenting influences on youth substance use. *Journal of Pediatric Psychology, 32*(8), 877–887.

Legault, A. & Ducharme, F. (2009). Advocating for a parnet with dementia in a long-term care facility: The process experienced by daughters. *Journal of Family Nursing, 15*(2), 198–219.

Lipsmeier, G. (1999). Standard oder Fehler? Einige Eigenschaften von Schätzungsverfahren bei komplexen Stichprobenplänen und aktuelle Lösungsansätze. *ZA-Information, 44*, 96–117.

Marshall, M. N. (1996). Sampling for qualitative research. *Family Practice, 13*(6), 522–525.

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 11*(3). Retrieved October 4, 2013, from http://www.qualitative-research.net/index.php/fqs/article/view/1428.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Thousand Oaks: Sage.

Peterson, J. A., Schacht Reisinger, H., Schwartz, R. P., Mitchell, S. G., Kelly, S. M., Brown, B. S., et al. (2008). Targeted sampling in drug abuse research: A review and case study. *Field Methods, 20*(2), 155–170.

Safman, R. M. & Sobal, J. (2004). Qualitative sample extensiveness in health education research. *Health Education and Behavior, 31*(1), 9–21.

Scheuch, E. K. (1974) Auswahlverfahren in der Sozialforschung. *Handbuch der empirischen Sozialforschung, Bd. 3a Grundlegende Methoden und Techniken*. Stuttgart: Enke.

Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der empirischen Sozialforschung*. München: Oldenbourg.

The Rand Corporation. (1955). *A million random digits with 100 000 normal deviates*. Glencoe: Free Press.

Turgeman-Goldschmidt, O. (2005). Hackers' Accounts: Hacking as a Social Entertainment. *Social Science Computer Review, 23*(1), 8–23.

Watters, J. K. & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems, 36*(4), 416–430.

Wejnert, C. & Heckathorn, D. D. (2008). Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods and Research, 37*(1), 105–134.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur psychologischen Methodenlehre*. Göttingen: Hogrefe.

Winterich, J. A., Quandt, S. A., Grzywacz, J. G., Clark, P. E., Miller, D. P., Acuña, J., et al. (2009). Masculinity and the body: How African American and white men experience cancer screening exams involving the rectum. *American Journal of Men's Health, 3*(4), 300–309.

Datenerhebung

Nicola Döring, Jürgen Bortz

10.1 Beobachtung – 323

10.2 Interview – 356

10.3 Fragebogen – 398

10.4 Psychologischer Test – 430

10.5 Physiologische Messungen – 499

10.6 Dokumentenanalyse – 533

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_10, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Wissen, was man unter wissenschaftlicher Datenerhebung versteht und wie sie sich von nicht-wissenschaftlicher Datensammlung unterscheidet.
- Qualitative und quantitative Beobachtungsmethoden charakterisieren und anwenden können.
- Qualitative und quantitative Interviewtechniken erläutern und einsetzen können.
- Qualitative und quantitative selbstadministrierte Fragebogenmethoden differenzieren und entsprechende Fragebögen entwickeln können.
- Unterschiedliche Arten von projektiven und psychometrischen psychologischen Tests voneinander abgrenzen können und wissen, was bei Testanwendung und Testentwicklung zu beachten ist.
- Wichtige physiologische Messverfahren für unterschiedliche Organsysteme (z. B. Hirnaktivität, Herz-Kreislauf-Aktivität, elektrodermale Aktivität) in ihren Grundlagen beschreiben und ihrer Aussagekraft einschätzen können.
- Verschiedene Formen der qualitativen und quantitativen Dokumentenanalyse unterscheiden und dabei insbesondere das Vorgehen bei einer qualitativen und einer quantitativen Inhaltsanalyse vorgefundener Dokumente schildern können.
- Die Besonderheiten und Vor- und Nachteile der verschiedenen Datenerhebungsmethoden abwägen und die für ein konkretes Forschungsproblem passende(n) Datenerhebungsmethode(n) auswählen können.

Die **Datenerhebung** („data collection“) ist essenzieller Bestandteil jeder empirischen Studie. Innerhalb des Forschungsprozesses handelt es sich um eine mehr oder minder ausgedehnte Zeitperiode, in der systematisch numerisches und/oder nicht-numerisches Datenmaterial gesammelt wird, das der Beantwortung der Forschungsfragen bzw. der Entwicklung oder Prüfung der Forschungshypothesen dienen soll. Die Datenerhebung basiert auf spezifischen **Datenerhebungsmethoden**, die sich in ihrer Vorgehensweise, ihrem Aufwand sowie vor allem der Art und dem Informationsgehalt der resultierenden Daten deutlich unterscheiden. In der empirischen Sozialforschung differenzieren wir vor allem sechs Datenerhebungstechniken: die **Beobachtung** (▶ Abschn. 10.1), die mündliche Befragung bzw. das **Interview** (▶ Abschn. 10.2), die schriftliche Befragung bzw. die selbstadministrierte **Fragebogenmethode** (▶ Abschn. 10.3), den **psychologischen Test** (▶ Abschn. 10.4), die **physiologische Messung** (▶ Abschn. 10.5) und schließlich die **Dokumentenanalyse** (▶ Abschn. 10.6).

All diese Methoden des Datengewinns kommen auch im Alltag zum Einsatz. Was sie zu wissenschaftlichen Datenerhebungstechniken werden lässt und welche Unterformen jeweils existieren, wird ausführlich anhand aktueller Studienbeispiele dargestellt. Dabei wird immer auf Grenzen der Aussagekraft der Methoden, auf typische Fehler sowie auf den notwendigen zeitlichen, finanziellen und personellen Aufwand hingewiesen. Die Darstellung in diesem Lehrbuch soll dazu befähigen, vorliegende Studien, die mit den entsprechenden Methoden arbeiten, zu verstehen und zu bewerten und gleichzeitig auch den Grundstein für die praktische Umsetzung der Methoden in eigenen Untersuchungen legen. Dazu muss die Lektüre des Lehrbuchs freilich immer durch praktisches Üben ergänzt werden. Die verschiedenen Datenerhebungsmethoden werden hier vor allem aus **Anwendungsperspektive** betrachtet. Auf methodologische Aspekte und Ergebnisse der Methodenforschung (z. B. empirischer Vergleich unterschiedlicher Anwendungsvarianten einer Datenerhebungsmethode) wird jedoch verwiesen.

Mit Ausnahme der physiologischen Messung, die definitionsgemäß immer auf quantitative Messwerte hinausläuft, existieren alle Datenerhebungsmethoden sowohl in qualitativen als auch in quantitativen Varianten. **Qualitative Datenerhebungstechniken** sind nicht oder nur teilweise strukturiert, d. h. sie arbeiten mit nicht- oder teilstandardisierten Instrumenten und lassen somit den Forschenden wie den Untersuchungsteilnehmenden viel Raum, um jeweils fall- und situationsspezifisch mehr oder minder viele Details des untersuchten Sachverhaltes in die Datenerhebung einzubeziehen und vor allem auch flexibel zu entscheiden, welche Aspekte der Erfahrungswirklichkeit bzw. Lebenswelt (▶ Abschn. 2.3.1) einen wie großen Stellenwert für das Forschungsproblem haben. Die Offenheit des Vorgehens soll eine optimale Annäherung an den Untersuchungsgegenstand gewährleisten und die Chance erhöhen, dass sich auch neue und unerwartete Inhalte in den erhobenen Daten zeigen. So wird z. B. beim narrativen Interview als einer nicht-strukturierten qualitativen Interviewtechnik nur ein Erzählanstoß gegeben und die Befragten können sich dann frei in eigenen Worten äußern. Beim Leitfadeninterview als einer teilstrukturierten qualitativen Interviewtechnik wird ein Fragenkatalog zugrunde gelegt, dieser kann und soll jedoch auf die Befragungsperson abgestimmt werden, d. h. es können spontan auftretende neue Fragen integriert und auch Änderungen der Reihenfolge der behandelten Themen vorgenommen werden.

Bei **quantitativen Datenerhebungsmethoden** ist dagegen der Strukturierungsgrad sehr hoch und es wird mit vollstandardisierten Instrumenten gearbeitet. Im vollstrukturierten quantitativen Interview sind z. B. nicht nur Anzahl, Reihenfolge und Wortlaut aller Fragen exakt vorgegeben, sondern auch die Antwortvorgaben ausformuliert, so dass die Befragten nur die auf sie am besten

zutreffenden Antworten auszuwählen haben. Die Aussagekraft einer Studie, die mit quantitativen Datenerhebungsmethoden arbeitet, steht und fällt mit der **Güte des standardisierten Messinstruments**. Welche Eigenschaften und Gütekriterien standardisierte Messinstrumente haben und wie sie zu entwickeln und zu prüfen sind, nimmt deswegen im Zusammenhang mit der Darstellung quantitativer Datenerhebungsmethoden großen Raum ein. Insofern bestehen in der quantitativen Forschung zwischen der **Operationalisierung** der im Rahmen des Forschungsproblems interessierenden Variablen (► Kap. 8) und der Vorbereitung der Datenerhebung durch entsprechende Instrumentenentwicklung (z. B. vollstandardisierter Beobachtungsplan oder vollstandardasierter Fragebogen) enge Verbindungen. Die Aussagekraft der gewonnenen quantitativen Daten steht und fällt mit dem Einsatz theoretisch und empirisch fundierter Messinstrumente, die nachweislich die Ausprägungen der interessierenden Merkmale erfassen. Sie sind von ungeprüften Ad-hoc-Instrumenten zu unterscheiden.

Mit einzelnen quantitativen Datenerhebungstechniken sind teilweise auch **spezifische Stichprobenvarianten** (► Kap. 9) verknüpft: Wird in einer Studie z. B. die mündliche Befragungstechnik des vollstrukturierten Telefoninterviews eingesetzt, so existieren unterschiedliche Vorgehensweisen, um eine bevölkerungsrepräsentative Telefonstichprobe zu ziehen. Die Wahl und Umsetzung der Datenerhebungsmethode ist schließlich auch auf das **Untersuchungsdesign** (► Kap. 7) abzustimmen. So sind bestimmte Datenerhebungsmethoden – z. B. viele physiologische Messverfahren – auf eine Laborumgebung angewiesen und können nicht in einer Feldstudie zum Einsatz kommen.

Neben dem Strukturierungsgrad, der qualitative und quantitative Datenerhebungsmethoden voneinander abgrenzt, gibt es noch zahlreiche weitere Kriterien, nach denen Datenerhebungsmethoden klassifiziert werden. Diese Kriterien werden im Zusammenhang mit den einzelnen Methoden jeweils behandelt. So unterscheiden wir z. B. zwischen reaktiven und nonreaktiven Methoden. Bei **reaktiven Datenerhebungsmethoden** wissen die untersuchten Personen, dass sie an einer Studie teilnehmen. Die im Zuge der Datenerhebung generierten Daten unterliegen somit unterschiedlichen Verzerrungen (z. B. Interviewdaten zur Internetnutzung können durch sozial erwünschtes Antworten verzerrt sein). Bei **nonreaktiven Datenerhebungsmethoden** findet jedoch kein Eingriff in das natürliche Verhalten der Untersuchungspersonen statt, etwa weil – im Rahmen der Forschungsethik (► Abschn. 4.1) – verdeckte Beobachtungen durchgeführt oder nur Verhaltensspuren erhoben werden (z. B. Logfiles dokumentieren die Internetnutzung unverzerrt).

Im Zuge der Digitalisierung wächst fortlaufend die Menge an Datenspuren, die jeder Mensch im Alltag hinterlässt. Viele dieser Daten werden unternehmensintern zu kommerziellen Zwecken ausgewertet. Aktuell ist noch weitgehend offen, ob und wie die empirische Sozialforschung **„Big Data“** verantwortungsvoll zum Zweck des wissenschaftlichen Erkenntnisgewinns nutzen kann und darf. Dieses Beispiel illustriert, dass wissenschaftliche Datenerhebungsmethoden historischem Wandel unterliegen und aktuell stark geprägt sind durch einen **Trend zur computer- und internetgestützten Datenerhebung**.

10.1 Beobachtung

Lernziele

- Wissen, was eine wissenschaftliche Beobachtung von einer Alltagsbeobachtung unterscheidet.
- Verschiedene Typen wissenschaftlicher Beobachtung im Hinblick auf Herangehensweise und Anwendungsfelder voneinander abgrenzen können.
- Potenzielle Beobachtungsfehler und Gegenmaßnahmen kennen.
- Eine qualitative sowie eine quantitative Beobachtung mit geringem Komplexitätsgrad planen können.
- Die jeweiligen Vor- und Nachteile qualitativer und quantitativer Beobachtung abwägen können.

In diesem Kapitel wird zunächst erklärt, wodurch sich eine wissenschaftliche Beobachtung als empirische Datenerhebungsmethode (► Abschn. 10.1.1) auszeichnet. Dann werden jeweils drei wichtige Varianten der qualitativen Beobachtung (► Abschn. 10.1.2) sowie der quantitativen Beobachtung (► Abschn. 10.1.3) erläutert und an Beispielen illustriert.

10.1.1 Wissenschaftliche Beobachtung

Um zu verstehen, was eine wissenschaftliche Beobachtung als Datenerhebungsverfahren auszeichnet und worin sie anderen Datenerhebungsmethoden – wie z. B. der wissenschaftlichen Befragung (► Abschn. 10.2; ► Abschn. 10.3) – überlegen ist, wenden wir uns in diesem Abschnitt zunächst ihrer **Definition und Indikation** zu. Anschließend wird erläutert, worin sich die wissenschaftliche Beobachtung von der **Alltagsbeobachtung** unterscheidet. Um unterschiedliche Formen der wissenschaftlichen Beobachtung zu differenzieren, stehen fünf **Klassifikationskriterien** zur Verfügung – von der Strukturiertheit (qualitative vs. quantitative Beobachtung) bis zur Transparenz des Vorgehens (offene vs. verdeckte Beobachtung). Schließlich werden noch typische **Beobachtungsfehler** und entsprechende Gegenmaßnahmen behandelt.

Studienbeispiel

Qualitative und quantitative Beobachtung

Ein ethnografischer Feldforscher beobachtet Jugendliche aus der „Schwarzen Szene“ bei einem Konzert. Er macht Fotos, um Körperinszenierungen, Kleidung und Schmuck festzuhalten und erstellt in eigenen Worten Feldnotizen zu typischen Gesten, Getränken, Gesprächsthemen etc. Auf diese Weise erhebt er **qualitative Beobachtungsdaten** und zwar sowohl **visuelle Daten (Fotos)** als auch **verbale Daten (Feldnotizen)**.

Eine Versuchsleiterin beobachtet die Untersuchungspersonen während eines Laborexperiments und protokolliert die Häufigkeit und Intensität bestimmter Reaktionen (z. B. verbaler und non-verbaler Ausdruck von Freude oder Ärger) während der Nutzung eines Computerspiels in einem vorher ausgearbeiteten Beobachtungsbogen. Im Zuge dieser strukturierten quantitativen Beobachtung werden **quantitative bzw. numerische Beobachtungsdaten (Messwerte)** erhoben (z. B. Häufigkeiten und Intensitäten von bestimmten Emotionsäußerungen).

Definition und Indikation der wissenschaftlichen Beobachtung

Wissenschaftliche Beobachtung – Unter einer wissenschaftlichen Beobachtung („scientific observation“) versteht man die zielgerichtete, systematische und regelgeleitete Erfassung, Dokumentation und Interpretation von Merkmalen, Ereignissen oder Verhaltensweisen mithilfe menschlicher Sinnesorgane und/oder technischer Sensoren zum Zeitpunkt ihres Auftretens.

In den Natur- und Ingenieurwissenschaften sowie in der Medizin ist die Beobachtung mittels **technischer Geräte** (Mikroskop, Teleskop, Spektrometer, Ultraschallgerät etc.) oft der einzige Zugang zu den interessierenden Sachverhalten, insbesondere wenn diese sehr kleine (Nano- und Mikrokosmos) oder sehr große Dimensionen haben (Makrokosmos). In den Sozialwissenschaften, in denen die Untersuchung menschlichen Verhaltens und somit Sachverhalte mittlerer Dimension (Mesokosmos) eine zentrale Rolle spielen, kommen zu Beobachtungszwecken zuweilen ebenfalls technische Hilfsmittel zum Einsatz (z. B. eine Software registriert das Internetnutzungsverhalten von Untersuchungspersonen mittels automatisch erstellter Protokolldateien). Typischer ist in den Sozialwissenschaften jedoch die **direkte** (anstatt technikvermittelte) Beobachtung allein auf der Basis **menschlicher Sinnesorgane**. Dabei ist je nach Fragestellung sowohl a) ein nicht-strukturiertes oder teilstrukturiertes bzw. qualitatives Vorgehen als auch b) ein strukturiertes bzw. quantitatives Vorgehen möglich (► Studienbeispiel „Qualitative und quantitative Beobachtung“).

Da bei der wissenschaftlichen Beobachtung die interessierenden Daten zum Zeitpunkt ihres Auftretens sozusagen „live“ erfasst werden, überfordert die Fülle und Frequenz der Informationen menschliche Beobachter potenziell. Deswegen wird der zu beobachtende Verhaltensstrom samt Umgebungsbedingungen oftmals teilweise oder ganz **medial aufgezeichnet** (z. B. Foto-, Audio-, Videoaufnahme), sofern dies forschungsethisch zu vertreten ist. Anhand dieser Verhaltens-Aufzeichnungen ist eine zeitversetzte Analyse und Interpretation möglich.

Wissenschaftliche Beobachterinnen und Beobachter greifen primär auf ihren Sehsinn zurück, ergänzt um den Hörsinn. Um der Körperlichkeit menschlichen Erlebens und Verhaltens gerecht zu werden, kann es sinnvoll sein, weitere **Sinnesmodalitäten** zu nutzen (z. B. zur Erfassung von Temperaturen, Gerüchen, Texturen am Schauplatz der Beobachtung). Auch die körperlichen Reaktionen der beobachteten Personen selbst können von wissenschaftlichem Interesse sein. Während manche Körperreaktionen (z. B. Schwitzen, Erröten) visuell beobachtbar sind, ist der größte Teil der physiologischen Prozesse für das menschliche Beobachterauge unsichtbar. Deswegen kommen technische Geräte zum Einsatz, um Blutdruck, Muskelspannung oder Hirnaktivitäten der beobachteten Personen zu Forschungszwecken zu messen. Obwohl es sich hier ebenfalls um quantitative Beobachtungen handelt, werden diese speziellen technologiegestützten Datenerhebungsverfahren zur Hervorhebung ihrer Besonderheiten in der empirischen Forschung typischerweise als **„physiologische Messungen“** bezeichnet und gesondert behandelt ► Abschn. 10.5.

Bei der sozialwissenschaftlichen Beobachtung anderer Personen und ihrer Verhaltensweisen wird von den Forschenden eine **Außenperspektive** eingenommen. Demgegenüber soll die Befragung (► Abschn. 10.2 und ► Abschn. 10.3) auch einen Zugang zur Innenwelt (Gedanken, Gefühle etc.) der Untersuchungsteilnehmenden ermöglichen. Die mündliche oder schriftliche Befragung ist die in der Sozialforschung gebräuchlichste Methode der Datenerhebung. Es gibt jedoch eine Reihe von Gründen, die im konkreten Untersuchungsfall dafür sprechen können, die Beobachtung als Datenerhebungsmethode einer Befragung vorzuziehen. Folgende Umstände sprechen für eine besondere Eignung bzw. **Indikation der Beobachtungsmethode** als Datenerhebungsverfahren im Rahmen eines empirischen Forschungsprozesses (◘ Abb. 10.1):

1. Eine Befragung ist immer daran gebunden, dass Personen hinsichtlich der interessierenden Sachverhalte verbal auskunftsfähig bzw. -freudig sind und wahrheitsgemäß antworten. Säuglinge und Kleinkinder können z. B. aufgrund **eingeschränkter Verbalisierungsfähigkeiten** nur bedingt befragt, in ihrem Spiel- und Interaktionsverhalten jedoch beobachtet werden. Ebenso ist in der Forschung mit Tieren die Beobachtungsmethode unverzichtbar.

2. Über **automatisierte und unbewusste Verhaltensweisen** (z. B. Mimik, Gestik) lassen sich mittels Beobachtung, nicht jedoch mittels Befragung Erkenntnisse gewinnen.
3. Auch Themen oder Zielgruppen, bei denen die **Auskunftswilligkeit beschränkt** ist, können mittels Beobachtung untersucht werden (z. B. Drogenhandel im Bahnhofsviertel, Verhalten von Polizei und Demonstrierenden bei gewalttätigen Auseinandersetzungen). Geringe Auskunftswilligkeit besteht nicht nur bei normverletzenden Verhaltensweisen, sondern oft schlicht aus Zeitmangel (z. B. erklären sich viel beschäftigte Fachkräfte manchmal eher bereit, sich bei der Arbeit beobachten zu lassen als an Interviews über ihren Arbeitsalltag teilzunehmen, die einen zeitlichen Zusatzaufwand bedeuten würden).
4. Die Befragung ist immer eine reaktive Methode, die auf die aktive und bewusste Mitwirkung der Befragten angewiesen ist und somit Daten produziert, die mehr oder minder deutliche Verzerrungen (z. B. durch Selbstdarstellungsverhalten) beinhalten. Eine Beobachtung kann dagegen **non-reaktiv** (d. h. nichtteilnehmend und verdeckt) stattfinden, so dass nicht in die natürlichen Abläufe eingegriffen wird (z. B. erfragte Selbstauskunft zum umweltbewussten bzw. abfallvermeidenden Verhalten versus beobachtete tatsächliche Menge und Zusammensetzung des Hausmülls; Selbstauskunft des Service-Personals zur Qualität der Kundenbetreuung versus Beobachtung der Dienstleistung).
5. Nicht zuletzt kann die Befragung immer nur zusammenfassende und punktuelle Aussagen über Verhaltensweisen erfassen, während die Beobachtung das **Verhalten im Zeitverlauf** kontinuierlich abbilden kann (z. B. Veränderungen des Schülerverhaltens über den Verlauf einer gesamten Unterrichtsstunde oder Klassenreise).

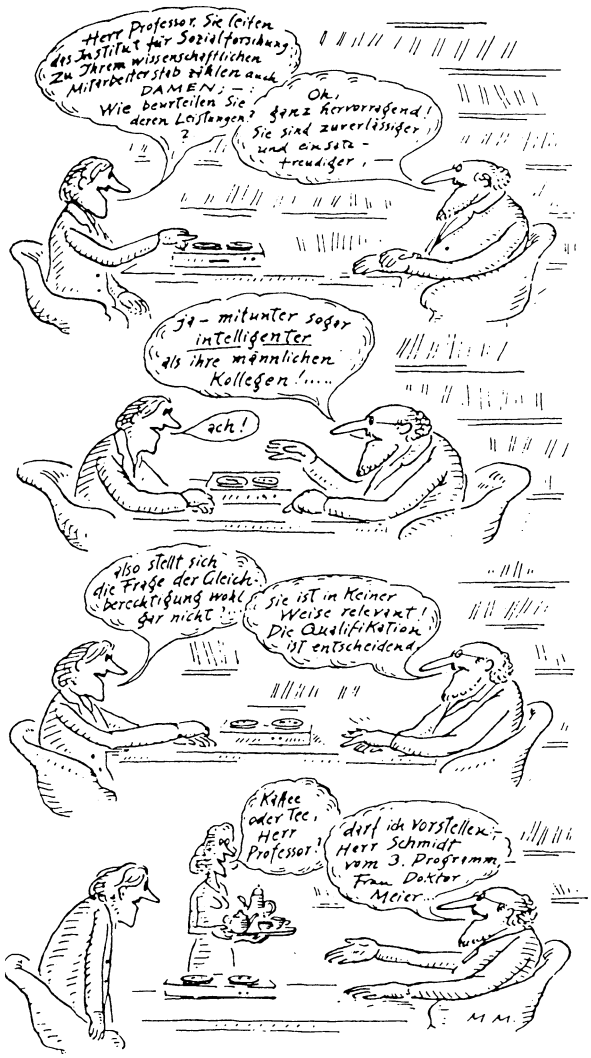


Abb. 10.1 Mit eigenen Augen: Warum die Feldbeobachtung der Befragung überlegen sein kann. (Aus Marcks, 1974. Mit freundlicher Genehmigung von Tai M. Lüdicke)

Es ist wichtig, die Vorzüge der Beobachtungsmethode gegenüber Befragungstechniken im Blick zu behalten, um nicht routinemäßig auf Interview oder Fragebogen zurückzugreifen. Die Beobachtung stellt zuweilen eine echte Alternative zur Befragung dar. Zudem ist sie fast immer eine sinnvolle Ergänzung. So profitiert die Datenerhebung per Interview oder Fragebogen in der Regel davon, dass die Forschenden das jeweils interessierende Handlungsfeld zumindest explorativ schon einmal beobachtet haben. Andererseits ist zu beachten, dass die Datenerhebung mittels Beobachtung **drei Nachteile** aufweist:

1. Die Beobachtung ist in der Regel **zeit- und kostenaufwändiger** als die Befragung.
2. Viele subjektive Erlebensphänomene sind einer Fremdbeobachtung nicht zugänglich (**Grenzen der Beobachtbarkeit**) und müssen erfragt werden.
3. Durch die starke Dominanz der Befragungsmethode in den Sozialwissenschaften ist die Beobachtungsmethode **auf methodologischer Ebene weniger ausgearbeitet**.

Abgrenzung der wissenschaftlichen Beobachtung von der Alltagsbeobachtung

Wenn im Alltag und in den Massenmedien soziale Sachverhalte diskutiert werden, stützt man sich häufig auf Beobachtungen:

Studienbeispiel

Anekdotische Evidenz: „Neun Monate nach Schneechaos und Stromausfall: Babyboom im Münsterland“.

Das Hamburger Abendblatt berichtete am 26. August 2006: MÜNSTER. Es war im November des vorigen Jahres, als ein verheerendes Schneechaos über das Münsterland hereinbrach. 250 000 Menschen in den Kreisen Steinfurt, Borken und Coesfeld waren tagelang ohne Strom, weil 50 Hochspannungsmasten unter der Eis- und Schneelast umgestürzt waren. Doch der Wintereinbruch hatte auch seine gute Seite: Bei Kerzen und Kaminfeuern rückten viele Paare offenbar enger zusammen. Das Ergebnis zeigt sich derzeit auf den Entbindungsstationen. Auf den Standesämtern werden mehr Geburten als sonst gemeldet. So rechnet auch die Kreisstadt Steinfurt im September mit deutlich mehr Neugeborenen als im Vorjahr. „Wir haben monatlich etwa 50 Geburten. In den nächsten Wochen werden es wahrscheinlich 65 sein. Das lag wohl am Stromausfall“, sagt die Standesbeamtin Gudrun Frahling. Einen vergleichbaren Baby-Zuwachs habe es in den vergangenen acht Jahren nicht gegeben. (dpa)

- Beispielsweise wird eine **Lehrerin** zitiert, die auf dem Schulhof beobachtet: „Es wird heutzutage immer öfter weitergeschlagen und -getreten, auch wenn das Opfer kampfunfähig am Boden liegt. Früher war das anders.“
- Oder ein **Psychotherapeut** schlägt Alarm, weil er bei seinen Patienten „eine starke Verunsicherung der männlichen Geschlechtsidentität“ beobachtet.
- Eine **Hebamme** teilt mit, dass neun Monate nach einem regionalen Stromausfall in ihrer Klinik viel mehr Babys als sonst zur Welt gekommen sind.

Derartige Alltagsbeobachtungen haben aus wissenschaftlicher Sicht den Charakter **anekdotischer Evidenz** („anecdotic evidence“; ► Studienbeispiel „Anekdotische Evidenz: ‚Neun Monate nach Schneechaos und Stromausfall: Babyboom im Münsterland‘“). Sie basieren auf realen Erfahrungen und wirken überzeugend. Im Unterschied zur wissenschaftlichen Beobachtung erfolgen sie jedoch vollkommen **unsystematisch** und **subjektiv**. Sie sind grundsätzlich nicht geeignet, um fundierte generalisierbare Aussagen zu treffen.

Die höhere Geburtenzahl neun Monate nach einem Stromausfall, die es seit Jahrzehnten immer wieder in die Schlagzeilen schafft, erweist sich aus wissenschaftlicher Sicht schlicht als eine „urbane Legende“. Wird nicht die **willkürlich herausgegriffene Einzelbeobachtung** in einer einzelnen Klinik oder in einem einzigen Standesamt zugrunde gelegt, sondern werden systematisch die Geburtenzahlen aus mehreren Orten der betroffenen Region einbezogen und mit mehreren früheren Zeitpunkten verglichen, ist keinerlei Effekt beobachtbar (Izenman & Zabell, 1981; Udry, 1970). Die Geburtenzahlen unterliegen seit jeher, unabhängig von Schneechaos und Stromausfällen, saisonalen Schwankungen. Auch die Beobachtung der Lehrerin an ihrer Schule ist zu hinterfragen: Sind die Schulhof-Prügeleien in der Vergangenheit wirklich anders abgelaufen? Hat sie diese so genau beobachtet und kann sich so genau erinnern? Fallen ihr vielleicht nur jetzt einzelne Negativbeispiele stark auf, weil Jugendgewalt so oft in der Presse diskutiert wird? Wie ist die Situation an anderen Schulen? Derartige Fragen bleiben bei der unsystematischen Alltagsbeobachtung offen.

Die wissenschaftliche Beobachtung („scientific observation“) erfolgt zielgerichtet und systematisch im Kontext eines empirischen Forschungsprozesses: Wenn man sich für den Einfluss von Stromausfällen auf die Geburtenrate interessiert, wird im Vorfeld der Datenerhebung festgelegt, wann und wo und über welchen Zeitraum hinweg welche Beobachtungsdaten erhoben und in die Auswertung einbezogen werden sollen. Diese Festlegungen betreffen die Operationalisierung, das Untersuchungsdesign und die Stichprobenziehung bei einer Beobachtungsstudie und damit die **Selektivität des Beobachtungsprozesses**: Es kann – auch in der Wissenschaft – niemals alles vollständig beobachtet werden, was für das Forschungsproblem von Bedeutung wäre. Wie die Auswahl getroffen wird, muss jedoch überzeugend begründet und sorgfältig geplant sein. Bei der Auswahl sind – neben der Beobachterrolle (z. B. teilnehmend, nichtteilnehmend) – folgende Aspekte zu beachten:

1. **Beobachtungsorte** („observation sites“): Wo soll beobachtet werden? Im Labor oder im Feld? Wo genau im Feld, an wie vielen unterschiedlichen oder ähnlichen Orten?
2. **Beobachtungszeiten** („observation times“): Wann und wie lange soll beobachten werden?
3. **Beobachtungsobjekte** („observation objects“): Welche Personen (z. B. Selbst oder Andere), welche Gruppen, Ereignisse oder Gegenstände sollen beobachtet werden?
4. **Beobachtungseinheiten** („observation units“): Welche konkreten Aspekte der Beobachtungsobjekte (z. B. welche Eigenschaften, Verhaltensweisen) sollen hinsichtlich ihrer Qualität, Häufigkeit oder Intensität beobachtet werden? Umgekehrt: Was ist für die Studie unwesentlich und kann ignoriert werden?

Weiterhin ist zu entscheiden, ob alle Beobachtungen unter denselben natürlichen Bedingungen erfolgen (nichtexperimentelle Studie) oder verschiedene künstlich hergestellte Bedingungen kontrastiert (quasi-/experimentelle Studie) werden sollen. Durch die Entscheidung all die-

Studienbeispiele

Wissenschaftliche quantitative Beobachtungen

In einer quasi-experimentellen Beobachtungsstudie zum Flirtverhalten wurde eine Stichprobe von $n = 283$ **Versuchspersonen** jeweils **7 Minuten im Labor** einer deutschen Universität beobachtet (Penke & Asendorpf, 2008, S. 1125). Insgesamt ergab sich also ein Beobachtungsmaterial von 283 × 7 Minuten = **33 Beobachtungsstunden**. Es wurde überprüft, ob sich Versuchspersonen mit unterschiedlicher sozio-sexueller Orientierung (restriktiv: Sex bevorzugt in festen Beziehungen vs. nicht-restriktiv: Sex auch wünschenswert außerhalb fester Beziehungen) in ihrem Flirtverhalten unterscheiden. Tatsächlich flirteten Untersuchungspersonen mit nicht-restriktiver sozio-sexueller Orientierung intensiver.

Demgegenüber wurden in einer nicht-experimentellen Studie zur Handynutzung am Steuer Beobachtungsdaten im Feld erhoben und zwar jeweils über **1.5 Stunden an verschiedenen Straßenkreuzungen**. Berücksichtigt wurden Kreuzungen in zwei israelischen Großstädten, zwei Kleinstädten und zwei Dörfern in jeweils zwei vergleichbaren Bezirken. Dabei wurden alle an einem klaren Novembertag zwischen 8.00 und 10.30 Uhr morgens passierenden Fahrzeuge beobachtet. In den insgesamt 2 (Anzahl der Orte pro Ortsgröße) × 3 (Ortsgröße: Großstadt, Kleinstadt, Dorf) × 2 (Bezirke pro Ort) × 1.5 Stunden = **18 Beobachtungsstunden** wurden Daten von $n = 1\,062$ **Autofahrern und Autofahrerinnen** erfasst (Rosenbloom, Ben-Eliyahu, Nemrodov, Biegel, & Perlman, 2009, S. 216). In dieser Studie kam es auf den Vergleich der drei Ortsgrößen an: Es zeigte sich, dass manche Verkehrsregeln (z. B. Anschnallpflicht, Beachtung von Vorfahrtszeichen) auf dem Dorf signifikant häufiger missachtet wurden, bei der Handynutzung am Steuer jedoch keine überzufälligen Unterschiede zwischen den Ortsgrößen beobachtbar waren.

ser Fragen wird das **Design der Beobachtungsstudie** konkretisiert, das sehr unterschiedlich ausfallen kann (zu Untersuchungsdesign ▶ Kap. 7), ebenso wird die Stichprobe gebildet (▶ Kap. 9).

In den in ▶ Studienbeispiele „Wissenschaftliche quantitative Beobachtungen“ präsentierten Untersuchungen erfolgte die Datenerhebung mit unterschiedlichen Stichprobenverfahren:

- **Zeitstichprobe** („time sample“): Handelt es sich bei der interessierenden Beobachtungseinheit um ein zeitlich ausgedehntes Verhalten (z. B. Flirtverhalten), so ist eine Zeitstichprobe indiziert: In regelmäßigen Abständen (z. B. alle 30 Sekunden) erfassen die Beobachtenden die gerade aktuelle Ausprägung (z. B. Intensität) des Verhaltens (hier: Flirtverhalten) auf einer Ratingskala (systematische Zeitstichprobe). Zeitstichproben stellen hohe Anforderungen an die Konzentrationsfähigkeit: Eine durchgängige Beobachtungszeit von mehr als 30 Minuten dürfte selbst geschulte Beobachter überfordern. Sinnvoll ist es deswegen, mit Videoaufzeichnungen zu arbeiten. Damit sind auch andere Varianten der Zeitstichprobe möglich (z. B. Zufallsauswahl von Zeitpunkten aus einem sehr langen Verhaltensstrom). Bei wenig gleichförmig verlaufenden Verhaltenssequenzen (z. B. bei denen Anfangs- und Schlussphase besonders wichtig sind) empfiehlt sich statt einer Zufallsstichprobe eine systematische Stichprobe, die den Besonderheiten der Verhaltenssequenz Rechnung trägt (z. B. enger getaktete Messzeitpunkte in der Anfangs- und Schlussphase).
- **Ereignisstichprobe** („event sample“): Handelt es sich bei der interessierenden Beobachtungseinheit um Einzelereignisse, einzelne Objekte oder kurze Verhaltenssequenzen, so kann eine Ereignisstichprobe gezogen werden. Auf das obige Beispiel bezogen heißt dies: Bei jedem die Kreuzung passierenden Auto wird erfasst, ob die am Steuer sitzende Person gerade per Handy telefoniert oder nicht (interessierendes Ereignis).

Alltagsbeobachtungen sind in ihrer generalisierenden Aussagekraft nicht nur deswegen sehr beschränkt, weil es sich um unsystematische Einzelbeobachtungen handelt, sondern auch, weil **subjektive Wertungen und Überzeugungen** unkontrolliert einfließen. Was der genannte Psychotherapeut unter „Verunsicherung in der männlichen Geschlechtsidentität“ versteht, ist nicht im Einzelnen bekannt. Möglich wäre es, dass andere Therapeutinnen und Therapeuten die Verhaltensweisen derselben Patienten nach einer Beobachtung ganz anders beschreiben, vielleicht als „gesteigerte Reflexion der eigenen Geschlechtsrolle“.

Die wissenschaftliche Beobachtung ermöglicht durch ein regelgeleitetes Vorgehen die Gewinnung **objektiver (d. h. intersubjektiv nachvollziehbarer) Erkenntnisse**: Es wird transparent gemacht und begründet, welche konkreten beobachteten Verhaltensweisen in welcher Weise interpretiert werden. Wenn diese **Beobachtungsregeln** bekannt und – bei quantitativen Beobachtungen – in ein entsprechend überprüftes standardisiertes Beobachtungsschema gebracht sind, sollten unterschiedliche Forschende zu denselben Daten und Ergebnissen kommen, wenn sie dieselben Phänomene beobachten. Diese Festlegungen betreffen die Perspektivität, Selektivität und Konstruktivität/Konstruiertheit des Beobachtungsprozesses. Durch Beobachtung ist es weder in den Natur- noch den Sozialwissenschaften möglich, vollkommen voraussetzungs- und theorielos einfach festzustellen, „wie die Realität beschaffen ist“. In jeden Beobachtungsprozess gehen Vorannahmen ein, die das Ergebnis beeinflussen. Diese müssen bei einer wissenschaftlichen Beobachtung als solche reflektiert und kontrolliert werden. Die gründliche Schulung von wissenschaftlichen Beobachtenden, der parallele Einsatz mehrerer Beobachtender sowie die Überprüfung der

Überstimmung ihrer Beobachtungsdaten sichern gegen Beobachtungsfehler ab. Eine seriöse Beobachtungsstudie benennt und problematisiert die Grenzen ihrer Aussagekraft.

Eine entsprechend kritische Reflexion von Alltagsbeobachtungen ist ebenfalls wünschenswert und nützlich. Angenommen, ein Sozialarbeiter stellt fest, dass in dem von ihm betreuten Jugendtreff Sprache und Benehmen der Heranwachsenden angesichts zunehmenden Pornografiekonsums durch „sexuelle Verrohung" geprägt sind. Wie ist diese anekdotische Alltagsbeobachtung zu bewerten?

- **Perspektivität der Beobachtung:** Mit welchem Ziel, aus welcher Perspektive, von welchem Standpunkt aus wird beobachtet? Sieht der Sozialarbeiter „seine Jugendlichen" v. a. als Opfer der Medienwelt? Traut er ihnen eigene Beurteilungsmaßstäbe und Aneignungsformen zu? Welche negativen und/oder positiven Erfahrungen hat er selbst mit pornografischen Mediendarstellungen gemacht? Welche Vorstellungen vom richtigen und angemessenen sexuellen Ausdruck von Mädchen und Jungen liegen seiner Beobachtung einer „Verrohung" zugrunde?
- **Selektivität der Beobachtung:** Was geht in die Beobachtung ein, und was geht nicht ein? Was der Sozialarbeiter im Jugendtreff beobachtet ist nur ein kleiner Ausschnitt des Verhaltens. Wie verhalten sich die Jugendlichen, wenn er oder andere Erwachsene nicht dabei sind? Wie verhalten sich die Jugendlichen in geschlechterheterogenen und -homogenen Gruppen, wie in intimen Situationen? Wie verhalten sich Jugendliche aus verschiedenen Szenen, Milieus, Kulturen, die den betrachteten Jugendtreff gar nicht aufsuchen?
- **Konstruktivität/Konstruiertheit der Beobachtung:** Wie werden die Beobachtungen registriert, dokumentiert und interpretiert? Mit welchen Begriffen und Konzepten wird auf das beobachtete Geschehen Bezug genommen? So ist „Verrohung" verknüpft mit Konzepten wie „Barbarei", „Brutalität", „Kaltblütigkeit", „Geschmacklosigkeit" und „Unkultur". Würde man zur Beschreibung des beobachteten Verhaltens der Jugendlichen andere Begriffe verwenden (z. B. „Sexualisierung", „Enthemmung", „Provokation", „Parodie" etc.), würden wiederum ganz andere Sinngebungen naheliegen.

Alltagsbeobachtung – Für die Alltagsbeobachtung („everyday observation") sind willkürliche Einzelbeobachtungen, die subjektiv interpretiert und bewertet werden, typisch (sog. anekdotische Evidenzen).

Wissenschaftliche Beobachtung – Bei der wissenschaftlichen Beobachtung („scientific observation") erfolgt die Datenerhebung im Rahmen eines empirischen Forschungsprozesses in geplanter und dokumentierter Weise. Perspektivität, Selektivität und Konstruiertheit des Beobachtungsprozesses werden offengelegt, theoretisch reflektiert und kontrolliert (z. B. gleichzeitiger Einsatz mehrerer Beobachtender und Vergleich der Ergebnisse), um Beobachtungsfehler zu minimieren.

Klassifikationskriterien für wissenschaftliche Beobachtungsmethoden

In der empirischen Sozialforschung werden verschiedene Arten der Beobachtung anhand von **sechs zentralen Klassifikationskriterien** unterschieden. Diese Kriterien sind heranzuziehen, um den wissenschaftlichen Beobachtungsprozess einer Studie einzuordnen sowie eine eigene Datenerhebung mittels Beobachtungsmethode zu planen.

1. **Strukturierungsgrad der Beobachtung**
   Wir unterscheiden wissenschaftliche Beobachtungen gemäß dem Grad ihrer Strukturierung. Der Grad der Strukturierung bezieht sich dabei auf die Beobachtungssituation, der **Grad der Standardisierung** dagegen auf das verwendete Beobachtungsinstrument. Bei der **unstrukturierten Beobachtung** („unstructured observation") werden visuelle und verbale Beobachtungsdaten zum interessierenden Beobachtungsgegenstand ohne jegliche Vorgabe von Beobachtungsrichtlinien erhoben. Diese „offene" oder „freie" Beobachtung ist das Pendant zum nicht-strukturierten offenen Interview, dem weder ein vorher festgelegter Fragenkatalog noch vorgegebene Antwortalternativen zugrunde liegen. Bei der **teil-strukturierten Beobachtung** („semi-structured observation") existieren konkrete Forschungsfragen oder theoretische Konzepte, an denen sich die Erfassung der Beobachtungsdaten orientiert. Die teil-strukturierte Beobachtung entspricht in der Welt der Befragungsmethoden dem Leitfadeninterview, bei dem die Fragen vorgegeben, jedoch die Antwortmöglichkeiten offen sind. Bei der **(voll)strukturierten Beobachtung** („structured observation") schließlich werden für vorher genau festgelegte Variablen mit festgelegten Ausprägungen anhand eines Beobachtungschemas numerische Messwerte erhoben. Die vollstrukturierte Beobachtung ist das Pendant zur vollstrukturierten Befragung. Die strukturierte Beobachtung wird auch als systematische oder geschlossene Beobachtung bezeichnet. Der Beobachtungsbogen als standardisiertes Erhebungsinstrument bei der strukturierten Beobachtung entspricht dem standardisierten Fragebogen bei der Befragung. In der Praxis hat es sich eingebürgert, nicht-strukturierte und teil-strukturierte Beobachtungsmethoden zusammenfassend als **qualitative Beobachtung** („qualitative observation") zu etikettieren und sie der **quantitativen Beobachtung** („quantitative observation"), also der (voll)strukturierten Beobachtung, gegenüberzustellen.
2. **Gegenstand der Beobachtung**
   Beobachtet werden in der qualitativen wie quantitativen Forschung meist Merkmale und Verhaltensweisen anderer Personen (**Fremdbeobachtung**, „external observation"). Die fokussierten Beobachtungseinhei-

ten können auf einzelne Verhaltensaspekte beschränkt (qualitative oder quantitative Beobachtung mit geringerem Komplexitätsgrad) oder komplexer sein und z. B. Muster untersuchen (ethnografische Feldbeobachtung; strukturierte Verhaltensbeobachtung). Auch das eigene Verhalten der Forscherin bzw. des Forschers kann zum Forschungsgegenstand werden (**Selbstbeobachtung**, „introspection"). Hierzu existiert in der qualitativen Forschung die Methode der Autoethnografie („autoethnography"). In der quantitativen Forschung sind Selbstbeobachtungen der Forschenden kaum relevant. Selbstbeobachtungen der Untersuchungspersonen spielen in der quantitativen wie qualitativen Forschung eine wichtige Rolle. Sie werden in der Forschung mittels mündlicher und schriftlicher Befragungsmethoden (z. B. biografisches Interview, Persönlichkeitstest, lautes Denken, Tagebuchmethode) erfasst.

3. **Direktheit der Beobachtung**
   Neben der direkten Fremd- und Selbstbeobachtung existiert die **indirekte Beobachtung** („indirect observation") bzw. **Beobachtung von Verhaltensspuren** („observation of behavioral traces"). Bei der indirekten Beobachtung kommt es zu keinerlei Kontakt zwischen Beobachtenden und Beobachteten, deswegen wird auch von **non-reaktiver Beobachtung** („non-reactive observation") gesprochen. Die non-reaktive Beobachtung bezieht sich auf **quantifizierbare Verhaltensspuren** (z. B. Umfang und Zusammensetzung des Hausmülls; öffentlich einsehbare, nutzerspezifische Daten über die Häufigkeit des Postens von Beiträgen in einem Onlineforum). Die Erhebung reichhaltiger, interpretativ auswertbarer „Spuren" bzw. Dokumente menschlichen Handelns (z. B. Briefe, Akten, Familienfotos, Wohnungseinrichtungen, Kunstwerke) wird traditionell nicht der Beobachtung, sondern der Dokumentenanalyse (► Abschn. 10.6) zugeordnet. Die entsprechenden Dokumente oder Artefakte (im Unterschied zu einfachen Verhaltensspuren) eignen sich weniger, um konkrete Verhaltensabläufe im Alltag nachzuvollziehen, sondern lassen sich eher als Ausdruck überdauernder Einstellungen, Werte, Beziehungen etc. interpretieren.
4. **Ort der Beobachtung**
   Beobachtungen können im Lebensumfeld (**Feldbeobachtung**; „field observation") oder in kontrollierten Laborsituationen (**Laborbeobachtung**; „laboratory observation") stattfinden. Zudem kann das Verhalten in physischen Umgebungen (**Offline-Beobachtung**; „offline observation") oder auch in virtuellen bzw. Online-Umgebungen beobachtet werden (**Online-Beobachtung**; „online observation").
5. **Involviertheitsgrad der Beobachterrolle**
   Bei der Fremdbeobachtung ist zu unterscheiden, ob Mitglieder des Forschungsteams sich auf reine Beobachtung aus der Außenperspektive beschränken (**nicht-teilnehmende Beobachtung**; „non-participant observation") oder ob sie am beobachteten sozialen Geschehen partizipieren und mit den beobachteten Personen in Kontakt treten (**teilnehmende Beobachtung**; „ participant observation"). Bei der teilnehmenden Beobachtung lassen sich unterschiedliche Grade der Involvierung voneinander abgrenzen: So können Forschende primär beobachten und passiv am Feldgeschehen teilnehmen (**passive Teilnahme**; „observer as participant") oder sie nehmen primär am Feldgeschehen teil und beobachten nebenbei (**aktive Teilnahme**; „participant as observer"). Die Wahl der Vorgehensweise hängt vom Erkenntnisinteresse sowie den Merkmalen des Beobachtungsfeldes ab: So mag bei der teilnehmenden Beobachtung einer Unterrichtsstunde eine passive Teilnehmerrolle genügen, während die teilnehmende Beobachtung einer Expedition eine deutlich aktivere Beteiligung am Feldgeschehen verlangt.
6. **Transparenz der Beobachtung**
   Bei der Fremdbeobachtung stellt sich – unabhängig vom Involvierungsgrad der Beobachterrolle – noch die Frage nach der Transparenz des Vorgehens: Wissen die beobachteten Personen, dass ihr Verhalten gerade wissenschaftlich untersucht wird (**offene Beobachtung**; „overt observation") oder wird die Beobachtung heimlich durchgeführt (**verdeckte Beobachtung**; „covert observation")? Eine verdeckte Beobachtung kann stattfinden, indem sich die Beobachtenden im Feld nicht als Forschende zu erkennen geben, sondern als vermeintlich reguläre Feldteilnehmende auftreten. Möglich sind auch Videoaufzeichnungen oder die Verwendung von Untersuchungsräumen mit Einwegscheiben, so dass sich Beobachtete und Beobachtende nicht direkt begegnen. Bei der verdeckten Beobachtung ist die Einhaltung ethischer Grundsätze wichtig (zur Forschungsethik ► Abschn. 4.1). Es ist zu beachten, dass die Bezeichnung „offene" Beobachtung völlig unterschiedliche Bedeutungen hat: Einerseits kann sie auf eine nicht-strukturierte Beobachtung (siehe oben Punkt 1) verweisen. Andererseits steht die Bezeichnung für eine transparente Beobachtung, bei der die beobachteten Personen wissen, dass ihr Verhalten wissenschaftlich erfasst wird.

Nicht alle Ausprägungen der genannten sechs Kriterien sind sinnvoll miteinander kombinierbar. Stattdessen haben sich bestimmte Konstellationen etabliert: So finden

Tabelle 10.1 Verbreitete Formen der wissenschaftlichen Beobachtung und ihre Klassifikation

Wissenschaftliche Beobachtung						
Grad der Strukturierung	**Keine oder geringe Strukturierung: Qualitative Beobachtung**			**Starke Strukturierung: Quantitative Beobachtung**		
Gegenstand der Beobachtung	Fremdverhalten	Eigenes Verhalten	Verhaltensspuren	Fremdverhalten	Eigenes Verhalten	Verhaltensspuren
Formen der wissenschaftlichen Beobachtung	– Qualitative Beobachtung mit geringem Komplexitätsgrad – Ethnografische Feldbeobachtung	– Autoethnografie	–	– Quantitative Beobachtung mit geringem Komplexitätsgrad – Strukturierte Verhaltensbeobachtung		– Nonreaktive Beobachtung von Verhaltensspuren

nicht-strukturierte bzw. qualitative Beobachtungen in der Regel als Feldbeobachtungen unter passiver oder aktiver Beteiligung der Forschenden statt. Demgegenüber sind strukturierte bzw. quantitative Beobachtungen meistens nicht-teilnehmend ausgelegt und können sowohl im Feld als auch im Labor angesiedelt sein.

Welche Beobachtungsformen in diesem Kapitel näher erläutert werden, ist noch einmal in Tab. 10.1 dargestellt: Entscheidend sind der Grad der Strukturierung (geringe vs. hohe Strukturierung) sowie der Gegenstand der Beobachtung (Fremdverhalten, eigenes Verhalten, Verhaltensspuren).

Beobachtungsfehler und Gegenmaßnahmen

Um die Voraussetzungen und Grenzen des wissenschaftlichen Beobachtungsprozesses noch präziser und grundlegender zu beschreiben, ist auf drei Forschungsfelder zurückzugreifen:

1. Die Leistungsfähigkeit von Menschen als Beobachtenden (und somit als menschlichen Erhebungs- bzw. Messinstrumenten) ist Gegenstand der **Psychologie** (v. a. Wahrnehmungs-, Kognitions- und Gedächtnis- sowie Sozialpsychologie). Beim Einsatz von menschlichen Beobachtenden ist aufgrund diverser psychologischer Prozesse mit Verzerrungen, sog. Beobachter- oder **Beobachtungsfehlern („observer error“)**, zu rechnen. Insbesondere Wahrnehmungs-, Erinnerungs- und Deutungsfehler können auftreten (Greve & Wentura, 1997, S. 60; Häder, 2006, S. 303ff.). Diese werden im Rahmen wissenschaftlicher Beobachtungen teilweise kontrolliert und reduziert sowie bei der Interpretation von Ergebnissen ausdrücklich berücksichtigt.
2. Die Leistungsfähigkeit von technischen Geräten und Sensoren, die zu Beobachtungszwecken eingesetzt werden, ist Gegenstand der **Ingenieurwissenschaft** (v. a. Informatik, Messtechnik, Biomedizintechnik etc.). Auch hier ist mit mehr oder minder starken Messfehlern zu rechnen.
3. Die Möglichkeiten und Grenzen, mittels Beobachtung überhaupt zu gültigen wissenschaftlichen Erkenntnissen zu gelangen, werden nicht zuletzt innerhalb der **Philosophie** bearbeitet (v. a. im Rahmen der Erkenntnis- und Wissenschaftstheorie; ▶ Kap. 2).

Um im Zuge einer wissenschaftlichen Beobachtung zu möglichst aussagekräftigen Daten zu gelangen, sind Beobachtungsfehler zu vermeiden. Dabei ist nach Beobachtungsfehlern zu differenzieren, die a) auf die Beobachtungssituation und b) auf die Person des Beobachters zurückzuführen sind (Häder, 2006, S. 303ff.).

Fehler durch die Beobachtungssituation

Bei der Planung einer Beobachtungsstudie ist darauf zu achten, hinsichtlich der **Stichprobenauswahl** (Beobachtungsorte, -zeiten, -objekte und -einheiten) sinnvolle Ausschnitte der empirischen Wirklichkeit zu wählen, etwa besonders typische, normale oder bewusst extreme Konstellationen (Saunders, Lewis, & Thornhill, 2003). Dies setzt theoretische und/oder empirische Vorkenntnisse über den Gegenstand voraus. Grob lassen sich drei Fehler durch die Beobachtungssituation unterscheiden:

1. Die **Reaktivität der Beobachtungssituation**, d. h. die nicht-intentionale Beeinflussung des Verhaltens der beobachteten Personen durch den Beobachtungsprozess, ist gering zu halten. Eine Änderung des Verhaltens kann dadurch entstehen, dass Beobachtungspersonen a) bei einer nicht-verdeckten Beobachtung wissen, dass sie unter Beobachtung stehen und/oder b) bei einer teilnehmenden Beobachtung vom Verhalten der Beobachtenden direkt beeinflusst werden. Reaktivität nimmt in der Regel nach einer kurzen Eingewöhnungsphase ab, so dass es sinnvoll ist, erst da-

nach mit der Datenerfassung zu beginnen (Sedlmeier & Renkewitz, 2007). Bei der teilnehmenden Feldbeobachtung sollten die Forschenden ihre mögliche Beeinflussung des Beobachtungsfeldes reflektieren, in Feldnotizen dokumentieren und im Team diskutieren. Wird die Reaktivität der Beobachtungssituation bei der Datenerhebung gänzlich ausgeschlossen (etwa indem nicht das Verhalten direkt, sondern nur Verhaltensspuren beobachtet werden), so spricht man von einer **nicht- oder nonreaktiven Beobachtungsmethode** (▶ Abschn. 10.1.3 „Nonreaktive Beobachtung von Verhaltensspuren“).

2. Insbesondere bei Laborbeobachtungen kann die **Künstlichkeit der Beobachtungssituation** verfälschen. Etwa wenn Versuchspersonen sich in einem Laborraum anders verhalten als im Alltag (Problem der ökologischen Validität). Bei welchen konkreten Forschungsfragen oder Beobachtungspersonen die Künstlichkeit der Situation zu nennenswerten Verfälschungen führt, ist vorab theoretisch und/oder empirisch zu klären. So kann die Beobachtung von Kleinkindern im Labor kontrainduziert sein, wenn diese die ungewohnte Umgebung als beängstigend empfinden.
3. Nicht zuletzt können **Fehler bei der praktischen Durchführung von Beobachtungen** dadurch entstehen, dass in bestimmten Situationen die zur Dokumentation notwendige Videokamera versagt, Feldnotizen verloren gehen oder sich das Beobachtungsinstrument als unzureichend erweist und unerwartete Geschehnisse nicht erfasst. Entsprechende Fehler sind durch Pretests der Beobachtungsinstrumente sowie Erprobung der technischen Hilfsmittel unter Realbedingungen zu vermeiden.

■ Fehler durch die Person des Beobachters

Die Beobachtenden fungieren bei der Beobachtung selbst als Datenerhebungsinstrumente: Sie müssen zunächst die die interessierenden Aspekte der beobachteten sozialen Wirklichkeit entdecken und wahrnehmen, sie kognitiv verarbeiten und beurteilen, mehr oder minder lange im Gedächtnis behalten und schließlich als qualitatives oder quantitatives Datenmaterial angemessen aufzeichnen (z. B. Feldnotizen im Anschluss an eine ethnografische Feldbeobachtung oder Ausfüllen eines standardisierten Beobachtungsbogens während einer strukturierten Verhaltensbeobachtung im Labor). Für die Bewältigung dieser Aufgaben sind kognitive Leistungen erforderlich, die fehleranfällig sind. Grob lassen sich vier Typen von Fehlern durch die Beobachtenden differenzieren, die prinzipiell sowohl bei nicht- oder teilstrukturierten als auch bei vollstrukturierten Beobachtungen auftreten können (Greve & Wentura, 1997, S. 60; Häder, 2006, S. 303ff.).

1. **Wahrnehmungsfehler**
   Die Grundprobleme der Beobachtung, nämlich ihre Konstruiertheit und Perspektivität spiegeln sich in Wahrnehmungsfehlern wider. So bemerken Beobachtende im Sinne selektiver Wahrnehmung v. a. jene Aspekte des Beobachtungsgeschehens, die sie aufgrund von Vorannahmen erwarten, die sie persönlich interessant finden oder die dem Gesamteindruck entsprechen (Halo-Effekt), während andere Aspekte evtl. ausgeblendet bleiben. Gerade bei Feldstudien entstehen Wahrnehmungsfehler durch den Grad der Vertrautheit der Forschenden mit dem Forschungsgegenstand. Gehören die Forschenden z. B. selbst zu der sozialen Gruppe, die sie beobachten, besteht die Gefahr eines Verlusts an wissenschaftlicher Distanz zum Forschungsthema („going native“; z. B. verstärkte Wahrnehmung von positiven Aspekten), gehören sie indessen nicht selbst zu der beobachteten sozialen Gruppe, so besteht das erhöhte Risiko von unreflektierten Abgrenzungen („othering“; z. B. verstärkte Wahrnehmung von negativen Aspekten). Weiterhin können Beobachtende durch Ermüdung oder mangelnde Motivation gedanklich von den zu fokussierenden Beobachtungseinheiten abdriften bzw. die bei der Beobachterschulung erlernten Prinzipien der Wahrnehmung vernachlässigen („observer drift“).
2. **Interpretations- bzw. Urteilsfehler**
   Wenn wahrgenommene Merkmale oder Verhaltensweisen bei der strukturierten Beobachtung auf Ratingskalen einzuschätzen sind, kommen eine Reihe möglicher Urteilsfehler zum Tragen (Cranach & Frenz, 1969; Sedlmeier & Renkewitz, 2007), etwa die Tendenz zur Mitte (Vermeidung von Extremwerten an den Endpunkten der Skala), Ankereffekte (Abhängigkeit des Urteils von der Verankerung der Skalenenden z. B. durch verbale Etiketten), aber auch persönliche Urteilstendenzen (z. B. Tendenz mancher Beobachter, stets besonders streng oder besonders milde über andere Personen zu urteilen). Bei nicht-strukturierten Beobachtungen können implizite Theorien oder Vorurteile sowie auch die Identifikation mit den Beobachteten dazu führen, dass das Beobachtungsgeschehen in verzerrter Weise interpretiert wird. Entsprechende Fehlertendenzen sind durch die Auswahl geeigneter Beobachter/innen sowie Beobachterschulungen zu reduzieren.
3. **Erinnerungsfehler**
   Menschliche Gedächtniskapazitäten sind begrenzt. So werden Ereignisse, die von der Normalität abweichen, die häufig auftreten, die ganz am Anfang oder ganz am Ende einer Beobachtungsperiode stattfinden, besser erinnert als andere. Erinnerungsfehlern ist u. a. durch

Foto-, Audio- und Videoaufzeichnungen des Geschehens, eine Verkürzung von Beobachtungszeiten, eine parallele Protokollierung während des Beobachtens sowie durch Beobachterschulungen entgegen zu wirken.

4. **Wiedergabefehler**
   Auch Beobachtungsfakten, die korrekt wahrgenommen, interpretiert und erinnert werden, können im Beobachtungsschema oder in Feldnotizen verzerrt wiedergegeben werden, etwa weil Beobachter/innen vermeiden möchten, dass die Beobachtungspersonen in schlechtem Licht dastehen, weil sie sich Arbeit sparen möchten oder weil für die Dokumentation zu wenig Zeit zur Verfügung steht. Auch hier sind ggf. Beobachterschulungen sowie ggf. Beobachterkontrollen während der Studie indiziert und vielleicht auch die Modalitäten der Protokollierung zu optimieren (z. B. Vereinfachung von Beobachtungsbögen).

10

Maßnahmen gegen Beobachtungsfehler

Maßnahmen gegen Beobachtungsfehler wurden vereinzelt bereits angesprochen. Sie umfassen vor allem drei Ansätze: 1.) die **Auswahl geeigneter Personen für die Beobachterrolle** (z. B. Vorerfahrungen, Motivation, Zuverlässigkeit), 2.) die Durchführung gründlicher **Beobachterschulungen** sowie 3.) bei umfassenderen Studien **zwischenzeitliche Überprüfungen** der Qualität der Arbeit der Beobachtenden (z. B. durch temporären Einsatz von Co-Beobachtenden und Prüfung der Übereinstimmung ihrer Beobachtungsdaten).

Beobachterschulungen können unterschiedlich aufgebaut sein. Zum einen können sie den Charakter einer Informationsveranstaltung haben, zum anderen können sie eine Kombination aus Informationselementen und praktischen Übungen darstellen. Reine Informationsveranstaltungen, die über die oben genannten Beobachtungsfehler aufklären, haben nur begrenzten Nutzen, da sie in geringem Umfang verhaltensrelevante Hinweise enthalten (Obermann, 2006). Vielversprechender sind Schulungen, die den Blick für relevante Beobachtungseinheiten schärfen und die Fähigkeit zur Differenzierung zwischen verschiedenen Beobachtungsdimensionen verbessern. Häufig eingesetzt werden Beobachterschulungen in der Praxis zur Vorbereitung der Beobachtenden von Assessment-Centern, die der Personalauswahl dienen, aber auch für Forschungszwecke eingesetzt werden. Im Folgenden sind typische Elemente einer Beobachterschulung und Methoden zu deren Realisierung aufgeführt (Obermann, 2006; Schnell, Hill, & Esser, 2008):

- Vermittlung von Informationen über die Absicht der Beobachtungsstudie (Vortrag).
- Kennenlernen der Methode und der eigenen Wahrnehmungstendenzen durch Probebeobachtungen (z. B. anhand von Videoaufzeichnungen).
- Bei strukturierten Beobachtungen: Vorstellung des Beobachtungsschemas mit seinen Dimensionen und Items; bei nicht-strukturierten Beobachtungen: Informationen zur Gestaltung von Feldnotizen (Vortrag).
- Durchführung von mehreren Probebeobachtungen anhand von Rollenspielen oder Videoaufzeichnungen (Erfahrungsaustausch, individuelles Feedback an die Beobachtenden).
- Beobachtungen unter „Ernstbedingungen“ (z. B. Durchführung von Beobachtungen im Labor oder im Feld, Erfahrungsaustausch, individuelles Feedback an die Beobachtenden).
- Prüfung der Beobachterübereinstimmung, bei qualitativen Beobachtungen interpretativ, bei quantitativen Beobachtungen statistisch.

Je nach Ziel der Studie kann es sinnvoll sein, die Beobachtenden über das Erkenntnisinteresse der Beobachtungsstudie im Unklaren zu lassen. Für eine solche Entscheidung sind Vorüberlegungen zu den Auswirkungen der Informationen auf den Beobachtungsprozess notwendig. Während bei strukturierten Beobachtungen eine erfolgreiche Vorabschulung ausreichend ist, sind bei umfassenderen qualitativen Beobachtungen im Zeitverlauf wiederholte Rücksprachen mit den Beobachtenden im Team sinnvoll, da hier der Forschungsprozess dynamischer ist.

10.1.2 Qualitative Beobachtung

Das offene, **nicht-strukturierte Vorgehen** bei der qualitativen Beobachtung ermöglicht es, Phänomene besonders detailreich zu erfassen (meist in Form verbaler, visueller bzw. audiovisueller Daten) sowie im Verlauf des Forschungsprozesses unerwartete Aspekte entdecken zu können. Demgegenüber konzentriert sich die quantitative Beobachtung auf die präzise Messung vorher definierter Merkmale oder Verhaltensweisen, wobei eine deutlich größere Menge an **Beobachtungseinheiten** untersucht werden kann als beim qualitativen Vorgehen. Die qualitative Beobachtung ist typischerweise eine **Feldbeobachtung**, sie findet in der natürlichen Umwelt der beobachteten Akteure statt und nicht in Laborsituationen. Dabei kann es sich bei dem untersuchten Beobachtungsfeld um **Offline-Kontexte** handeln (z. B. Beobachtung im Schulunterricht, in einem Schnellrestaurant, in einer Spielhalle, auf einer Baustelle) oder auch um **Online-Kontexte** (z. B. Beobachtung in einem Online-Konferenzraum, in einem Online-Mehrpersonenspiel, in einem Online-Chat).

Es ist zu beachten, dass eine völlig unfokussierte offene Beobachtung aufgrund mangelnder Zielorientierung nicht wissenschaftlich wäre. Vielmehr ist es typisch für

den qualitativen Forschungsprozess, sich spätestens im Verlauf einer längeren Feldbeobachtungsstudie geeignete theoretische Konzepte zu erarbeiten, um die Konstruiertheit des Beobachtungsprozesses reflektieren und diskutieren zu können und nicht den impliziten Theorien und Vorannahmen der Forschenden zu überlassen. In der Praxis besteht ein fließender Übergang zwischen der nichtstrukturierten und der teilstrukturierten qualitativen Beobachtung.

Qualitative Beobachtung – Bei der qualitativen Beobachtung („qualitative observation") werden verbale, visuelle bzw. audiovisuelle Daten erhoben, die den jeweiligen Gegenstand der Beobachtung sehr detailreich repräsentieren. Besonders einschlägig sind verbale Daten in Form von Beobachtungsprotokollen. Qualitative Beobachtung erfolgt häufig als (aktiv oder passiv) teilnehmende Beobachtung („participant observation") im Feld.

Wofür ist die qualitative Beobachtung besonders geeignet (Indikation)? Wenn qualitativ beobachtet wird, liegt das Augenmerk auf den **sozialen Interaktionen zwischen Handelnden in ihrer jeweiligen Lebenswelt**, die hinsichtlich ihrer Muster und Bedeutungen rekonstruiert werden sollen. In Wissenschaftsdisziplinen und Forschungsfeldern, die sich für zwischenmenschliche Interaktionen im Feld interessieren, spielt die qualitative Beobachtung eine wichtige Rolle (z. B. Mikrosoziologie, Sozialpsychologie, Arbeits- und Organisationspsychologie, Geschlechterforschung, Erziehungswissenschaft, Pflegeforschung). Auch in der Technikforschung ist die qualitative Beobachtung nützlich. Denn es zeigt sich, dass Menschen sich nicht nur oft gemeinsam mit Technologien befassen, sondern dass sie technische Artefakte (z. B. Computer) ihrerseits oft als soziale Akteure behandeln, mit ihnen sprechen und ihnen Emotionen zuschreiben.

Da mittels qualitativer Beobachtung anstrebt wird, Daten zu erheben, auf deren Basis die beobachteten sozialen Interaktionen dann detailliert und sinnverstehend rekonstruiert werden können, müssen die Beobachtenden selbst den Sinn des Beobachteten entschlüsseln. Die qualitative Beobachtungsmethode und ihre Gestaltung und Umsetzung ist an eine **theoretische Reflexion** dessen gebunden, was soziale Interaktionen und deren Bedeutung ausmacht. Die zentrale Frage der qualitativen Beobachtung sozialer Situationen und Interaktionen lautet somit: „Was geht hier eigentlich vor?" (Goffman, 1996, S. 16). Diese Frage stellt sich gemäß der **Theorie des symbolischen Interaktionismus** für die Handelnden selbst, die sich gegenseitig beobachten und ihr Handeln wechselseitig aufeinander abstimmen. Und sie stellt sich dementsprechend auch aus Sicht der qualitativ Forschenden bzw. Beobachtenden, die diese sozialen Prozesse sinnverstehend nachvollziehen möchten (Blumer, 1973; Goffman, 1971, 1996). Entwickelt wurde der symbolische Interaktionismus in den 1930er-Jahren am soziologischen Institut der Universität Chicago in den USA. Der Soziologe **Herbert Blumer** (1973, S. 81) fasste die drei Grundannahmen der Theorie folgendermaßen zusammen:

1. Menschen handeln Dingen gegenüber aufgrund der Bedeutung, die diese Dinge für sie haben.
2. Diese Bedeutung entsteht in einem Interaktionsprozess.
3. Die Bedeutung wird von der Person in Auseinandersetzung mit den Dingen selbst interpretiert, daraufhin entsprechend gehandhabt und geändert.

Der wichtigste Vertreter des symbolischen Interaktionismus ist der ebenfalls der **Chicagoer Schule** zuzurechnende Soziologe **Erving Goffman**, der in seinem Werk ein sehr differenziertes Begriffsinstrumentarium zur Beschreibung sozialer Interaktionen entwickelt hat, das bis heute vielfach angewendet wird (Goffman, 1971, 1996). So schlägt Goffman (1969) zur Beobachtung und Beschreibung menschlichen Verhaltens u. a. die **Metapher des Theaters** vor: Akteure in sozialen Situationen agieren demnach wie Darstellende auf einer Bühne; sie nehmen zur Situation passende Rollen ein, die sie in weniger beobachteten Momenten – auf der Hinterbühne – ablegen bzw. vorbereiten (z. B. Verhalten des Pflegepersonals im Krankenzimmer: Vorderbühne versus im Personalraum: Hinterbühne). Kenntnisse der Theorie des symbolischen Interaktionismus (und/oder anderer Theorien, die soziale Interaktionen betreffen, z. B. psychoanalytische Theorien) stellen ein wichtiges Hilfsmittel für qualitativ Beobachtende dar, weil sie **allgemeine Deutungskategorien** liefern, mit denen das Beobachtete während des Beobachtens interpretiert werden kann. Ein gänzlich theoriefreies Beobachten ist schließlich nicht möglich (zum Prinzip der relativen theoretischen Offenheit qualitativer Forschung ► Abschn. 2.3.2).

Welche Formen der qualitativen Beobachtung lassen sich unterscheiden? Zur Differenzierung sind die bereits vorgestellten Dimensionen zur Klassifikation von Beobachtungsverfahren heranzuziehen: Gegenstand der Beobachtung, Direktheit der Beobachtung, Ort der Beobachtung, Involviertheitsgrad der Beobachterrolle und Transparenz der Beobachtung. Besonders wichtig ist dabei die Differenzierung nach dem Beobachtungsgegenstand, da Selbst- und Fremdbeobachtung eine jeweils andere Vorgehensweise bei der qualitativen Datenerhebung erfordern. Zudem ist bei der Fremdbeobachtung noch nach der Komplexität der Beobachtungseinheiten zu unterscheiden. Im Folgenden werden dementsprechend drei zentrale und häufig umgesetzte Varianten der qualitativen Beobachtung vorgestellt:

1. die **qualitative Beobachtung mit geringem Komplexitätsgrad**, die sich als Fremdbeobachtung auf einzelne Aspekte des sozialen Geschehens konzentriert,
2. die **ethnografischen Feldbeobachtung**, die als Fremdbeobachtung das Ziel verfolgt, komplexe Interaktionsabläufe und soziale Welten umfassend zu rekonstruieren und
3. die **Autoethnografie**, die nicht die Fremdbeobachtung sondern die Selbstbeobachtung der Forschenden in den Mittelpunkt rückt.

Spezialformen der qualitativen Beobachtung, wie z. B. die psychoanalytische Kinderbeobachtung, werden hier nicht behandelt. Dazu sei auf einschlägige Fachpublikationen (z. B. die Fachzeitschrift *Infant Observation*) verwiesen.

Qualitative Beobachtung mit geringem Komplexitätsgrad

Ziel der qualitativen Beobachtung ist es, soziales Geschehen sinnverstehend und möglichst ganzheitlich zu erfassen. Dies setzt in der Regel längere Beobachtungszeiträume voraus, wie sie in der ethnografischen Feldforschung sowie bei der Autoethnografie gegeben sind. Zuweilen kommt es aber auch vor, dass für ein konkretes Forschungsproblem nur Einzelaspekte des beobachteten Geschehens wichtig sind. In diesem Fall ist eine qualitative Beobachtung mit geringem Komplexitätsgrad indiziert: Man konzentriert sich von vorne herein auf bestimmte Ausschnitte des beobachteten Geschehens.

Qualitative Beobachtung mit geringem Komplexitätsgrad – Die Qualitative Beobachtung mit geringem Komplexitätsgrad („qualitative observation with a lesser//lower degree of complexity") ist eine teilstrukturierte Beobachtung des Handelns anderer Personen (Fremdbeobachtung), die sich auf vorher festgelegte einzelne Aspekte des Verhaltens konzentriert. Die interessierenden sozialen Sachverhalte werden von den Beobachtenden im Beobachtungsprotokoll in eigenen Worten beschrieben. Die qualitative Beobachtung mit geringem Komplexitätsgrad findet häufig als passiv teilnehmende Feldbeobachtung statt, sie kann aber auch im Labor durchgeführt werden.

Indem man einzelne Aspekte, auf die sich die Beobachtung konzentrieren soll, vorher festlegt, erfolgt keine unstrukturierte, sondern eine teilstrukturierte Datenerhebung. Im Unterschied zur vollstrukturierten Beobachtung wird jedoch der interessierende Teilaspekt nicht nur in seiner Auftretenshäufigkeit ausgezählt oder in seiner Intensität bewertet, sondern in seinen qualitativen Merkmalen beschrieben. Das ►Studienbeispiel „Qualitative Beobachtung mit geringem Komplexitätsgrad" verdeutlicht die Vorgehensweise.

Ethnografische Feldbeobachtung

Bei der ethnografischen Feld**forschung**, deren zentraler Bestandteil die ethnografische Feld**beobachtung** ist, handelt es sich um eine sehr anspruchsvolle Methode, die über eine reine Datenerhebungstechnik hinausgeht und eine eigene **qualitative Forschungsstrategie** darstellt. Es liegen zahlreiche Lehr- und Handbücher zur ethnografischen Feldforschung bzw. Feldbeobachtung vor (Atkinson, Coffey, & Delamont, 2007; Beer, 2008; Girtler, 2001; Hammersley & Atkinson, 2007; O'Reilly, 2005). Die Kenntnis der Methode ist wichtig, um entsprechende Studien bei der Lektüre verstehen und einordnen sowie selbst qualitativ forschen zu können. Zudem können explorative Feldbeobachtungen, die z. B. quantitative Studien vorbereiten oder Teil von Mixed-Methods-Studien sind, auf der Basis methodischer Prinzipien der Feldforschung besser geplant und reflektiert werden.

Ethnografische Feldbeobachtung – Die ethnografische Feldbeobachtung („ethnographic field observation") ist eine nicht-strukturierte Beobachtung der sozialen Verhaltensweisen anderer Personen im natürlichen Umfeld, wobei kontextgebundene, **komplexe Beobachtungseinheiten** (z. B. Interaktionsmuster anstelle einzelner Verhaltensweisen) in den Blick genommen und von den Beobachtenden in eigenen Worten in Feldnotizen beschrieben werden. Kennzeichnend für die ethnografische Feldbeobachtung ist ein umfassendes Eintauchen der Forschenden in das zu beobachtende Feld, teilweise für längere Zeitperioden (teilnehmende Feldbeobachtung, „participant observation", und oft in der Form der aktiven Teilnahme: „participant-as-observer").

Ethnografische Feldforschung – Da es bei der ethnografischen Feldbeobachtung um die Rekonstruktion komplexer Beobachtungseinheiten geht, wird sie meist durch **Feldgespräche** sowie die **Sammlung von Dokumenten und Artefakten** ergänzt und zusammenfassend als ethnografische Feldforschung („ethnographic field research") bezeichnet.

Erste ethnografische Feldbeobachtungen wurden Anfang des 20. Jahrhunderts durchgeführt und bis heute hat dieser Forschungsansatz große Bedeutung in der empirischen Sozialforschung (►Historische Entwicklung der ethnografischen Feldforschung).

■ Ablauf einer ethnografischen Feldbeobachtung

Zu einem ethnografischen Feldforschungsprojekt gehören typischerweise sechs Phasen: 1.) Planung und Vorbereitung, 2.) Feldzugang und Forscherrolle, 3.) Beobachten und Agieren im Feld, 4.) Dokumentation der Feldtätigkeit, 5.) Ausstieg aus dem Feld, 6.) Auswertung und Ergebnisbericht.

Planung und Vorbereitung Neben **organisatorischen Vorbereitungen** (Finanzierung, Zeitplan, ggf. Zusammenstellung eines Forschungsteams) ist inhaltlich die **Präzisierung des Forschungsproblems** (►Abschn. 5.1)

Studienbeispiel

Qualitative Beobachtung mit geringem Komplexitätsgrad

Forschungsfrage

Der Soziologe und Kriminalitätsforscher Patrick Gerkin interessierte sich für das **Verhalten von Täter/innen während des sog. Täter-Opfer-Ausgleichs** (Gerkin, 2009). Beim Täter-Opfer-Ausgleich geht es darum, dass (geringfügige) Straftaten nicht anonym durch den Staat bestraft werden, sondern die Täter/innen sich bei ihren Opfern direkt um eine Wiedergutmachung bemühen, sofern die Opfer mit diesem Vorgehen einverstanden sind. Der Täter-Opfer-Ausgleich soll die Verantwortungsübernahme auf Täterseite begünstigen und Wiederholungstaten eher verhindern als dies die klassische Strafjustiz vermag.

Durchführung der qualitativen Beobachtung

Gerkin nahm in den USA Kontakt mit einer Behörde auf, die Täter-Opfer-Ausgleich organisiert und beobachtete eine Reihe solcher Sitzungen. Dabei agierte er **offen als Forscher** (alle Teilnehmenden einer Sitzung erklärten vorher schriftlich ihr Einverständnis mit der wissenschaftlichen Beobachtung) und nahm eine passiv teilnehmende Rolle ein (er saß im selben Raum, aber nicht mit den Verhandlungspartnern zusammen am Tisch, sondern separiert am Rand und mischte sich nicht in das Geschehen ein). Da zum Sozialverhalten der Täter/innen während eines Täter-Opfer-Ausgleichs keine wissenschaftlichen Ergebnisse und Messinstrumente existierten, wurde ihr Sozialverhalten durch den Beobachter ohne vorgefertigtes Beobachtungsschema offen **in eigenen Worten** beschrieben. Dabei konzentrierten sich Beobachtung und Beschreibung gemäß Forschungsproblem auf **ausgewählte Aspekte des Täter-Opfer-Ausgleichs** (teilstrukturierte Beobachtung), nämlich speziell auf die Art und den Grad der aktiven Teilnahme des Täters oder der Täterin a) am Ablauf der Sitzung und b) an der Aushandlung der Wiedergutmachungsleistungen.

Forschungsergebnis

In der Auswertung der Studie verdichtete der Forscher die beobachteten Verhaltensweisen zu drei Arten der Täterpartizipation (gering – mittel – hoch), die er qualitativ anhand von typischer Körpersprache, Blickkontakt, Gesprächsverhalten etc. beschrieb: Typisch für geringe Partizipation waren z. B. Vermeiden des Blickkontaktes, Starren auf den Tisch oder aus dem Fenster, fehlende Redebeiträge, bei Fragen keine Reaktion oder Ein-Wort-Antworten. Die qualitative Beobachtungsstudie liefert Anhaltspunkte, unter welchen Bedingungen (Personenzusammensetzung, Opferverhalten, Mediatorverhalten) die unterschiedlichen Formen des Täterverhaltens auftreten und wie die eigentlich angestrebte aktive Partizipation des Täters oder der Täterin am Täter-Opfer-Ausgleich gefördert werden könnte.

Vorteil der Beobachtungsstudie gegenüber einer Befragungsstudie

Es ist davon auszugehen, dass eine Befragungsstudie der Täter/innen zu ihrem Erleben des Täter-Opfer-Ausgleichs zur Klärung der Forschungsfrage weniger geeignet gewesen wäre. Insbesondere die Gruppe der Täter/innen mit geringem Partizipationsgrad hätte vermutlich in einem wissenschaftlichen Interview nur wenig Auskünfte geben können/wollen. Der Forscher dagegen arbeitete auf der Basis seiner eigenen Beobachtungen heraus, dass in den von ihm untersuchten Fällen der Täter-Opfer-Ausgleich oft darauf hinauslief, dass von Seiten der Opfer (oder begleitenden Angehörigen der Opfer) den (in der Regel allein auftretenden) Täter/innen eine Art „Moralpredigt" gehalten wurde, die Täter/innen selbst sich aber gar nicht äußerten und somit die eigentlich angestrebte aktive Verantwortungsübernahme kaum stattfinden konnte. Zudem beschreibt der Forscher seine Beobachtung, dass die anwesenden Mediatoren eine aktive Partizipation der Täter meist nicht förderten, sondern deren Passivität und Schweigen hinnahmen. Eine Befragung der Opfer und/oder der Mediatoren zum Täterverhalten während des Täter-Opfer-Ausgleichs hätte die vom Forscher beobachtete Problematik vermutlich nicht in dieser Deutlichkeit zutage gefördert, weil Opfer und Mediatoren jeweils eine andere Perspektive auf das Geschehen haben und stärker mit sich selbst beschäftigt sind.

Abgrenzung der wissenschaftlichen Beobachtung von der Alltagsbeobachtung

Was diese qualitative Beobachtung von einer Alltagsbeobachtung unterscheidet, ist das zielgerichtete, systematische und regelgeleitete Vorgehen, das im Artikel nachvollziehbar dokumentiert ist, inklusive einer Diskussion von Perspektivität (Theorie des Täter-Opfer-Ausgleichs, Fokus auf Partizipation des Täters), Selektivität (Begründung der Auswahl der beobachteten Täter-Opfer-Ausgleichsfälle sowie Diskussion ihrer Verallgemeinerbarkeit) und Konstruiertheit der Beobachtung (wiederum im Rahmen theoretischer Annahmen zum Täter-Opfer-Ausgleich).

Abgrenzung der qualitativen Beobachtung mit geringem Komplexitätsgrad von der ethnografischen Feldbeobachtung

Von einer ethnografischen Feldbeobachtung bzw. Feldforschung unterscheidet sich diese einfache teilstrukturierte qualitative Beobachtung insofern, als der Forscher nicht tiefgehend und langfristig in die soziale Wirklichkeit der beobachteten Täter/innen eintaucht, sondern gezielt das Partizipationsverhalten während des Täter-Opfer-Ausgleichs herausgreift. Er beobachtet z. B. nicht, ob und wie die Beteiligten sich auf den Täter-Opfer-Ausgleich vorbereiten, ob und mit wem sie nachher darüber sprechen, inwiefern sich ihr Auftreten beim Täter-Opfer-Ausgleich von ihrem Verhalten in anderen Situationen unterscheidet etc.

für den Erfolg des Projektes entscheidend. Gerade weil natürliche Lebensumwelten, auf die sich Feldforschung einlässt, eine schier unerschöpfliche Fülle von Merkmalen, Ereignissen, Akteuren und Abläufen bieten, sind gezielte, aufmerksamkeitsstrukturierende Forschungsfragen wichtig. Das **Prinzip der theoretischen Offenheit**, das eine wichtige wissenschaftstheoretische Grundlage im qualitativen Paradigma empirischer Sozialforschung darstellt, gestattet es jedoch, Forschungsfragen und forschungsleitende theoretische Konzepte im Laufe eines zirkulär strukturierten Forschungsprozesses zu reformulieren, um sich dem Untersuchungsgegenstand schrittweise immer stärker anzunähern (► Abschn. 2.3.2). Da ein völlig theorie- und voraussetzungsloses Beobachten

Historische Entwicklung der ethnografischen Feldforschung

Zum Verständnis der ethnografischen Feldbeobachtung und ihrer Indikation ist ein Blick zurück auf die **historische Entwicklung** hilfreich, die hier mit Fokus auf die **Fremdheit bzw. Exotik des Beobachtungsfeldes** anhand prominenter Studien skizziert wird:

- Die ethnografische Feldbeobachtung als wissenschaftliche Datenerhebungsmethode im Kontext der Feldforschung wurde in der Völkerkunde **(Ethnologie)** entwickelt und zunächst für die Erforschung **„fremder" Völker und Kulturen** eingesetzt: Informationen über außereuropäische Völker erhielten Ethnologen im 19. Jahrhundert zunächst nur durch Gegenstände und Berichte, die Missionare, Reisende und Kolonialbeamte z. B. aus Übersee mitbrachten. Erst zu Beginn des 20. Jahrhunderts begannen Ethnologen verstärkt, selbst vor Ort Beobachtungsdaten zu sammeln. Sie fuhren in ferne Länder, lebten z. T. jahrelang unter den „Wilden", erlernten ihre Sprache und nahmen am Alltag teil, etwa an der Essenszubereitung. Als Vater der ethnografischen Feldforschung gilt der polnische Sozialanthropologe **Bronislaw Malinowski**, der sich von 1914 bis 1918 auf den Trobriand-Inseln im Südpazifik (Papua-Neuguinea) aufhielt und die Kultur der Trobriander erforschte, u. a. auch ihr – nach westlichen Maßstäben der damaligen wie heutigen Zeit – besonders freizügiges Sexualleben im Jugendalter (Malinowski, 1922, 1929). Dass ein völliges „Eintauchen" in eine fremde Kultur nicht so einfach möglich ist und auch wissenschaftlich geschulte Ethnologinnen und Ethnologen Gefahr laufen, das Leben der „Eingeborenen" jeweils durch die „Brille" ihrer eigenen Kultur zu sehen (Beobachterfehler), ist mittlerweile vielfach belegt worden und unterstreicht die Bedeutung der Kontrolle und kritischen Reflexion des Beobachtungsprozesses in der Wissenschaft.
- Nicht fremde Länder und Kulturen, sondern **Minderheiten und Subkulturen im eigenen Land** (insbesondere in der Großstadt) wurden von der **Chicagoer Schule der Soziologie** zu Beginn des 20. Jahrhunderts in den Blick genommen (z. B. Obdachlose, Straßengangs, Prostituierte). Eine klassische Studie dieser Ära stammt von dem amerikanischen Soziologen **William F. Whyte,** der das italienische Viertel von Chicago untersuchte und dazu im Rahmen seiner Feldforschung für mehr als drei Jahre (von 1937 bis 1940) bei einer italienischen Familie im Viertel wohnte, Italienisch lernte, Freizeitvereine besuchte, Mitglied in diversen Cliquen wurde und auf diese Weise die typischerweise an Straßenecken sich aufhaltenden Gangs und ihre sozialen Regeln genau beobachten konnte (Whyte, 1943). Die Chicagoer Schule ist bis heute richtungsweisend im Bereich der Feldforschung, da sie nicht nur zahlreiche ausgesprochen aufwändige Beobachtungsstudien hervorgebracht hat, sondern mit dem symbolischen Interaktionismus auch den einflussreichsten Theorierahmen für qualitative Beobachtungen entwickelt hat.
- Jenseits fremder Kulturen oder Subkulturen sind auch ganz **„normale" Schauplätze des Alltagslebens** Gegenstand ethnografischer Feldforschung geworden, seien es Arbeitersiedlungen, Krankenhäuser, Schulen, Fabriken oder Spielhallen. Nicht mehrere Jahre, aber durchaus mehrere Wochen verbrachten Mitglieder des Forschungsteams um den österreichischen Soziologen **Paul Lazarsfeld** zwischen 1931 und 1932 in der Arbeitersiedlung Marienthal bei Wien, um die Auswirkungen der dort durch Werksschließungen massiv ausgebrochenen Arbeitslosigkeit zu untersuchen. Im Rahmen der Feldforschung sammelten die Forschenden z. B. qualitative Beobachtungsdaten bei Familienbesuchen, in Ärztesprechstunden, bei einem Mädchenturnkurs und einem Schnittzeichenkurs. Die **Marienthal-Studie** gehört zu den international anerkannten Klassikern der empirischen Sozialforschung, da neben teilnehmenden Feldbeobachtungen weitere qualitative und quantitative Datenerhebungsmethoden genutzt und im Sinne eines Mixed-Methods-Designs miteinander kombiniert wurden (Jahoda, Lazarsfeld, & Zeisel, 1933). Eine vergleichbare Studie wurde im Verbundprojekt „Social Capital im Umbruch europäischer Gesellschaften – Communities, Familien, Generationen" unter Leitung des Soziologen Heinz Bude von 2007 bis 2009 in Wittenberge in Brandenburg durchgeführt und löste ein großes Medienecho aus (http://www.ueberlebenimumbruch.de).
- Während das Forschungsteam um Lazarsfeld immerhin noch eine halbstündige Bahnfahrt auf sich nehmen musste, um von Wien in ihr Beobachtungsfeld „Arbeitersiedlung Marienthal" zu gelangen, konzentrieren sich andere Forschende auf Beobachtungsfelder, die buchstäblich vor der Tür ihres Arbeitszimmers liegen und auf den ersten Blick **banal und kaum untersuchungswürdig erscheinen** mögen. So studierte der deutsche Soziologe **Stefan Hirschauer** (1999) in einer viel zitierten Studie das **Fahrstuhlfahren** am Beispiel von Fahrstühlen an seiner damaligen Universität in Bielefeld. Mit dem Begriffsinstrumentarium des symbolischen Interaktionismus rekonstruierte er minutiös, wie soziale Prozesse des geordneten Ein- und Aussteigens, der Positionierung und Kommunikation bzw. aktiven Kommunikationsvermeidung unter den Passagieren im Fahrstuhl ablaufen. Die Daten wurden dabei über sechs Jahre hinweg im Rahmen eines Lehrprojektes von Studierenden erhoben, so dass letztlich teilnehmende Feldbeobachtungen von mehreren Tausend Fahrstuhlfahrten in die Studie eingingen.
- In jüngerer Zeit mehren sich **Feldstudien aus Konflikt- und Krisenregionen** (z. B. im Kontext der Politikwissenschaft oder der Friedensforschung), wobei hier die besonderen Gefahren für die Forschenden sowie für die Forschungsteilnehmenden und daraus resultierende Sicherheitsmaßnahmen verstärkt diskutiert werden (Ferrell & Hamm, 1998; Sriram, King, Mertus, & Martin-Ortega, 2009). Ebenso gewinnen im Zuge der Internationalisierung und Globalisierung teilnehmende **Beobachtungen in interkulturellen Situationen** an Bedeutung (z. B. im Rahmen von Schüleraustausch; Weigand & Hess, 2007). Ein weiterer relativ junger Bereich ist die ethnografische **Online-Feldforschung**, kurz: Online-Ethnografie, bei der die Forschenden über längere Zeit ausgewählte Online-Communitys (etwa in Online-Foren oder Online-Mehrpersonenspielen) teilnehmend beobachten (Hine, 2000, 2005; Kozinets, 2009).

nicht möglich ist, sollte man die **eigene Perspektive** auf den Forschungsgegenstand reflektieren und transparent machen. Zur Strukturierung und Beschreibung des beobachteten Verhaltens lassen sich oft theoretische Konzepte aus dem symbolischen Interaktionismus fruchtbar machen.

Feldzugang und Forscherrolle Schauplätze des Alltagslebens lassen sich nach ihrer **Zugänglichkeit** in offene (z. B. Straße, Bahnhof), halboffene (z. B. Geschäfte, Hochschulen) und geschlossene (z. B. Wohnzimmer, Therapieraum, Zeitungsredaktion) Schauplätze unterteilen. Um das Geschehen an geschlossenen Schauplätzen mittels Feldbeobachtung zu untersuchen, müssen Forschende in das Feld eingeführt werden, d. h., sie brauchen die Erlaubnis zur Anwesenheit und müssen eine für die anderen Feldakteure akzeptable Rolle einnehmen (z. B. können sie sich für eine Feldbeobachtung im Betrieb als Forschende zu erkennen geben: offene Beobachtung oder eine die Forschung verheimlichende Rolle etwa als Praktikant annehmen: verdeckte Beobachtung). Selbst für teilnehmende Beobachtung an offenen Schauplätzen ist meist eine offizielle Einführung in das Feld erforderlich. So treffen sich Straßengangs zwar auf offen zugänglichen Straßen oder Plätzen; dennoch können sich Forschende nicht einfach ungefragt unter sie mischen (wenn sie ein Gangmitglied persönlich kennen, kann dieses sie jedoch mitnehmen und in die Gruppe einführen).

Zugang zu einer zu erforschenden Lebenswelt erhält man bei offenen und halboffenen Schauplätzen durch bloße Anwesenheit, durch Teilnahme und Interesse an Aktivitäten, Ansprechen von Feldakteuren sowie Ausbau von Alltagskontakten (z. B. Gespräche mit Nachbarn, Gastwirten, Verkäufern). Bei geschlossenen Schauplätzen greift man meist auf sog. **Gatekeeper** (Schlüsselpersonen) zurück. Das sind Einzelpersonen, die zum Feld gehören und sich bereit erklären, das Forschungsprojekt zu unterstützen, indem sie die Forschenden mit Informationen versorgen und ihre Integration ins Feld fördern. Am erfolgversprechendsten ist es, sich um die Unterstützung von Personen zu bemühen, die im Feld hohe Autorität genießen; in Institutionen sind die Hierarchieebenen zu beachten.

Aus der Perspektive der Forschenden lassen sich Beobachtungsfelder nicht nur nach Zugänglichkeit, sondern auch nach dem **Grad der Vertrautheit** differenzieren: So können z. B. Feldstudien über Jugendszenen aus der Insider-Perspektive durchgeführt werden (z. B. Forschende gehören selbst der Szene an, die sie untersuchen) oder aus der Perspektive von Outsidern (z. B. Forschende befassen sich speziell für das Forschungsprojekt erstmals mit der betreffenden Szene). **Forschung aus der Insider-Perspektive** erlaubt eher eine aktive Teilnahme am Feldgeschehen, erschließt häufig zusätzliche Kontakte und Datenquellen, fördert ein Verständnis der Sichtweisen der Feldmitglieder (die von feldzugehörigen Forschenden nicht im Zuge des **„Othering“** als „die Anderen“ betrachtet werden) und eröffnet zudem mehr Möglichkeiten der Selbstbeobachtung bis hin zur systematischen Autoethnografie. Gleichzeitig besteht bei großer Identifikation die Gefahr distanzloser Nähe zum Forschungsgegenstand (**Going Native**), welche dann durch verstärkte methodische Reflexion und bewusste Distanzierung (z. B. „sich künstlich dumm stellen“, um das vertraute Geschehen mit Abstand zu betrachten) abgefangen werden muss. Welches Spannungsfeld sich ergibt, wenn Forschende aus der Insider-Perspektive Feldforschung betreiben, inwiefern sie dabei ihre Rolle einerseits im Beobachtungsfeld und andererseits in der Scientific Community behaupten müssen (Labaree, 2002), wird z. B. bei ethnografischen Feldbeobachtungen in Polizeidienststellen deutlich, die von einem als Sozialforscher tätigen ehemaligen Polizisten durchgeführt wurden (Behr, 2002).

Forschung aus der **Outsider-Perspektive** impliziert Distanz und verhindert vollständige Teilnahme am Feldgeschehen (z. B. weil Forschende sich auf die im Feld typischen Verhaltensweisen nicht einlassen können oder wollen, etwa wenn es um Sport- oder Drogenszenen geht). Es ist für eine Feldbeobachtung aber gar nicht immer nötig, als vollwertiges Mitglied des Feldes anerkannt zu werden. Auch als randständiges Feldmitglied, etwa als Gast, Besucher oder Freund eines Informanten, kann man entsprechend „mitlaufen“ und das Feldgeschehen beobachten (▶ Studienbeispiel „Feldzugänge bei der Feldbeobachtungsstudie ‚Tea Room Trade‘“). Hinsichtlich der **aktiven Teilnahme am Feldgeschehen** ist also anstelle einer dichotomen Trennung (aktiv versus passiv) von vielen graduellen Übergängen und Nuancen der Involvierung in das Feldgeschehen auszugehen.

Bei Forschung aus der mehr oder minder distanzierten Outsider-Perspektive ist besonders darauf zu achten, dass aufgrund der Ferne zum Geschehen nicht Informationen und Bedeutungsgehalte übersehen werden, es zu Fehlinterpretationen kommt oder die Forschenden ihre eigenen Vorurteile auf die ihnen fremden Feldmitglieder projizieren.

Ziel der ethnografischen Feldbeobachtung ist es, das Feldgeschehen möglichst unverzerrt zu erfassen. **Beobachterfehler**, die sich z. B. aus zu großer Nähe oder zu großer Distanz zum Gegenstand ergeben, sollen durch eine Reflexion des Beobachtungsprozesses reduziert werden. Auch die Arbeit im Forschungsteam, die es erlaubt, Beobachtungen zu diskutieren und individuelle Sichtweisen kritisch zu hinterfragen, bietet eine Strategie zur Sicherung der Qualität wissenschaftlicher Feldbeobachtung.

Studienbeispiel

Feldzugänge bei der Feldbeobachtungsstudie „Tea Room Trade"

Geschickt fand der Promotionsstudent **Laud Humphreys** (1970, 1972) eine passende Rolle, um Ende der 1960er-Jahre in den USA als verdeckt teilnehmender Beobachter das Phänomen der **Klappensexualität** (Klappe = öffentliche Toilette; im Englischen in Anspielung auf die optische Ähnlichkeit von Tee und Urin: „tea room") von Männern zu untersuchen. Seine Informanten rieten ihm, als **Feldzugang zu den öffentlichen Toiletten** die **Rolle des Aufpassers** zu übernehmen, also den Eingang und die Fenster im Auge zu behalten und die anderen zu warnen, falls Fremde sich nähern oder eine Polizeikontrolle droht, denn homosexuelle Kontakte waren damals in den USA kriminalisiert. Als Beobachtungsorte wählte Humphreys mehrere öffentliche Toiletten in Parks einer Großstadt und stellte im Zuge einer Ereignisstichprobe ein Sample von 100 sexuellen Interaktionen („tea room encounters") zusammen.

Ein Ergebnis der Feldstudie (das später durch andere Studien bestätigt wurde, z. B. Desroches, 1990) bezieht sich auf das verbale Kommunikationsverhalten der Beteiligten, das der Feldbeobachter aus nächster Nähe mitverfolgte:

» There is another aspect of the tearoom encounter that is crucial. I refer to the silence of the interaction. Throughout most homosexual encounters in public restrooms, nothing is spoken. One may spend many hours in these buildings and witness dozens of sexual acts without hearing a word. Of 50 encounters on which I made extensive notes, only in 15 was any word spoken. Two were encounters in which I sought to ease the strain of legitimizing myself as lookout by saying, „You go ahead – I'll watch." Four were whispered remarks between sexual partners, such as „Not so hard!" or „Thanks." One was an exchange of greetings between friends. The other eight verbal exchanges were in full voice and more extensive, but they reflected an attendant circumstance that was exceptional. When a group of us were locked in a restroom and attacked by several youths, we spoke for defense and out of fear. This event ruptured the reserve among us and resulted in a series of conversations among those who shared this adventure for several days afterward. Gradually, this sudden unity subsided, and the encounters drifted back into silence. (Humphreys, 1972, S. 261)

Ein weiterer Befund der vielzitierten „Tea Room Trade"-Studie (Humphreys, 1970, 1972) bezog sich auf die Lebenssituation der Männer, die auf öffentlichen Toiletten Sex mit anderen Männern suchten. Es handelte sich nämlich zu einem großen Teil um Familienväter, was angesichts der damaligen Kriminalisierung von homosexuellen Kontakten in den USA ein besonderes Skandalon war. Der Feldforscher war an Informationen zur Lebens- und Familiensituation der beobachteten Männer gekommen, indem er sich die Kennzeichen der vor den Klappen parkenden Wagen notierte, die zugehörigen Adressen recherchierte und die Betreffenden unter einem Vorwand zu Hause aufsuchte, wo er dann ihre Ehefrauen und Kinder antraf. Diese Vorgehensweise, die stark in die Privatsphäre der beobachteten Personen eingriff und sie der Gefahr eines ungewollten Outings aussetzte, wurde im Nachhinein als gravierende **Verletzung der Forschungsethik** bewertet und führte dazu, dass der Entzug des durch die Studie erlangten Doktortitels von der Universität Washington gefordert, letztlich aber nicht umgesetzt wurde (zur Forschungsethik ▶ Abschn. 4.1). Der gewählte Feldzugang zu den öffentlichen Toiletten als Beobachtungsort war legitim, der **Feldzugang zum häuslichen Umfeld** war es in der hier beschriebenen Weise jedoch nicht, auch wenn Humphreys, der sich später als schwul outete, mit seiner soziologischen Forschung niemandem schaden, sondern ein gesellschaftliches Tabu aufdecken wollte.

Neben Beobachterfehlern stellt allein die **Anwesenheit von Forschenden eine potenzielle Störungsquelle** dar, etwa wenn Feldmitglieder sich beobachtet und bewertet fühlen. So ist zu befürchten, dass Forschende in ihrer Beobachterrolle als „Fremdkörper" das Verhalten der Feldsubjekte beeinflussen und verzerrte Informationen bekommen. Um dieser Verzerrung entgegen zu wirken, wird teilweise verdeckt bzw. getarnt beobachtet: Bei der **verdeckten Feldforschung** geben sich Forschende nicht als solche zu erkennen, sondern spielen den Patienten, mimen den Praktikanten oder lassen sich scheinbar als Mitglied einer Sekte anwerben, die sie untersuchen möchten. Damit wird jedoch auch Einfluss auf das Feldgeschehen genommen. Zudem kollidiert diese Vorgehensweise mit der ethischen Anforderung informierter Einwilligung aller Untersuchungsteilnehmenden in die jeweilige Studie (▶ Abschn. 4.1.1). Verdeckte Feldbeobachtung ist dennoch unter bestimmten Umständen gerechtfertigt. Insbesondere die nicht-teilnehmende Feldbeobachtung an öffentlich zugänglichen Online- oder Offline-Schauplätzen unterliegt keinen Restriktionen, sofern der Ergebnisbericht anonymisiert wird.

Die praktische Erfahrung zeigt, dass in vielen Fällen eine offene (d. h. nicht-verdeckte) Feldbeobachtung (insbesondere durch szenevertraute Insider) hinsichtlich befürchteter Verfälschungen problemloser ist als oft erwartet, weil die Feldakteure schnell zur Tagesordnung übergehen: Bei längeren Feldaufenthalten wird die **Beobachterrolle weitgehend neutralisiert**, weil sich das Feld an die Präsenz der Feldforschenden gewöhnt und sie somit genauso unscheinbar bzw. „unsichtbar" werden wie die Forschenden, die in die Rolle von „normalen" Feldmitgliedern schlüpfen. Aber auch bei kurzen Beobachtungssequenzen lassen sich Feldmitglieder überraschend wenig von ihren eigentlichen Aktivitäten abbringen (z. B. lenkt eine Redaktionsbeobachtung als offene Feldbeobachtung die beobachteten Journalist/innen meist allenfalls einige Minuten ab, bevor sie die mit Notizblock oder Lap-

top am Rand sitzenden Forschenden ignorieren und zu ihrem üblichen Arbeitsalltag zurückkehren).

Ein weiterer Grund für verdeckte Beobachtung kann darin liegen, dass man davon ausgeht, dass im Beobachtungsfeld grundsätzlich wissenschaftliche Forschung unerwünscht ist. Dies mag insbesondere für Beobachtungsfelder am Rande der Legalität gelten.

Beobachten und Agieren im Feld In diesem Beobachtungsprotokoll beschreibt eine studentische Beobachterin erste Eindrücke im Forschungsfeld „illegaler Drogenstrich" in Frankfurt am Main:

> » An diesem Tag waren viele Leute auf dem Platz, da sich dort einige Teilnehmer einer Sportveranstaltung aufhielten. Ansonsten standen wenige Frauen (weniger als von mir erwartet) an den gegenüber parkenden Autos, kleine Grüppchen von „Junkies" saßen, schwer von uns einsehbar, hinter den Autos. Dass dies hier als ein Teil des illegalen Straßenstrichs gilt, war zunächst nicht für mich ersehbar. Auf der Straße herrschte reger Verkehr. Auffällig wurde nach einigen Minuten eine Anzahl von Pkws, die im Schritttempo die Einbahnstraße abfuhren und dabei den Verkehr aufhielten. Es standen immer mal wieder Frauen hinter den parkenden Autos auf, andere liefen umher und verschwanden. Ich wusste zunächst gar nicht, auf was, wen oder welche Stelle ich mich bei meiner Beobachtung konzentrieren sollte. Außerdem war ich teilweise mit mir und meinen Gedanken und Gefühlen beschäftigt. [. . .] Gleich zu Anfang wurde mir bewusst, dass ich mich zu orientieren versuchte und ein Alibi für meinen Aufenthalt in dieser Gegend suchte. Ich fragte mich, was die Leute wohl von uns dachten. Versuchten sie auch uns einzuschätzen, so wie ich sie einer Gruppe zuordnete? Wir schienen nicht in diese Gegend zu passen – langsam dahin schlendernd und uns umschauend. (Langer, 2007, S. 96f.)

Es wird deutlich, dass die Beobachterin einerseits ihre eigene Rolle reflektiert, andererseits Schwierigkeiten hat, überhaupt etwas Forschungsrelevantes zu „sehen" („Dass dies hier als ein Teil des illegalen Straßenstrichs gilt, war zunächst nicht für mich ersehbar"). Sie nähert sich dem Beobachtungsfeld offenbar mit bestimmten **Erwartungen** (z. B. wie Prostituierte und Freier aussehen oder sich verhalten) und da diese Erwartungen nicht erfüllt werden, sieht sie erst einmal „nichts". Erst nach mehrmaligen Feldaufenthalten zeigt sich, dass es gerade typisch für den illegalen Drogenstrich ist, dass Prostituierte und Kunden für Außenstehende kaum zuzuordnen sind und die Kontaktaufnahme ungeordnet wie eine Zufallsbegegnung abläuft. Im Rahmen des von der Erziehungswissenschaftlerin Antje Langer (2007) geleiteten Projekts erwiesen sich **theoretische Konzepte** des symbolischen Interaktionismus (Goffman, 1969) als hilfreich, um zu beschreiben, wie Prostituierte und Freier unter Bedingungen von Illegalisierung und Stigmatisierung sowie der Überlagerung verschiedener „Szenen" (Drogenszene, Straßenstrich) miteinander kommunizieren. Nach einer anfangs breit gestreuten Beobachtung des Feldgeschehens folgte eine Fokussierung anhand der theoretischen Konzepte, womit ein **Übergang von der nicht-strukturierten zur teilstrukturierten Beobachtung** verbunden ist.

Trotz genauer Beobachtung und geeigneter theoretischer Konzepte können Feldforschende (insbesondere in der Rolle von Outsidern) nicht alles, was sie sehen und wahrnehmen, richtig verstehen und deuten. Sie sind auf Erklärungen von „Insidern" angewiesen. Dabei sind Befragung und Beobachtung oftmals parallel einzusetzen, indem man z. B. beim Beobachten im Feld Fragen einstreut, um das Geschehen einordnen zu können. Diese Fragetechnik steht dem Alltagsgespräch näher als dem strukturierten Interview, weshalb statt von einem „ethnografischen Feldinterview" meist vom **„Feldgespräch"** als einer Variante des unstrukturierten Interviews die Rede ist (► Abschn. 10.2.2 „Unstrukturiertes Interview"). Die Befragungspersonen sollten nicht in die Defensive geraten und sich eher animiert fühlen, ausführliche Erläuterungen abzugeben. Dazu ist es empfehlenswert, „Warum-Fragen" zu vermeiden und stattdessen **deskriptive Fragen** zu stellen, die mit „wie", „wann", „wo" oder „was" beginnen (Jorgensen, 1990). Erfolgt die Feldbeobachtung nicht-verdeckt, d. h. die Forschenden sind als solche im Feld sichtbar, lassen sich ergänzend zu den Feldbeobachtungen und Feldgesprächen auch weitere reaktive Datenerhebungsmethoden einsetzen (z. B. psychometrische Tests oder Fragebögen). Günstig ist es, Feldstudien im Team durchzuführen und insbesondere Teilaufgaben zu delegieren; ggf. können auch einzelne Feldmitglieder gebeten werden, bestimmte Informationen zu beschaffen.

Die **Reflexion der eigenen Forscherrolle** sowie der ethisch verantwortungsvolle Umgang mit Beziehungen und Kontakten zu Feldakteuren sind während des gesamten Feldforschungsprozesses relevant. Inwiefern Feldforschende mit dramatischen Situationen und ethischen Dilemmata konfrontiert werden (z. B. Beobachtung von Straftaten im Feld), hängt von dem Forschungsproblem und vom Beobachtungsfeld ab. So berichtet der österreichische Soziologe Roland Girtler (2001) eine Reihe heikler Episoden, die sich in Folge seiner oft unkonventionellen Feldstudien unter „Sandlern", „Ganoven", „Dirnen" und „Zuhältern" ergaben – vom Übernachtungswunsch an Girtlers Haustür bis zum Diebstahl in seinem Forschungsinstitut. Ebenso bleibt das fortwährende Changieren zwi-

schen Eintauchen in das Feld einerseits und analytischer Distanz andererseits ein Thema, das sich in einem Forschungsteam am besten bearbeiten lässt. Eine Strategie, um **Perspektivenvielfalt** sicherzustellen, besteht darin, multiple Rollen im Feld wahrzunehmen. Dies taten z. B. Douglas, Rasmussen und Flanagan (1977), welche die kalifornische Nudistenkultur der 1960er- und 1970er-Jahre untersuchten und sowohl am örtlichen Nacktbaden als auch an Bürgerprotesten gegen das Nacktbaden teilnahmen. Allerdings konnte diese Taktik nur verhältnismäßig kurze Zeit unbemerkt durchgehalten werden.

Wie oft, wie lange und an welchen Orten die Feldbeobachtung stattfindet, ob man bestimmte Zeitspannen abdecken möchte (z. B. eine komplette Woche) oder bestimmte Ereignisse anzielt, ist im Sinne einer **bewussten Stichprobenbildung** („purposive sample"; ▶ Abschn. 9.2) zu beantworten. In der Praxis der Feldforschung erfolgt die Anwesenheit im Feld oft weniger nach einem vorher detailliert ausgearbeiteten Plan, sondern ergibt sich aus den aktuellen Gelegenheiten und den Empfehlungen von Gatekeepern.

10

Dokumentation der Feldtätigkeit Zur Dokumentation von Feldaufenthalten können Foto-, Audio- und Videoaufzeichnungen oder andere Registriermethoden eingesetzt werden, soweit dies nicht Irritationen schafft. Konzentrierte Beobachtungsfähigkeit und ein gutes Gedächtnis sind für Feldforschende dennoch unabdingbar. In regelmäßigen Abständen sind alle wesentlichen Ereignisse und Informationen in einem **Feldtagebuch** („field journal") zu notieren. Solche **Feldnotizen** („field notes") können stichpunktartig im Feld erfolgen (z. B. Notieren von Namen, Schlüsselbegriffen, Abfolgen von Ereignissen) und sollten unmittelbar nach Verlassen des Feldes ausformuliert werden. Ein Raster für ein **teilstandardisiertes Beobachtungsprotokoll** ist ◘ Tab. 10.2 zu entnehmen.

Als Faustregel gilt, dass auf **1 Stunde Beobachtung im Feld ca. 14 Stunden Dokumentationsarbeit** folgen (Berg, 1989, S. 73). Es ist empfehlenswert, in den Feldnotizen die äußeren Umstände (Räumlichkeiten, Gegenstände, anwesende Personen) genau zu beschreiben. Ereignisse, Äußerungen von Feldmitgliedern sowie subjektive Empfindungen und Gedanken der Forschenden sollten bei der Protokollierung nicht vermischt werden. Weitere gesammelte Dokumente ergänzen den Materialbestand. In diesem Prozess sind Datenerhebung und Dateninterpretation verwoben. Neue Fragestellungen und Hypothesen, die sich während der Feldarbeit bilden, werden durch weiteres Sammeln von Informationen untermauert oder widerlegt.

Ausstieg aus dem Feld Je besser man im Feld integriert war, desto problematischer kann sich der Ausstieg gestalten. Persönliche Bindungen zu Feldmitgliedern sind entstanden, man hat sich an das Leben im Feld gewöhnt und weiß, dass außerhalb des Feldes die mühsame Auswertungsarbeit beginnt. Sofern das Beenden der Studie nicht durch äußere Umstände erzwungen wird („Enttarnung" bei verdeckter Beobachtung durch unvorhergesehene Vorkommnisse, Abgabetermin für den Projektbericht, Auslaufen der Projektstelle oder der Finanzierung), ist ein **schrittweiser Rückzug** aus dem Feld empfehlenswert, in dessen Verlauf Feldaufenthalte seltener und kürzer werden und so auf beiden Seiten eine „Entwöhnung" stattfindet. Bei verdeckten Feldstudien benötigt man plausible Erklärungen, um nicht im Nachhinein die Tarnung zu lüften.

Im Umgang mit Feldmitgliedern, beim Kennenlernen und Aufbauen von Freundschaften sowie beim Verabschieden können sich für die Forschenden Unsicherheiten und Irritationen ergeben, etwa wenn sie mit fremden Milieus konfrontiert sind, mit eigenen Vorurteilen zu kämpfen haben oder sich als Helfer zum Eingreifen animiert fühlen. Dass teilnehmende Beobachtung mitunter persönlich stark beanspruchen kann, wie mit diesen Belastungen umzugehen ist und welchen Einfluss sie auf die Datenerhebung haben, wird in der Literatur vielfach angesprochen (Girtler, 2001; Legewie, 1987). Diese Bilanz unterstreicht die Notwendigkeit von Feldforschung im Team sowie von Angeboten der kollegialen Inter- oder Supervision in der Forschungspraxis.

Auswertung und Ergebnisbericht Wenn man das Feld verlassen hat, besitzt man neben persönlichen Eindrücken und Erfahrungen ein umfangreiches Ton-, Bild- und Textmaterial. Einige Schritte der Analyse und Interpretation des Materials wurden schon während der Feldphase durchgeführt und im Feldtagebuch festgehalten. Eine erschöpfende Auswertung erfolgt erst nach Verlassen des Feldes; sie sollte zeitnah durchgeführt werden, damit wenige Informationen vergessen werden. Zudem kann man unmittelbar nach dem Feldausstieg zur Klärung von Fragen noch einmal auf die Informanten im Feld zurückkommen, die zu späteren Zeitpunkten meist schwer erreichbar sind. Da die Ergebnisse einer Feldforschungsstudie meist textförmig vorliegen (Gesprächs- und Beobachtungsprotokolle, Feldnotizen mit eigenen Gedanken und Gefühlen etc.), sind zur Auswertung vor allem die unterschiedlichen Methoden der qualitativen Daten- und insbesondere der qualitativen Dokumentenanalyse indiziert (▶ Abschn. 12.1; ▶ Abschn. 10.6.2).

Tabelle 10.2 Aufbau eines teilstandardisierten Beobachtungsprotokolls bestehend aus offenen Fragen. (Aus Przyborski & Wohlrab-Sahr, 2009, S. 63)

Ort, Zeit	Beobachtungen	Kontextinformationen	Methodische und Rollen-Reflexionen	Theoretische Reflexionen
Wo befinde ich mich zu welchem Zeitpunkt?	Wie sieht das Feld aus?	Durch welche Rahmenbedingungen z. B. finanzieller, familiärer, rechtlicher, politischer Art oder durch welche vor dem Untersuchungszeitraum liegenden Abläufe wird das Feld mitbestimmt?	Wie ist meine Rolle als Forscher im Feld?	Wie lässt sich das bisher Beobachtete in vorläufiger Weise theoretisch fassen?
	Welche genauen Abläufe gibt es?		Haben Beobachtungen im Feld bestimmte methodische Konsequenzen?	Welche Zusammenhänge deuten sich an?
	Wer tut was und wie mit wem?			
	Gibt es Routinen?			
	Gibt es besondere Ereignisse?			
	Welche Konstellationen gibt es?			
	Gibt es hervorgehobene Personen mit höherer Kontakthäufigkeit, besonderen Befugnissen?			
	Gibt es Personen, die kaum/nicht kontaktiert werden?			
	Wie ist die Art des Kontakts?			
	Gibt es Gruppenbildungen und Grenzziehungen?			
	Gibt es Hinweise auf relevante Beziehungen zu Personen/Einrichtungen außerhalb des unmittelbaren Feldes?			

Autoethnografie

Die **Autoethnografie** („autoethnography“) als qualitative Datenerhebungsmethode im Rahmen der Selbstbeobachtung der Forschenden wird häufig eingesetzt, um Ereignisse und Verhaltensweisen zu untersuchen, die mit **starken Emotionen** verbunden sind (z. B. Krankheit, Trennung, Tod von Angehörigen). Metaphorisch kann die persönliche Konfrontation mit einem kritischen Lebensereignis (z. B. einer lebensbedrohlichen Erkrankung) für die autoethnografisch arbeitenden Forschenden das Eintauchen in eine fremde Kultur oder Subkultur (z. B. Leben im Krankenhaus, Gemeinschaft mit anderen Kranken) bedeuten, analog herkömmlicher ethnografischer Feldforschung. Nur dass in diesem Fall das Beobachtungsfeld nicht frei gewählt und auch nicht aus freien Stücken verlassen werden kann. Generell hat die mit jeder Selbstbeobachtung verbundene **Insider-Perspektive** den Vorteil, dass die Feldaufenthalte in der Regel ausgedehnter sind und mehr Daten gesammelt werden können (z. B. jahrebis jahrzehntelange tägliche Erfahrung mit einer eigenen chronischen Erkrankung). Zudem erlaubt die Selbstbeobachtung den reflektierenden Zugang zu allen Gedanken, Gefühlen, Träumen etc., die mit dem Beobachtungsgeschehen einhergehen und die bei einer Fremdbeobachtung nicht direkt zugänglich wären (z. B. autoethnografische Studien zur langjährigen Teilnahme am Extrem-Ausdauersport: Drummond, 2010; zu einem beruflichen Aufstieg in die Managerposition: Kempster & Stewart, 2010; zur Teilnahme an der Swinger-Szene: Wagner, 2009).

Autoethnografie – Bei der Autoethnografie („autoethnography“) handelt es sich um eine ethnografische Feldbeobachtung, bei der die Forscherin oder der Forscher als vollständiger Teilnehmer bzw. Insider („complete participant“, „member researcher“) im untersuchten Beobachtungsfeld agiert und der Fokus der Datenerhebung auf der **Selbstbeobachtung** und **Selbstreflexion** liegt. Per Autoethnografie werden vor allem menschliche Ausnahme- und Extremsituationen (z. B. Trennung oder Tod von Angehörigen, Krankheit und Sterben, außergewöhnliche Hobbys und Tätigkeiten) untersucht. In ihrem wissenschaftstheoretischen Verständnis und ihrer Vorgehensweise unterscheidet sich die **evokative/emotionale Autoethnografie** („evocative/emotional autoethnography“), die mit tradierten Kriterien der Wissenschaftlichkeit bricht und die Gültigkeit ihrer Befunde vor allem von emotionaler Nachvollziehbarkeit abhängig macht, von der **analytischen Autoethnografie** („analytical autoethnography“), die auf theoretische Analyse des Datenmaterials und intersubjektive Nachvollziehbarkeit setzt.

Die autoethnografische Methode wirkt der Gefahr des „Othering“ entgegen, also der Tendenz, dass Forschen-

de bei umstrittenen Themen ihre Untersuchungsobjekte als „die Anderen" konstruieren und implizit sich selbst bzw. der Gruppe der Forschenden eine Sonderstellung unter den Menschen zuweisen. Nicht zuletzt bietet der autoethnografische Forschungsprozess durch die starke persönliche Involvierung den Forschenden Chancen der Selbsterkenntnis. Nicht jede Forscherin und jeder Forscher steht dieser von üblichen Konventionen der Wissenschaft abweichenden Selbstoffenbarung aufgeschlossen gegenüber. Zudem stößt der Ansatz dort an Grenzen, wo Themen untersucht werden sollen, zu denen Forschende mangels eigener Betroffenheit keinen Insiderzugang haben (können).

Da bei der Autoethnografie die persönlichen Erfahrungen der Forschenden im Mittelpunkt stehen, stellt sich erkenntnistheoretisch die Frage, wie auf dieser Basis generalisierbare wissenschaftliche Erkenntnisse gewonnen werden können. Eine Strömung innerhalb der qualitativen Forschung propagiert gemäß einem **postmodernen Wissenschaftsverständnis** die sog. **evokative** oder **emotionale Autoethnografie** („evocative/emotional autoethnography") als grundlegende methodologische Innovation, die wesentlich auf die **amerikanische Soziologin Carolyn Ellis** zurückgeht (Ellis, 1991, 2004; Ellis & Bochner, 2000): Intersubjektive Nachvollziehbarkeit soll nicht durch regelgeleitetes Vorgehen, sondern durch einen Bruch mit wissenschaftlichen Normen erreicht werden. Entscheidend ist, dass das Lesepublikum **mitfühlen** kann, was insbesondere durch einen sehr lebendigen impressionistischen Erzählstil und sogar durch künstlerische Darstellungsformen (z. B. Gedichte) erreicht werden soll. Valide sind die Daten und Schlussfolgerungen, wenn sie beim Lesen glaubwürdig wirken. Die Generalisierbarkeit wird nicht an der Anzahl der untersuchten Personen festgemacht, sondern an der Anzahl der Personen, die die Studie lesen und emotional nachvollziehen können. Evokative autoethnografische Studien sind mittlerweile sowohl in etablierten qualitativen Fachzeitschriften (z. B. *Journal of Contemporary Ethnography; Symbolic Interaction; Qualitative Sociology*) als auch in neueren Journals (z. B. *Qualitative Inquiry; Cultural Studies <=> Critical Methodologies*) zu finden.

Von der evokativen Autoethnografie lässt sich die **analytische Autoethnografie** („analytical autoethnography") abgrenzen, die erkenntnistheoretisch stärker in der Tradition der klassischen ethnografischen Forschung steht (Anderson, 2006). Diese will willkürliche Subjektivität vermeiden und legt Wert darauf, dass a) neben den persönlichen Erfahrungen des Forschers oder der Forscherin auch die Erfahrungen anderer Feldakteure erhoben werden und dass b) der Umgang mit den autoethnografischen Daten eng mit theoretischer Analyse verbunden wird (im Unterschied zu einer primär emotionalisierenden Präsentation des Materials in der evokativen Autoethnografie).

10.1.3 Quantitative Beobachtung

Die **quantitative** bzw. **(voll)strukturierte Beobachtung** zielt auf die präzise Messung einiger weniger vorher genau definierter Merkmale oder Verhaltensweisen im Zuge des Beobachtungsprozesses ab. Der hohe Strukturierungsgrad ermöglicht bei Bedarf die Untersuchung großer Mengen an Beobachtungseinheiten, wobei sowohl mit Zeit- als auch mit Ereignisstichproben operiert wird. Eine quantitative Beobachtung kann in der natürlichen Umgebung als quantitative Feldbeobachtung stattfinden, oft wird sie jedoch im Labor eingesetzt. Sie ist von Nutzen, wenn ausdrücklich keine Selbstauskunftsdaten, sondern Verhaltensreaktionen erfasst werden sollen, etwa im Rahmen experimenteller Studien.

Quantitative Beobachtung – Bei der quantitativen bzw. vollstrukturierten Beobachtung („quantitative observation", „fully structured observation") werden numerische Beobachtungsdaten über das Verhalten anderer Personen (Fremdbeobachtung) erhoben, die einzelne Aspekte des beobachteten Geschehens in ihrer Häufigkeit, Dauer oder Intensität messen. Als Datenerhebungsinstrument fungiert ein standardisierter Beobachtungsplan.

Die quantitative Beobachtung dient der Fremdbeobachtung. Quantitative Selbstbeobachtungen von Forschenden – analog der qualitativen Autoethnografie – haben sich nicht etabliert. Wenn man an quantitativen Selbstbeobachtungen interessiert ist, so werden anstelle der Daten einer einzelnen Forscherin bzw. eines einzelnen Forschers lieber ganze Stichproben von Untersuchungsteilnehmenden zur Selbstbeobachtung angeleitet (z. B. schriftlich mittels standardisierter Tagebuchmethode oder mündlich mittels der Methode „Lautes Denken").

Welche Formen der quantitativen Beobachtung lassen sich unterscheiden? Zur Differenzierung sind die vorgestellten Dimensionen zur Klassifikation von Beobachtungsverfahren heranzuziehen:

- Bei der **quantitativen Beobachtung mit geringem Komplexitätsgrad** werden nur wenige und einfache Merkmale oder Verhaltensweisen herausgegriffen und in ihrer Häufigkeit, Dauer oder Intensität erfasst.
- Bei der **strukturierten Verhaltensbeobachtung** geht es um komplexere Merkmale und Verhaltensweisen, es kommen umfangreichere Beobachtungsinstrumente zum Einsatz.
- Bei der **nonreaktiven Beobachtung** geht es um eine quantitative Beobachtung von materiellen oder virtuellen Verhaltensspuren.

Beobachtungsbogen Handynutzung am Steuer.

Ort: 98693 Ilmenau, Langewiesener Straße.

Zeit: Montag, 4. Oktober 2010; 10.00 – 11.00 Uhr.

Beobachterin: L. Sanders

Fahrzeug	Handynutzung		Fahrzeugtyp			Geschlecht	
Nr.	nein	ja	PKW	LKW	Krad	weibl.	männl.
1	x		x			x	
2	x		x			x	
3		x	x				x
4	x				x		x
5	x			x			x
6		x	x			x	
7	x		x				x
8	x		x				x
9	x		x			x	
10		x	x			x	
...							

Abb. 10.2 Ausgefüllter vollstandardisierter Beobachtungsbogen (Ausschnitt) zur Handynutzung am Steuer für eine quantitative Beobachtung mit geringem Komplexitätsgrad (Ereignisstichprobe, fiktive Daten)

Quantitative Beobachtung mit geringem Komplexitätsgrad

Die **Verkehrsforschung** ist eines der Gebiete, in denen quantitative Feldbeobachtungen mit geringem Komplexitätsgrad, d. h. Beobachtungen, bei denen nur wenige einfache Merkmale strukturiert beobachtet werden, oft zu finden sind. So befassen sich eine Reihe von Studien mit Verbreitung und Folgen der Handynutzung während des Autofahrens. Hierbei wird z. B. der Frage nachgegangen, ob in Staaten mit Handyverbot tatsächlich am Steuer weniger zum Mobiltelefon gegriffen wird als in Staaten ohne Handyverbot. Oder es wird beobachtet, welches Geschlecht und Alter Handytelefonierende im Auto haben (z. B. Townsend, 2006). Auch Zusammenhänge zwischen Automarke und Handynutzung oder Unterschiede zwischen Dorf, Kleinstadt und Großstadt hinsichtlich des Handygebrauchs im Auto lassen sich mittels einfacher quantitativer Beobachtung erfassen (z. B. Rosenbloom, et al., 2009).

Quantitative Beobachtung mit geringem Komplexitätsgrad – Bei der quantitativen Beobachtung mit geringem Komplexitätsgrad („quantitative observation with a lesser/lower degree of complexity“) werden einige **wenige und einfache Aspekte** des beobachteten Geschehens in ihrer Häufigkeit, Dauer und/oder Intensität gemessen, so dass numerische Beobachtungsdaten resultieren. Quantitative Beobachtungen mit geringem Komplexitätsgrad finden sowohl im Feld als auch im Labor statt. Sie basieren auf einem einfachen standardisierten Beobachtungsinstrument, das oft selbst konstruiert wird.

Da es sich bei Straßen um rund um die Uhr öffentlich zugängliche Orte handelt, können für die Beobachtungsstudie die geeigneten **Beobachtungsorte** (z. B. Kreuzungen in unterschiedlichen Stadtteilen oder Städten) und **Beobachtungszeiten** (z. B. Wochentage und Wochenende, Tag und Nacht, geringes und hohes Verkehrsaufkommen) prinzipiell frei gewählt werden. Die Auswahl der Orte und Zeiten für die Beobachtung entspricht der **Stichprobenziehung** (▶ Kap. 9) und sollte von inhaltlichen und methodischen Kriterien geleitet sein, um die Aussagekraft der Studie zu optimieren.

Die Datenerhebung erfolgt bei der Verkehrsbeobachtung, indem am Straßenrand postierte Forschende ihre Beobachtungen der einzelnen Fahrzeuge (Beobachtungseinheiten) in einem dafür vorbereiteten **Beobachtungsbogen** („observation sheet“) notieren, der als standardisiertes Datenerhebungsinstrument fungiert (Abb. 10.2). Der Beobachtungsbogen operationalisiert die interessierenden Variablen und sollte einem Pretest unterzogen werden. Ebenso sollten die Beobachtenden vor der Haupterhebung

in einem Probedurchlauf geschult werden, um Beobachtungsfehler zu minimieren. Da während der Beobachtungszeit jedes passierende Fahrzeug einzeln beobachtet wird, handelt es sich um eine **Ereignisstichprobe**.

Vor der Konstruktion des standardisierten Beobachtungsbogens müssen zunächst die relevanten Kategorien bzw. Variablen und ihre Ausprägungen identifiziert werden (z. B. 1) Handynutzung am Steuer: nein/ja; 2) Fahrzeugtyp: Pkw/Lkw/Kraftrad; 3) Geschlecht der Person am Steuer: weiblich/männlich). Der Beobachtungsbogen sollte logisch und grafisch so gestaltet sein, dass Beobachtende möglichst schnell, unkompliziert und eindeutig ihre Beobachtungen in Form von Strichen oder Kürzeln eintragen können.

Der standardisierte Beobachtungsbogen operationalisiert die manifesten nominalskalierten bzw. dichotomen Variablen Handynutzung, Fahrzeugtyp und Geschlecht. Inwiefern genügt dieses standardisierte Beobachtungsinstrument den wissenschaftlichen Gütekriterien Objektivität, Reliabilität und Validität (▶ Abschn. 10.4.1)?

1. Die **Objektivität** (Beobachterunabhängigkeit, „objectivity“) ist durch die klare Definition der Kategorien samt ihrer Ausprägungen, d. h. die **Standardisierung des Instruments**, gegeben, das – spätestens nach einem Beobachtertraining – alle Beobachtenden in gleicher Weise anwenden sollten.
2. Die **Reliabilität** (Messgenauigkeit, „reliability“) kann empirisch überprüft werden anhand der Beobachterübereinstimmung. Hierzu müssen mindestens 2 Beobachter/innen eine Mindestzahl an Fällen parallel (d. h. hier: zur selben Zeit am selben Ort dieselben Fahrzeuge) beobachten und protokollieren. Für jede einzelne Variable kann dann die **Beobachterübereinstimmung** („inter-observer agreement“, „inter-observer reliability“, „inter-rater agreement“, „concordance“) berechnet werden, die als Schätzung der Messgenauigkeit gilt (▶ Beobachterübereinstimmung).
3. Die **Validität** (Gültigkeit, „validity“) adressiert als Gütekriterium die Frage, ob die Kategorien im Beobachtungsbogen tatsächlich die interessierenden Merkmale inhaltlich angemessen erfassen. Dies muss durch die **Konzeptspezifikation** dargelegt werden (Verweise auf anerkannte Theorien, Definitionen und Operationalisierungen in der Fachliteratur), auch eine Überprüfung der Inhaltsvalidität der Kategorien durch **Urteil von Fachexperten** („expert review“) ist möglich. Im vorliegenden Fall sind – wie bei Beobachtungen mit geringem Komplexitätsgrad üblich – die interessierenden Merkmale allerdings so eindeutig, dass die Validität der verwendeten Kategorien angenommen werden kann (zur Problematik der Operationalisierung von Geschlecht ▶ Abschn. 8.5.1).

Quantitative Beobachtungen mit geringem Komplexitätsgrad, die sich auf das Verhalten im Straßenverkehr beziehen, können u. a. dazu dienen, sozialpsychologische und soziologische Hypothesen experimentell zu überprüfen, etwa wenn es darum geht, menschliches **Hilfeverhalten im Feld** zu untersuchen (z. B. lässt sich beobachten, unter welchen von den Forschenden geschaffenen Bedingungen Anhalter/innen mitgenommen werden oder bei Autopannen geholfen wird). Auch hier enthält der Beobachtungsbogen typischerweise nur wenige manifeste Variablen wie z. B. Automarke oder Geschlecht. Quantitative Beobachtungen mit geringem Komplexitätsgrad werden nicht nur im Feld, sondern auch im **Labor** durchgeführt, etwa um **aggressive Verhaltensweisen** nach Nutzung gewalthaltiger Medienangebote (Bandura, Ross, & Ross, 1961; Ferguson et al., 2008) oder Flirtverhalten zwischen Unbekannten zu erfassen.

Da es beim **Flirtverhalten** auf viele Details nonverbaler Kommunikation ankommt, erfolgt eine Beobachtung in der Regel nicht live, sondern anhand von Videomaterial, das von den Forschenden im Nachgang mehrfach genau betrachtet werden kann. Da systematische Videoaufzeichnungen des Flirtverhaltens im Feld aus ethischen Gründen nicht möglich sind, finden entsprechende Studien im Labor statt. Die Testpersonen wissen nicht, dass es um ihr Flirtverhalten geht, sondern interagieren in vermeintlichen Wartepausen mit anderen Testpersonen oder Konfidenten (Mitwissern) des Versuchsleiters. Diese Interaktionen sind Gegenstand der folgenden Untersuchung (Penke & Asendorpf, 2008): Es wurden siebenminütige Interaktionssequenzen zwischen einer wartenden Versuchsperson und einem Konfidenten über zwei Videokameras aufgezeichnet (Ganzkörperaufnahme von der Seite und Nahaufnahme des Gesichts). Anschließend beurteilen geschulte Beobachter global auf einer Ratingskala, wie intensiv gerade geflirtet wird („Wie stark flirtet die Person mit dem Mitwisser?“ 1: nicht sehr stark bis 7: sehr stark).

Bei der Beobachtung von Flirtverhalten geht es nicht um separate Einzelereignisse, sondern um einen Verhaltensstrom, so dass eine **Zeitstichprobe** indiziert ist: Während der Sichtung der Beobachtungsvideos erklingt alle 30 Sekunden ein Signalton, zu dem die Beobachtenden die wahrgenommene Intensität des Flirtverhaltens auf der Ratingskala notieren (◘ Abb. 10.3). Die Intensität des Flirtverhaltens in einer 5-minütigen Episode ergibt sich als Summe der Ratings (z. B. bei der ersten Versuchsperson: $1 + 1 + 2 + 2 + 1 + 3 + 2 + 1 + 1 + 1 = 15$) relativiert an der Anzahl der Messzeitpunkte (hier: 10): Für die erste Versuchsperson bedeutet dies: $15/10 = 1.5$. Der durchschnittliche Skalenwert von 1.5 besagt, dass die betreffende Versuchsperson Nr. 1 in der beobachteten Sequenz kaum Flirtverhalten zeigte.

VP	30 s	30 s	30 s	30 s	30 s	30 s	30 s	30 s	30 s	30 s
1	1	1	2	2	1	3	2	1	1	1
2	4	5	5	6	4	5	4	3	4	3
3	2	2	1	2	1	2	1	1	1	1
4	3	1	2	3	3	3	2	3	2	3
5	1	2	3	3	1	3	2	2	1	1
6	2	3	4	4	5	3	4	2	2	4
7	1	1	1	2	1	2	2	1	1	1
8	1	2	3	3	4	3	2	1	1	1
9	3	3	4	5	4	6	4	4	3	2
10	1	1	2	2	3	3	4	3	1	1

Abb. 10.3 Ausgefüllter Beobachtungsbogen (Ausschnitt) zum Flirtverhalten für eine quantitative Beobachtung mit geringem Komplexitätsgrad (Zeitstichprobe); Ratings der Intensität des Flirtverhaltens (1: nicht sehr stark bis 7: sehr stark). Videoaufzeichnung der Interaktionssequenz zwischen Versuchsperson und Mitwisser; 5 Minuten mit 10 Beobachtungszeitpunkten (alle 30 Sekunden). Fiktive Daten. *Vp* Versuchsperson

Will man die Messgenauigkeit (Reliabilität) dieser intervallskalierten Beobachtungsdaten zum Flirtverhalten überprüfen, so muss mindestens ein weiterer geschulter Beobachter dasselbe Videomaterial kodieren. Die Beobachterübereinstimmung wird bei kardinalskalierten Daten mithilfe der **Intra-Class-Correlation ICC** berechnet (► Beobachterübereinstimmung). Objektivität wird aufgrund der Standardisierung des Vorgehens und Validität aufgrund der theoretisch fundierten Konzeptspezifikation (► Abschn. 8.2) und inhaltlichen Eindeutigkeit vorausgesetzt und nicht empirisch überprüft.

Um zusammenfassend zu beschreiben, wie intensiv eine bestimmte Person im Beobachtungszeitraum flirtet, können die Ratings a) der einzelnen Erfassungszeitpunkte, b) der beiden Kameraaufzeichnungen sowie c) der verschiedenen Beobachter zu einem Gesamtwert aggregiert werden. Auf diese Weise werden **Beobachtungsfehler „herausgemittelt“.** Sinnvoll ist diese Aggregierung nur, wenn die Daten auf Basis der beiden Kameras und der verschiedenen Beobachtenden gemäß Intra-Klassen-Korrelation gut übereinstimmen.

Will man das Flirtverhalten noch differenzierter beobachten, so kann man die geschulten Beobachter/innen anstatt der Globalbewertung auf einer Ratingskala auch einzelne Aspekte des nonverbalen Kommunikationsverhaltens in der zeitlichen Dauer (z. B. Lächeln, Lachen) oder in der Auftretenshäufigkeit (z. B. Haare zurückwerfen, Berührung des Gegenübers) in Beobachtungsbögen eintragen lassen und diese Werte wiederum **zu Indexwerten zusammenfassen** (Penke & Asendorpf, 2008, S. 1125). An dieser Stelle würde die einfach gehaltene quantitative Beobachtung in eine komplexere strukturierte Verhaltensbeobachtung übergehen, da ein umfassenderes Beobachtungsschema erstellt werden muss.

Strukturierte Verhaltensbeobachtung

Das **Beobachtungsinstrument** der strukturierten Verhaltensbeobachtung wird synonym als Beobachtungsplan („observation plan“), Beobachtungssystem („observation system“; „observational coding system“), Beobachtungsschema („observation scheme“) oder Beobachtungsinventar („observation inventory“) bezeichnet. Aufgrund der höheren Komplexität des Beobachtungsinstrumentes ist folglich auch das Beobachtertraining sehr viel wichtiger und fällt umfangreicher aus als bei der quantitativen Beobachtung mit geringem Komplexitätsgrad. Ebenso muss bei der Entwicklung des Datenerhebungsinstruments mehr Aufwand betrieben werden, um Validität und Reliabilität sicherzustellen (► Beobachterübereinstimmung; Abb. 10.4).

Strukturierte Verhaltensbeobachtung – Bei der strukturierten Verhaltensbeobachtung („structured behavioural observation“) werden **zahlreiche und komplexe Aspekte** des beobachteten Geschehens in ihrer Häufigkeit, Dauer und/oder Intensität gemessen, so dass numerische Beobachtungsdaten entstehen. Strukturierte Verhaltensbeobachtungen finden sowohl im Feld als auch im Labor statt. Sie basieren auf mehr oder minder umfassenden standardisierten Beobachtungsinstrumenten („observation plan/system/scheme/inventory“), deren Gütekriterien (v. a. Validität und Reliabilität) geprüft und mindestens zufriedenstellend ausgeprägt sind. Oft werden etablierte Beobachtungssysteme eingesetzt.

Die strukturierte Verhaltensbeobachtung spielt in der klinischen Forschung (Patientenverhalten) sowie in der pädagogischen Forschung (kindliche Entwicklung, Schüler-

Beobachterübereinstimmung

Bei der strukturierten Verhaltensbeobachtung kommt ein standardisiertes Beobachtungssystem zum Einsatz, mit dessen Hilfe die Beobachtenden das Beobachtete auf unterschiedlichen Kategorien in vorgegebenem Format (z. B. trifft zu/trifft nicht zu; Intensität auf einer Ratingskala etc.) einordnen. Ein standardisiertes Beobachtungssystem gilt als messgenaues Datenerhebungsinstrument, wenn die **Messgenauigkeit jeder einzelnen Kategorie** mit einem angemessen hohen **Beobachterübereinstimmungs- bzw. Reliabilitätskoeffizienten** nachgewiesen wurde. Wenn die Beobachterübereinstimmung („observer agreement", „inter-observer reliability") bei allen Kategorien hoch ist, wird dies als Hinweis darauf gedeutet, dass das Beobachtungssystem insgesamt problemlos auf die zu beobachtenden Fälle anwendbar ist und zu messgenauen Daten führt. Geringe Beobachterübereinstimmung deutet darauf hin, dass das Beobachtungssystem unklar bzw. messungenau ist und hinsichtlich einzelner Kategorien überarbeitet werden muss und/oder dass mindestens ein Beobachter verzerrte Beurteilungen abgibt. Die Gründe für die Messungenauigkeit sind zu klären und zu beseitigen (z. B. Überarbeitung des Beobachtungssystems, erneute Schulung der Beobachtenden etc.). Unreliable Kategorien, auf denen geschulte Beobachtende bei der Beobachtung derselben Fälle völlig unterschiedliche Urteile abgeben, dürfen bei der späteren Datenauswertung nicht berücksichtigt werden. Die Beobachterüberstimmungs- bzw. Reliabilitätskoeffizienten der einzelnen Kategorien des Beobachtungssystems werden üblicherweise in einem **Pretest** geprüft und im Untersuchungsbericht dokumentiert (sowohl einzeln für jede Kategorie sowie zusammenfassend z. B. als durchschnittliche Reliabilität des gesamten Instruments). Beim Pretest wird eine Stichprobe von Fällen aus der interessierenden Population **durch mindestens zwei geschulte Beobachter/innen parallel erfasst** (vergleichbar ist die Vorgehensweise bei der Entwicklung und Prüfung eines standardisierten Kategoriensystems für eine quantitative Inhaltsanalyse; ► Abschn. 10.6.3). Werden in der Hauptuntersuchung mehrere Beobachtende bzw. Kodierende parallel eingesetzt, können zusätzlich auch für die Daten der Hauptuntersuchung Reliabilitätskoeffizienten bestimmt werden (oftmals wird man ein im Pretest geprüftes Instrument in der Haupterhebung aus forschungsökonomischen Gründen dann aber jeweils nur von einem einzigen Beobachter/Kodierer pro Fall anwenden lassen). Es ist darauf zu achten, dass bei der Reliabilitätsprüfung ein Mindeststichprobenumfang nicht unterschritten wird. Idealerweise nutzt man zur Reliabilitätsprüfung **optimale Stichprobenumfänge**, die ausreichende Teststärke gewährleisten, um anhand der Stichprobendaten die „wahre" Reliabilität präzise genug schätzen zu können (allgemein zum Konzept des „optimalen Stichprobenumfangs" ► Abschn. 14.3; zur Bestimmung von optimalen Stichprobenumfängen speziell für unterschiedliche Reliabilitätskoeffizienten s. z. B. Gwet, 2012). Reliabilitätsangaben, die auf zu kleinen Stichproben basieren, sind nicht aussagekräftig. Zur Interpretation von Reliabilitätsschätzungen liegen **Beurteilungsstandards („benchmarks") für die unterschiedlichen Übereinstimmungskoeffizienten** vor (z. B. ein Cohens-Kappa-Koeffizient um $\kappa = .50$ wird als mittelmäßige Messgenauigkeit eingestuft), die jedoch nur als sehr grobe Richtlinien gelten dürfen (Gwet, 2012, Kap. 6). Bei der Interpretation der Höhe von Reliabilitätskoeffizienten müssen immer auch die Beschaffenheit der Kategorien (einfache manifeste Merkmale sind messgenauer zu erfassen als komplexere Konstrukte) sowie die Anzahl der Merkmalsausprägungen berücksichtigt werden (so liegt bei einem zweifach gestuften Merkmal die zufällige Beobachterübereinstimmung bereits bei 50 %, so dass für substanzielle Beobachterübereinstimmung deutlich höhere Werte anzusetzen sind).

Maße der Beobachterübereinstimmung (Reliabilität) existieren für nominal-, ordinal- und kardinalskalierte Kategorien in einem standardisierten Beobachtungssystem (■ Tab. 10.34).

Skalenniveau der Kategorie im Beobachtungssystem	Beobachterübereinstimmungskoeffizient (Auswahl)
Nominalskalierte Kategorien	– Prozentuale Übereinstimmung („Holsti-Formel") – Cohens Kappa (Cohens κ)
Ordinalskalierte Kategorien	– Cohens Weighted Kappa (Cohens Weighted κ) – Spearmans Rho (Spearmans ρ) – Kendalls Tau (Kendalls τ)
Kardinalskalierte (intervall- oder verhältnisskalierte) Kategorien	– Bivariate Korrelation *r* – Intra-Class-Correlation (ICC)

Beobachterübereinstimmung bei nominalskalierten Kategorien

Die einfachste Form der Berechnung der Beobachterübereinstimmung bei nominalskalierten Kategorien ist die **prozentuale Übereinstimmung**. Hier werden alle zwischen zwei Beobachtenden übereinstimmenden Kategorisierungen an der Gesamtzahl der Beobachtungen relativiert. Die prozentuale Übereinstimmung (Wertebereich 0 %–100 %) sollte bei einfachen manifesten Kategorien nahe 100 % liegen, um von hoher Beobachterübereinstimmung bzw. Realiabilität auszugehen. Bei mehr als zwei Beobachtenden sind die prozentualen Übereinstimmungen paarweise zu berechnen und arithmetisch zu mitteln (durchschnittliche prozentuale Übereinstimmung).

Die prozentuale Übereinstimmung überschätzt die tatsächliche Beobachterübereinstimmung, da auch Zufallsübereinstimmungen enthalten sind. Diese werden bei der Berechnung durch **Cohens Kappa** explizit berücksichtigt (Cohen, 1960). Cohens Kappa hat einen Wertebereich von −1 bis maximal +1, wobei in der Regel geringere mögliche Maximalwerte in Abhängigkeit von der Höhe der Zufallsübereinstimmungen vorliegen. Je höher der Kappa-Wert, desto besser ist die Beobachterübereinstimmung. Werte über .75 gelten nach konventionellen Standards als sehr gut, Werte zwischen .60 und .75 werden als gut eingestuft und Werte zwischen .40 und .60 als mittelmäßige bzw. gerade noch ausreichende Messgenauigkeit eingeordnet. Angesichts variabler Maximalwerte des Kappa-Koeffizienten und seiner Beeinflussung durch die Anzahl der Kategorien sowie der Kategorienbesetzungen ist eine inhaltlich differenzierte Interpretation notwendig und eine schematische Einteilung gemäß konventioneller Standards unzureichend (Sim & Wright, 2005; Gwet, 2012, Kap. 6). Zu beachten ist, dass aus testtheoretischer Sicht Reliabilitäten nur positive Werte haben können (► Abschn. 10.4.4); ein rechnerisch negativer Kappa-Koeffizient zeigt mangelnde Beobachterüberstimmung an.

Bei mehr als zwei Beobachtenden sind die Übereinstimmungen paarweise zu berechnen und anschließend ist aus den einzelnen Cohens Kappa-Werten der Median zu bilden (alternativ kann bei mehr als zwei Beobachtenden die Übereinstimmung auch über den **Kappa-Koeffizienten von Fleiss** berechnet werden). Cohens Kappa befindet sich in der Statistiksoftware SPSS im Menü

Analysieren → Deskriptive Statistiken → Kreuztabellen. Hier muss das Aktionsfeld „Statistiken“ ausgewählt und ein Häkchen beim Feld „Kappa“ gesetzt werden (für ein Berechnungsbeispiel der Kodierer-Übereinstimmung mittels Cohens Kappa ▶ Abschn. 10.6.3 „Quantitative Inhaltsanalyse“). Hinweise zum optimalen Stichprobenumfang für die Reliabilitätsprüfung mittels Cohens Kappa liefern z. B. Cantor (1996), Sim und Wright (2005) sowie Gwet (2012). Für Stichprobenanforderungen zum Vergleich von Kappa-Koeffizienten s. Donner (1998).

Beobachterübereinstimmung bei ordinalskalierten Kategorien

Handelt es sich um ordinalskalierte Beobachtungsdaten, so kann die Beobachterübereinstimmung mit **Cohens Weighted Kappa** (Cohen, 1968) berechnet werden (maximaler Wertebereich −1 bis +1). Cohens Weighted Kappa ist standardmäßig nicht in SPSS integriert. Alternativ können auch **Spearmans Rho** oder – bei kleinen Stichproben ($N < 30$) – **Kendalls Tau** berechnet werden (z. B. Bortz & Lienert, 2003). Diese beiden Koeffizienten sind im Gegensatz zu Cohens Weighted Kappa nicht zufallskorrigiert. Sie stellen Maße für die Stärke des ordinalen Zusammenhangs dar. Spearmans Rho zählt dabei zu den gebräuchlichsten Maßen bei der Berechnung von Korrelationen zwischen ordinalskalierten Urteilen (Wirtz & Caspar, 2002; Gwet, 2012). In SPSS sind die genannten ordinalen Übereinstimmungs- bzw. Zusammenhangsmaße unter dem Menü Analysieren → Korrelation → Bivariat zu finden. Hier kann jeweils ein Häkchen bei den Feldern „Kendall-Tau-b“ und „Spearman“ gesetzt werden.

Beobachterübereinstimmung bei kardinalskalierten Kategorien

Im einfachsten Fall kann die **bivariate Korrelation** r mit dem Wertebereich −1 bis +1 berechnet werden. Sie überschätzt jedoch teilweise die Beobachterübereinstimmung. Die **Intra-Class-Correlation** (ICC) wird als Reliabilitätsmaß bei Variablen mit Intervall- oder Verhältnisskalenniveau (bzw. zusammenfassend: Kardinalskalenniveau) deswegen bevorzugt. Sie stellt ein Maß für die Stärke des Zusammenhangs der Urteile zweier Beobachter dar, die dieselben Ereignisse beobachtet und im Beobachtungsbogen beurteilt haben. Die ICC weist einen Wertebereich von −1 bis +1 auf. Ein Wert von 0 bedeutet, dass kein Zusammenhang zwischen den Urteilen der Beobachtenden besteht. Ebenso muss ein negativer Wert im Rahmen einer Beobachterübereinstimmung als Hinweis auf fehlende Reliabilität verstanden werden. Je stärker sich der Intra-Klassen-Korrelationskoeffizient dem Wert 1 nähert, desto enger ist der Zusammenhang und desto höher ist auch die Reliabilität der Urteile. ICC-Werte über .70 gelten meist als hohe Reliabilität. Es gibt verschiedene Möglichkeiten, die ICC zu berechnen und zu interpretieren (Wirtz & Caspar, 2002; Gwet, 2012). Die Intra-Klassen-Korrelation befindet sich in SPSS unter dem Menü Analysieren → Skalierung → Reliabilitätsanalyse. Nach Öffnen des Untermenüs „Statistiken“ kann im Feld „Korrelationskoeffizient in Klassen“ ein Häkchen gesetzt werden. Es kann nun entweder die **justierte Intra-Klassen-Korrelation** berechnet werden (Typ: „Konsistenz“), die individuelle Tendenzen von Beobachtern zulässt (z. B. wenn ein Beobachter grundsätzlich immer etwas strenger als der andere urteilt, aber das Urteilsmuster beider Beobachter übereinstimmt, liegt Konsistenz vor), oder die **unjustierte Intra-Klassen-Korrelation** (Typ: „absolute Übereinstimmung“), die jegliche Urteilstendenz der Beobachtenden als Fehler wertet. In der Praxis wird oft die justierte Intra-Klassen-Korrelation als Reliabilitätsmaß für kardinalskalierte Kategorien in standardisierten Beobachtungssystemen verwendet (für ein Berechnungsbeispiel der Kodierer-Übereinstimmung mittels Intra-Class Correlation ▶ Abschn. 10.6.3 „Quantitative Inhaltsanalyse“). Hinweise zum optimalen Stichprobenumfang für die Reliabilitätsprüfung mittels ICC liefern z. B. Shoukri, Asyali und Donner (2004).

Beobachterübereinstimmung und Validität

Zuweilen wird angemerkt, bei der Prüfung standardisierter Datenerhebungsinstrumente wie eben bei den hier betrachteten Beobachtungssystemen (aber auch bei standardisierten Fragebögen ▶ Abschn. 10.3.3 oder standardisierten inhaltsanalytischen Kategoriensystemen ▶ Abschn. 10.6.3) stünde in der quantitativen Forschung die Reliabilität zu sehr im Vordergrund – auf Kosten der Validität. Tatsächlich ist die Validität das gewichtigere Gütemerkmal. Eine empirische Sicherung der Beobachterüberstimmung im Sinne einer Reliabilitätsprüfung des Erhebungsinstrumentes ist also kein Ersatz oder keine Alternative, sondern eine notwendige Ergänzung zu der im quantitativen Forschungsprozess **vorgelagerten Klärung der Validität des Instrumentes**. Dass die Kategorien im Beobachtungssystem tatsächlich das messen, was sie messen sollen, muss im Zuge der Operationalisierung durch eine gründliche **Konzeptspezifikation** und eine sorgfältige theoretische Ableitung, Definition und Begründung der Kategorien sichergestellt werden (▶ Abschn. 8.2). Insbesondere bei der Entwicklung von komplexen Datenerhebungsinstrumenten ist ergänzend zu der theoretischen Vorarbeit und Orientierung an etablierten und geprüften Instrumenten oft auch eine Prüfung der **Inhaltsvalidität** des jeweiligen Erhebungsinstrumentes durch **Beurteilung von Fachexperten** („expert review“) empfehlenswert.

verhalten, Lehrerverhalten) eine wichtige Rolle. Sie dient auf Gruppenebene der Beantwortung von Forschungsfragen bzw. Überprüfung von Hypothesen. Sie kann aber auch auf Individuumsebene in der **Diagnostik** eingesetzt werden, etwa um gezielt verhaltensauffällige Kinder zu identifizieren, die möglicherweise unter einer Aufmerksamkeitsstörung leiden. Insbesondere der diagnostische Einsatz von Beobachtungsinstrumenten **(„behavior assessment“)** verlangt aufgrund der möglichen Folgen für die Betroffenen, dass Gütekriterien wie Validität und Reliabilität genau geprüft werden bzw. dass nur entsprechend geprüfte und etablierte Instrumente zum Einsatz kommen und keine Ad-hoc-Instrumente.

Aufgrund des hohen Aufwandes, der mit der Entwicklung psychometrischer Beobachtungsinstrumente verbunden ist, werden viele von ihnen – ebenso wie psychometrische Tests – nur kostenpflichtig abgegeben (z. B. Analyse psychischer Anforderungen und Belastungen in der Bü-



Abb. 10.4 Spontane Beobachterübereinstimmung im Alltag. (Aus Goscinny & Sempé 1975. Mit freundlicher Genehmigung von Diogenes)

roarbeit – Das RHIA/VERA-Büro-Verfahren von Leitner, Lüders, Greiner, Niedermeier, & Volpert, 1993; Autism Diagnostic Observation Schedule ADOS von Rühl, Bölte, Feineis-Matthews, & Poustka, 2004). Andere psychometrische Beobachtungsinstrumente sind kostenfrei der Literatur zu entnehmen (z. B. SYMLOG – System for the Multiple Level Observation of Groups von Bales, 2001; Bales & Cohen, 1979; Kategoriensystem für Partnerschaftliche Interaktion KPI von Hahlweg, 2004; The Strange Situation Protocol von Main & Hesse, 1990).

Als Beispiel für ein psychometrisches Beobachtungssystem sei das **REDSOCS** („revised edition of the school observation coding system“ von Bagner, Boggs, & Eyberg, 2010) vorgestellt. Dieses Instrument dient der Verhaltensbeobachtung von Vorschul- und Grundschulkindern. Es soll im Schulunterricht eingesetzt werden und konzentriert sich auf störendes Schülerverhalten, das theoriebasiert in drei Dimensionen erfasst wird:

1. **Unangemessenes Verhalten** („inappropriate behavior“): Hierzu zählen Verhaltensweisen eines Kindes, die im Kontext einer Unterrichtsstunde als unangemessen bezeichnet werden. Dies sind beispielsweise Schreien, aggressives Verhalten wie Treten und Schlagen, andere Kinder stören, indem z. B. wiederholt gegen deren Stuhl getreten wird, oder Abschreiben.
2. **Unkooperatives Verhalten** („noncompliant behavior“): Hier werden Verhaltensweisen zusammengefasst, die in Reaktion auf eine verbale Anweisung einer Lehrkraft gezeigt werden. Kooperatives Verhalten wäre gegeben, wenn ein Kind einer Anweisung innerhalb von fünf Sekunden folgt. Als unkooperatives Verhalten wird kodiert, wenn ein Kind einer Anweisung nicht folgt, es auch nicht versucht, oder wenn das Kind sich in einer Art und Weise verhält, die den Aufforderungen zuwiderläuft. Als Beispiel für unkooperatives Verhalten wäre zu nennen, wenn ein Kind der Aufforderung aufzuräumen nicht folgt, sondern stattdessen weiter mit einem Spielzeugauto spielt.
3. **Unaufmerksames Verhalten** („off-task behavior“): Unaufmerksames Verhalten bedeutet in diesem Zusammenhang, dass sich die beobachteten Kinder nicht auf die zu bearbeitende Aufgabe und die zur Ver-

fügung stehenden Materialien konzentrieren, keine entsprechenden Handlungen ausführen (z. B. Schreiben oder Rechnen) oder nicht in angemessener Weise um Unterstützung bitten. Beispiele für unaufmerksames Verhalten in einem Klassenzimmer sind lautes Reden, Aufstehen oder Tagträumen während des Unterrichts.

Diese und weitere Verhaltensweisen sind im Manual des Beobachtungssystems detailliert in einem **Codebuch** dargestellt, definiert und mit klaren Kodieranweisungen versehen. Die Beobachtungskategorien im standardisierten Beobachtungssystem stellen **Operationalisierungen einzelner theoretischer Konstrukte im Rahmen von Theoriemodellen** dar. Sie messen also im wissenschaftstheoretischen Verständnis des Kritischen Rationalismus (▶ Abschn. 2.2.3) nicht, wie das Verhalten der Schulkinder in Wirklichkeit „ist", sondern wie es im Licht der verwendeten Theorien und Konstrukte erscheint (▶ Abschn. 8.3.4). Die Selektivität, Perspektivität und Konstruiertheit des Beobachteten ist also bei der strukturierten Verhaltensbeobachtung (wie bei jeder anderen quantitativen oder qualitativen Beobachtungstechnik) ausdrücklich zu reflektieren (▶ Abschn. 10.1.1). Die theoretische Reflexion und Meta-Reflexion der Voraussetzungen und Konsequenzen des Einsatzes eines bestimmten standardisierten Beobachtungssystems muss sowohl am Anfang des Forschungsprozesses bei der Wahl des Beobachtungsinstrumentes (Soll das REDSOCS oder ein anderes Instrument verwendet werden?) als auch am Ende des Forschungsprozesses bei der Ergebnisinterpretation (Welche Perspektive auf das beobachtete Verhalten liefert das REDSOCS mit seinem spezifischen Theoriehintergrund? Welche anderen Theorien wären anwendbar? Wo liegen dementsprechend die Grenzen der Aussagekraft der erhobenen Beobachtungsdaten?) erfolgen.

Angenommen, eine Beobachtung mit dem REDSOCS soll während des Mathematikunterrichtes stattfinden. Wichtig ist, dass die beobachteten Kinder in einer Unterrichtsstunde klar definierte Regeln zu befolgen und Aufgaben zu bearbeiten haben (sonst ist das System nicht sinnvoll anwendbar). Erst die klare Regelstruktur einer Unterrichtsstunde ermöglicht eine eindeutige Feststellung von unaufmerksamem oder unkooperativem Verhalten, was z. B. in einer Pausensituation nicht der Fall wäre. Die Beobachtungen im Klassenzimmer können von 20 Minuten bis zu einer Stunde dauern. Für jede Beobachtung wird das gezeigte Verhalten pro Kind insgesamt 10 Minuten lang kodiert, wobei nach einer Kodierminute immer eine 1- bis 3-minütige Pause für die Beobachter eingehalten wird. Innerhalb der zu kodierenden Minute wird das Verhalten des Kindes alle 10 Sekunden protokolliert (**Zeitstichprobe**). Die Beobachter bzw. Kodierer (denn sie protokollieren ihre Beobachtungen in vorgegebenen Kategorien) nutzen typischerweise einen Signalgeber, der zu Beginn jedes 10-Sekunden-Intervalls innerhalb der Kodierminute ein Tonsignal ausgibt. Üblicherweise werden die Verhaltensweisen eines Kindes zu drei verschiedenen Zeitpunkten innerhalb von 14 Tagen beobachtet. Auf diese Weise werden die natürliche Variabilität im kindlichen Verhalten berücksichtigt und genauere und stabilere Ergebnisse über einen gewissen Zeitraum hinweg gewonnen. Der Kodierbogen des REDSOCS-Beobachtungssystems ist in ◘ Abb. 10.5 dargestellt.

Nonreaktive Beobachtung von Verhaltensspuren

Nonreaktive Verfahren der Datenerhebung wurden durch das klassische Werk „Unobstrusive Measures" von Webb, Campbell, Schwartz und Sechrest (1966) in der empirischen Sozialforschung bekannt gemacht (für ein aktuelles Lehr- und Handbuch siehe Lee, 2000). Zu den nonreaktiven Verfahren zählen die Erhebung und Analyse sowohl qualitativer als auch quantitativer Spuren bzw. Dokumente menschlichen Verhaltens und Handelns: Hinweisschilder, Hausordnungen, Autoaufkleber, T-Shirt-Sprüche, Verkaufsstatistiken, Bestenlisten, Verträge, Urkunden, Sitzungsprotokolle etc. lassen Rückschlüsse auf menschliches Erleben und Verhalten zu, ohne dass Forschende durch das Sammeln und Auswerten dieser Verhaltensspuren in das Geschehen eingreifen oder das Verhalten und Erleben der Handelnden beeinflussen. Ein Großteil dieser Verfahren wird – insbesondere wenn es sich um reichhaltige verbale oder visuelle Dokumente handelt, die interpretativ auszuwerten sind – heute der qualitativen **Dokumentenanalyse** (▶ Abschn. 10.6.2) zugeordnet: Interessiert man sich für den Wandel des Männer- und Frauenbildes in der Gesellschaft, so lassen sich z. B. die Männer- und Frauendarstellungen auf den **Titelseiten populärer Publikumszeitschriften** in den letzten 30 Jahren auswerten. Ein solches nonreaktives Vorgehen bei der Datenerhebung wäre der Dokumentenanalyse zuzuordnen, das visuelle Datenmaterial wäre in entsprechenden **Archiven oder Datenbanken** zu finden (zur Dokumentenanalyse ▶ Abschn. 10.6).

Nonreaktive Verfahren der Datenerhebung – Nonreaktive Verfahren der Datenerhebung („nonreactive/nonintruding/unobstrusive measures") umfassen allgemein alle Datenerhebungstechniken, bei denen das Verhalten und Erleben der Untersuchungsteilnehmenden nicht durch Interaktion mit den Forschenden beeinflusst werden kann. Die erhobenen Daten sind somit unverzerrt von möglichen Reaktionen der Untersuchten auf den Forschungsprozess selbst.

Um den Grad der Politisierung auf einem Hochschulcampus zu erfassen oder aktuelle politische Kontroversen zu ermitteln, können im Rahmen einer nonreakti-

School Observation Sheet

Observer: _____
Date: _____
Child: _____

Status: (*circle one*) Primary or Reliability
School: _____

	Minute 1							Minute 2							Minute 3							Minute 4							Minute 5						
	1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6	
Appropriate																																				
Inapproriate*																																				
Comply																																				
Noncomply																																				
No Compliable Command Given																																				
On Task																																				
Off Task																																				
Not Applicable																																				

	Minute 6							Minute 7							Minute 8							Minute 9							Minute 10						
	1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6		1	2	3	4	5	6	
Appropriate																																				
Inapproriate*																																				
Comply																																				
Noncomply																																				
No Compliable Command Given																																				
On Task																																				
Off Task																																				
Not Applicable																																				

* **Inappropriate behavior includes:** whine, tantrum, destructive, disruptive, talks out, cheating, self-stimulation, demanding, yell, cry, negativism, out of area

Notes:

Abb. 10.5 Beobachtungsbogen des REDSOCS („revised edition of the school observation coding system"). (Aus Bagner et al., 2010. Mit freundlicher Genehmigung der Autoren)

ven Datenerhebung **Auto- und Wandaufkleber, Plakate und Graffitis** mit politischen Bezügen gesammelt und ausgewertet werden. Hierzu wäre es notwendig, dass menschliche Beobachtende systematisch das Campusgelände durchstreifen und alle relevanten Informationen registrieren. Ein entsprechendes Vorgehen wäre der **nonreaktiven Beobachtung** zuzuordnen. Nicht nur findet die Forschungsaktivität im Feld statt (anstelle bloßer Archivrecherche), auch die Verknüpfung der Verhaltensspuren mit dem Liveverhalten von Menschen ist enger: So kann

bei der Erhebung von Autoaufklebern live beobachtet werden, wo das entsprechende Fahrzeug parkt. Während das Auszählen von Graffitis auf Häuserwänden als nonreaktive Beobachtung im Feld einzuordnen ist und den beschriebenen „Livecharakter" aufweist, wäre eine Analyse von Fotoaufnahmen von Graffitis aus einem Foto-Archiv eher dem Bereich der Dokumentenanalyse zuzuordnen.

Nonreaktive Beobachtung – Die nonreaktive Beobachtung („non-reactive observation") als Spezialform der nonreaktiven Datenerhebung bezieht sich auf die Erfassung von quantifizierbaren Verhaltens**spuren** im Feld mithilfe menschlicher Sinnesorgane und/oder technischer Sensoren zum Zeitpunkt (oder im Zeitraum) ihres Auftretens. Nonreaktive Beobachtung kann sich dabei auf Offline-Verhaltensspuren („non-reactive offline observation") sowie auf Online-Verhaltensspuren („non-reactive online observation") beziehen.

Die nonreaktive Beobachtung als einfache strukturierte Feldbeobachtung von Verhaltensspuren erfordert wie die strukturierte Verhaltensbeobachtung eine Festlegung der Beobachtungseinheiten, der Beobachtungsorte und -zeiten. Da einzelne Verhaltensspuren gesichert werden, ist mit **Ereignisstichproben** zu operieren. Besonders geeignet ist die nonreaktive Beobachtung, um menschliche Verhaltensweisen frei von Einflüssen durch den Forschungsprozess zu erfassen. Vollständige **Nonreaktivität** hat folgende Voraussetzungen (zum Kontinuum nonreaktiver Verfahren s. Fritsche & Linneweber, 2006):

1. Das Beobachtungssetting ist nicht durch die Forschenden hergestellt, sondern ist Teil der natürlichen Lebenswelt.
2. Es findet keinerlei direkter Kontakt zwischen den Forschenden und den beobachteten Personen statt (es werden nur Verhaltensspuren erfasst). Somit wissen die indirekt beobachteten Personen nicht, dass ihr Verhalten wissenschaftlich untersucht wird.
3. Dementsprechend kann es nicht vorkommen, dass das Verhalten der beobachteten Personen durch Kenntnis der Beobachtung, der Forschungsfragen, der Hypothesen oder der Beeinflussbarkeit des Instruments verzerrt wird.

Die Bedeutung nonreaktiver Beobachtungsmethoden wächst mit der Veralltäglichung moderner **Informations- und Kommunikationstechnologien**, die menschliches Verhalten im Feld detailliert digital aufzeichnen und fortlaufend **digitale Verhaltensspuren** erzeugen. Ein Großteil dieser Daten unterliegt dem Datenschutz (z. B. Aufenthaltsort von Handynutzenden registriert über die Funkzelle; Bewegungsmuster von Autos mit GPS-Empfängern etc.). Nur ein Teil ist für Forschungszwecke zugänglich. Dies betrifft insbesondere die Online-Forschung, weshalb im Folgenden die nonreaktive Beobachtung von Offline- und Online-Spuren getrennt behandelt werden.

▪ Nonreaktive Beobachtung von Offline-Verhaltensspuren

Eine spezielle nonreaktive Beobachtungsmethode von Offline-Verhaltensspuren („non-reactive offline observation") zur Erfassung von **Einstellungen gegenüber Organisationen** stellt die **Lost-Letter-Technik** dar (Milgram, Mann, & Harter, 1965). Hierbei werden adressierte und frankierte Briefe in der Weise im öffentlichen Raum ausgelegt, dass sie den Eindruck erwecken, verloren worden zu sein. Die Adressaten der Briefe sind jeweils bestimmten Organisationen zuzuordnen (z. B. politische Partei, Religionsgemeinschaft). Das Einwerfen der Briefe in den Briefkasten wird indirekt durch den Posteingang beim Forschungsteam beobachtet (die Briefe gelangen über eine Postfachadresse an die Forschenden, nicht an die namentlich genannten Adressaten) bzw. der Posteingang stellt in diesem Fall die Verhaltensspur der untersuchten Personen dar. Man nimmt an, dass eine hohe Rücklaufquote für eine positive Einstellung gegenüber der betreffenden Organisation spricht, da man sich für positiv bewertete Organisationen eher die Mühe macht, den gefundenen Brief in einen Briefkasten einzuwerfen. Wenn die Lost-Letter-Technik für einen Vergleich zwischen Organisationen eingesetzt wird, handelt es sich um ein quasi-experimentelles Untersuchungsdesign. Die Nonreaktivität dieses Verfahrens ist insofern eingeschränkt, als das Auslegen der Briefe an bestimmten Stellen z. B. eines Wohngebietes einen Eingriff der Forschenden in das Feld bedeutet (Fritsche & Linneweber, 2006). Eine ganz andere indirekte Methode zur Messung von Einstellungen ist der Implicit-Association-Test (IAT), der den Testverfahren zugeordnet ist und mit Reaktionszeiten arbeitet (► Abschn. 10.4.1).

Wie an der Lost-Letter-Technik erkennbar, ist nicht jede Beobachtung von Verhaltensspuren als vollständig nonreaktiv aufzufassen. Umgekehrt kann auch eine einfache direkte Verhaltensbeobachtung letztlich kaum Reaktivität auslösen und somit als nonreaktive Methode gelten, etwa wenn Verkehrsbeobachter am Straßenrand von den Autofahrern gar nicht wahrgenommen werden.

Welche Arten von Offline-Verhaltensspuren lassen sich für wissenschaftliche Zwecke beobachten? Die **Zusammensetzung des Hausmülls** gibt Auskunft über das Umweltbewusstsein (Mülltrennung), aber auch über Ernährungs- und Konsumgewohnheiten. Stark **abgewetzter Teppich** vor bestimmten Exponaten im Museum deutet darauf hin, dass diese Werke besonders viel Publikum anziehen. Um die Beliebtheit von Radiosendern in einer Region zu erfassen, kann man in lokalen Autowerkstätten beobachten, welche **Radiosender in den Autoradios** der zur Inspektion oder Reparatur gebrachten Wagen gerade eingestellt sind. Das ► Studienbeispiel „Nonreaktive Beobachtung von Offline-Verhaltensspuren" zeigt die Erfassung des Händewaschens.

Studienbeispiel

Nonreaktive Beobachtung von Offline-Verhaltensspuren

Einer Forschergruppe gelang es, einen Monat lang das **Händewaschverhalten** von rund 200 000 Personen auf englischen Autobahnraststätten im Zuge einer nonreaktiven Beobachtung zu erfassen. Dazu installierten sie Infrarotsensoren, die sowohl das Betreten der Waschräume als auch die Benutzung des Seifenspenders registrierten: Bei 65 % der Besucherinnen der Damentoilette gegenüber 31 % der Besucher der Herrentoilette wurde Händewaschen nonreaktiv beobachtet (Judah et al., 2009). Die Verwendung von zwei unabhängigen **Infrarotsensoren an der Tür und am Seifenspender** erlaubte eine völlig anonyme Datenerfassung, die ethisch vertretbar ist, während die Installation von Videokameras in den Toilettenräumen sicher nicht akzeptabel gewesen wäre. Die Methode kann auch eingesetzt werden, um im Rahmen einer Evaluationsstudie die Wirksamkeit von unterschiedlichen Kampagnen zu überprüfen, die zum Händewaschen animieren sollen (z. B. Plakat über dem Waschbecken, das auf die Gefahr der Übertragung von Krankheitserregern hinweist).

10

Nonreaktive Beobachtung von Online-Verhaltensspuren

Im Rahmen der Online-Forschung können automatisch registrierte Verhaltensspuren unterschiedlichster Art verwendet werden, die als nonreaktive Beobachtung von Online-Verhaltensspuren einzuordnen sind („non-reactive online observation“). Dabei besteht einerseits die Möglichkeit, dass Forschende im Rahmen von Online-Studien entsprechende Protokolldateien selbst anfertigen oder auf öffentlich zugängliche Tools und Datenbanken zurückzugreifen. So gibt es z. B. die Möglichkeit, sich mit **Google Trends** (http://www.google.com/trends) anzeigen zu lassen, welche Suchbegriffe in welchen Ländern in welchen Zeitperioden wie häufig in die Suchmaschine eingegeben wurden. Hier werden mutmaßlich alle Suchanfragen einbezogen. Allerdings kann man sich keine absoluten Häufigkeiten ausgeben lassen, sondern nur relative Angaben zur Verbreitung des Suchbegriffs im Vergleich zu anderen Suchbegriffen oder im Zeitverlauf. Entsprechende Daten können genutzt werden, um zu beschreiben, welche Themen bei Internetnutzenden welcher (Bundes-)Länder gerade besonders populär sind. Mögliche Forschungsfragen könnten lauten: Inwiefern strahlen regionale Ereignisse (z. B. Kommunalwahlen) bundesweit aus und beeinflussen das Online-Suchverhalten auch in anderen Bundesländern? Wie stark spiegeln sich internationale Großereignisse (z. B. Weltklimagipfel) im nationalen Online-Suchverhalten? Werden sexualbezogene Suchanfragen im Zuge einer mutmaßlich wachsenden Sexualisierung der Gesellschaft häufiger oder sinkt ihr relativer Anteil im Zuge einer Anpassung der Online-Population an die breite Gesellschaft?

Dem **Alexa-Ranking** (http://www.alexa.com) ist zu entnehmen, welches die jeweils meistbesuchten Websites unterschiedlicher Länder sind. Hier zeigt sich z. B., dass unter den Top-100-Websites der meisten Länder ca. fünf sexualbezogene Angebote auf eher mittleren Plätzen zu finden sind. Entsprechende Online-Verhaltensspuren sind indessen kritisch zu reflektieren. So nehmen am Alexa-Ranking nur Internet-Nutzende teil, die sich freiwillig und ausdrücklich dazu bereit erklärt haben, ihr Web-Nutzungsverhalten automatisch beobachten zu lassen (durch Installation der Alexa-Toolbar), was die Nonreaktivität der Datensammlung einschränkt und einen starken Freiwilligen-Bias nach sich zieht.

Über andere Tools und Datenbanken lässt sich beobachten, wo gerade Microblogging-Beiträge geschrieben werden oder über welche Themen aktuell in der Blogosphäre diskutiert wird. Auch für Online-Posts in einem öffentlichen Webforum oder auf einer Content-Plattform veröffentlichte Videos ist als Online-Verhaltensspur u. a. jeweils die **Anzahl der Abrufe** sichtbar – hier handelt es sich sozusagen um das virtuelle Pendant des abgenutzten Teppichs im Museum vor beliebten Ausstellungsstücken.

Lernquiz

1. Nennen Sie fünf Vorzüge der Beobachtung gegenüber der Befragung mit jeweils einem Beispiel.
2. Was ist unter Perspektivität, Selektivität und Konstruiertheit der Beobachtung zu verstehen?
3. In welchem Zusammenhang fällt bei der qualitativen und qantitativen Beobachtung jeweils besonders viel Arbeit an?
4. Wie überprüfen Sie die Beobachterübereinstimmung wenn a) zwei Beobachter registrieren, ob Autofahrer angeschnallt sind oder nicht; b) zwei Beobachter auf Ratingskalen bewerten, wie intensiv Versuchspersonen flirten, c) zwei Beobachter protokollieren, ob Schulkinder im Unterricht unangemessenes Verhalten zeigen und dabei den Schweregrad des Verhaltens einstufen (leichtes, mittelschweres und schweres Fehlverhalten)?
5. Welche Varianten der Stichprobenziehung sind für quantitative Beobachtungsstudien relevant und wann werden sie eingesetzt?
6. Welche Aussage/n stimmen?
   a) Eine ethnografische Feldbeobachtung ist immer teilnehmend.
   b) Ohne die Zustimmung des Gatekeepers darf eine ethnografische Feldbeobachtung nicht stattfinden.

c) Während einer ethnografischen Feldbeobachtung können Notizen, Fotos und Videos erstellt werden.
d) Eine ethnografische Feldbeobachtung ist verdeckt durchzuführen.

7. Was ist eine strukturierte Verhaltensbeobachtung, wie und wo wird sie durchgeführt?
8. Was ist eine nonreaktive Beobachtung, wie und wo kommt sie zum Einsatz?
9. In einem Schnellrestaurant wird von einem Sitzplatz in der Nähe des Verkaufstresens aus beobachtet, wie oft und von wem das Aktionsangebot der Woche bestellt wird und wie oft das Personal ausdrücklich auf das Angebot hinweist. Was trifft auf diese Studie zu?
   a) Es handelt sich um eine ethnografische Feldbeobachtung.
   b) Es handelt sich um eine nicht-teilnehmende Beobachtung.
   c) Es handelt sich um eine strukturierte Verhaltensbeobachtung.
   d) Es handelt sich um eine Zeitstichprobe.
   e) Es handelt sich um eine quantitative Beobachtung mit geringem Komplexitätsgrad.
10. Was versteht man unter der Lost-Letter-Technik?

Übungsaufgaben

1. Oft wird behauptet, dass „Sex“ das beliebteste Suchwort im Internet ist. Überprüfen Sie diese Behauptung mittels nonreaktiver Beobachtung des Suchverhaltens in Deutschland für den Zeitraum der letzten Jahre. Nutzen Sie dazu die Tools der Suchmaschine Google (z. B. „Google Trends“). Kontrastieren Sie „Sex“ als Suchbegriff mit anderen Suchbegriffen (z. B. „Facebook“, „Weihnachten“, „Wetter“).
2. Führen Sie eine teilnehmende ethnografische Feldbeobachtung des Fahrstuhlfahrens durch. Suchen Sie sich dazu ein Gebäude mit Fahrstuhl und absolvieren und dokumentieren Sie mehrere Fahrstuhlfahrten (schriftliche Feldnotizen). Konzentrieren Sie sich dabei auf alle Phasen des Vorgangs: 1) das Warten auf den Fahrstuhl, 2) das Einsteigen und einen Platz Einnehmen, 3) das Verhalten während der Fahrt inklusive verbaler und nonverbaler Kontakte unter Mitfahrenden sowie 4) das Aussteigen aus dem Fahrstuhl. Welche Erkenntnisse haben Sie gewonnen (Auswertung und Interpretation der Feldnotizen)? Vergleichen Sie Ihre Ergebnisse mit der Fahrstuhl-Studie von Hirschauer (1999).
3. Untersuchen Sie im Rahmen einer quantitativen Beobachtung mit geringem Komplexitätsgrad Ausmaß und Art des Multitasking-Verhaltens von Studierenden im Hörsaal während einer Vorlesungssitzung. Beobachtete Nebenbei-Beschäftigungen könnten sein: Private Aktivitäten auf dem Laptop, Handynutzung, Gespräche mit Sitznachbarn, Zeitungslektüre etc. Erstellen Sie im Vorfeld der Beobachtung einen standardisierten Beobachtungsbogen und füllen Sie diesen aus.

Weiterführende Literatur

- Bernard, H. R. (1995). *Research methods in anthropology. Qualitative and quantitative approaches* (2nd ed.). Walnut Creek, CA: AltaMira Press.
- Dewalt, K. M. & Dewalt, B. R. (2002). *Participant observation: A guide for fieldworkers*. Walnut Creek, CA: AltaMira Press
- Faßnacht, G. (2007). *Systematische Verhaltensbeobachtung*. Stuttgart: UTB.
- Gehrau, V. (2002). *Die Beobachtung in der Kommunikationswissenschaft. Methodische Ansätze und Beispielstudien*. Konstanz: UVK.
- Greve, W. & Wentura, D. (1997). *Wissenschaftliche Beobachtung: Eine Einführung*. Weinheim: Beltz/PVU.
- Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters* (3rd ed.). Gaithersburg: Advanced Analytics, LLC.
- Harrigan, J. A., Rosenthal, R., & Scherer, K. (Eds.). (2005). *The new handbook of methods in nonverbal behavior research*. Oxford: Oxford University Press.

Literatur

Anderson, L. (2006). Analytic autoethnography. *Journal of Contemporary Ethnography, 35*(4), 373–395.

Atkinson, P., Coffey, A., & Delamont, S. (Eds.). (2007). *Handbook of Ethnography*. Los Angeles: Sage.

Bagner, B., Boggs, S., & Eyberg, S. (2010). Evidence-based school behavior assessment of externalizing behavior in young children. *Education and Treatment of Children, 33*(1), 65–83.

Bales, R. F. (2001). *Social interaction systems: Theory and measurement*. New Brunswick: Transaction Publishers.

Bales, R. F. & Cohen, S. P. (1979). *SYMLOG: a system for the multiple level observation of groups*. London: Collier Macmillan.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology, 63*(3), 575–582.

Beer, B. (Ed.). (2008). *Methoden der ethnologischen Feldforschung*. Berlin: Reimer.

Behr, R. (2002). Lebenswelt Polizei: Ein ethnografischer Zugang zur Berufsidentität von Polizeibeamten. *Forum Qualitative Sozialforschung, 3*(1), Art. 13. Abgerufen 6. September, 2013, unter http://www.qualitative-research.net/index.php/fqs/article/view/877/1908.

Berg, B. (1989). *Qualitative research methods for the social sciences.* Boston: Allyn & Bacon.

Blumer, H. (1973). Der methodologische Standort des symbolischen Interaktionismus. In Arbeitsgruppe Bielefelder Soziologen (Ed.), *Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit* (Bd. 1, S. 80–148). Opladen: Westdeutscher Verlag.

Bortz, J. & Lienert, G. A. (2003). *Kurzgefasste Statistik für die Klinische Forschung: Leitfaden für die verteilungsfreie Analyse kleiner Stichproben* (2. Aufl.). Heidelberg: Springer.

Bryman, A. (2008). *Social Research Methods* (3rd. ed.). Oxford: Oxford University Press.

Cantor, A. B. (1996). Sample size calculations for Cohen's Kappa. *Psychological Methods, 1*(2), 150–153.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, J. (1968). Weighted Kappa. Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Cranach, V. M. & Frenz, H.-G. (1969). Systematische Beobachtungen. In C. F. Graumann (Hrsg.), *Handbuch Psychologie. Band 7 Sozialpsychologie* (S. 269–330). Göttingen: Hogrefe.

Desroches, F. (1990). Tearoom trade: A research update. *Qualitative Sociology, 13*(1), 39–61.

Donner, A. (1998). Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine, 17*(10), 1157–1168.

Douglas, J. D., Rasmussen, P. K., & Flanagan, C. A. (1977). *The nude beach.* Beverly Hills: Sage.

Drummond, M. (2010). The natural: An autorethnography of a masculinized body in sport. *Men and Masculinities, 12*(3), 374–389.

Ellis, C. (1991). Sociological introspection and emotional experience. *Symbolic Interaction, 14*(1), 23–50.

Ellis, C. (2004). *The ethnographic I: A methodological novel about autoethnography*. Walnut Creek: AltaMira Press.

Ellis, C. & Bochner, A. B. (2000). Autoethnography, personal narrative, reflexivity: Researcher as subject. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 733–768). Thousand Oaks: Sage.

Ferguson, C. J., Rueda, S. M., Cruz, A. M., Ferguson, D. E., Fritz, S., & Smith , S. M. (2008). Violent video games and aggression. Causal relationship or byproduct of family violence and intrinsic violence motivation? *Criminal Justice and Behavior, 35*(3), 311–332.

Ferrell, J. & Hamm, M. S. (1998). *Ethnography at the edge: Crime, deviance, and field research*. Boston: Northeastern University Press.

Fritsche, I. & Linneweber, V. (2006). Nonreactive Methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 189–203). Washington: American Psychological Association.

Gerkin, P. (2009). Participation in victim-offender mediation: Lessons learned from observations. *Criminal Justice Review, 34*(2), 226–247.

Girtler, R. (2001). *Methoden der Feldforschung.* Wien: Lit Verlag.

Goffman, E. (1969). *Wir alle spielen Theater. Die Selbstdarstellung im Alltag*. München: Piper.

Goffman, E. (1971). *Verhalten in sozialen Situationen. Strukturen und Regeln der Interaktion im öffentlichen Raum*. Gütersloh: Bertelsmann.

Goffman, E. (1996). *Rahmen-Analyse. Ein Versuch über die Organisation von Alltagserfahrungen*. Frankfurt: Suhrkamp.

Goscinny, R. & Sempé, J.-J. (1975). *Der kleine Nick und die Schule*. Zürich: Diogenes.

Greve, W. & Wentura, D. (1997). *Wissenschaftliche Beobachtung: Eine Einführung*. Weinheim: Beltz/PVU.

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters* (3rd ed.). Gaithersburg: Advanced Analytics, LLC.

Häder, M. (2006). *Empirische Sozialforschung*. Wiesbaden: VS Verlag.

Hahlweg, K. (2004). Kategoriensystem für partnerschaftliche Interaktion (KPI): Interactional Coding System (ICS). In P. K. Kerig & D. H. Baucom (Eds.), *Couple observational coding systems* (S. 127–142). Mahwah: Erlbaum.

Hammersley, M. & Atkinson, P. (2007). *Ethnography: Principles in Practice.* (Vol. 3). London: Taylor & Francis.

Hine, C. (2000). *Virtual Ethnography*. Thousand Oaks: Sage.

Hine, C. (2005). *Virtual Methods*. Oxford: Berg.

Hirschauer, S. (1999). Die Praxis der Fremdheit und die Minimierung von Anwesenheit. Eine Fahrstuhlfahrt. *Soziale Welt, 50*(3), 221–246.

Humphreys, L. (1970). *Tea room trade*. Chicago: Aldine.

Humphreys, L. (1972). Tearoom trade: Impersonal sex in public places. In W. Feigelman (Ed.), *Sociology full circle; contemporary readings on society*. New York: Praeger.

Izenman, A. J. & Zabell, S. L. (1981). Babies and the blackout: The genesis of a misconception. *Social Science Research, 10*(3), 282–299.

Jahoda, M., Lazarsfeld, P. F., & Zeisel, H. (1933). *Die Arbeitslosen vom Marienthal. Ein soziographischer Versuch*. Leipzig: Hirzel.

Jorgensen, D. L. (1990). *Participant observation. A methodology for human studies*. Thousand Oaks: Sage.

Judah, G., Aunger, R., Schmidt, W.-P., Michie, S., Granger, S., & Curtis, V. (2009). Experimental pretesting of hand-washing interventions in a natural setting. *American Journal of Public Health, 99*(22), 405–411.

Kempster, S. & Stewart, J. (2010). Becoming a leader: A co-produced autoethnographic exploration of situated learning of leadership practice. *Management Learning, 41*(2), 205–219.

Kozinets, R. (2009). *Netnography. Doing ethnographic research online.* Thousand Oaks: Sage.

Labaree, R. (2002). The risk of 'going observationalist': negotiating the hidden dilemmas of being an insider participant observer. *Qualitative Research, 2*(1), 97–122.

Langer, A. (2007). Mit Goffman auf dem Drogenstrich. Rückblick auf eine „gelungene Forschungsbeziehung". *Österreichische Zeitschrift für Soziologie, 32*(2), 90–104.

Lee, R. M. (2000). *Unobtrusive methods in social research*. Buckingham: Open University Press.

Legewie, H. (1987). *Alltag und seelische Gesundheit. Gespräche mit Menschen aus dem Berliner Stephansviertel*. Berlin: Psychatrie-Verlag.

Leitner, K., Lüders, E., Greiner, B., Niedermeier, A., & Volpert, W. (1993). *RHIA/VERA. Analyse psychischer Anforderungen und Belastungen in der Büroarbeit – Das RHIA/VERA-Büro-Verfahren*. Göttingen: Hogrefe.

Main, M. & Hesse, E. (1990). Parents' unresolved traumatic experiences are related to infant disorganized attachment status. Is frightened and/or frightening parental behavior the linking mechanism? In M. Greenberg, D. Cicchetti & E. M. Cummings (Eds.), *Attachment in*

the preschool years: Theory, research and intervention. Chicago: University of Chicago Press.
Malinowski, B. (1922). *Argonauten des westlichen Pazifik.* New York: Dutton.
Malinowski, B. (1929). *Das Geschlechtsleben der Wilden in Nordwest-Melanesien.* Leipzig: Grethlein.
Marcks, M. (1974). *Weißt du, daß du schön bist?* Starnberg: Raith.
Milgram, S., Mann, L., & Harter, S. (1965). The lost-letter technique: A tool of social science research. *Public Opinion Quarterly, 29*(3), 437–438.
O'Reilly, K. (2005). *Ethnographic methods*. London: Taylor & Francis.
Obermann, C. (2006). *Assessment Center. Entwicklung, Durchführung und Trends.* Wiesbaden: Gabler.
Penke, L. & Asendorpf, J. (2008). Beyond global sociosexual orientations: A more differentiated look at sociosexuality and its effects on courtship and romantic relationships. *Journal of Personality and Social Psychology, 95*(5), 1113–1135.
Przyborski, A. & Wohlrab-Sahr, M. (2009). *Qualitative Sozialforschung. Ein Arbeitsbuch.* München: Oldenbourg.
Rosenbloom, T., Ben-Eliyahu, A., Nemrodov, D., Biegel, A., & Perlman, A. (2009). Committing driving violations: An observational study comparing city, town and village. *Journal of Safety Research, 40*(3), 215–219.
Rühl, D., Bölte, S., Feineis-Matthews, S., & Poustka, F. (2004). *ADOS. Diagnostische Beobachtungsskala für autistische Störungen. Deutsche Fassung der Autism Diagnostic Observation Schedule von Catherine Lord, Michael Rutter, Pamela C. Dilavore und Susan Risi.* Göttingen: Hogrefe.
Saunders, M. N. K., Lewis, P., & Thornhill, A. (2003). *Research methods for business students*. London: Pearson.
Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der empirischen Sozialforschung*. München: Oldenbourg
Sedlmeier, P. & Renkewitz, F. (2007). *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson.
Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13*(4), 251–271.
Sim, J. & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268.
Sriram, C. L., King, J., Mertus, J., & Martin-Ortega, O. (2009). *Surviving field research: Working in violent and difficult situations*. New York: Routledge.
Townsend, M. (2006). Motorists' use of hand held cell phones in New Zealand: An observational study. *Accident Analysis and Prevention, 38*(4), 748–750.
Udry, J. R. (1970). The effect of the great blackout of 1965 on births in New York City. *Demography, 7*(3), 325–327.
Wagner, B. (2009). Becoming a sexual being: Overcoming constraints on female sexuality. *Sexualities, 12*(3), 289–311.
Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Ubstrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.
Weigand, G. & Hess, R. (2007). *Teilnehmede Beobachtung in interkulturellen Situationen.* Frankfurt/Main: Campus.
Whyte, W. F. (1943). *Street corner society: The social study of an Italian slum*. Chicago: University of Chicago Press.
Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliablität.* Göttingen: Hogrefe.

10.2 Interview

Lernziele

- Wissen, was die Interviewtechnik als wissenschaftliche Datenerhebungsmethode auszeichnet.
- Verschiedene Typen wissenschaftlicher Interviews im Hinblick auf Herangehensweise und Anwendungsfelder voneinander abgrenzen können.
- Potenzielle Interviewfehler und Gegenmaßnahmen kennen.
- Eine qualitative sowie eine quantitative Interviewstudie planen und durchführen können.
- Die jeweiligen Vor- und Nachteile qualitativer und quantitativer Interviews abwägen können.

In diesem Kapitel wird zunächst erklärt, wodurch sich eine wissenschaftliche mündliche Befragung bzw. ein wissenschaftliches Interview als empirische Datenerhebungsmethode auszeichnet (▶ Abschn. 10.2.1). Dann werden verschiedene Varianten des qualitativen (▶ Abschn. 10.2.2) sowie des quantitativen Interviews (▶ Abschn. 10.2.3) erläutert und an Studienbeispielen illustriert.

10.2.1 Wissenschaftliches Interview

Um zu verstehen, was eine wissenschaftliche mündliche Befragung als Datenerhebungsverfahren kennzeichnet und welche Vor- und Nachteile sie im Vergleich zu anderen Datenerhebungsmethoden – z. B. der Beobachtung oder der schriftlichen Befragung – aufweist, wenden wir uns zunächst ihrer **Definition und Indikation** zu. Anschließend wird erläutert, worin sich das wissenschaftliche Interview vom **Alltagsgespräch** unterscheidet. Um unterschiedliche Formen wissenschaftlicher Interviews zu differenzieren, stehen verschiedene **Klassifikationskriterien** zur Verfügung – etwa der Strukturierungsgrad (qualitatives vs. quantitatives Interview), die Anzahl der gleichzeitig befragten Personen (Einzel- oder Gruppeninterview) oder die Art des Interviewkontaktes (persönliches, telefonisches oder Online-Interview). Schließlich werden noch typische **Interviewfehler** und Gegenmaßnahmen behandelt.

Definition und Indikation des wissenschaftlichen Interviews

Die Befragung ist generell die in den empirischen Sozialwissenschaften am häufigsten eingesetzte Datenerhebungsmethode. Im qualitativen Forschungsansatz stellt das unstrukturierte oder halbstrukturierte Interview die wichtigste Datenerhebungstechnik dar. Im quantitativen Forschungsansatz dominiert die Fragebogenmethode (▶ Abschn. 10.3), allerdings werden auch vollstrukturierte mündliche Interviews eingesetzt. Quantitative mündliche und schriftliche Befragungen werden zusammenfassend als **Umfragen** („surveys“) bzw. Umfrageforschung („survey research“) bezeichnet. Bei **Meinungsumfragen** spricht man auch von „opinion polls“.

Wissenschaftliche mündliche Befragung bzw. wissenschaftliches Interview – Unter einer wissenschaftlichen mündlichen Befragung bzw. einem wissenschaftlichen Interview („research interview“/„scientific interview“) verstehen wir die zielgerichtete, systematische und regelgeleitete Generierung und Erfassung von verbalen Äußerungen einer Befragungsperson (Einzelbefragung) oder mehrerer Befragungspersonen (Paar-, Gruppenbefragung) zu ausgewählten Aspekten ihres Wissens, Erlebens und Verhaltens in mündlicher Form. Interviews können im persönlichen Face-to-Face-Kontakt sowie telefonisch (fernmündlich) oder online (schriftliche Mündlichkeit) durchgeführt werden. Die dem Interview zugrunde liegenden verbalen Fragen werden den Befragungspersonen („interviewees“, „respondents“) in einer Interaktion jeweils von einer Interviewerin oder einem Interviewer („interviewer“) gestellt. Die Antworten werden dokumentiert und systematisch analysiert. Die vier zentralen Elemente der mündlichen Befragung sind a) die Befragungspersonen, b) die Interviewerin bzw. der Interviewer, c) die Interviewsituation und d) die Interviewfragen.

Qualitative und quantitative Interviews liefern Datenmaterial für unterschiedliche Fachdisziplinen bzw. Forschungsfelder, z. B. für die Psychologie, die Soziologie, die Kommunikationswissenschaft, die Pädagogik, die Gesundheitsforschung, die Wirtschaftswissenschaften, die Sprachforschung, die Volkskunde (z. B. Interviews mit Gewährsleuten, „oral history“), die Geschichtswissenschaft (Interviews mit Zeitzeugen, z. B. Alltags- und Technikgeschichte), die Geschlechterforschung, die Kriminologie oder die Markt- und Meinungsforschung.

Unter welchen Bedingungen sollte man sich im jeweiligen Forschungsfeld für die Interviewtechnik anstatt z. B. für die Fragebogentechnik oder die Beobachtungsmethode entscheiden (**Indikation**)?

- Die Interviewtechnik als Befragungsform hat gegenüber der Beobachtung den Vorteil, dass zum einen Aspekte des **subjektiven Erlebens** zugänglich werden, die generell nicht der Beobachtung zugänglich sind (z. B. Gefühle, Meinungen, Überzeugungen, Glaubensinhalte) und dass zum anderen auch **nicht direkt beobachtbare Ereignisse und Verhaltensweisen** erfasst werden können (z. B. Erlebnisse, die sich auf private oder in der Vergangenheit liegende Situationen beziehen).
- Gegenüber der schriftlichen Befragung mittels selbst ausgefülltem Fragebogen, die von den Befragungspersonen entsprechende Sehfähigkeit, Konzentration sowie Lese- und Schreibkompetenzen erfordert, handelt es sich bei der Interviewtechnik um eine **niedrigschwellige alltagsnahe Methode**, mit der auch

Befragungspersonen angesprochen werden, für die eine schriftliche Befragung ungeeignet wäre (z. B. Menschen mit eingeschränkten Lese- und Schreibfähigkeiten).
- Die mündliche Befragung findet in einer **Live-Situation** statt (meist als persönliches Interview, aber auch als Telefon- oder Online-Interview), während Fragebögen oft zeitversetzt und ohne direkten Kontakt zum Forschungsteam ausgefüllt werden. Dadurch erhält man mittels Interviewtechnik **Hintergrundinformationen** über die Befragungspersonen und die Befragungssituation und kann die Datenqualität besser einschätzen (z. B. wie zügig wird geantwortet, wo gibt es Rückfragen, wie kooperativ wirkt die Befragungsperson etc.).
- Die Live-Situation des Interviews sorgt für eine **persönlichere Atmosphäre** als das Ausfüllen eines Fragebogens. Dies kann wünschenswert sein, wenn man Befragungspersonen im Zuge einer Studie individuell ansprechen und auf ihre Antworten eingehen möchte. Diese Option wird insbesondere bei qualitativen Interviews ausgeschöpft.
- Wenn Befragungspersonen sich mündlich anstatt schriftlich äußern, können sie zu einer einzelnen Frage **in kürzerer Zeit viel mehr Informationen** liefern. Insbesondere ausführliche Schilderungen komplexer Zusammenhänge oder zeitlich lang ausgedehnter Prozesse erhält man von den meisten Menschen nicht schriftlich, sondern nur mündlich im Zuge qualitativer Interviews. Will man eine **größere Zahl von Fragen** stellen und/oder handelt es sich um **kompliziertere Fragen**, so ist das Interview der Fragebogentechnik vorzuziehen, da umfangreiche und komplizierte schriftliche Fragebögen auf geringe Akzeptanz stoßen.

Neben den Vorzügen der Interviewtechnik sind auch ihre **Nachteile** zu beachten:
- Durch die Notwendigkeit, alle Befragungspersonen einzeln persönlich zu kontaktieren und mit ihnen live die komplette Befragung durchzuführen, entsteht bei Interviews ein sehr viel **höherer Zeit- und Kostenaufwand pro Befragungsperson** als bei der Fragebogentechnik.
- Die Interviewsituation stellt für die Interviewerinnen und Interviewer eine besondere Herausforderung dar: Sie müssen sich einerseits individuell auf die soziale Interaktion mit unterschiedlichen Befragungspersonen einstellen und andererseits dafür sorgen, dass das Vorgehen im Sinne der Wissenschaftlichkeit abläuft, wobei die Anforderungen bei qualitativen und quantitativen Interviews unterschiedlich sind. Es werden **belastbare und sorgfältig geschulte Interviewende** benötigt, die für schriftliche Fragebogenstudien nicht notwendig sind.
- Die **geringere Anonymität der Live-Situation** des Interviews – im Unterschied zur Fragebogentechnik – kann für die Interviewten ein ehrliches Beantworten heikler Fragen erschweren (z. B. verstärkte Antwortverfälschungen im Sinne sozial erwünschten Antwortens, um vor dem Interviewer gut dazustehen).
- Die **Reaktivität** der Interviewmethode (Befragte wissen, dass sie an einem Interview teilnehmen) birgt im Unterschied zu nonreaktiven Methoden der Datenerhebung (z. B. Beobachtung von Verhaltensspuren, Dokumentenanalyse) die Gefahr **zahlreicher Verzerrungen**. Es handelt sich bei Interviewdaten um aktiv durch den Forschungsprozess generiertes Datenmaterial.

Abgrenzung des wissenschaftlichen Interviews vom Alltagsgespräch

Während im **Alltagsgespräch** die Beteiligten wechselseitig Fragen stellen und Antworten geben, handelt es sich beim **Interview** um eine asymmetrische Kommunikationssituation mit klarer Rollenverteilung zwischen der Person, die Fragen stellt bzw. Informationen erlangen möchte (Interviewer/in) und der Person, die Antworten gibt (Befragungsperson).

Asymmetrische Interviewsituationen werden nicht nur als wissenschaftliche Interviews in der Forschung realisiert, sondern spielen auch in vielen anderen Lebensbereichen eine Rolle. Man denke z. B. an klinische bzw. diagnostische Interviews (Schneider & Margraf, 2008), an journalistische Interviews, an Vernehmungen oder an Vorstellungsinterviews. Auch diese Interviews erfolgen, genau wie das wissenschaftliche Interview, zielgerichtet und systematisch, auch hier werden die Antworten der Befragten in der Regel dokumentiert und ausgewertet.

Was das wissenschaftliche Interview von anderen Interviewformen indessen grundlegend unterscheidet, ist seine konsequente Orientierung an **Regeln der Wissenschaftlichkeit** mit dem Ziel, zu **generalisierbaren Aussagen** über die Erfahrungswirklichkeit zu gelangen. Demgegenüber laufen andere Interviewtypen meist auf Beurteilungen einzelner Personen hinaus.

Im Sinne der Wissenschaftlichkeit muss bei Interviewstudien die **Auswahl der Befragungspersonen** methodischen **Regeln der Stichprobenziehung** (▶ Kap. 9) genügen, sofern keine Vollerhebung vorgenommen wird. Bei qualitativen Interviews greift man z. B. auf Stichproben von theoretisch besonders relevanten Fällen zurück (z. B. Auswahl typischer oder extremer Fälle). Bei quantitativen persönlichen oder telefonischen Interviewstudien kommen teilweise echte Zufallsstichproben aus der jewei-

ligen Zielpopulation (z. B. Bevölkerung einer Stadt) zum Einsatz. In jedem Fall muss die Auswahl der Interviewten sorgfältig geplant, begründet und hinsichtlich der Grenzen der Aussagekraft der Studie kritisch diskutiert werden. So genannte **Blitzumfragen** unter spontan herausgegriffenen Passanten auf der Straße, wie sie im Fernsehen oft zu sehen sind, haben keinerlei generalisierende Aussagekraft, da die Stichprobenziehung in diesem Fall nicht regelgeleitet erfolgt und nicht kritisch hinterfragt wird.

Weiterhin muss bei einer wissenschaftlichen Interviewstudie darauf geachtet werden, dass die ausgewählten Befragungspersonen **aussagekräftige Daten** liefern. Dies betrifft die Gestaltung der Interviewsituation, die Merkmale und Verhaltensweisen der Interviewerinnen und Interviewer sowie die konkrete Formulierung, Reihenfolge und Präsentation der Fragen und Antwortmöglichkeiten (Interviewinstrument: z. B. Leitfaden bei einem qualitativen Interview, vollstandardisierter Fragebogen bei einem quantitativen Interview). All diese Faktoren sind auf die befragte Zielgruppe abzustimmen. Entsprechende **Regeln der Instrument-Entwicklung und Interview-Durchführung** werden einerseits theoretisch (in der Umfrageforschung auf der Basis von Theorien des Antwortverhaltens, z. B. Bachleitner, Weichbold, & Aschauer, 2010) und andererseits aus empirischen Methodenstudien abgeleitet, die unterschiedliche Vorgehensweisen einem praktischen Vergleich hinsichtlich der erzielten Datenqualität unterziehen (z. B. in der Umfrageforschung der experimentelle Vergleich von Interview-Fragebögen mit unterschiedlicher Fragen-Reihenfolge; vgl. Petersen, 2002).

Während bei einem quantitativen Interview die Hauptarbeit in der Konstruktion des standardisierten Interview-Instruments liegt, fällt bei qualitativen Interviews die Hauptarbeit bei der Analyse der Interviewmitschriften (Transkripte) an. Denn dann muss im Sinne der Wissenschaftlichkeit gemäß **Regeln der Interviewauswertung** sichergestellt werden, dass aus den umfangreichen Interviewäußerungen intersubjektiv nachvollziehbar tragfähige Schlussfolgerungen gezogen werden und nicht beliebige Zitate herausgegriffen werden.

Klassifikationskriterien für wissenschaftliche Interviews

Der Variantenreichtum wissenschaftlicher Interviewtechniken ist enorm und kann in einem einzigen Kategoriensystem nur unvollständig zum Ausdruck gebracht werden. Wir konzentrieren uns hier auf die **sechs wichtigsten Klassifikationskriterien** für die Forschungspraxis:

1. **Grad der Strukturierung der Interviewsituation** mit den drei Stufen unstrukturiert/nicht-strukturiert versus halbstrukturiert/teilstrukturiert versus vollstrukturiert. Der Grad der Strukturierung bezieht sich auf die Interviewsituation, der **Grad der Standardisierung** auf das verwendete Interviewinstrument. Dagegen sind **Offenheit bzw. Geschlossenheit** Merkmale der einzelnen Fragen (Atteslander, 2010; Tab. 10.3). Zuweilen werden diese Ebenen jedoch vermischt, so dass auch Begriffe wie „offenes Interview“ (für ein unstrukturiertes Interview) oder „halbstandardisiertes Interview“ (für ein halbstrukturiertes Interview) zu finden sind. Das **unstrukturierte Interview** („unstructured interview“) basiert auf **keinem vorher entwickelten Interviewinstrument**. Stattdessen äußern sich die Befragungspersonen nach einem Intervieweinstieg völlig frei (narratives Interview; Methode des lauten Denkens) oder es werden spontan offene Fragen gestellt, die sich aus der Situation im Feld ergeben (ethnografisches Feldinterview/Feldgespräch). Unstrukturierte Interviews zum selben Forschungsthema verlaufen sehr unterschiedlich und sind nicht direkt vergleichbar. Da die Interviewenden die Interviewsituation unstrukturiert lassen (indem sie gerade kein Fragenraster vorgeben), haben die Befragten die Möglichkeit, ihrerseits Strukturierungen vorzunehmen, etwa indem sie ausführlich darüber sprechen, was ihnen wichtig ist und aus ihrer Sicht unwichtige Punkte auslassen oder abkürzen. Nicht-strukturierte Interviews (z. B. narrative Interviews) werden per Audioaufzeichnung vollständig dokumentiert und anschließend für die Auswertung verschriftet (transkribiert). Zuweilen verzichtet man auf eine Audioaufnahme und beschränkt sich auf Notizen der wichtigsten Punkte (z. B. bei ethnografischen Feldgesprächen), um das Feldgeschehen nicht zu stören.
   Das **halbstrukturierte Interview** („semi-structured interview“) basiert auf einem **Interview-Leitfaden**. Der Interview-Leitfaden besteht aus einem Katalog offener Fragen, zu denen sich die Befragungspersonen in eigenen Worten äußern sollen. Der Leitfaden gibt die Fragen und ihre Reihenfolge vor, erlaubt aber den Interviewenden je nach Interviewsituation in gewissen Grenzen individuelle Anpassungen. So können Fragen vorgezogen, übersprungen oder vertieft werden, um den Interviewfluss am Laufen zu halten. Häufig ist von „teilstandardisierten“ Interviewleitfäden die Rede, damit ist gemeint, dass die Fragen oder Themenblöcke (grob) vorgegeben sind, jedoch keine Antwortalternativen angeboten werden, sondern sich die Befragten in eigenen Worten äußern. Das halbstrukturierte Interview wird typischerweise per Audioaufnahme vollständig dokumentiert und anschließend wortwörtlich verschriftet.
   Das **(voll)strukturierte Interview** („structured interview“; auch: „survey interview“ oder „interviewer-administered survey“) basiert auf einem

Tabelle 10.3 Drei Interviewformen nach dem Grad ihrer Strukturierung

Interviewform (Grad der Strukturierung)	Interviewinstrument (Grad der Standardisierung)	Interviewfragen (Offenheit/Geschlossenheit)
Unstrukturiertes Interview = nicht-strukturiertes Interview	Kein Instrument	Offene Fragen: *Erinnern Sie sich an den Tag, als Sie die Diagnose bekommen haben?* *Wie ist das damals gewesen, und wie sind die folgenden Tage verlaufen?*
Halbstrukturiertes Interview = teilstrukturiertes Interview	Halbstandardisiertes = teilstandardisiertes Instrument: Interview-Leitfaden	Offene Fragen: *Welche Symptome hatten Sie?* *Wie haben Ihre Kinder auf die Krankheit reagiert?*
Vollstrukturiertes Interview = strukturiertes Interview	Vollstandardisiertes = standardisiertes Instrument: Interview-Fragebogen	Geschlossene Fragen/Aussagen mit Antwortvorgaben: *Nehmen Sie momentan Medikamente ein?* *ja/nein* *Bewerten Sie Ihren aktuellen Gesundheitszustand auf einer Schulnotenskala!* *1/2/3/4/5/6*

(voll)standardisierten Fragebogeninstrument. Bei einem solchen Interviewfragebogen sind Anzahl, Abfolge und Wortlaut der geschlossenen Fragen sowie der Antwortmöglichkeiten exakt vorgegeben und für die Interviewenden verbindlich. Im Zuge des vollstrukturierten Interviews werden die Fragen und Antwortalternativen gemäß Fragebogen der Reihe nach vorgelesen. Die von der Befragungsperson jeweils gewählten Antwortalternativen werden vom Interviewer im Instrument markiert. Es kann sich um einen Paper-Pencil-Interview-Fragebogen („paper and pencil interview", „paper assisted personal interview", PAPI) oder um einen elektronischen Fragebogen handeln, der vom Interviewer am Computer ausgefüllt wird, sei es bei einem mündlichen Interview („computer-assisted personal interview", CAPI) oder einem telefonischen Interview („computer-assisted telephone interview", CATI).

Unstrukturierte und halbstrukturierte Interviews werden den qualitativen Befragungsmethoden zugerechnet (**qualitatives Interview**) und laufen auf eine interpretative Auswertung der Transkripte hinaus (zur qualitativen Datenanalyse ▶ Abschn. 12.1). Vollstrukturierte Interviews stellen eine quantitative Datenerhebungstechnik dar (**quantitatives Interview**) und münden in statistische Analysen der von den Befragungspersonen ausgewählten standardisierten Antwortvorgaben.

2. **Anzahl der gleichzeitig interviewten Befragungspersonen** mit den zwei Formen Einzelbefragung versus Paar- oder Gruppenbefragung. Typischerweise erfolgen qualitative wie quantitative Interviews als dyadische Kommunikation zwischen je einer Interviewerin bzw. einem Interviewer und einer Befragungsperson (**Einzelinterview**, „individual interview"). Im Mittelpunkt steht die Gewinnung von Informationen über Aspekte des Verhaltens und Erlebens der konkreten Befragungsperson. Insbesondere in der qualitativen Forschung werden auch **Paar- und Gruppenbefragungen** durchgeführt. Eine besonders wichtige Variante ist die **Fokusgruppen-Diskussion** („focus group discussion", „focus group interview"). Hier steht ein Diskussionsthema von allgemeinem Interesse im Fokus, zu dem sich alle Mitglieder der Gruppe äußern. Dass die Gruppenmitglieder die Äußerungen der anderen hören und darauf mit Zustimmung oder Ablehnung reagieren, ist bei dieser Form der Gruppenbefragung gewünscht. Ziel der Fokusgruppenmethode ist es u. a., die Variationsbreite und Überzeugungsstärke einzelner Meinungen und Einstellungen zu einem Befragungsthema zu erkunden, die im Einzelinterview verborgen blieben.
3. **Art des Interviewkontaktes** bzw. **Interviewmodus** („interview mode") in den drei Hauptvarianten persönlich versus telefonisch versus online. Mündliche Befragungen werden im Regelfall als **persönliches** bzw. **direktes Interview** im **Face-to-Face-Kontakt** durchgeführt, entweder in der häuslichen oder beruflichen Umgebung der Befragungspersonen bzw. vor Ort im Feld (Feldinterview) oder im Forschungsinstitut (Teststudio-Interview). Der Befragungsort beim Face-to-Face-Interview beeinflusst u. a. die Kosten der Studie. Insbesondere für qualitative Interviews wird der persönliche Kontakt als wichtig für eine vertrauensvolle Befragungssituation und die Erfassung von Hintergrundinformationen erachtet. Es mehren sich jedoch die wissenschaftlichen Interviews, die **fernmündlich** als **Telefoninterviews** durchgeführt werden, was z. B. Zeit und Kosten für die Anreise reduziert. Angesichts der Veralltäglichung der Online-

Kommunikation haben sich mittlerweile computervermittelte bzw. **Online-Interviews** etabliert, die kostengünstig mit international verstreuten Befragungspersonen möglich sind. Textbasierte computervermittelte Interviews (z. B. per Online-Chat, Messenger, E-Mail) basieren auf sog. **schriftlicher Mündlichkeit** (Oraliteralität), denn die Ausdrucksweise im Live-Chat ähnelt mündlichen Äußerungen stärker als z. B. dem Aufsatzschreiben.

4. **Anzahl der Interviewenden** mit den Varianten ein Interviewer (herkömmliches Interview), zwei Interviewer (Tandem-Interview, Kincaid & Bright, 1957) und mehr als zwei Interviewer („Hearing" bzw. Board-Interview, Oldfield, 1951, S. 117). Wenn mehrere Interviewende eine Person befragen, können sie sich beim Fragen abwechseln und die Interviewsituation gemeinsam besser überblicken. Tandem- und Board-Interviews werden in der Praxis jedoch selten eingesetzt, da dieses Vorgehen aufwändig ist und auf die Befragungspersonen einschüchternd wirken kann.
5. **Art der Befragungspersonen** mit den Hauptvarianten Experten-Interview versus Laien- bzw. Betroffenen-Interview. Üblicherweise werden Befragungspersonen als Betroffene befragt, um deren persönliche Erfahrungen und Sichtweisen es geht. Zuweilen werden **Experteninterviews** durchgeführt, in denen Fachexpertise ausgewiesener Spezialisten erhoben wird, meist im Rahmen von Leitfadeninterviews.
6. **Art der Interviewtechnik** mit sehr vielen Varianten insbesondere im Bereich der qualitativen Einzelbefragung (Mey & Mruck, 2007, 2010). In Abhängigkeit von den theoretischen Vorannahmen zum Interviewprozess und den inhaltlichen Zielen der Befragung lassen sich unterschiedliche Varianten von unstrukturierten Interviews (z. B. narratives Interview, ethnografisches Interview) und halbstrukturierten Interviews (z. B. problemzentriertes Interview, fokussiertes Interview) unterscheiden. Je nach Interviewvariante und Art der Fragen werden dabei von den Befragungspersonen unterschiedliche Textsorten generiert (z. B. Erzählungen, Beschreibungen, Bewertungen, Erklärungen etc.). Nicht alle Ausprägungen der genannten Kriterien sind sinnvoll miteinander kombinierbar. Stattdessen haben sich bestimmte Konstellationen etabliert (Tab. 10.4).

Interviewfehler und Gegenmaßnahmen

Wie ein perfekt fehlerfreies Interview aussehen sollte, das maximal aussagekräftige qualitative oder quantitative Daten hervorbringt, lässt sich trotz intensiver Methodenforschung zur Interviewtechnik nicht pauschal beantworten. Wichtige Quellen von Interviewfehlern und entsprechende Gegenmaßnahmen lassen sich unterscheiden hinsichtlich a) der Interviewenden, b) der Befragungspersonen, c) des Interviewinstrumentes und Interviewablaufs sowie d) der Interviewdokumentation und Interviewauswertung.

Fehler durch die Interviewenden

Es ist unstrittig, dass die Person, die ein Interview durchführt, das Ergebnis entscheidend beeinflussen kann („interviewer-related error"), etwa durch Suggestivfragen, ein die Befragungsperson einschüchterndes Auftreten oder das Überspringen von Fragen. Durch bewusste Auswahl (1) und Schulung (2) sollen Interviewerinnen und Interviewer auf ihre Aufgabe vorbereitet werden. Zudem wird das Verhalten der Interviewenden teilweise einer Kontrolle (3) unterzogen.

1. **Auswahl von Interviewenden**. Wer ist als „gute Interviewerin" bzw. „guter Interviewer" für eine mündliche Befragungsstudie einzusetzen? Oft werden hierfür in der akademischen Forschung studentische Hilfskräfte engagiert oder wissenschaftliche Mitarbeiter/innen eingesetzt, während die Markt- und Meinungsforschung mit (semi-)professionellen Interviewer/innen operiert. Generell sollten Interviewende eine **hohe Kommunikations- und Sozialkompetenz** mitbringen sowie die Fähigkeit zur Selbstreflexion. Diese wird benötigt, um mit verschiedensten Personen eine angenehme Gesprächsatmosphäre herstellen und auch bei Komplikationen aufrecht erhalten zu können, ohne dabei die Regeln der Wissenschaftlichkeit aus dem Blick zu verlieren. Besonderes Augenmerk ist dabei auch auf die verwendete Sprache bzw. Wortwahl zu legen, die für die Zielgruppe verständlich sein muss.
   Hinsichtlich **soziodemografischer Merkmale der Interviewenden** („interviewer characteristics") scheint oft eine möglichst geringe Distanz zu den Befragungspersonen empfehlenswert. So werden z. B. bei sexualwissenschaftlichen Befragungen gleichgeschlechtliche Interviewende eingesetzt („matching of interviewer and respondent"). Die Wechselwirkungen zwischen Befragungsperson, Befragungsthema und Merkmalen der Interviewenden können jedoch unvorhergesehen sein. So mag die angenommene Präferenz, intime gesundheits- oder sexualbezogene Fragen lieber gleichgeschlechtlichen Interviewenden zu offenbaren, nicht generell gelten, sondern in Abhängigkeit von Geschlecht, Alter und sexueller Orientierung der Befragungspersonen unterschiedlich ausfallen. Sofern es im Rahmen größerer Studien möglich ist, sollte man Befragungspersonen bei heiklen Themen daher selbst entscheiden lassen, ob sie z. B. lieber von einem Mann oder einer Frau interviewt werden möchten. Insgesamt ist festzuhalten, dass nur punktuelle und teilweise widersprüchliche

Tabelle 10.4 Verbreitete Varianten des Forschungsinterviews und ihre Klassifikation

Mündliche Befragung							
	Qualitative Befragung					**Quantitative Befragung**	
Grad der Strukturierung	**Unstrukturiertes Interview**		**Halbstrukturiertes Interview**			**Vollstrukturiertes Interview**	
Anzahl der gleichzeitig interviewten Befragungspersonen	Einzelinterview	Einzel- und/oder Gruppeninterview	Einzelinterview	Einzel- und/oder Gruppeninterview	Gruppeninterview	Einzelinterview	Gruppeninterview
Interviewvarianten	– Narratives Interview – Methode des lauten Denkens	– Ethnografisches Feldinterview	– Leitfaden-Interview – Telefonisches Leitfaden-Interview – Online-Leitfaden-Interview – Experten-Interview – Problemzentriertes Interview	– Fokussiertes Interview	– Unstrukturiertes bzw. halbstrukturiertes Paar- oder Gruppeninterview – Fokusgruppen-Diskussion	– Vollstrukturiertes persönliches Interview – Vollstrukturiertes Telefon-Interview – Vollstrukturiertes Online-Interview	–

Tabelle 10.5 Umgang mit schwierigen Befragungspersonen. (Mod. nach Saunders, Lewis, & Thornhill, 2009, S. 340)

Schwieriges Verhalten der Befragungsperson	Reaktionsmöglichkeiten der Interviewenden
Befragungsperson ist wortkarg, gibt einsilbige Ja/Nein-Antworten	Zunächst sind Zeitdruck und mangelnde Anonymität (Anwesenheit Dritter) als Ursachen für geringe Auskunftsbereitschaft auszuschließen. Dann sollten die Fragen so offen wie möglich gestellt und Pausen ausgehalten werden, um zu signalisieren, dass man mehr hören möchte.
Befragungsperson ist redselig, schweift wiederholt vom Thema ab	Zunächst kann man die Befragungsperson bitten, zu pausieren, um sich zum bisher Gesagten Notizen zu machen (dies wirkt weniger konfrontativ als direktes Unterbrechen). Anschließend kann man wieder auf themenrelevante Aspekte zu sprechen kommen.
Befragungsperson fängt an, ihrerseits den Interviewer zu befragen	Man sollte sich für das Interesse bedanken und die Fragen auf das Nachgespräch verschieben. Schließlich wolle man für die wissenschaftliche Studie zunächst die Sichtweise der Befragungsperson erfahren.
Befragungsperson präsentiert sich als Methodenexperte und kritisiert die Interviewtechnik oder die Zielsetzung der Studie	Man sollte die methodischen Prinzipien des eigenen Vorgehens (z. B. Art der Stichprobenauswahl, Interviewtechnik) benennen und die Rückmeldungen der Befragungsperson dankend als Anregung notieren.
Befragungsperson zeigt emotionale Belastung, beginnt z. B. zu weinen	Zunächst sollte man signalisieren, dass emotionale Reaktionen in Ordnung sind und der Befragungsperson ausreichend Zeit lassen, sich wieder zu beruhigen. Im Zweifelsfall kann angeboten werden, die entsprechende Frage zu überspringen. Im Nachgespräch sollte man sich rückversichern, dass es der Befragungsperson wieder gut geht und ggf. über Beratungs- und Unterstützungsmöglichkeiten informieren.

Befunde dazu vorliegen, unter welchen Bedingungen und in welcher Weise Geschlecht, Alter, Ethnizität oder sozialer Status der Interviewenden zu Fehlern und Verzerrungen in den erhaltenen Antworten führen (Davis, Couper, Janz, Caldwell, & Resnicow, 2009).

Generell sollten Interviewende ein der Interviewsituation angemessenes **Erscheinungsbild** abgeben und glaubhaft die jeweilige Forschungseinrichtung vertreten können. Vorsicht ist geboten, wenn **äußere Merkmale** der Interviewenden tatsächlich oder vermeintlich auf ihre Einstellungen zu den erfragten Themen hindeuten. Führt etwa eine Rollstuhlfahrerin Interviews über „Behindertenfeindlichkeit" durch, ist mit stärkeren Interviewereffekten im Sinne sozial erwünschten Antwortens zu rechnen, als wenn dieselbe Interviewerin Fragen zum „Umweltbewusstsein" stellt. In der qualitativen Forschung wird der Einfluss von Auftreten, Ausstrahlung und Erscheinungsbild der Interviewenden auch als deren spezifischer „Reizwert" bezeichnet (Mey & Mruck, 2007, S. 273).

Um die unbewusste Beeinflussung von Versuchspersonen durch Versuchsleitende im Experiment zu reduzieren, werden teilweise „blinde" Versuchsleiter eingesetzt, die über die Forschungshypothesen nicht informiert sind (zur Verblindung ▶ Abschn. 7.6.1). Ein entsprechendes Vorgehen zur Verhinderung von Interviewereffekten ist nicht praktikabel, da die Interviewsituation es erfordert, dass die Interviewenden Sinn und Zweck der Studie genau kennen, etwa um bei einem Leitfaden-Interview den Interview-LeitfadenAblauf des Interviews und die Formulierung der Fragen optimal der Situation anpassen zu können. Wichtig ist daher eine gründliche Interviewerschulung, die Verzerrungen entgegenwirkt.

2. **Schulung von Interviewenden** („interviewer training"). Im Zuge des Interviewertrainings werden die ausgewählten Interviewer/innen zunächst mit Hintergründen und Zielsetzungen der Studie vertraut gemacht. Anschließend lernen sie die konkrete Interviewtechnik und das Interviewinstrument kennen. Im nächsten Schritt führen sie in Rollenspielen Probe-Interviews durch. Hierbei kann es hilfreich sein, in das Interviewgeschehen typische Probleme einzubauen (die Befragungsperson redet zu viel, schweigt zu lange, verweigert Antworten etc.; Tab. 10.5), um den Umgang mit schwierigen Befragungssituationen zu üben. Die Probeinterviews sollten mit einer Videokamera aufgezeichnet werden, so dass die Möglichkeit besteht, Fehler und Schwächen in der Interviewführung im Nachhinein aufzuarbeiten. Im Zuge der Probe-Interviews sollte auch die Dokumentation der Antworten (z. B. Einsatz des Audioaufzeichnungsgerätes, Interviewnotizen, Ausfüllen des standardisierten Interviewfragebogens) geübt werden. Bei längeren Studien und großer Arbeitslast der einzelnen Interviewer kann eine **Zwischenschulung** sinnvoll sein, damit das

Gelernte aufgefrischt und z. B. eine immer schnellere und oberflächlichere Abwicklung vollstrukturierter Interviews vermieden wird (Olson & Peytchev, 2007).

3. **Kontrolle von Interviewenden.** Markt- und Meinungsforschungsinstitute, die eine große Zahl von Interviewern beschäftigen, kontrollieren bei quantitativen Umfragen meist stichprobenartig, ob die Interviewenden tatsächlich die angegebenen Haushalte aufgesucht haben. Denn hier tritt teilweise das Problem der **Interviewfälschung** in dem Sinne auf, dass die Interviewer selbst die Fragebögen ausfüllen, um Zeit und Arbeit zu sparen. Bei telefonischen Interviewstudien besteht für die Studienleitung bei Verwendung eines Telefonlabors die Möglichkeit, sich in laufende Interviews hineinzuschalten und den korrekten Interviewverlauf zu prüfen. Ebenso ist bei mündlichen Interviews eine Zwischenkontrolle und Rückmeldung anhand der Audioaufzeichnungen oder der Transkripte der Interviews möglich, etwa wenn man feststellt, dass ein Interviewer zu viele oder zu wenige Rückfragen stellt oder bestimmte Antworten suggeriert. Je umfangreicher und zeitlich ausgedehnter eine Studie ist und je mehr verschiedene Interviewende zum Einsatz kommen, desto wichtiger werden Zwischenkontrollen zur Qualitätssicherung (dies ist auch in der ISO-Norm 20252 zur Markt-, Meinungs- und Sozialforschung festgelegt).

Fehler durch die Befragungspersonen

Die methodisch fundierte Stichprobenkonstruktion ist ein wesentlicher Garant der Wissenschaftlichkeit einer Interviewstudie. Selbst wenn ein Stichprobenplan vorliegt, können jedoch eine Reihe von Problemen zu verzerrten Ergebnissen führen, die besonders bei jenen quantitativen Befragungen, die auf Zufallsstichproben basieren, eine große Rolle spielen: Dazu zählen die mangelnde Erreichbarkeit der Befragungspersonen (1), die Interviewverweigerung (2), die Ablehnung einzelner Fragen (3) und diverse, auf sozialen und kognitiven Ursachen basierende Antwortverfälschungen (4). Entsprechende Effekte sind teils auf Merkmale der Befragungspersonen selbst zurückzuführen, teils auf Wechselwirkungen mit der Person des Interviewers, mit dem Instrument oder mit der Interviewdurchführung. Sie werden im Abschnitt zu quantitativen Interviews (▶ Abschn. 10.2.3) sowie zu quantitativen Fragebögen (▶ Abschn. 10.3.3) detailliert behandelt.

Bei qualitativen Studien rekrutiert man meist nur **auskunftswillige Personen**, so dass Probleme der Erreichbarkeit und Teilnahmeverweigerung eine geringe Rolle spielen und auch nicht in Prozentangaben quantifizierbar sind. Die Ablehnung einzelner Fragen sowie verfälschende Antworten durch die Befragungspersonen sollen bei qualitativen Interviews durch eine individuell zugeschnittene, flexible und entspannte Interviewsituation vermieden werden. Agieren Befragungspersonen trotzdem wortkarg, ausweichend oder widersprüchlich, so sollten entsprechende Merkmale des Antwortverhaltens in den Interviewnotizen vermerkt und bei der Auswertung berücksichtigt werden. Da es im qualitativen Ansatz oft genau darauf ankommt, die subjektiven Sichtweisen der Befragten zu rekonstruieren, sollte eine vorschnelle Einteilung der Antworten in „wahr“ oder „falsch“ unterbleiben. Berichtet etwa ein Jugendlicher in offensichtlich übertriebener Weise über sportliche Bestleistungen und Ehrungen, so ist er nicht als „Aufschneider“ zu entlarven, sondern es gilt zu ergründen, welche Bedeutung sein „Lebensmärchen“ für ihn hat (Reinders, 2005, S. 28f.).

Generell sind Interviewtechniken auf die verbalen Kommunikationsmöglichkeiten der jeweiligen Zielgruppen abzustimmen. Bei **qualitativen Interviews mit Kindern** ist es empfehlenswert, die Situation kindgerecht zu gestalten (Trautmann, 2009), etwa durch Einbeziehung von Spielzeugfiguren, mit denen Kinder Szenen darstellen (Charlton & Neumann, 1990) oder unter Verwendung einer Handpuppe als Interviewer (Paus-Haase, 2000). Auch bei **qualitativen Interviews mit Jugendlichen** sind Besonderheiten dieser Altersgruppe zu beachten, um aussagekräftige Antworten in Interviews zu erzielen (Reinders, 2005; Walther, Preckel, & Mecklenbräuker, 2010). Ältere Befragungspersonen müssen ihrerseits adäquat angesprochen werden. Bei **Interviews mit tauben oder hörbeeinträchtigten Personen** sind wiederum besondere Vorkehrungen zu treffen (Balch & Mertens, 1999).

Sowohl bei quantitativen als auch bei qualitativen Interviews wird man Fälle, bei denen trotz sorgfältiger Interviewdurchführung berechtigte Zweifel an der Glaubwürdigkeit und Interpretierbarkeit der Antworten bestehen bleiben und/oder von denen nur sehr unvollständige Antworten vorliegen, mit entsprechender Begründung von der Auswertung ausschließen. Derartige Maßnahmen der Datenaufbereitung (▶ Kap. 11) müssen sorgfältig dokumentiert werden, damit nicht der Eindruck einer unethischen Manipulation des Datenmaterials in Richtung auf Wunschegebnisse entsteht (▶ Abschn. 4.2.2)

Fehler durch Interviewinstrument und Interviewdurchführung

Bei qualitativen Interviews wird entweder ohne Instrument (unstrukturiertes Interview) oder mit einem Leitfaden (halbstrukturiertes Interview) gearbeitet. Bei quantitativen Interviews wird ein vollstandardisierter Fragebogen mit geschlossenen Fragen und Antwortalternativen zugrunde gelegt (zur Fragebogenkonstruktion ▶ Abschn. 10.3.3). Die theoriebasierte Konstruktion des jeweiligen Instruments und ein gründlicher Pretest tragen dazu bei,

Fehlerquellen im Instrument zu korrigieren (z. B. unverständliche Fragen, zu hoher oder zu geringer Detaillierungsgrad von Fragen, zu große oder zu geringe Menge an Fragen). Ein gutes Instrument sollte so ausgelegt werden, dass die Befragungspersonen alle Fragen auf Anhieb ohne große Belastung beantworten können und Antwortverfälschungen oder Interviewabbrüche vermieden werden.

Das Instrument ist das zentrale Hilfsmittel der Interviewenden zur Steuerung der Interviewsituation. Hier spielt die Darstellungsform eine entscheidende Rolle: Ist der ausgedruckte Leitfaden übersichtlich formatiert und in großer Schrift gestaltet, so dass die Interviewenden sich jederzeit gut orientieren und überblicken können, welche Fragen bereits gestellt wurden und welche als nächstes anstehen? Standardisierte quantitative Interviews finden heute überwiegend computergestützt statt, d. h., die Interviewenden sehen den Fragenverlauf inklusive Filterführung auf dem Computermonitor und geben die Antworten elektronisch ein. Auf diese Weise schließt man Fehler im Interviewablauf aus (z. B. der Interviewer überspringt Fragen, stellt Fragen doppelt oder vergisst eine Filterführung).

Interviewfehler kommen auch durch **ungünstige Rahmenbedingungen** zustande. Etwa wenn durch die Wahl des Termins (z. B. in der Mittagspause) Zeitdruck entsteht und die Antwortbereitschaft sinkt. Oder wenn durch die Wahl des Ortes (z. B. im heimischen Wohnzimmer, wo andere Familienmitglieder mithören) Sorgen um die Privatsphäre entstehen. Hier sollten nach Möglichkeit die Bedürfnisse der Befragungspersonen berücksichtigt werden. Ansonsten entscheiden in der Praxis forschungsökonomische Erwägungen über die Wahl von Orten und Zeiten für Interviews.

Schließlich ist insbesondere bei qualitativen Interviews das Verhältnis zwischen Interviewer/in und Befragungsperson, das **Arbeitsbündnis** („interview rapport“) von entscheidender Bedeutung für reibungslose und aussagekräftige Interviews. Angestrebt ist, dass die Beteiligten sich beide auf das Interview und die wechselseitige Verständigung einlassen, so dass eine vertrauensvolle (allerdings nicht zu vertrauliche: „overrapport“) Atmosphäre entsteht (▶ Abschn. 2.3.4). Durch sorgfältige Auswahl und Schulung der Interviewenden soll dies begünstigt werden, ebenso durch Wahlmöglichkeiten der Befragten hinsichtlich für die Qualität der Interaktion möglicherweise wichtiger Interviewereigenschaften (z. B. kultureller Hintergrund, Geschlecht, Alter). Wie genau ein optimales Arbeitsbündnis im Interview beschaffen sein muss, ist theoretisch und empirisch schwer zu fassen (Goudy & Potter, 1975). Wenn es bei manchen Interviews nicht oder nur bedingt gelingt, eine fruchtbare Kommunikationssituation zu schaffen, so können entsprechende Notizen der Interviewenden zur Atmosphäre bei der Interviewsituation bei der Interpretation der Daten berücksichtigt werden bis hin zum Ausschluss von Befragten, deren Äußerungen aus triftigen Gründen für unglaubwürdig gehalten werden.



Abb. 10.6 Während des Interviews sind die Antworten der Befragungsperson in geeigneter Weise festzuhalten. (Goldmanns großer Cartoonband 1989. Schweine mit Igeln. München: Goldmann, S. 190. Mit freundlicher Genehmigung von Renate Alf)

Fehler bei der Dokumentation und Auswertung von Interviews

Der gravierendste Fehler dieser Rubrik betrifft die **unvollständige Dokumentation** der Interview-Antworten, die in Form von Gesprächsnotizen, Audio- bzw. Videoaufzeichnungen (qualitatives Interview) oder mittels papier- oder computergestützter Erfassung der standardisierten Antworten (quantitatives Interview) erfolgt (Abb. 10.6). Dieser Fehler geht oft auf technische Pannen (fehlerhafte Bedienung des Aufzeichnungsgerätes, nicht genügend Batterien etc.) zurück und kann verhindert werden, indem die Dokumentation im Rahmen der Interviewerschulung geübt wird. Weitere Fehler können sich bei der qualitativen und quantitativen Interviewauswertung ergeben. Diese sind nicht interviewspezifisch, sondern können analog auch bei der Analyse von Beobachtungs- oder Fragebogendaten auftreten. Sie

betreffen die generelle Orientierung der Datenanalyse an wissenschaftlichen Auswertungsmethoden (▶ Kap. 12) und Gütekriterien (▶ Kap. 3).

10.2.2 Qualitatives Interview

Die Besonderheit qualitativer Interviewtechniken liegt darin, dass der Interviewverlauf weniger von den Interviewenden strukturiert und dafür stärker **von den Interviewten gesteuert** wird. Denn bei einem qualitativen Interview werden **offene Fragen** gestellt, auf die die Befragungspersonen umfassend **in eigenen Worten antworten**. Darüber hinaus wird der Interviewablauf flexibel gestaltet, so dass auch neue Fragen spontan aufgeworfen und weiterverfolgt werden können, wenn sich das aus der Interviewsituation ergibt. Da im Rahmen des qualitativen Interviews individuell vertiefend auf angesprochene Sachverhalte eingegangen werden kann, wird es in der Literatur auch oft als **Tiefeninterview** („in-depth interview") bezeichnet. Im engeren Sinne spricht man von Tiefeninterviews, wenn sich Interviewführung und -auswertung an psychoanalytischen bzw. tiefenpsychologischen Theorien und Konzepten orientieren und z. B. Übertragungsphänomene zwischen Interviewenden und Interviewten thematisiert werden (Bock, 1992).

Während bei strukturierten mündlichen Befragungen die Person des Interviewers in den Hintergrund tritt, da jeweils ein standardisiertes Interview-Instrument (Fragebogen mit geschlossenen Fragen und Antwortvorgaben) abgearbeitet wird, fungieren die Interviewenden in qualitativen Befragungen selbst als **Erhebungsinstrumente**: Sie gehen auf die Befragungsperson ein und stellen spontan Rückfragen. Ihre eigenen Gedanken, Gefühle und Reaktionen auf die Befragten werden notiert und können in die Analysen einfließen. So könnte eine Interviewerin eigene Gefühle der Langeweile während des Gesprächs hypothetisch als Indiz dafür werten, dass der Interviewpartner keine authentische Erlebnisschilderung abgibt, sondern nur eine vorher zurechtgelegte Geschichte präsentiert. Solche Deutungen müssen jedoch ihrerseits gut begründet und intersubjektiv nachvollziehbar sein.

Qualitative Interviews – Qualitative Interviews („qualitative interviews") arbeiten mit offenen Fragen, so dass sich die Befragten mündlich in eigenen Worten äußern können. Zudem wird der Gesprächsverlauf weniger von den Interviewenden und ihren Fragen vorstrukturiert, sondern stärker von den Befragten mitgestaltet. Auf diese Weise sollen die individuellen Sichtweisen der Befragten nicht nur oberflächlich, sondern detailliert und vertieft erschlossen werden. Qualitative Interviews existieren in vielen verschiedenen methodischen Varianten.

Im Folgenden wird zunächst der Ablauf eines qualitativen Interviews dargestellt. Anschließend werden drei Arten von qualitativen Interviews vorgestellt: 1.) das nichtstrukturierte Interview, 2.) das halbstrukturierte Interview sowie 3.) das qualitative Gruppeninterview (für eine Klassifikation unterschiedlicher qualitativer Interviewvarianten s. Mey & Mruck, 2007).

Ablauf eines qualitativen Interviews

Bevor wir auf einzelne Interviewtechniken eingehen, wollen wir zunächst den Gesamtablauf einer qualitativen Befragungsstudie schildern, der sich in **zehn Arbeitsschritte** einteilen lässt.

1. **Inhaltliche Vorbereitung.** Zur inhaltlichen Planung einer qualitativen Interviewstudie zählen die Festlegung des Befragungsthemas und der Forschungsfragen, Überlegungen zur Auswahl der Befragungspersonen und der Interviewenden, Wahl der geeigneten qualitativen Befragungstechnik sowie die Zusammenstellung und Erprobung von Interviewfragen. Nach Abschluss der inhaltlichen Vorbereitungen sollte klar sein, zu welchem Thema wer wie zu interviewen ist.
2. **Organisatorische Vorbereitung.** Alle Interviewerinnen und Interviewer sollten im Vorfeld geschult werden und Interviews in Form von Rollenspielen geprobt haben. Zudem sollte Kontakt zu den Befragungspersonen hergestellt werden. Sofern diese nicht ohnehin im Feld anzutreffen sind (ethnografisches Feldinterview), müssen sie gemäß einem begründeten Stichprobenplan rekrutiert werden. Hierfür kann man sich im Bekannten- und Freundeskreis umhören, Aushänge an einschlägigen Orten platzieren, Zeitungsanzeigen schalten etc. Oftmals wird ein sog. Rekrutierungsfragebogen (Screener) eingesetzt, um die vorab definierten Rekrutierungskriterien zu erheben (z. B. soziodemografische Variablen, Krankheiten etc.). Nach dem Erstkontakt und der Terminvereinbarung ist mit Ausfällen und Terminverschiebungen zu rechnen, so dass man entweder von Anfang an großzügig potenzielle Befragungspersonen rekrutiert oder bei Bedarf nachrekrutiert bis die gewünschte Stichprobengröße und Stichprobenzusammensetzung erreicht sind bzw. im Laufe der Interviews eine inhaltliche Sättigung eingetreten ist (zu qualitativen Stichprobenarten wie z. B. dem „theoretical sampling" ▶ Abschn. 9.2.1). Das sorgfältige Zusammenstellen des Interviewmaterials gehört ebenfalls zur organisatorischen Vorbereitung (Audiorekorder, Speichermedien, Ersatzbatterien, Interview-Leitfaden, Visitenkarte, Prospekt oder Informationsmaterial über das Forschungsprojekt etc.). Die organisatorische Vorbereitung ist beendet, wenn die Interviewenden startklar sind und wissen, wann, wo und wie die Interviews durchzuführen sind.
3. **Gesprächsbeginn.** Sind Interviewerin bzw. Interviewer und Befragungsperson am verabredeten Ort

(meist in der Wohnung oder am Arbeitsplatz der Befragten) zusammengetroffen, sollte durch gegenseitiges Vorstellen und ein wenig Smalltalk eine entspannte Atmosphäre hergestellt werden (Warming-up), bevor das Untersuchungsanliegen dargestellt und damit das Interview eingeleitet wird (Interviewereröffnung). Bei einem persönlichen Interview sollte das Sitzarrangement eine **angenehme Nähe-Distanz-Regulation** zulassen (z. B. einander im Abstand von ca. einem Meter schräg gegenüberstehende Stühle). Da bei der Auswertung von qualitativen Interviews eigentlich nur selten auf eine Tonaufzeichnung verzichtet werden kann, sind Akzeptanzprobleme auf Seiten der Befragten möglichst im Vorfeld abzubauen. Dabei geht es um psychologische Barrieren beim Sprechen vor einem Mikrofon, aber auch um Datenschutzbedenken. Eine schriftliche Vereinbarung über die Einhaltung genau umschriebener Maßnahmen zum Datenschutz erhöht die Sicherheit der Befragten und gehört zur Selbstverpflichtung der Forschenden im Sinne der Forschungsethik (s. ▶ Abschn. 4.1).

Ein Nachteil der expliziten Auseinandersetzung mit den **Modalitäten von Audio-bzw. Videoaufzeichnungen** und Datenschutzproblemen besteht darin, dass dies möglicherweise bei manchen Auskunftspersonen erst Bedenken erzeugt, die vorher gar nicht bestanden. Aus pragmatischer Sicht wird deswegen manchmal empfohlen, das Aufzeichnungsgerät einfach ganz selbstverständlich auf dem Tisch aufzubauen und zu den inhaltlichen Fragen überzugehen. Vor Beginn des Interviews sollten die Funktionsfähigkeit des Audio- oder Videorekorders und die Ton- bzw. Bildqualität geprüft werden.

4. **Durchführung und Aufzeichnung des Interviews.** Die Hauptaufgabe des Interviewers ist die Steuerung des Gesprächsablaufs, d. h., eigene Reaktionen und auch das nonverbale Verhalten der Befragten sollten aufmerksam verfolgt werden. Zudem sind Interviewende bei qualitativen Interviews meist gefordert, aus dem Gespräch heraus weiterführende und vertiefende Fragen zu formulieren und dafür zu sorgen, dass die Befragungsperson beim Thema bleibt. Die der Befragungsperson wie dem Interviewer bei offenen Befragungen zugestandenen Gestaltungsspielräume bergen besondere Risiken und Probleme: Beim freien Generieren von Fragen während des Gesprächsverlaufs sollten Interviewende Vorsicht walten lassen, um die Befragungspersonen nicht versehentlich in eine Ecke zu drängen oder durch Emotions- und Meinungsäußerungen zu verunsichern. Heikle Fragen können im Zweifelsfall für den letzten Gesprächsteil aufgehoben werden.

   Freies Erzählen, wie es von den Befragungspersonen bei vielen offenen Interviews gefordert wird, liegt nicht jeder und jedem. Interviewende sollten sich darauf gefasst machen, ebenso mit wortkargen Interviewpartnern konfrontiert zu sein wie mit äußerst redseligen. Während in strukturierten Befragungen solche persönlichkeitsbedingten Unterschiede durch das reglementierende Fragenkorsett und durch standardisierte Antwortvorgaben nivelliert werden, liegt es bei offenen Befragungen an den Interviewenden, die richtige Balance zwischen Eingreifen (**direktiverer Stil** zur Förderung einer halbstrukturierten Interviewsituation) und Laufenlassen (**nondirektiver Stil** zur Förderung der Authentizität) zu finden und eine angemessene Interviewdauer einzuhalten (◘ Tab. 10.5).

5. **Gesprächsende.** Dem offiziellen Ende des qualitativen Interviews schließt sich in der Regel eine Phase des informellen Gesprächs an, die äußerlich der Begrüßungsphase ähneln mag, inhaltlich aber meist nicht viel mit Smalltalk zu tun hat. Interviewende sollten jetzt trotz eigener Erschöpfung besonders aufmerksam sein, da Befragungspersonen oftmals gerade nach Abschluss des Interviews wichtige Informationen nachliefern und die Gesprächssituation kommentieren. Es wird empfohlen, das Aufzeichnungsgerät nicht zu frühzeitig auszuschalten, sondern auch das Nachgespräch aufzuzeichnen oder – falls das Gerät schon abgeschaltet wurde – Inhalte des Nachgesprächs zumindest zu notieren. Wenn substanzielle Informationen erst im Nachgespräch geliefert werden, kann dies zuweilen aber auch auf Schwächen der Interviewgestaltung hindeuten (Mey & Mruck, 2007, S. 268). Zudem kann es ethisch problematisch sein, Äußerungen zu verwenden, die Befragte aus ihrer Sicht „off the records", d. h. nach dem offiziellen Ende des Interviews gegeben haben. Ergänzende Fragebögen zu soziodemografischen Merkmalen oder psychometrische Tests können sich an das Interview anschließen.
6. **Verabschiedung.** Bei der Verabschiedung sollten nach Möglichkeit eine Visitenkarte und/oder Informationsmaterialien über das Forschungsprojekt hinterlassen werden. Bei Interesse kann eine kurze Ergebnismitteilung an die Untersuchungsteilnehmenden angekündigt werden. Insbesondere bei Befragungen zu belastenden und bedrohlichen Themen ist ein Angebot zur Nachbetreuung bereitzustellen (z. B. Telefonnummer, Beratungsgespräch, zweites Interview etc.). Eine qualitative Befragung ist nicht nur eine Datenerhebungsmethode, sondern kann – wie jede andere Form der reaktiven Datenerhebung – ungeplant auch als Intervention wirken, etwa wenn durch die

Gesprächsnotizen (Postskriptum) zu einem Leitfadeninterview

Im Rahmen einer qualitativen Interviewstudie zu HIV wurde für einen Interviewpartner folgendes Postskriptum angefertigt (Stassen & Seefeldt, 1991, S. 198):

Alex

(Die Angaben zu Ort, Zeit, Dauer des Interviews etc. werden hier nicht wiedergegeben.)

„Nach den obligatorischen Erläuterungen zum Datenschutz und einführenden Erklärungen zur Thematik erzählt Alex seine Vorgeschichte kurz und knapp. Hin und wieder wird sein Erzählstil durch Satzabbrüche begleitet, wodurch der Sinn des Gemeinten nicht immer klar zu erfassen ist. Alex stellt klar, dass das Thema Tod kein Thema ist, mit dem er sich befassen möchte. Alex nimmt im Verlauf des Interviews die Rolle des Fachmannes ein und erklärt z. B. Krankheitsverlauf und medizinische Behandlung der HIV-Infektion.

Das Interview mit Alex verläuft kooperativ, wobei häufige Störungen durch die geschäftlichen Tätigkeiten Alex' entstehen. Der Grund für die Teilnahme am Interview ist die Zielsetzung Alex', sich für Aidskranke zu engagieren, zumal er davon ausgeht, dass diese Krankheit auch noch in 20 Jahren aktuell sein wird.

Unklar blieb:

- Auseinandersetzung mit der Homosexualität,
- Familienhintergrund."

im Interview angeregte Reflexion über die eheliche Arbeitsteilung schwelende Konflikte plötzlich stärker zutage treten und Trennungsabsichten virulent werden. Nicht zuletzt sollte den Befragungspersonen für ihre Unterstützung der Studie gedankt werden.

7. **Gesprächsnotizen.** Nach Abschluss des Interviews sind Gesprächsnotizen anzufertigen, meist werden sie als **Postskriptum** bezeichnet. Solche Gesprächsnotizen beinhalten Beschreibungen der Interviewten (äußere Erscheinung, seelische Verfassung, Gesundheitszustand etc.) sowie deren Räumlichkeiten und dokumentieren die Gesprächsatmosphäre, die Verfassung der Interviewenden und Unterbrechungen (Telefonate, Hereinkommen der Kinder etc.). Auch scheinbar offensichtliche Nebensächlichkeiten (z. B. Datum und Uhrzeit der Befragung) sollten notiert werden. Das Postskriptum wird bei späteren Validitätsbeurteilungen des Materials herangezogen (► Gesprächsnotizen (Postskriptum) zu einem Leitfadeninterview).
8. **Transkription.** Die Audioaufzeichnungen müssen vor einer interpretativen Auswertung – je nach Erkenntnisinteresse und Studiendesign – entweder vollständig oder auszugsweise wortwörtlich verschriftet werden („verbatim transcription"; zur Datenaufbereitung ► Abschn. 11.3.2). Da die Audioaufzeichnungen heute meist digital vorliegen, erfolgt die Transkription am Computer mit entsprechender **Transkriptions-Software**. Hierzu gibt es mittlerweile ein recht umfangreiches Angebot. Einige bewährte und häufig eingesetzte Programme stehen im Internet kostenlos zum Download bereit (z. B. f4 für PC und f5 für Mac, Express Scribe sowie EXMARaLDA für das Transkribieren von Interviews in mehreren Sprachen; zur Übersicht http://www.audiotranskription.de). Die entsprechenden Programme übernehmen die Aufgaben eines Tonwiedergabegerätes, erlauben ein komfortables Rück- und Vorspulen über Tastenkürzel oder Fußpedale, ein Verlangsamen der Abspielgeschwindigkeit und das Einfügen von Zeitmarken und Textbausteinen über Tastenkombination.

Ein Transkript enthält nicht nur den Interviewtext, sondern informiert auch über prägnante Merkmale des Gesprächsverlaufs (Tonhöhe, Pausen, Lachen, gleichzeitiges Sprechen etc.), die für die spätere Interpretation von Bedeutung sein können. In machen Fällen wird auf eine vollständige Transkription verzichtet (z. B. aus forschungsökonomischen Gründen und wenn die Interviews nur eine Datenquelle neben anderen darstellen). Alternativ arbeitet man dann nur mit den Interview-Notizen sowie mit weiteren Notizen, die beim Abhören der Audioaufzeichnungen generiert werden (für ein entsprechendes sechsstufiges Dokumentations- und Auswertungsverfahren s. Halcomb & Davidson, 2006).

Verschriftete Gespräche wirken durch unvollständige Sätze, verschluckte Silben, umgangssprachliche Wendungen und Füllwörter oft holprig und schlecht formuliert. Inwieweit man hier beim Transkribieren glätten darf, hängt vom theoretischen Interesse und vom verwendeten **Transkriptionssystem** und dessen Regeln ab. Im Zweifelsfall sollte das Transkript lieber zu viele als zu wenige Informationen über den Gesprächsverlauf konservieren. Übertriebener Eifer nach „Messgenauigkeit" ist aber fehl am Platz, da die Messung von Pausen in Hundertstelsekunden oder die Differenzierung zwischen 15 verschiedenen Formen des therapeutischen „Hms" wohl nur in Spezialfällen (linguistische Analyse) zu neuen Einsichten verhilft (Flick, 1995, S. 162). Zur Kennzeichnung nonverbaler und paraverbaler Äußerungen (z. B. Lachen, Husten, Sprechpausen) werden üblicherweise festgelegte **Transkriptionszeichen** verwendet, von denen ◘ Tab. 10.6 eine Auswahl verdeutlicht (ausführlicher hierzu Kuckartz, 2009; MacWhinney, 2000; Rehbein, Schmidt, Meyer, Watzke, & Herkenrath, 2004; Selting, 2001; Selting et al., 2009). Es ist zu beachten, dass unterschiedliche Transkriptionssysteme existieren und dass die Relation zwischen der leibgebun-

Tabelle 10.6 Einige Transkriptionszeichen nach Selting et al. (2009)

Transkriptionszeichen	Bedeutung
montag kam er ins krankenhaus	Interviewtext (nur Kleinschreibung!)
MONtag kam er ins krankenhaus	Betonung von Silben durch Großschreibung
MONtag kam er (-) ins krankenhaus	Geschätzte Kurzpause durch (-)
MONtag kam er (- - -) ins krankenhaus	Geschätzte längere Pause durch (- - -)
MONtag kam er (2.0) ins krankenhaus	Gemessene Pause mit Längenangabe in Sekunden mit einer Stelle hinter dem Punkt
MONtag kam er (2.0) ins kranken\	Abbruch eines Wortes oder Satzes durch \
MONtag kam er (2.0) in_s kranken\	Wortverschmelzung durch _
MONtag kam er (2.0) in_s krankn\	Ausgefallene Buchstaben werden ausgelassen
MONtag ka:m er (2.0) in_s krankn\	Dehnung durch :
MONtag ka:m er (2.0) in_s krankn\ ((weinen))	Para- und außersprachliche Handlungen und Ereignisse in ((. . .))
MONtag ka:m er (2.0) in_s ≪ t krankn>\ ((weinen))	Tonhöhe fallend ≪ *t* > (steigend: ≪ *h* . . . >)
I: [Wann] A: [MONtag] ka:m er (2.0) in_s < krank'n\ ((weinen))	Gleichzeitiges Reden von Interviewer (I) und Befragungsperson (hier: A) markiert durch [. . .]

denen Gesprächssituation einerseits und dem auf Basis einer Audioaufzeichnung nach einem bestimmten System erstellten Transkript andererseits theoretischer und methodologischer Reflexion bedarf (Dresing & Pehl, 2010). Das **Transkript** als **schriftliches Dokument** eines qualitativen Interviews ist ein forschungsgeneriertes Artefakt, das nicht mit der ursprünglichen Gesprächssituation gleichzusetzen ist (zur Definition und Einordnung von Dokumenten ▶ Abschn. 10.6.1). Beim Abfassen von Transkripten sind zusätzlich zu den Transkriptionszeichen einige **Richtlinien der Textgestaltung** zu beachten:

- ca. 50–55 Zeichen pro Zeile (erlaubt Randbemerkungen),
- Text in einfachem Zeilenabstand,
- bei jedem Sprecherwechsel eine Leerzeile einfügen,
- ebenso bei Themenwechsel eine Leerzeile einfügen,
- Sprecher durch Großbuchstaben und Doppelpunkt kennzeichnen,
- den gesamten Text seitenweise sowie zeilenweise durchnummerieren.

Um ein Vielfaches aufwändiger als die Verschriftung von Audioaufnahmen ist die **Transkription von Videoaufzeichnungen**, zumal wenn die Aktionen mehrerer Personen erfasst werden sollen. Hierzu sind einzelne Handlungssegmente zu definieren und hinsichtlich der beteiligten Akteure, ihrer Position zueinander, ihrer Mimik und Gestik etc. zu beschreiben. Videoaufzeichnungen sind vor allem bei Gruppeninterviews empfehlenswert, um die verschiedenen Sprecher besser differenzieren zu können. Angesichts des hohen Zeitaufwandes für die Transkription wird diese bei größeren Studien an entsprechend **geschulte Transkripteure** delegiert. Bei kleineren Studien und Abschlussarbeiten wird die Transkription indessen von den Forschenden selbst vorgenommen. Dies hat den Vorteil, dass man im Zuge der Transkription das Interviewmaterial sehr genau kennenlernt und somit die anschließende Auswertung fundierter erfolgen kann. Neben der Transkription sind oft auch weitere Schritte der **Datenaufbereitung** (Sortierung des Materials, Anonymisierung des Materials etc.) notwendig (▶ Abschn. 11.3), bevor die qualitative Datenanalyse beginnen kann.

9. **Analyse der Transkripte**. Bei qualitativen Interviews werden verbale Daten erhoben, die als Transkripte aufbereitet zur Analyse zur Verfügung stehen. Typischerweise wird man qualitative Interviews mithilfe **qualitativer Datenanalyseverfahren** auswerten (▶ Abschn. 12.1). Bei größeren qualitativen Interviewstudien werden ergänzend meist auch einige **quantitative Auswertungen** vorgenommen, so dass

man etwa die Prozentzahl der Befragten angibt, die im Interview eine bestimmte Meinung vertreten oder über eine spezifische Erfahrung berichtet haben. Die Aussagekraft entsprechender Quantifizierungen hängt dabei maßgeblich von der Art der Stichprobenbildung ab (► Kap. 9). Bei der qualitativen Analyse von Interviewtranskripten ist auf ein methodisch angeleitetes Vorgehen großen Wert zu legen, dabei können entsprechende Software-Tools (z. B. MAXQDA, HyperRESEARCH, ATLAS.ti) hilfreich sein (► Abschn. 12.1.3). Das beliebige Herausgreifen einzelner Zitate missachtet die Kriterien der Wissenschaftlichkeit.

Es ist zu beachten, dass die **Zirkularität** ein zentrales Prinzip qualitativer Sozialforschung ist (► Abschn. 2.3.3). Somit durchlaufen Datenerhebung und Datenanalyse innerhalb einer Studie idealerweise mehrere Zyklen, d. h. auf der Basis von Zwischenergebnissen der Datenauswertung werden gezielt zusätzliche, das Gegenstandsverständnis fördernde Fälle ausgewählt.

10. **Archivierung des Materials.** Bei einer qualitativen Befragung entsteht eine Fülle von Material, das sorgfältig zu archivieren, personenweise zu nummerieren und vor fremdem Zugriff zu schützen ist (**Datenschutz**). Pro Befragungsperson können vorliegen:
    - Audio- bzw. Videoaufzeichnung,
    - Transkript (elektronisch als Datei und Ausdruck auf Papier),
    - Angaben zur Textentstehung (Ort, Zeit, Interviewer, Interviewpartner, Transkripteur),
    - Postskript (Notizen über die Gesprächssituation und den Interviewpartner),
    - ggf. weitere Materialien zum Interview (soziodemografischer Fragebogen, Fotos, Zeichnungen, psychologische Tests).

    Interviewäußerungen sind persönliche Daten, zu deren Schutz die Forschenden gemäß Forschungsethik (► Abschn. 4.1) verpflichtet sind. Folgende Maßnahmen sind im Sinne des Datenschutzes zu ergreifen (und ggf. vorher schriftlich mit den Interviewten zu vereinbaren):

    Das Interviewmaterial muss verschlossen und für Unbefugte unzugänglich aufbewahrt werden. Interviewende müssen über die von ihnen durchgeführten Interviews Stillschweigen bewahren oder Erzählungen so allgemein halten, dass kein Rückschluss auf die Befragungspersonen möglich ist. Identifizierende Merkmale der Interviewten (Name, Wohnort, Beruf, Alter etc.) sind im archivierten Material, aber auch in späteren Ergebnispräsentationen, zu vermeiden oder, wenn sie inhaltlich relevant sind, geeignet zu modifizieren. Identifizierbarkeit entsteht häufig erst durch die Kombination von Merkmalen (z. B. Wohnort und Beruf), so dass einige relevante Merkmale (z. B. Alter, Beruf) unverändert bleiben können, wenn andere modifiziert werden (z. B. Wohnort). Identifizierende Merkmale dritter Personen (von Interviewten genannte Kollegen, Freunde, Familienangehörige etc.) sind ebenfalls unkenntlich zu machen. Identifizierendes Audio- und Videomaterial sollte nur in Ausnahmefällen längerfristig aufbewahrt werden, etwa wenn es in Forschung und Lehre auch nach Abschluss des Projektes benötigt wird. Ebenso wie quantitative Datensätze lassen sich auch qualitative Datensätze (anonymisierte Interviewtranskripte) öffentlich zugänglichen Forschungs-Archiven zur Verfügung stellen (Medjedovic & Witzel, 2010), so dass sie von der Scientific Community für **Sekundäranalysen** (► Abschn. 7.4) nutzbar sind.

Unstrukturiertes Interview

Bei einem unstrukturierten bzw. nicht-strukturierten Interview kommt es darauf an, die Untersuchungsteilnehmenden in ihren Äußerungen nicht durch ein vorgefertigtes Fragenkorsett einzuengen, sondern ihnen die Möglichkeit zu geben, sich zu den für sie wichtigen Aspekten des Gesprächsgegenstandes völlig frei zu äußern. In der qualitative Forschung werden verschiedene Varianten unstrukturierter Interviews unterschieden.

Unstrukturiertes Interview – Dem nicht-strukturierten bzw. unstrukturierten Interview („unstructured interview“) liegt kein vorgefertigtes Interviewinstrument zugrunde und der Interviewverlauf wird auch nicht durch die Interviewenden in Form eines Wechselspiels von Fragen und Antworten geleitet. Stattdessen liefern Interviewende mit einer einzelnen biografischen Frage einen Erzählanstoß **(narratives Interview),** fordern zur Verbalisierung handlungsbegleitender Denkprozesse auf **(Methode des lauten Denkens)** oder stellen im Zuge einer ethnografischen Feldbeobachtung spontane Verständnisfragen zum Feldgeschehen **(ethnografisches Feldinterview)**. Weil die Interviewsituation durch die Forschenden nicht aktiv vorstrukturiert wird, kommen Strukturierungen durch die Befragten zustande (z. B. anhand ihrer Relevanzstrukturen, subjektiven Theorien etc.).

Während das narrative Interview, die Methode des lauten Denkens und das ethnografische Feldinterview etablierte nicht-strukturierte Interviewtechniken sind, die im Zuge qualitativer Befragungsstudien zum Einsatz kommen, werden **nicht näher charakterisierte unstrukturierte bzw. „offene“ Interviews** oft im Rahmen von Voruntersuchungen durchgeführt. Dabei richtet man erkundende Fragen an Betroffene oder Experten, um herauszufinden, welche Aspekte des Themas überhaupt relevant sind und weiterverfolgt werden sollten. Solche Vorstudien-Interviews werden meist nicht vollständig verschriftet und auch nicht systematisch ausgewertet, sondern dienen der Ideenfindung und Themenzuspitzung (zur Eingrenzung des Forschungsproblems einer Studie ► Abschn. 5.1).

Narratives Interview

Entwickelt wurde die narrative Interviewtechnik in den 1970er-Jahren maßgeblich durch den deutschen Soziologen Fritz Schütze (1976a, 1976b, 1977a, 1977b, 1983, 1984, s. auch Bernart & Krapp, 1997; Lucius-Hoene & Deppermann, 2002; Wengraf, 2001). Etabliert hat sich das narrative Interview als Einzelinterview besonders in der Biografieforschung. Es wird auch als **autobiografisch-narratives Interview** bezeichnet und thematisiert z. B. Berufs- oder Krankheitsbiografien (Perleberg, Schütze, & Heine, 2006). Eine narrativ-biografische Befragung kann einerseits der Forschung wichtige Daten liefern, andererseits in der psychosozialen Beratungspraxis nützlich sein, weil die Befragten beim Erzählen und Reflektieren ihrer Lebensgeschichte zu neuen Einsichten gelangen, die ihnen bei der Lösung anstehender privater oder beruflicher Probleme helfen können (Rosenthal, 2002; Völzke, 2005).

10

Narratives Interview – Das narrative Interview („narrative interview") wird nach einer kurzen Erklärung des Vorgehens durch einen Erzählanstoß der Interviewenden eingeleitet, der sich auf autobiografische Erfahrungen bezieht und eine Stegreiferzählung der Befragungsperson auslösen soll. Im Zuge des freien Erzählens ergeben sich für die Befragten von allein Erzählzwänge, die zum Weitererzählen animieren. An die Haupterzählung schließt sich eine vom Interviewer geleitete Nachfrage- und Bilanzierungsphase an.

Der Ablauf eines narrativen Interviews kann in **fünf Phasen** eingeteilt werden (für ein Arbeitsbuch zur Methode s. Küsters, 2009):

1. **Erklärungs-Phase:** Zielsetzung und Ablauf des narrativen Interviews werden erklärt.
2. **Einleitungs-Phase:** Ein Erzählanstoß bzw. eine Erzählaufforderung sollen zur Stegreiferzählung führen. Ein Beispiel für einen **Erzählanstoß** im narrativen Interview wäre folgende Aufforderung: „Frau M., Sie sind vor zwei Jahren in Rente gegangen, erzählen Sie doch einmal, wie das gewesen ist! Wie war Ihr letzter Tag in der Firma?" Fragen nach der Befindlichkeit, nach Meinungen oder Gefühlen (z. B. „Frau M., Sie sind nun Rentnerin, wie fühlen Sie sich dabei?") wären hingegen keine Anstöße zum Erzählen, sondern zum Beschreiben. Das im Zentrum stehende Thema sollte für die Befragungspersonen relevant sein und ihnen das Gefühl geben, quasi als Alltagsexperten zum Thema gehört zu werden.
3. **Erzähl-Phase:** Im Hauptteil des narrativen Interviews erzählen die Befragten eine Geschichte zum Befragungsthema, die nicht durch inhaltliche Kommentare seitens der Interviewenden unterbrochen werden sollte. Die Interviewenden bemühen sich um eine angenehme Gesprächsatmosphäre, indem sie den Befragten Interesse und Verständnis signalisieren und Schweigepausen zulassen. Ein zu häufiges Nicken oder ähnliche Formen der Zustimmung, die bei Befragten den Eindruck erwecken könnten, die Interviewenden wüssten ohnehin schon alles, sind möglichst zu unterlassen.
   Narrative Interviews sind besonders informativ, wenn während des Erzählens von selbst **Erzählzwänge** entstehen, die zum Weitererzählen motivieren, ohne dass Rückfragen notwendig sind. Dabei unterscheidet man (Schütze, 1977a, 1977b):
   - einen **Detaillierungszwang** (die Erzählenden merken, dass Aspekte ihrer Geschichte unklar geblieben sind und Details nachgeliefert werden müssen),
   - einen **Gestaltschließungszwang** (Teile der Erzählung werden von den Erzählenden als noch nicht abgeschlossen empfunden und zu einer abgerundeten Geschichte vervollständigt) sowie
   - einen **Zwang zur Kondensierung** und **Relevanzfestlegung** (die Erzählenden sehen sich vor die Aufgabe gestellt, aufgrund der begrenzten Zeit nur die wichtigen Handlungsstränge komprimiert zu erzählen und als irrelevant empfundene Nebenaspekte zu kürzen oder zu überspringen).

   Im Laufe des freien Erzählens werden aufgrund dieser Erzählzwänge im narrativen Interview oftmals mehr Informationen offenbart als bei direktem Nachfragen, das auf Widerstände, Misstrauen oder Verschlossenheit treffen kann
4. **Nachfrage-Phase:** Ist die Haupterzählung aus Sicht der Befragten beendet, können in einer Nachfragephase offen gebliebene Hintergründe, Details und Widersprüchlichkeiten geklärt werden.
5. **Bilanzierungs-Phase:** Schließlich kann man in einer Bilanzierungsphase die Befragungsperson durch direkte Fragen zu einer abschließenden Bewertung der Geschichte anregen: „Welche Konsequenzen hatten diese Erlebnisse für Ihr weiteres Leben?". Derartige Fragen nach Bewertungen und Begründungen sollten die Befragten aber nicht in die Enge treiben oder Rechtfertigungsdruck erzeugen. Das Auseinanderhalten der Erzählphase von einer Bewertungsphase ist wichtig, weil jeweils unterschiedliche Aspekte des subjektiven Erlebens angesprochen werden: Beim Erzählen geht es um konkrete Episoden, beim Bewerten und Argumentieren um Theorien bzw. mentale Modelle.

Methode des lauten Denkens

Die Methode des lauten Denkens ist insofern den mündlichen Befragungsmethoden zuzuordnen, als die Untersuchungspersonen ihre Gedanken in Worte fassen und

mündliche verbale Daten generiert werden, die ohne den Forschungsprozess nicht existieren würden. Im Unterschied zu anderen qualitativen Interviewvarianten spielt jedoch beim lauten Denken die Interaktion zwischen Auskunftspersonen und Interviewenden keine Rolle, vielmehr ergeben sich die Äußerungen als **Kommentare zu einem selbst absolvierten Handlungsablauf**. Die untersuchte Handlung wird von den Forschenden ausgewählt und findet oft im **Forschungslabor** statt.

Methode des lauten Denkens – Bei der Methode des lauten Denkens („think-aloud method", „think-aloud protocol" [TAP], „think-aloud test") werden Untersuchungsteilnehmende aufgefordert, all ihre Gedanken in Worte zu fassen, die während einer bestimmten Aktivität (z. B. einen Text übersetzen, eine Entscheidung treffen, ein Computerprogramm nutzen) auftreten. Die interessierende Aktivität zusammen mit dem lauten Denken wird in der Regel per Video aufgezeichnet und anschließend verschriftet.

Hinsichtlich des Zeitpunktes der Handlungskommentierung werden zwei Varianten der Methode des lauten Denkens differenziert: Werden die Verbalisierungen der eigenen Gedanken direkt live während der interessierenden Aktivität abgegeben, so handelt es sich um die **Methode des periaktionalen lauten Denkens** („concurrent reports").

Beispiel: Im Usability-Labor sollen Untersuchungspersonen mittels des bereitgestellten Computers in einem Online-Shop eine Bestellung auslösen und sprechen dabei fortlaufend ihre Gedanken aus, während sie sich durch den Shop „durchklicken". Sofern der Handlungsablauf sehr beanspruchend ist, kann dieser auch auf Video aufgezeichnet werden, so dass die Befragungspersonen ihre Gedanken dann nachträglich beim Anschauen des Videos verbalisieren können (**Methode es postaktionalen lauten Denkens**, „retrospective report"). Die Methode des lauten Denkens wird in der Psychologie, der Kognitionsforschung, vor allem auch der **Produkt- und Usability-forschung** eingesetzt, etwa um Anhaltspunkte darüber zu gewinnen, an welchen Stellen ein Webangebot oder eine Software missverständlich gestaltet sind und im Sinne der Gebrauchstauglichkeit verbessert werden könnten (Bowles, 2010; Ericsson & Simon, 1980, 1993).

■ Ethnografisches Feldinterview

Ethnografische Feldinterviews werden nicht separat, sondern im Kontext der ethnografischen Feldforschung durchgeführt. Sie gehen also mit ethnografischen Feldbeobachtungen Hand in Hand (► Abschn. 10.1.2 „Ethnografische Feldbeobachtung"). Indem die Feldforschenden sich über längere Zeit vor Ort im Untersuchungsfeld aufhalten und die Feldmitglieder bei ihren Aktivitäten und Gesprächen beobachten, können sie spontane Rückfragen zum Geschehen stellen (► Studienbeispiel „Ethnografisches Feldinterview"). Solche ethnografischen Interviews oder Feldgespräche werden im Sinne des nicht-strukturierten Vorgehens nicht durch ein vorbereitetes Fragenraster organisiert, sondern ergeben sich durch die natürlichen Abläufe im Feld. Feldinterviews lassen sich nicht nur offline, sondern auch online durchführen, etwa wenn Forschende im Zuge der Feldbeobachtung einer Internet-Community mit einzelnen Mitgliedern Feldgespräche per Online-Chat führen (Garcia, Standlee, Bechkoff & Cui, 2009).

Studienbeispiel

Ethnografisches Feldinterview

In einer dänischen Feldforschungsstudie wurde das **Ernährungsverhalten von Jugendlichen in zwei Sportzentren** über mehrere Wochen hinweg nicht-teilnehmend beobachtet. Zudem wurden Feldinterviews mit **einzelnen Jugendlichen** sowie mit **natürlichen Gruppen** (Freundesgruppen, die in der Cafeteria an einem Tisch zusammensaßen) durchgeführt, in denen es darum ging, warum die Jugendlichen welche Speisen konsumieren. Da die Forschenden mit den Jugendlichen zusammen in der Cafeteria saßen und beobachteten, wer welche Speisen kauft und wie sie verzehrt werden, konnten sie Detailfragen zum Geschehen formulieren und sich auch in Gespräche einklinken, die die Jugendlichen selbst über ihre Cafeteria-Bestellungen führten. Es zeigte sich u. a., dass Pommes Frites besonders beliebt waren, weil sie sich von den häuslichen Mahlzeiten abheben, man sie im Freundeskreis leicht teilen kann und weil sich Jugendliche damit von den jüngeren Kindern abgrenzen können, für die warme Mahlzeiten außerhalb der Familie sowohl unerschwinglich als auch von Seiten der Eltern meist untersagt waren (Sylow & Holm, 2009).

Ethnografische Feldinterviews – Im Rahmen ethnografischer Feldforschung werden neben Feldbeobachtungen häufig ethnografische Feldinterviews („ethnographic interview", „field interview") als Einzel- oder Gruppeninterviews geführt. Sie folgen keinem festen Leitfaden, sondern gehen spontan auf die Situation im Feld ein und operieren oft mit W-Fragen (Wer? Wie? Wo? Wozu? Wie oft?). Um ihre Nähe zum Alltagsgespräch hervorzuheben, werden sie auch als **Feldgespräche** oder **ethnografische Forschungsgespräche** bezeichnet.

Ethnografische Feldinterviews gewinnen u. a. dadurch ihre Wissenschaftlichkeit, dass die Auswahl der Informanten reflektiert wird, dass die Äußerungen dokumentiert, im Kontext einschlägiger Theorien sowie vorliegender Forschungsergebnisse systematisch ausgewertet und z. B. im Forschungsteam diskutiert werden. Der Vorteil des nicht-strukturierten Feldinterviews liegt darin, dass es ohne Vorbereitung spontan geführt werden kann. Der Informationsnutzen mag jedoch zuweilen begrenzt sein und die systematische Auswertung eine Herausforderung darstellen.

Halbstrukturiertes Interview

In Abgrenzung zum unstrukturierten Interview, das nicht durch einen vorgefertigten Katalog von Fragen organisiert ist, basiert das halbstrukturierte Interview auf einem solchen Fragenkatalog, dem Interview-Leitfaden. Ein Leitfaden-Interview ist insofern halb- oder teilstrukturiert, als der Leitfaden grob vorgibt, welche Fragen die Interviewenden den Befragten in welcher Reihenfolge stellen. Im Unterschied zum vollstrukturierten Interview, bei dem sämtliche Fragen, deren Reihenfolge sowie die Antwortmöglichkeiten exakt vorgegeben sind, bestehen beim Leitfaden-Interview mehr Freiheitsgrade: Die Interviewenden dürfen spontan vom Leitfaden abweichen, um Vertiefungs- und Zusatzfragen zu stellen, die sich aus dem Gesprächsverlauf ergeben. Auch dürfen sie die Fragen im Wortlaut dem Gegenüber anpassen. Die Befragten wiederum beantworten die Fragen aus dem Leitfaden in eigenen Worten.

Teilstrukturiertes bzw. halbstrukturiertes Interview – Dem teilstrukturierten bzw. halbstrukturierten Interview („semi-structured interview") liegt ein Interview-Leitfaden („interview guide") als Liste offener Fragen (d.h. Fragen ohne festgelegte Antwortmöglichkeiten) zugrunde. Der Interview-Leitfaden bietet ein Grundgerüst, das für eine Vergleichbarkeit der Interviews sorgt. Er kann flexibel an die jeweilige Interviewsituation angepasst werden. Das halbstrukturierte Interview wird synonym auch als Leitfaden-Interview bezeichnet.

Im Folgenden werden wir zunächst die Grundprinzipien jedes halbstrukturierten bzw. Leitfaden-Interviews behandeln und dann auf fünf spezifische Varianten eingehen: auf das telefonische Leitfaden.Interview, das Online-Leitfaden-Interview, das Experten-Interview, das problemzentrierte Interview und das fokussierte Interview.

Leitfaden-Interview

Durch den **Interview-Leitfaden** und die darin angesprochenen Themen erhält man beim halbstrukturierten Interview ein Gerüst für Datenerhebung und Datenanalyse, das die Ergebnisse unterschiedlicher Interviews vergleichbar macht. Dennoch lässt das Leitfaden-Interview genügend Spielraum, spontan aus der Interviewsituation heraus neue Fragen und Themen einzubeziehen (zur Formulierung des Interview-Leitfadens bzw. interviewerad-ministrierten teilstandardisierten Fragebogens ▶ Abschn. 10.3.2 zum teilstandardisierten Fragebogen).

Die Fragen sind im Leitfaden nicht wortwörtlich vorgegeben, sondern können und sollten von den Interviewenden passend zur Situation und zum Gegenüber formuliert werden. Die Verständigung wird dadurch gesichert, dass die Interviewenden jeweils so oft nachfragen, bis sie die Bedeutung der Antworten verstanden haben. Der Interview-Leitfaden wird anhand der inhaltlichen Themen und Fragestellungen der Interviewstudie entwickelt, in **Probe-Interviews** (entsprechend dem Pretest in der quantitativen Forschung) überprüft und bei Bedarf überarbeitet (z. B. hinsichtlich Verständlichkeit und Vollständigkeit der Fragen und der Dauer des Interviews). Der **Ablauf der Fragen** wird meist so gestaltet, dass **biografische Grundinformationen** (Alter, Ausbildung, Beruf etc.) am Anfang erfragt werden (ggf. auch mit einem schriftlichen Fragebogen), damit sich die Interviewenden ein Bild vom Gegenüber machen und individuell auf die Befragungsperson eingehen können. Es schließen sich **allgemeine Fragen** zum Untersuchungsthema an, die im Verlauf des Interviews durch **detaillierte Fragen** ergänzt werden. **Heikle oder intime Fragen** stehen am Ende des Interviews, damit zuvor ein entsprechender Rapport zwischen Interviewer und Interviewpartner hergestellt werden kann, aber auch, damit eventuelle Irritationen nicht den gesamten Interviewverlauf überschatten.

Leitfaden-Interview – Das halbstrukturierte bzw. Leitfaden-Interview („semi-structured interview") basiert auf einer geordneten Liste offener Fragen (Interview-Leitfaden). Leitfaden-Interviews können persönlich, telefonisch (telefonisches Leitfaden-Interview) und online (Online-Leitfaden-Interview), mit Laien und mit Experten (Experten-Interview) durchgeführt werden. Sie können sich auf bestimmte Problemlagen konzentrieren (problemzentriertes Interview) oder auf konkrete Medienangebote fokussieren (fokussiertes Interview). Typisch sind bei den verschiedenen Varianten des Leitfaden-Interviews jeweils Einzelinterviews; es existieren aber auch halbstrukturierte Paar- und Gruppeninterviews.

Interview-Leitfaden – Der Interview-Leitfaden („interview guide") ist ein oft als „halbstandardisiert" bezeichnetes Instrument der Datenerhebung, weil zwar im Vorfeld ein Fragenkatalog festgelegt wird, aber keinerlei Antwortvorgaben existieren. Zudem kann und soll der Leitfaden flexibel der Interviewsituation angepasst werden (z. B. hinsichtlich Wortwahl, Reihenfolge der Fragen, Zusatzfragen). Die „Standardisierung" des Leitfadens besteht also nicht in exakt vorgegebenen Fragenformulierungen, sondern in der inhaltlichen Vergleichbarkeit des Fragenkatalogs.

Üblicherweise umfasst ein **Interview-Leitfaden 1–2 Seiten mit 8–15 Fragen** (Gläser & Laudel, 2004). Diese lassen sich in Hauptfragen und Differenzierungsfragen einteilen und können wörtlich und/oder stichpunktartig im Leitfaden stehen (▶ Studienbeispiel „Leitfadeninterview").

Wurden biografische und demografische Angaben nicht oder nur teilweise am Anfang des qualitativen Interviews erhoben, so werden sie gebündelt am Ende mündlich oder schriftlich erfasst (Nationalität, Bildungsstand, Beruf, Einkommen etc.). Je nach Forschungsfrage können mit einem abschließenden Kurzfragebogen weitere relevante Aspekte (z. B. Erkrankungen, Hobbies) erhoben werden, allerdings gilt bei persönlichen Daten das Gebot der Sparsamkeit.

Steht ein durch Probe-Interviews überprüfter Interview-Leitfaden zur Verfügung, wurden die Befragungspersonen gemäß einem inhaltlich aussagekräftigen und

Studienbeispiel

Leitfadeninterview

Der **Kernreaktorunfall** vom 26. April 1986 im ukrainischen **Tschernobyl** galt bis zum Jahr 2011 (Reaktorunfall im japanischen Fukushima) als schwerste nukleare Havarie mit weitreichenden gesundheitlichen, ökologischen und sozioökonomischen Folgen. Im Rahmen einer psychologischen Studie sollten wenige Jahre nach der Katastrophe von Tschernobyl die psychologischen Folgen in Deutschland untersucht werden, wobei folgender Interview-Leitfaden zum Einsatz kam (Legewie et al., 1990, S. 61):

Hauptfragen

1. Können Sie sich noch an die Zeit unmittelbar nach dem Unfall erinnern? Erzählen Sie, wie Sie davon erfahren und wie Sie darauf reagiert haben.
2. Wie ging es dann weiter bis heute? Wie hat sich Tschernobyl auf Ihr Leben ausgewirkt?

Detaillierungsfragen

3. Genaue Beschreibung der Stimmungen, Gedanken, Gefühle, Ängste und Hoffnungen;
4. Phantasievorstellungen oder Träume im Zusammenhang mit Tschernobyl;
5. Änderungen der Lebensgewohnheiten, besondere Handlungsweisen;
6. Reaktion der Mitmenschen und Auswirkungen auf den Interviewten;
7. Bedeutung der Information durch die Medien;
8. Einfluss auf wichtige Lebensentscheidungen;
9. Zusammenhang von Atomkraftwerken und Atomwaffen;
10. Vergleich mit früheren Lebensereignissen (ausführliche Erzählung!);
11. Konsequenzen für die persönliche Zukunft;
12. Einflussmöglichkeiten auf zukünftige gesellschaftliche Entwicklung;
13. Einstellung zu eigenem politischen Engagement und
14. Bedeutung für den Sinn des eigenen Lebens.

Anmerkung

Dieser Leitfaden arbeitet mit stichpunktartigen Detaillierungsfragen, deren genaue Bedeutung und Anwendung den Interviewenden durch die Interviewerschulung bekannt war. Als Stichpunktliste vermittelt dieser Leitfaden schnellen Überblick in der Interviewsituation. Häufig werden in der Praxis aber auch Leitfäden mit ausformulierten Fragen genutzt.

forschungsökonomisch vernünftigen Stichprobenverfahren ausgewählt und alle Interviewerinnen und Interviewer sorgfältig geschult, so kann die eigentliche Datenerhebung beginnen. Studien mit Leitfaden-Interviews umfassen oft Stichproben im Umfang von rund **10–20 Personen**, teilweise existieren jedoch auch qualitative Interviewstudien mit größeren Stichproben von bis zu 200 und mehr Befragten. Üblicherweise dauern Leitfadeninterviews **zwischen 1–2 Stunden**, pro Interviewstunde werden **5–10 Stunden Transkriptionszeit** veranschlagt

Exemplarische Abschätzung des Aufwandes für die Datenerhebung mittels Leitfadeninterview

Zeitaufwand für die Interviewdurchführung

- Stichprobenumfang für qualitative Befragungsstudien: 10–20 Befragungspersonen
- Länge von Leitfadeninterviews: 1–2 Stunden
- Beispiel: 15 Interviews à 1.5 h = 22.5 Netto-Interviewstunden plus Fahrzeiten sowie Vor- und Nachbereitung
- In Summe ca. 40 Arbeitsstunden für die Durchführung von 15 Interviews

Zeitaufwand für die Transkription

- Zeitaufwand für 1 Interviewstunde: 5–8 Stunden Transkriptionszeit
- Beispiel: 22.5 Interviewstunden à 6 Stunden Transkriptionszeit
- In Summe ca. 135 Arbeitsstunden für die Transkription von 15 Interviews

Umfang des verbalen Datenmaterials

- 1 Interviewstunde entspricht ca. 20–30 Seiten Transkript
- Beispiel: 22.5 Interviewstunden à 25 Seiten
- In Summe ca. 565 Seiten Transkript (verbales Datenmaterial) für 15 Interviews

Zeitaufwand für die qualitative Datenanalyse

Der Zeitaufwand für die qualitative Datenanalyse hängt von der gewählten Analysemethode ab (▶ Abschn. 12.1). Er ist normalerweise mindestens ebenso groß wie der gemeinsame Zeitaufwand für Interviewdurchführung und Transkription. In diesem Beispiel sind für Interviewdurchführung und Transkription zusammen ca. 22 achtstündige Arbeitstage (175 Arbeitsstunden bzw. rund ein Arbeitsmonat) anzusetzen. Hinzu käme mindestens ein Monat Auswertungszeit für 15 Interviews.

(▶ Exemplarische Abschätzung des Aufwandes für die Datenerhebung mittels Leitfadeninterview).

Nach der Durchführung jedes Interviews werden die Einverständniserklärung der Befragten, die Audioaufzeichnung sowie Postskriptum (z. B. zum Eindruck von der Befragungsperson, ihrer Auskunftsbereitschaft, der Interviewatmosphäre etc.) gemäß Regeln des Datenschutzes archiviert. Die Audioaufzeichnungen werden schließlich zusammen mit den soziodemografischen Angaben und Interviewnotizen verschriftet und dabei anonymisiert.

Ein Ausschnitt (Basistranskript) aus einem nach dem gesprächsanalytischen Transkriptionssystem 2 (GAT 2, Selting, et al., 2009) transkribierten Gespräch ist in ◘ Tab. 10.7 zu finden.

▪ Telefonisches Leitfaden-Interview

Interviews werden in der Regel mündlich im Face-to-Face-Kontakt durchgeführt, teilweise aber auch telefo-

Tabelle 10.7 Ausschnitt aus einem nach GAT 2 transkribierten Gespräch (Stock, 1994, S. 49–51) „Der widerliche Kerl"

01	S1:	ja:; (.) die VIERziger generation so;=	20		und (.) die ältere tochter is AUCH in berlin gewesen? (1.1)
02		=das=s: !WA:HN!sinnig viele die sich da ham [SCHEIden	21		und (.) der KE::RL,
03	S2:	[ja;	22		das war aber ein peneTRANter: !WI!derling.=also (1.0)
04	S1:	lasse[n.=	23		der hat (.) äh sein GARten wie (.) !PIK! AS (- -) gePFLEGT,=
05	S2:	[hm,	24		=ne,
06	S1:	=oder scheiden lassen ÜBERhaupt.	25		!KEIN! BLÄTTchen,=
07	S2:	hm,	26		=und NICHTS,=
08		(- -)	27		=englischer Rasen, (- -)
09	S1:	heute noch-	28		un:dh: bei !JE!der gelegenheit hat er die poliZEI
10		(2.1)	29		gerufen,
11		s=is der UMbruch.	30		und sich mit den NACHbarn ange[legt,=ne, (1.2)
12	S2:	n besonders GUtes beispiel das warn mal unsere NACHbarn.	31	S1:	[phhh hohoho
13		(1.0)	32	S2:	un wenn da: einmal: jemand zum abschied geHUPT hat,
14		ähm (1.0)	33		da war der in NULL komma nix draußen;
15		DREISsig jahre verheiratet, (- -)	34		und hat da RUMgeschrien;=
16		das letzte kind (.) endlich aus m HAUS,	35		=ich hol die poliZEI: und [so-
17		zum stuDIERN, (- -)	36	S1:	[das GIBS doch wohl nich.
18		WEGgegangen,=ne,	37	S2:	ja; V:OLLkommen WIderlich.=ne, (- -)
19		nach berLIN, (- -)			

nisch oder computervermittelt. Die Art des Interviewkontakts wird als Interviewmodus bezeichnet. Während in der quantitativen Umfrageforschung die Unterschiede zwischen verschiedenen Interviewmodi als sog. **Methoden- oder Modus-Effekte** („mode effects") sehr intensiv untersucht werden, wird dieser Aspekt in der qualitativen Methodenforschung nur selten diskutiert.

Telefonisches Leitfaden-Interview – Das telefonische Leitfaden-Interview] („semi-structured telephone interview") ist eine medienvermittelte Variante des Leitfaden-Interviews. Hierbei findet der Interviewkontakt telefonisch statt.

Einer der **Vorteile telefonischer Leitfaden-Interviews** liegt in der Forschungsökonomie: Der zeitliche, personelle und finanzielle Aufwand für Reisen reduziert sich und der Kreis der Befragten kann geografisch ausgeweitet werden. Ein weiterer Vorteil ist die größere Anonymität: Befragungspersonen fällt es zuweilen leichter, am Telefon über heikle, schambesetzte Themen zu sprechen, als Auge in Auge mit dem Interviewer. Auch mögliche unbewusste Vorbehalte aufgrund des äußeren Erscheinungsbildes sowohl der Befragungsperson als auch des Interviewers sind beim Telefonkontakt reduziert, wenngleich die wahrnehmbar bleibenden Personenmerkmale (Stimme, Dialekt etc.) die Eindrucksbildung beeinflussen. Nicht zuletzt bleibt beim Telefonkontakt die persönliche Sicherheit stärker gewahrt als beim persönlichen Zusammentreffen mit einem Interviewer, was z. B. für sozial ängstliche Befragungspersonen oder bei Forschungsinterviews in politischen Krisengebieten von Bedeutung sein

kann. Trotz dieser Vorteile des medienvermittelten Kontakts gilt in der Denkweise des qualitativen Ansatzes der persönliche Kontakt zwischen Interviewenden und Befragungspersonen oftmals als Königsweg der Datenerhebung, weil hier die leibgebundene Gesprächssituation mit allen Sinnen erlebt wird.

Dass Leitfadeninterviews telefonisch ohne nennenswerten Qualitätsverlust möglich sind, wird oft anekdotisch berichtet. Eine der wenigen **empirischen Methodenvergleichs-Studien** zeigte, dass telefonische Leitfadeninterviews etwas kürzer ausfallen als persönliche Leitfadeninterviews, die Befragten kürzere Redebeiträge liefern und sich häufiger rückversichern, ob ihre Antwort der Frage entspricht (Irvine, Drew, & Sainsbury, 2010). Eine Vergleichsstudie, welche die Erfahrungen von Interviewenden sowie Befragungspersonen bei leitfadengestützten Bewerbungsinterviews erhob, zeigte, dass Telefoninterviews von beiden Seiten akzeptiert wurden, wobei aufgrund der herausgefilterten Hintergrundinformationen die Befragungspersonen am Telefon einen positiveren Eindruck hinterließen als im persönlichen Interview (Straus, Miles, & Levesque 2001). Methodenvergleichsstudien, die persönliche, telefonische und Online-Leitfadeninterviews systematisch anhand unterschiedlicher Qualitätskriterien empirischer Forschung kontrastieren und praktische Empfehlungen ableiten, stehen weitgehend aus (Opdenacker, 2006).

Online-Leitfaden-Interview

Online-Leitfaden-Interviews sind indiziert, wenn man international Akteure aus dem Internet-Kontext befragen möchte (z. B. Betreiberinnen von Websites, Mitglieder von Online-Foren), die auf anderen Kommunikationswegen kaum erreichbar sind. Sowohl für die Forschenden als auch für die Befragungspersonen sind Online-Interviews mit geringerem logistischem Aufwand verbunden. Manche qualitative Studien gehen auch pragmatisch bei der Wahl des Kommunikationskanals vor und nutzen denselben Interview-Leitfaden, um einige ihrer Befragungspersonen mündlich, andere telefonisch und wieder andere per Online-Chat zu interviewen.

Online-Leitfaden-Interview – Das Online-Leitfaden-Interview („semi-structured online interview", „semi-structured Internet interview") ist eine medienvermittelte Variante des Leitfaden-Interviews. Hierbei findet der Interviewkontakt computervermittelt entweder zeitversetzt (z. B. E-Mail-Interview) oder zeitgleich (z. B. Chat-Interview, Webcam-Interview) statt. Es wird typischerweise dialogisch eine Frage gestellt, die Antwort abgewartet und dann die nächste Frage gestellt. Im Unterschied zum selbstadministrierten Online-Fragebogen kann bei einem Online-Leitfaden-Interview individuell auf die Antworten der Befragten eingegangen werden.

Das textbasierte (schriftliche Mündlichkeit) Online-Interview hat den Vorteil, dass es bei heiklen Themen einen besseren Schutz der Privatsphäre bietet als ein persönliches oder auch ein telefonisches Leitfadeninterview. Dafür ist eine Überprüfung der Identität der Befragten online oft kaum möglich. Beim Online-Interview liegen alle Interviewäußerungen automatisch digital vor, so dass keine Transkription notwendig ist und somit viel Arbeit gespart wird. Online-Interviewende sollten in computervermittelter Kommunikation geübt und geschult sein, etwa um bei synchronem Kontakt einen flüssigen Kommunikationsablauf zu gewährleisten und Besonderheiten der computervermittelten Kommunikation (z. B. Verwendung von Smileys und Emoticons) adäquat einzubeziehen. Qualitative Studien, die mit Online-Leitfaden-Interviews arbeiten, untersuchen so unterschiedliche Themen wie die Merkmale von Internet-Freundschaften (Henderson & Gilding, 2004), die Gründe für (un)safen Sex zwischen Männern (Adams & Neville, 2009) oder die Erfahrungen von Patienten mit künstlicher Ernährung (Silver, 2004; ▶ Studienbeispiel „Online-Leitfaden-Interview"). Auch Methodenbeiträge zum Online-Leitfaden-Interview liegen vor (z. B. Ayling & Mewse, 2009; Garcia et al., 2009; Salmons, 2009).

Experten-Interview

Die Methode des Experten-Interviews wurde im deutschsprachigen Raum maßgeblich durch einen Beitrag von Michael Meuser und Ulrike Nagel (1991) beeinflusst. Bei Experten-Interviews werden die Befragten nicht als Laien oder Betroffene, sondern als Fachleute für ein bestimmtes Thema angesprochen. Dabei können auch spezifische Fragentypen zum Einsatz kommen, etwa sog. zirkuläre Fragen, bei denen nach den Sichtweisen Dritter gefragt wird. Es können auch Meinungen anderer Experten präsentiert werden, um die Befragungsperson zu einer Reaktion zu bewegen (Gläser & Laudel, 2004).

Das mittels Experten-Interview erhobene Expertenwissen bezieht sich einerseits auf gedanklich und sprachlich leicht zugängliches **strukturelles Fachwissen**, andererseits aber auch auf **Praxis- und Handlungswissen**. Letzteres ist oft stark verinnerlicht und für Experten schwer in Worte zu fassen. Ist das Handlungswissen von zentraler Bedeutung, so kann das Experteninterview auch handlungsbegleitend im unmittelbaren Praxisfeld (z. B. am Arbeitsplatz, im Labor) erfolgen und gegebenenfalls mit der Beobachtungsmethode (▶ Abschn. 10.1) oder der Methode des lauten Denkens kombiniert werden. Anhand welcher Kriterien (z. B. formaler Abschluss, Praxiserfahrung, aktuelle berufliche Position) der Expertenstatus zugeschrieben wird, sollte genau durchdacht und erläutert werden. Dass Fachexperten je nach Themenfeld meist in **deutlich geringerer Zahl** als Nichtexperten zur Verfügung stehen und **oft schwer erreichbar sind**, muss bei der Konstruktion des Stichprobenplans (z. B. Schneeballverfahren) und bei der Rekrutierung beachtet werden

Studienbeispiel

Online-Leitfaden-Interview

Um **Patientenerfahrungen mit künstlicher Ernährung** („total parenteral nutrition“, TPN) zu erkunden, führte die amerikanische Medizinwissenschaftlerin Heidi Silver (2004) mit 6 Erwachsenen per Online-Chat Interviews durch, die auf folgendem Leitfaden basierten:

1. „Tell me about any fears you have related to receiving TPN.“
2. „Are there fears related to handling the TPN infusion?“
3. „Tell me how your life is different receiving TPN.“
4. „Tell me about changes you have made in your life because you are receiving TPN.“
5. „What are the greatest challenges you face receiving TPN?“
6. „How do you handle these kinds of challenges?“
7. „Tell me about how you interact with other people.“
8. „Tell me about your interactions with healthcare providers.“
9. „Let’s talk about eating; what can you tell me about that?“
10. „Tell me about the foods you eat and the ones you don’t.“
11. „How do you handle the diet restrictions you have?“
12. „Are there things you feel you are missing?“
13. „What else would you like to tell me about related to receiving TPN?“

Die Interviewtranskripte wurden qualitativ ausgewertet und auf dieser Basis wurden in zeitlichem Abstand zwei weitere Online-Interviews mit denselben Befragungspersonen durchgeführt. Die Ergebnisse der interpretativen Auswertung der Transkripte (6 Hauptthemen der betroffenen Patienten) wurden den Befragungspersonen mit Bitte um Feedback per E-Mail zugänglich gemacht, um die Gültigkeit der Schlussfolgerungen der Wissenschaftlerin zu überprüfen.

(z. B. besonders starke Orientierung an Expertenwünschen hinsichtlich Ort, Zeit, Kommunikationskanal des Interviews; ► Studienbeispiel „Experten-Interview“).

Experten-Interview – Das Experten-Interview („expert interview“) ist eine Variante des Leitfaden-Interviews, bei der die Befragungspersonen als fachliche Expertinnen und Experten zu einem Thema befragt werden und ihr Spezialwissen (strukturelles Fachwissen und/oder Praxis-/Handlungswissen) erschlossen werden soll. Die Definition und Rekrutierung der Experten stellt hier eine besondere Herausforderung dar. Zudem muss von den Interviewenden eine geeignete Rolle gewählt und eingenommen werden (z. B. als Co-Experte oder Laie).

Interviewende können gegenüber Fachleuten unterschiedliche Rollen einnehmen. Sie können z. B. als **Laien** auftreten, was die Experten anregt, ihr Wissen in vereinfachter Weise zu vermitteln. Durch entsprechende Vorbildung und/oder Vorbereitung können sie als **Co-Experten** auftreten, so dass der inhaltliche Austausch auf einem fachlich anspruchsvollen Niveau stattfindet. Welche Strategie zu wählen ist, hängt vom Thema ab und sollte reflektiert und begründet werden (Bogner, Littig, & Menz, 2005; Dorussen, Lenz, & Blavoukos, 2005). Das Experten-Interview weist Ähnlichkeiten zum **Eliten-Interview** auf, bei dem Personen aus der politischen oder wirtschaftlichen Elite befragt werden, die aufgrund ihrer Machtposition über spezifisches Expertenwissen verfügen (Littig, 2008).

Studienbeispiel

Experten-Interview

In sog. **Notfallplänen** wird für Organisationen und Behörden im Vorfeld festgelegt, wie in unterschiedlichen Gefahren- und Katastrophen-Situationen zu verfahren ist, um Leben, Gesundheit und Umwelt zu schützen. Inwiefern berücksichtigen Notfallpläne die spezifischen Bedürfnisse von Kindern? Um dieser Forschungsfrage im Rahmen einer qualitativen Interviewstudie nachzugehen, wurden Experten-Interviews durchgeführt (Woolsey & Bracy, 2010).

- **Definition der Experten:** Als Experten wurden Personen definiert, die entweder über eine mindestens 5-jährige Berufserfahrung im Bereich Katastrophenschutz verfügten oder zum Thema Katastrophenschutz wissenschaftlich publiziert hatten.
- **Erreichbarkeit der Experten:** Die Leitfaden-Interviews fanden telefonisch statt, da die Experten national und international geografisch weit verstreut waren. Insgesamt konnten 16 Experten überwiegend aus den USA rekrutiert werden.
- **Durchführung der Experten-Interviews:** Die telefonischen Leitfaden-Interviews wurden alle vom selben geschulten Interviewer in englischer Sprache durchgeführt und basierten auf einem selbstkonstruierten und per Probe-Interview überprüften Interview-Leitfaden (14 offene Fragen). Mit einer Ausnahme dauerten alle Interviews gut eine Stunde.
- **Dokumentation der Experten-Interviews:** Die Interviews wurden vollständig verschriftet. Die Transkripte wurden allen befragten Experten mit Bitte um notwendige Ergänzungen oder Korrekturen vorgelegt.
- **Auswertung der Experten-Interviews:** Die inhaltsanalytische Auswertung der Transkripte enthielt quantitative sowie qualitative Elemente (z. B. Häufigkeitsauszählungen, wörtliche Zitate). Das inhaltsanalytische Kategoriensystem wurde zunächst deduktiv erstellt und dann anhand der Transkripte induktiv ergänzt.
- **Ergebnisse der Studie:** Ein Befund der Studie war, dass die Mehrzahl der befragten Experten eine mangelnde Berücksichtigung der spezifischen physischen wie psychischen Bedürfnisse von Kindern in den Notfallplänen konstatierte.

■ Problemzentriertes Interview

Die Entwicklung des problemzentrierten Interviews (PZI; „problem-centered interview“) geht auf den deutschen Psychologen Andreas Witzel (1982, 1989, 1996, 2000) zurück, der einerseits Interviews ablehnte, die auf der Basis halb- oder vollstandardisierter Instrumente einem zu starren Ablauf isolierter Fragen und Antworten folgen, andererseits aber auch unstrukturierte narrative Interviews kritisierte, in denen die Befragungspersonen völlig frei erzählen und die Interviewer vornehmlich nur zuhören.

Das problemzentrierte Interview zielt deswegen auf die richtige **Balance zwischen zu starker und zu geringer Strukturierung** des Vorgehens ab und bettet die im Interview diskutierten gesellschaftsrelevanten Themen in den biografischen Kontext der Befragten ein (für einen Vergleich zwischen narrativem und problemzentriertem Interview s. Mey, 2000).

Das problemzentrierte Interview hat folgenden Ablauf mit sieben Phasen (Witzel, 2000):

1. **Erklärungsphase**: Zielsetzung und Ablauf des problemzentrierten Interviews werden der Befragungsperson erklärt. Insbesondere wird Wert darauf gelegt, den Befragungspersonen zu vermitteln, dass es nicht um „intellektuelle Leistungen" geht, sondern dass ihre individuellen Vorstellungen und Meinungen interessieren und unvoreingenommen akzeptiert werden.
2. **Kurzfragebogen zu Sozialdaten**: Vor Beginn des mündlichen Interviews wird von den Befragten ein schriftlicher Kurzfragebogen ausgefüllt, der einfache soziodemografische und biografische Fakten dokumentiert (Alter, Beruf der Eltern, Ausbildung etc.). Diese Hintergrundinformationen helfen dem Interviewer, sich auf sein Gegenüber einzustellen, und vermeiden eine Abfragesituation während des Interviews.
3. **Vorformulierte Einleitungsfrage als Erzählanstoß**: Ähnlich wie im narrativen Interview soll auch am Beginn des problemzentrierten Interviews eine Erzählaufforderung eine Stegreiferzählung zum Untersuchungsthema hervorlocken. Im Falle einer Interviewstudie zur Situation männlicher Grundschullehrer in dem frauendominierten Arbeitsfeld lautete die Einleitungsfrage: „Sie sind schon seit einigen Jahren Lehrer. Mich interessiert nun der Anfangspunkt Ihrer Beschäftigung, aber auch Ihre gegenwärtige Situation. Wie kam es denn dazu, dass Sie Grundschullehrer geworden sind?" (Baar, 2010, S. 418).
4. **Erzählungsgenerierende und verständnisgenerierende Kommunikationsstrategien im Anschluss an die Eingangserzählung**: Nach der Eingangserzählung folgt eine Interviewphase, die den roten Faden der Eingangserzählung weiterspinnt. **Allgemeine Sondierungen** vertiefen die von der Befragungsperson angesprochenen Aspekte und sollen die Erzählung weiterer Erfahrungsbeispiele hervorlocken (erzählungsgenerierende Strategie). Ergänzend sollen im Falle unklarer, ausweichender oder widersprüchlicher Aussagen **spezifische Sondierungen** sicherstellen, dass der Interviewer die Sichtweise der Befragungsperson richtig verstanden hat (verständnisgenerierende Strategie).
5. **Ad-hoc-Fragen anhand des Leitfadens**: Jene im Interview-Leitfaden auf der Basis theoretischer Vorüberlegungen festgelegten Themenbereiche, die von den Befragten im Zuge der Eingangserzählung nicht von selbst angesprochen wurden, werden in Form von Ad-hoc-Fragen flexibel durch den Interviewer eingebracht. Auch hier sollte ein mechanisches Frage-Antwort-Schema vermieden, sondern wiederum mit einerseits erzählungsgenerierenden und andererseits verständnisgenerierenden Kommunikationsstrategien operiert werden. Im Falle der bereits angesprochenen Grundschullehrer-Studie umfasste der Leitfaden z. B. noch Fragen zur didaktischen Gestaltung des Unterrichts, zum Verhältnis der männlichen Grundschullehrer zu ihren Schülern und Schülerinnen, zu den Kontakten zu anderen männlichen Kollegen, zu Kolleginnen sowie zur Schulleitung (Baar, 2010).
6. **Postskriptum**: Ergänzend zur fortlaufenden Audioaufzeichnung des problemzentrierten Interviews werden unmittelbar nach dem Interview so genannte Postskripte (Interviewnotizen) erstellt, das sind Anmerkungen zur Interviewsituation sowie zu thematischen Auffälligkeiten und Interpretationsideen.
7. **Transkription und Datenanalyse**: Die Interviewtranskripte werden einer qualitativen Datenanalyse unterzogen, wobei unterschiedliche Methoden zum Einsatz kommen können (▶ Abschn. 12.1). Grundsätzlich ist neben der **Einzelfallanalyse** eine **Typenbildung** sinnvoll, die auf systematisch kontrastierenden Fallvergleichen basiert.

Das problemzentrierte Interview gehört zum Grundkanon qualitativer Interviewtechniken (Mey, 2005). Es ist in der deutsch- und englischsprachigen Literatur vertreten (Scheibelhofer, 2008) und hat eine Reihe von Varianten hervorgebracht. Dazu zählen

- das **episodische Interview**, das narrativ-episodisches ebenso wie semantisches Wissen erschließen soll (Flick, 2002),
- das **themenzentrierte Interview**, das auch abgewehrte und latente Sinngehalte herausarbeiten möchte (Schorn, 2000) sowie
- das **personzentrierte Interview**, das sich an klientenzentrierter Gesprächsführung orientiert (Woelfer, 2000).

Es ist jedoch fraglich, ob es sich hier um so eigenständige und eindeutig abgrenzbare Varianten handelt, dass eine spezifische Etikettierung gerechtfertigt ist. Denn in der Logik des qualitativen Forschungsansatzes stellt jede Methoden*anwendung* immer auch eine gegenstandsbezogene Methoden*entwicklung* dar, ohne dass damit inflationär ständig neue Verfahren(svarianten) postuliert werden müssen (Mey, 2005).

Fokussiertes Interview

Die Technik des fokussierten Interviews wurde Anfang der 1940er-Jahre von dem **Soziologen Robert Merton** entwickelt und während des Zweiten Weltkriegs u. a. eingesetzt, um die Reaktionen von amerikanischen Soldaten und Zivilbevölkerung auf Propagandafilme und Radioprogramme zu erfassen (Merton, Fiske, & Kendall, 1990). Generell zielt das fokussierte Interview darauf ab, die Sichtweisen der Befragten zu einem fokussierten Objekt zu erfragen, das während des Interviews präsentiert wird (► Studienbeispiel „Fokussiertes Interview“).

Fokussiertes Interview – Das fokussierte Interview („focussed interview“, „focused interview“) ist eine Variante des Leitfaden-Interviews, bei der ein bestimmtes Anschauungsobjekt im Mittelpunkt steht bzw. es darum geht, die Reaktionen der Interviewten auf das **fokussierte Objekt** im Detail qualitativ zu ermitteln. Das fokussierte Objekt ist oft eine mediale Darstellung (z. B. Film, Fernseh- oder Radiosendung, Werbeanzeige) und wird zu Beginn oder im Verlauf des Interviews präsentiert. Ein fokussiertes Interview kann als Einzelinterview („individual focused interview“) oder als Gruppeninterview („focused group interview“) durchgeführt werden.

Ganz im Sinne heutiger Methodenkombinationen und **Mixed-Methods-Designs** sollte das fokussierte Interview als qualitative Datenerhebungstechnik in der Kommunikationsforschung in Kombination mit quantitativen Methoden eingesetzt werden: Weiß man etwa anhand quantitativer Daten aus Medienwirkungs-Experimenten, dass ein bestimmter Medieninhalt positive emotionale Reaktionen auslöst oder zu einer signifikanten Einstellungsänderung beim Publikum führt, so kann mittels fokussiertem Interview der Frage nachgegangen werden, welche konkreten Merkmale des komplexen medialen Stimulus-Materials (z. B. welche Filmszene, welche Äußerung in einem Radioprogramm) die Rezipienten in welcher Weise interpretieren. Das fokussierte Gruppeninterview gilt als **Vorläufertechnik der Fokusgruppen-Diskussion**, die heute in der Marktforschung stark verbreitet ist).

Vor der Durchführung eines fokussierten Interviews müssen a) eine **gründliche Inhaltsanalyse des fokussierten Objekts** durchgeführt und b) **Hypothesen über Bedeutung und Wirkung einzelner Aspekte des Objekts** abgeleitet werden (z. B. Hypothesen über die Interpretation einzelner Filmausschnitte). Auf der Basis dieser Hypothesen wird c) ein **Interview-Leitfaden** zusammengestellt, so dass bereits während der Interviews geprüft werden kann, ob die Äußerungen der Befragten die Hypothesen eher bestätigen oder widerlegen und welche neuen Erklärungsbeiträge die Interviewäußerungen liefern. Zur Durchführung eines fokussierten Interviews geben Merton, Fiske und Kendall (1990, S. 3ff.) folgende Ratschläge:

Studienbeispiel

Fokussiertes Interview

In einem mediensoziologischen Forschungsprojekt wurde mithilfe fokussierter Einzelinterviews der Frage nachgegangen, wie Jugendliche **Sex- und Erotikdarstellungen im nächtlichen TV-Programm** erleben und bewerten (Hoffmann, Krauß, & Gäbel, 2005).

Befragungspersonen: An der Studie nahmen 15 weibliche und 15 männliche Jugendliche im Alter von 16–18 Jahren aus dem Raum Potsdam teil.

Interviewende: Die Mädchen wurden von einer Interviewerin, die Jungen von einem Interviewer befragt, wobei die Altersdifferenz zwischen Befragten und Interviewten relativ gering war (7 bis maximal 11 Jahre).

Befragungsort: Die Interviews wurden an der Hochschule für Film und Fernsehen (HFF) Konrad Wolf in Potsdam-Babelsberg durchgeführt.

Interviewablauf: Die fokussierten Einzelinterviews dauerten zwischen 60 und 100 Minuten. Verschiedene Filmsequenzen wurden jeweils am Anfang sowie gegen Ende des Interviews gezeigt.

Fokussiertes Objekt: Den Jugendlichen wurden insgesamt zwölf verschiedene Filmsequenzen präsentiert, darunter ein Werbeclip für eine 0190er-Nummer, ein DSF-Sportclip und ein Ausschnitt aus dem Film „Showgirls“. Die Inhalte der Filmsequenzen wurden im Hinblick auf ein breites Spektrum an Sexualitätsdarstellungen ausgewählt, im Vorfeld analysiert und z. B. folgendermaßen beschrieben:

„Die dritte Sequenz stammt aus dem Erotikfilm Showgirls von Paul Verhoeven (1995) mit einer Altersfreigabe ab 16 Jahren. Er wurde hier stellvertretend für Filme gewählt, die erst ab 22.00 Uhr im Fernsehen gezeigt werden dürfen. Erzählt wird die Geschichte der attraktiven, jungen Nomi, die vor ihrer Vergangenheit flüchtet und ihre Chance am Las-Vegas-Strip sucht. Sie nimmt einen Job als erotische Tänzerin bzw. „Lap Dancer“ an. Die Sequenz (2:40 Minuten) zeigt die Tänzerin, wie sie, ausschließlich mit einem String und High Heels bekleidet, einen im Sessel sitzenden Mann durch ihren Tanz sexuell erregt. Der Lap Dance wird von einer Frau – vermutlich der des Mannes – beobachtet, die die Leistung auch letztlich bezahlt. Im Laufe ihrer ‚Arbeit‘ lässt die Tänzerin ihren String fallen. Man sieht, wie sie im Genitalbereich rasiert ist. Die Tänzerin hat ein vielfältiges Repertoire von Stimulanzen, leckt z. B. an ihrer Brustwarze, bewegt ihr Becken auf dem Schoß des bekleideten Mannes, reibt sich an ihm, um ihn so zum Höhepunkt zu bringen.“

Ergebnisse: Es zeigte sich, dass die Jugendlichen sexualisierte Darstellungen als weitgehend normal auffassten und differenziert bewerteten. Jungen störten sich weniger an der Quantität, sondern eher an der Qualität der Darstellungen. Mädchen kritisierten das ungleiche Geschlechterverhältnis, d. h. die Überzahl weiblicher Körper.

- Der Interviewer bzw. die Interviewerin sollte zwar den vorbereiteten Leitfaden nutzen, aber die Reaktionen der Befragten nicht beeinflussen. Die Interviewführung sollte **nondirektiv** sein und es den Befragten

ermöglichen, ihre persönlichen Interpretationen der Stimulus-Situation zu geben.
- Das fokussierte Interview sollte **situationsspezifisch** geführt werden. Wichtig ist es herauszufinden, welche Bedeutung die befragte Person einzelnen Teilen oder Elementen der untersuchten Situation beimisst oder welche Empfindungen sie bei ihr auslösen (Aufforderung zur **„retrospektiven Introspektion“** etwa durch die Frage „Wenn Sie zurückdenken, was war Ihre Reaktion bei diesem Teil des Films?“).
- Der Interviewablauf sollte für **unerwartete Reaktionen** Raum lassen und diese aufgreifen. Die vom Interview-Leitfaden abweichenden Gesprächsteile sind besonders geeignet, neue Hypothesen über die Wirkungsweise oder die Art der Situationsverarbeitung aufzustellen.
- Das fokussierte Interview sollte **„tiefgründig“** geführt werden: Der Interviewer sollte sich bemühen, über die Kennzeichnung affektiver Reaktionen als positiv oder negativ hinausgehend ein Höchstmaß an **„selbstenthüllenden“ Kommentaren** zu erhalten. Dies kann entweder durch direkte Fragen nach Gefühlen (z. B.: „Was empfanden Sie bei dieser Filmszene?“ oder „Wie ging es Ihnen dabei?“) oder durch die Wiederholung von Gefühlsäußerungen der Befragten durch die Interviewenden geschehen, die die Befragten implizit auffordern, weitere Emotionen zu äußern.

Qualitatives Gruppeninterview

Neben qualitativen Einzelinterviews sind auch qualitative Gruppenbefragungen durchführbar, die auf ökonomische Weise die Positionen mehrerer Befragungspersonen ermitteln und gleichzeitig Einblicke in die Gruppendynamik erlauben. Zu unterscheiden sind das unstrukturierte bzw. halbstrukturierte Paar- oder Gruppeninterview einerseits von der Methode der Fokusgruppen-Diskussion andererseits.

Unstrukturiertes bzw. halbstrukturiertes Paar- oder Gruppeninterview

Einzelinterviews sind üblicher als Gruppeninterviews. Selbst wenn in qualitativen Befragungsstudien natürliche Gruppen befragt werden (z. B. Freundespaare, Geschwisterpaare, Arbeitsteams), so finden die Interviews meist als Einzelinterviews separat voneinander statt, um eine wechselseitige Beeinflussung zu vermeiden und die Sichtweisen der Individuen unverzerrt zu erheben. Zuweilen ist man aber auch daran interessiert, die Sichtweisen mehrerer Beteiligter gleichzeitig zu hören, sei es in einem unstrukturierten oder halbstrukturierten Interview.

Unstrukturiertes bzw. halbstrukturiertes Paar- oder Gruppeninterview – Bei einem unstrukturierten bzw. halbstrukturierten Paar- oder Gruppeninterview („joint couple interview“, „paired interview“, „group interview“) werden zwei oder mehr Befragungspersonen gleichzeitig interviewt. Dabei kommt entweder kein Interviewinstrument (unstrukturiertes Interview) oder ein Leitfaden (halbstrukturiertes Interview) zum Einsatz.

Es gibt vor allem vier Gründe, aus denen ein qualitatives Paar- oder Gruppeninterview dem Einzelinterview vorgezogen werden kann:

1. **Natürlichkeit der Gesprächssituation**: Insbesondere bei unstrukturierten, ethnografischen Feldinterviews sind oftmals mehrere Personen vor Ort, die spontan gemeinsam mit der Feldforscherin oder dem Feldforscher interagieren. Man denke an die im Abschnitt zum ethnografischen Feldinterview zitierte dänische Studie zum Ernährungsverhalten von Jugendlichen in Sportzentren. Hier wurden die Jugendlichen typischerweise in ihren Freundesgruppen an den Cafeteria-Tischen angetroffen (Sylow & Holm, 2009). Es wäre künstlich gewesen, Individuen für ein Feldinterview zu separieren. In der Freundesgruppe ist die Gesprächsatmosphäre entspannter. Gleichzeitig lässt sich das Gruppenverhalten beobachten.
2. **Forschungsökonomie**: Will man im Zuge einer explorativen Vorstudie verschiedene Sichtweisen zum Thema sammeln, so kann ein nicht- oder halbstrukturiertes Gruppeninterview von Betroffenen oder Experten der Forschungsökonomie entgegenkommen, da im Vergleich zu Einzelinterviews Zeit gespart wird.
3. **Direktes Aufdecken von Widersprüchen**: Befragt man Elternpaare, wie sie sich bei der Betreuung ihres behinderten Kindes wechselseitig unterstützen, Freundespaare, welchen Stellenwert unterschiedliche Kommunikationskanäle für die Pflege ihrer Beziehung haben oder Ehepaare, welchen Anteil sie jeweils an den Hochzeitsvorbereitungen hatten, so ist durchaus mit Widersprüchen in den Aussagen der Individuen zu rechnen. Da es relativ mühsam ist, Diskrepanzen anhand der Transkripte separater Einzelinterviews aufzudecken und angemessen zu interpretieren, kann ein Paarinterview von Vorteil sein, in dem abweichende Auffassungen sofort zutage treten und von beiden Interviewpartnern kommentiert werden können. Während Widersprüche, die sich auf Fakten beziehen (z. B. exakte Datierung bestimmter Ereignisse) meist auflösbar sind, können Diskrepanzen in Einstellungen und Bewertungen nebeneinander bestehen bleiben.
4. **Erfassung gemeinsamer Sichtweisen:** Insbesondere wenn es darum geht zu untersuchen, welche dezidiert gemeinsame „Wir“-Perspektive ein Geschwister-

oder Liebespaar entwickelt hat, sind Paarinterviews indiziert. Paarinterviews können z. B. folgendermaßen eingeleitet werden (Bennett & McAvity, 1994, S. 94): „Jede Frage richtet sich an Sie beide. Sie selbst entscheiden, welche Fragen Sie jeweils beantworten möchten. Wenn nur einer von Ihnen etwas sagt, gehe ich davon aus, dass Sie übereinstimmen. Wenn Sie also anderer Auffassung sind, sagen Sie es bitte ausdrücklich."

Fokusgruppen-Diskussion

Bei der Fokusgruppen-Diskussion handelt es sich um eine spezielle Technik der qualitativen Gruppenbefragung, die sich aus dem fokussierten Interview (Merton, Fiske, & Kendall, 1990) heraus entwickelt, aber eine eigenständige Form angenommen hat. Fokusgruppen-Diskussionen werden sowohl in der akademischen Forschung als auch in der Marktforschung seit den 1980er-Jahren eingesetzt (Blank, 2007). Von der Methode der Fokusgruppen-Diskussion abzugrenzen sind weitere Varianten der Gruppendiskussion mit Wurzeln in der psychoanalytischen Sozialpsychologie (Leithäuser & Volmerg, 1988) oder verankert in der in der Soziologie etablierten dokumentarischen Methode (Bohnsack, Przyborski, & Schäffer, 2010).

Fokusgruppen-Diskussion – Eine Fokusgruppen-Diskussion („focus group interview", „focus group discussion") ist eine halbstrukturierte Gruppendiskussion zu einem konkreten Thema oder Produkt, das als Diskussionseinstieg präsentiert wird (Grundreiz). Die Teilnehmerzahl pro Gruppe liegt in der Regel bei 4–8 Personen. Pro Studie werden meist rund 4–8 Fokusgruppen mit einer Länge von jeweils 90–120 Minuten im Forschungslabor durchgeführt. Es wird einerseits eine Selbstläufigkeit der Gruppendiskussion angestrebt, gleichzeitig sorgt eine Moderatorin bzw. ein Moderator („moderator", „facilitator") dafür, dass die Gruppe beim Thema bleibt und vorher festgelegte Leitfragen (Diskussionsleitfaden) bearbeitet werden.

Ablauf und Besonderheiten der Fokusgruppen-Methode lassen sich anhand der **vier zentralen Arbeitsphasen** beschreiben (Blank, 2007; Bohnsack et al., 2010; Dammer & Szymkowiak, 2008; Krueger & Casey, 2008; Lamnek, 2005):

1. **Planung der Fokusgruppen-Diskussion**: Die Vorbereitung einer Fokusgruppen-Diskussion ist aufwändiger als die eines Einzelinterviews: Ein **geeigneter Grundreiz** muss ausgewählt, ein **stimmiger Diskussionsleitfaden** zusammengestellt und beides praktisch erprobt werden. Diskussionsleitende müssen ausgewählt und geschult werden. Vor allem aber müssen geeignete Teilnehmende rekrutiert und sinnvoll in Kleingruppen eingeteilt werden. Meist wird mit künstlich zusammengestellten **Ad-hoc-Gruppen** gearbeitet, teilweise mit **natürlichen Gruppen** einander bereits bekannter Personen (z. B. Studierende desselben Studienganges, Mitarbeitende derselben Firma). Beides kann Vor- und Nachteile haben: Unbekannte haben sich möglicherweise mehr zu erzählen, während unter Bekannten potenziell eine entspanntere Atmosphäre herrscht. Weiterhin ist bei der Gruppenzusammenstellung zu entscheiden, welche Personenmerkmale für das Thema relevant sind und inwiefern mit **homogenen** und/oder **heterogenen Gruppen** (z. B. hinsichtlich Geschlecht, Alter, Status) gearbeitet werden soll. Da Fokusgruppen im Labor stattfinden und die Teilnehmenden auf Einladung anreisen, ist mit einer gewissen Zahl von **Ausfällen** („no-shows") zu rechnen, so dass in der Regel jede Gruppe im Vorfeld etwas überbucht wird.
2. **Durchführung der Fokusgruppen-Diskussion**: Die Fokusgruppe beginnt mit einer **Einführung**, in der die Moderierenden sich und die Studie vorstellen, den Ablauf und die Spielregeln erklären (Datenschutz, Dauer der Fokusgruppe, kein gleichzeitiges Reden etc.). Zudem werden die Gruppenmitglieder einander bekannt gemacht, vorbereitete Namensschilder aufgestellt und ggf. Getränke und Snacks angeboten. Den eigentlichen Einstieg in die Thematik bietet der **Grundreiz** (provokante Aussage, Filmclip, Produktbeispiel, Infografik), der einen Diskussionsanstoß geben soll. Ziel ist eine **Selbstläufigkeit der Diskussion** ähnlich wie in alltäglichen Gruppengesprächen. Die Moderierenden nehmen eine neutrale Haltung ein. Sie greifen nur dann behutsam in das Kommunikationsgeschehen ein, wenn sich die Diskussion zu stark vom Thema entfernt oder wenn von der Gruppe bereits angesprochene Aspekte durch Beschreibungen und Erzählungen vertieft werden sollen. Die Diskussionsleitung kann die Diskussion steuern, indem sie Aussagen zusammenfasst, infrage stellt, überspitzt, nach weiteren oder anderen Meinungen fragt. Bei Bedarf können weitere **Reizargumente** ähnlich dem Grundreiz gegeben werden. Wenn die Gruppe das Thema erschöpfend diskutiert hat, werden in einer **Nachfrage-Phase** Themenaspekte von der Diskussionsleitung angesprochen, die noch nicht zur Sprache kamen, aber laut Diskussionsleitfaden relevant sind. Hier ist ein starres Frage-Antwort-Schema ebenso zu vermeiden wie das direkte Ansprechen einzelner Personen. Vielmehr bemüht man sich, erneut Diskussionen innerhalb der Gruppe anzustoßen. In der **Abschluss-Phase** kann die Gruppe mit Widersprüchen und Inkonsistenzen konfrontiert und nach weiteren, bisher vernachlässigten Aspekten gefragt werden. Die Fokusgruppe endet mit Dank und **Verabschiedung**. Entweder am Anfang oder am Ende kann man die Teilnehmenden noch einen **Kurzfragebogen** ausfüllen lassen, um ihre soziodemografischen Merkmale

sowie Vorerfahrungen mit dem Untersuchungsthema festzuhalten.

3. **Dokumentation der Fokusgruppen-Diskussion**: Fokusgruppen werden vollständig aufgezeichnet, wobei ergänzend zur **Audioaufnahme** (mehrere sehr gute Mikrofone) auch eine **Videoaufnahme** (ggf. mit zwei Kameras) sinnvoll ist, um die Wortbeiträge eindeutig den verschiedenen Personen zuordnen zu können. Pro Fokusgruppendiskussion ist mit ca. 60–100 Seiten **Transkript** zu rechnen. Bei einer Studie mit 10 Fokusgruppen kommen dementsprechend rund 600–1 000 Seiten an verbal-schriftlichem Datenmaterial zusammen. Entsprechender Zeit- und Kostenaufwand für die Transkription sind einzuplanen (für 1 Stunde Diskussionszeit ist mit ca. 8–10 Stunden Transkriptionszeit zu rechnen). Fokusgruppen finden in der Regel in **Forschungslabors mit Einwegscheiben** statt (diese sind von innen verspiegelt und von außen transparent), so dass der Diskussionsablauf von einem weiteren Mitglied des Forschungsteams (ggf. auch vom Kunden oder Auftraggeber) von außen live beobachtet und mit **Notizen** erfasst werden kann. Auf diese Weise wird die spätere Auswertung beschleunigt bzw. eine Vorabanalyse ermöglicht. Die Fokusgruppenteilnehmenden müssen aus forschungsethischen Gründen über diese Umstände aufgeklärt werden und ihr Einverständnis geben (► Abschn. 4.1.1).
4. **Auswertung der Fokusgruppen-Diskussion**: Die **qualitative Analyse** der Diskussionstranskripte läuft meist darauf hinaus, die **verschiedenen Erfahrungen und Meinungen** der Gruppenmitglieder zum fokussierten Thema zu sammeln. Seltener werden auch der Diskussionsprozess und die **Gruppendynamik** ausdrücklich analysiert (Farnsworth & Boon, 2010). Während in grundlagenwissenschaftlichen Studien divergierende und kontroverse Meinungen in der Gruppe als Ergebnis nützlich sein können, zielt der Einsatz in der Marktforschung stärker darauf ab, konkrete Entscheidungen (z. B. im Rahmen der Produktentwicklung) treffen zu können, so dass vor allem auch Konsenspositionen herauszuarbeiten sind.

Das **Spektrum der Themen**, zu denen Fokusgruppen durchgeführt werden, ist breit und umfasst z. B. die

- Arbeitsforschung (z. B. Stresserleben bei der Arbeit mit lernbehinderten Menschen; Raczka, 2005),
- Gesundheitsforschung (z. B. Barrieren bei der Behandlung von Männern mit Depression; Rochlen et al., 2010)
- Medienforschung (z. B. intensivere Nutzung von bewegungsfördernden „aktiven Videospielen“ bei Kindern; Dixon et al., 2010),
- Kriminalitätsforschung (z. B. Anforderungen straffälliger Mädchen an das Justizsystem; Garcia & Lane, 2009) oder
- Konsumforschung (z. B. Akzeptanz von als gesundheitsförderlich beworbenen „funktionellen Lebensmitteln“; Barrios, Bayarri, Carbonell, Izquierdo, & Costell, 2008).

Hinweise zur Durchführung von Fokusgruppen mit **besonderen Zielgruppen** wie Kindern und Jugendlichen (Clark, 2009) oder hörbeeinträchtigen Personen (Balch & Mertens, 1999) liegen vor.

Fokusgruppen können nicht nur mündlich, sondern auch computervermittelt als **Online-Fokusgruppen** („online focus groups“) durchgeführt werden, entweder zeitversetzt über Webforen oder zeitgleich via Online-Chat oder Online-Videokonferenz. Dadurch können Teilnehmende einbezogen werden, die für eine Live-Diskussion vor Ort nicht erreichbar sind. Der zeitversetzte Austausch über mehrere Tage und Wochen hinweg geht jedoch auf Kosten der Kohärenz der Diskussion, zudem sinkt im Laufe der Zeit die Teilnehmerzahl (Rezabek, 2000).

10.2.3 Quantitatives Interview

Vollstrukturierte Befragungen, mit deren Hilfe Selbstauskunftsdaten zu Verbreitung und Intensität einzelner Merkmale von **sehr großen Stichproben** gewonnen werden können, spielen in der Markt- und Meinungsforschung, der Wahlforschung sowie der amtlichen Statistik eine wichtige Rolle. Hier werden **bevölkerungsrepräsentative Daten** gebraucht, die nur durch entsprechend umfassende sowie nach Zufallsprinzipien ausgewählte Stichproben gewonnen werden können. Während der Stichprobenumfang bei qualitativen Interviewstudien meist im ein- bis zweistelligen Bereich rangiert, fällt er bei quantitativen Umfragen oft vier- bis fünfstellig aus. Auch basieren quantitative Interviewstudien häufiger auf **probabilistischen Stichprobenverfahren**, qualitative dagegen auf nicht-probabilistischen Samples (zu Stichprobenarten ► Kap. 9).

Quantitatives bzw. (voll)strukturiertes Interview – Das quantitative bzw. (voll)strukturierte Interview („quantitative interview“, „structured interview“) basiert auf einem (voll)standardisierten Interview-Fragebogen („interview questionnaire“, „survey questionnaire“, „interview schedule“), der den Befragungspersonen von den Interviewenden präsentiert wird, die auch die Antworten registrieren (intervieweradministrierter Fragebogen, „interviewer administered questionnaire“, „interviewer administered survey“). Der vollstandardisierte Interviewfragebogen besteht aus geschlossenen Fragen mit klaren Antwortvorgaben, die in exaktem Wortlaut und festgelegter Reihenfolge zu bearbeiten sind. Vollstandardisierte Interviews können persönlich, telefonisch oder online geführt

werden. Eine Ergänzung des standardisierten Instruments um einige offene Fragen ist möglich.

Vollstrukturierte bzw. quantitative Interviews eignen sich für **klar umgrenzte Themenbereiche**, über die man detaillierte Vorkenntnisse besitzt. Quantitative Interviews mit verschiedenen Personen sind direkt vergleichbar: Die jeweiligen Messwerte lassen sich angesichts der standardisierten Antwortvorgaben komfortabel zu Aggregatwerten zusammenführen und **deskriptiv- sowie inferenzstatistisch** auswerten.

Während bei qualitativen Interviews die Hauptarbeit in der interpretativen Auswertung der Transkripte besteht, muss bei quantitativen Interviews besonders viel Arbeitszeit in die sorgfältige **Entwicklung und Erprobung des vollstandardisierten Fragebogeninstruments** gelegt werden, damit die erhobenen Messwerte tatsächlich nachweislich die interessierenden Merkmalsausprägungen valide und reliabel erfassen. Zudem ist der Aufwand für die Datenerhebung bei quantitativen Interviewstudien proportional zum viel größeren Stichprobenumfang ebenfalls höher.

Vollstrukturierte Interviews entlasten die Befragungspersonen – im Vergleich zur Fragebogen-Erhebung – vom eigenständigen Lesen und Ausfüllen des standardisierten Fragebogens und erlauben im Falle von Verständnisproblemen Rückfragen. Die soziale Interaktion mit den Interviewenden reduziert im Vergleich zur schriftlichen Fragebogenmethode Abbruchquoten, führt zu sorgfältigerer Beantwortung, kann jedoch sozial erwünschtes Antworten steigern.

Im Folgenden werden zunächst der allgemeine Ablauf einer vollstrukturierten mündlichen Befragung beschrieben und die in der quantitativen Umfrageforschung relevanten Umfragefehler behandelt. Anschließend wird ein kursorischer Vergleich der verschiedenen Interview-Modi vorgenommen, bevor das persönliche, das telefonische und das computervermittelte vollstrukturierte Interview jeweils separat dargestellt werden.

Ablauf eines quantitativen Interviews

Die Arbeitsschritte bei quantitativen bzw. vollstrukturierten Interviews weichen von denen bei qualitativen bzw. nicht- oder halbstrukturierten Interviews teilweise deutlich ab. Fragen der Rekrutierung und Kontaktierung der Befragungspersonen sind besonders zu beachten, weil in der quantitativen Umfrageforschung häufig probabilistische Stichproben (► Abschn. 9.3.2) angestrebt werden:

1. **Inhaltliche Vorbereitung**: Wonach soll gefragt werden? Die Konstruktion des vollstandardisierten Interview-Fragebogens ist ein langwieriger Prozess, der mehrere Schleifen durchläuft und wesentlich arbeitsaufwändiger ist als die Zusammenstellung eines Leitfadens für ein halbstrukturiertes qualitatives Interview. Grundlage der Konstruktion des standardisierten Interview-Fragebogens sind Theorien zum Untersuchungsthema bzw. Forschungsproblem, explorative Voruntersuchungen und etablierte Fragebögen sowie die methodischen Regeln der Fragebogenkonstruktion. Da standardisierte Interview-Fragebögen sich nicht grundlegend von standardisierten Fragebögen unterscheiden, die von den Befragten selbst ausgefüllt werden, sei für Details zur Fragebogenkonstruktion auf ► Abschn. 10.3.3 verwiesen. Bei der Vorbereitung ist auch zu klären, wer befragt werden soll. Die Stichprobenziehung (► Abschn. 9.3) ist sorgfältig zu planen.
2. **Organisatorische Vorbereitung**: Hier sind Interviewerschulung sowie Vorbereitung und Wartung der technischen Hilfsmittel (Laptops, Telefonanlage, Software) angesprochen.
3. **Rekrutierung der Befragungspersonen**: In der quantitativen Umfrageforschung zielt man oft auf bevölkerungsrepräsentative Stichproben ab. Dabei werden unterschiedliche Auswahltechniken eingesetzt, je nachdem, ob man die Befragten für ein persönliches Interview zu Hause aufsuchen oder telefonisch per Festnetztelefon oder Handy interviewen möchte.
4. **Kontaktierung der Befragungspersonen**: Ein motivierender Erstkontakt ist besonders wichtig, um möglichst viele der angezielten Befragungspersonen tatsächlich für eine Teilnahme an der Studie zu gewinnen. Zudem spielt die Zeitplanung eine wichtige Rolle, so dass Terminvereinbarungen und Terminverschiebungen für die Befragten so komfortabel wie möglich zu gestalten sind.
5. **Gesprächsbeginn**: Nach der Begrüßung wird der Ablauf des Interviews (z. B. Zeitdauer) erläutert, auf die Bedeutung der Studie hingewiesen sowie Anonymität zugesichert. Auf das Ausschalten möglicher Störfaktoren wird hingewirkt (z. B. laufender Fernseher im Hintergrund).
6. **Durchführung und Aufzeichnung des Interviews**: Der Interviewverlauf folgt exakt dem standardisierten Fragebogeninstrument. Lediglich bei akustischen oder Verständnisproblemen wird die Frage im Wortlaut wiederholt. Eine Antwortverweigerung bei einzelnen Fragen wird zugelassen. Die von der Befragungsperson gewählten Antwortalternativen werden von den Interviewenden in den Fragebogen eingetragen, entweder auf Papier oder elektronisch.
7. **Gesprächsende**: Im Unterschied zum qualitativen Interview, an dessen offizielles Ende sich eine oft recht ausführliche Nachbesprechung anschließt, wird das standardisierte Interview nach der letzten Frage beendet. Allenfalls werden bei Interesse der Befra-

gungspersonen jetzt Hintergründe zur Studie erklärt („debriefing“) oder es wird versucht, die Befragungsperson für Folgestudien gewinnen.

8. **Verabschiedung**: Mit der Verabschiedung wird ein Dank für die Teilnahme verbunden. Die befragten Personen sollten in einer Stimmung verabschiedet werden, in der sie grundsätzlich zu weiteren wissenschaftlichen Interviews bereit sind.
9. **Gesprächsnotizen**: Während des Interviews unerwartet auftretende Störungen oder Beeinträchtigungen sowie Auffälligkeiten und Besonderheiten sind in einem Interviewprotokoll festzuhalten.
10. **Zusammenführung und Bereinigung des Datensatzes**: Insbesondere bei größeren vollstrukturierten Befragungsstudien, in denen mehrere Interviewende Daten erheben, müssen die verschiedenen Dateien mit den quantitativen Datensätzen zusammengeführt und bereinigt werden (z. B. Ausschluss von sehr unvollständig beantworteten oder abgebrochenen Interviews; zur Datenaufbereitung ► Kap. 11).
11. **Statistische Analyse**: Der bereinigte Datensatz wird einer deskriptivstatistischen Analyse unterzogen (z. B. Berechnung von Häufigkeitsverteilungen, Mittelwerten). Im Falle repräsentativer Stichproben wird eine statistische Schätzung von Populationsparametern mittels Konfindenzintervallen vorgenommen. Im Falle von Forschungshypothesen werden diese mithilfe statistischer Signifikanztests überprüft (zur quantitativen Datenanalyse ► Abschn. 12.2). Umfangreichere Interview-Studien können dabei Material für mehrere Auswertungen und Publikationen liefern (zur Ergebnispräsentation ► Kap. 13 und zur Präsentationsqualität ► Abschn. 3.1.2). Aus wissenschaftsethischen Gründen ist es wichtig, im Forschungsbericht zu dokumentieren, welche Variablen einer Interviewstudie in die jeweilige Auswertung eingehen und welche nicht (► Abschn. 4.2). Denn eine begründete thematische Auswahl muss abgrenzbar sein von dem manipulativen Herausgreifen der „besten“ Befunde bei gleichzeitigem Verschweigen von hypothesenkonträren Negativ-Ergebnissen im selben Datensatz.
12. **Archivierung des Materials**: Datensätze müssen intern archiviert werden, damit die Studie nachvollziehbar ist. Will man der Scientific Community darüber hinaus seinen Datensatz für Sekundäranalysen zur Verfügung stellen, so kann man ihn bei entsprechenden Archivdiensten anmelden bzw. einreichen (z. B. http://www.zpid.de; http://www.gesis.org; http://www.cessda.org). Zudem existieren Fachzeitschriften, die routinemäßig zusammen mit den Fachartikeln die der Studie zugrunde liegenden Datenerhebungsinstrumente und Datensätze publizieren (z. B. *Archives of Scientific Psychology*).

Umfragefehler und Gegenmaßnahmen

Quantitative Interviews können fehlerhafte Ergebnisse hervorbringen, etwa aufgrund von Merkmalen oder Verhaltensweisen der Interviewenden (z. B. mangelndes Zuhören) sowie der Befragungspersonen (z. B. Selbstdarstellungsverhalten). Weitere Fehlerquellen finden sich im Interviewinstrument (z. B. mangelnde Verständlichkeit der Fragen oder der Antwortalternativen im Interview-Fragebogen) sowie in der Interviewsituation (z. B. Störungen und Ablenkungen am Ort des Interviews, Zeitdruck). Diese **Interviewfehler** betreffen grundsätzlich sowohl qualitative als auch quantitative Interviews und können durch entsprechende Gegenmaßnahmen reduziert werden (z. B. Interviewerschulung, bewusste Wahl von Orten und Zeiten für Interviews etc.).

In der quantitativen Umfrageforschung steht die Zielsetzung im Mittelpunkt, die Ausprägung bestimmter Merkmale in einer Population (z. B. durchschnittliche Fernsehdauer, Parteipräferenz, Intensität von Ausländerfeindlichkeit) möglichst genau anhand der Daten einer repräsentativen Stichprobenbefragung zu schätzen. Das Konzept des **Total Survey Error** (Gesamtfehler einer quantitativen Umfrage) unterscheidet **fünf Arten von Umfragefehlern**, deren Ausmaß prinzipiell quantifizierbar ist und in die Qualitätsbewertung einer Umfrage eingehen sollte (Weisberg, 2005): 1. Abdeckungs-, 2. Nonresponse-, 3. Stichproben-, 4. Mess- und 5. Datenverarbeitungsfehler.

1. **Abdeckungsfehler** („coverage error“): Als **angezielte Grundgesamtheit** („target population“) bezeichnen wir alle Elemente, über die wissenschaftliche Aussagen getroffen werden sollen (z. B. volljährige Einwohnerinnen und Einwohner von Deutschland). Meist ist eine Vollerhebung nicht möglich, stattdessen wird eine Stichprobe gezogen. Dafür muss ein **Auswahlrahmen** („sample frame“) definiert werden, aus dem die Elemente für die Stichprobe per Zufall gezogen werden können (z. B. Zusammenstellung von Wohnadressen der Bevölkerung für persönliche Interviews, Zusammenstellung von Telefonnummern der Bevölkerung für Telefoninterviews). Alle Elemente im Auswahlrahmen bilden die **Auswahlgesamtheit** („frame population“). Idealerweise sollte die Auswahlgesamtheit exakt der angezielten Grundgesamtheit entsprechen bzw. diese genau abdecken.
   In der Praxis ist jedoch eine Abweichung typisch, d. h., meist enthält die Auswahlgesamtheit nicht alle Elemente der Grundgesamtheit, deckt diese also nur teilweise ab (**Unterabdeckungs-Fehler**, „undercoverage“). Dieser Unterabdeckungs-Fehler führt zu verzerrten Ergebnissen, da sich die nichterfassten Personen nicht zufällig verteilen, sondern ihre Charakteristiken einer Systematik folgen. Telefonumfragen,

die mit Festnetznummern als Auswahlrahmen arbeiten, führen z. B. zu einem Unterabdeckungs-Fehler bei allen Personen, die überhaupt keinen Festnetzanschluss (mehr) haben, sondern nur noch ein Mobiltelefon nutzen. Persönliche Umfragen, die mit einem Auswahlrahmen privater Wohnadressen arbeiten, müssen eine Nichtabdeckung für all jene Personen konstatieren, die über keine private Wohnadresse in Deutschland verfügen, etwa weil sie sich im Pflegeheim, im Gefängnis, im Kloster, in der Kaserne, im beruflichen Auslandsaufenthalt etc. befinden (Schnell, 1991).

Abdeckungsfehler können auch darin bestehen, dass Personen im Auswahlrahmen berücksichtigt werden, die gar nicht zur angezielten Grundgesamtheit gehören (**Überabdeckungs-Fehler**, „overcoverage").

Eine gute Umfrage definiert stets präzise die angezielte Grundgesamtheit und arbeitet mit einem Auswahlrahmen bzw. einer Auswahlgesamtheit, die sich durch möglichst vollständige Abdeckung auszeichnet. Unvermeidbare Abdeckungsfehler sollten in der Studie ausdrücklich benannt und in ihren Konsequenzen für das Studienergebnis diskutiert werden. Verbreitete Auswahlrahmen für persönliche Interviews und telefonische Interviews werden in den entsprechenden Abschnitten vorgestellt.

Da verschiedene Auswahlrahmen, die für unterschiedliche Interview-Modi genutzt werden, jeweils unterschiedliche Abdeckungsfehler mit sich bringen, versucht man, diese teilweise durch eine Kombination von Frames bzw. durch **Mixed-Mode-Befragungen** zu kompensieren. Damit ist gemeint, dass man Interviews mit unterschiedlichem Befragungs-Modus („survey modes") miteinander kombiniert, also einen Teil der Stichprobe z. B. mündlich oder postalisch (anhand von Wohnadressen als Auswahlrahmen), einen anderen Teil fernmündlich (anhand von Festnetz- und Mobilfunknummern als Auswahlrahmen) befragt, um eine möglichst breite Abdeckung der gesamten Zielpopulation zu gewährleisten. Hierbei ist jedoch durch entsprechende Techniken sicherzustellen, dass nicht dieselben Personen mehrfach ausgewählt werden bzw. die Auswahlwahrscheinlichkeiten alle Elemente der Grundgesamtheit sich nicht verschieben.

2. **Nonresponse-Fehler** („nonresponse error"): Selbst wenn ein weitgehend vollständiger und unverzerrter Auswahlrahmen für eine quantitative Umfrage vorliegt (z. B. Liste aller Adressen privater Haushalte einer Stadt aus dem Einwohnermeldeamtsregister), aus dem mit Zufallsverfahren eine Auswahl zu kontaktierender Haushalte getroffen wurde, so ist damit noch keine repräsentative Stichprobe gesichert. Denn Elemente der angezielten Stichprobe können zum einen nicht erreichbar sein (z. B. dem Interviewer wird die Haustür nicht geöffnet), zum anderen eine Umfrageteilnahme ausdrücklich verweigern (z. B. aufgrund von Zeitmangel oder fehlendem Vertrauen in das Forschungsinstitut). Die am Ende befragte Stichprobe (**Netto-Stichprobe**) weicht also von der ursprünglich geplanten Stichprobe (**Brutto-Stichprobe**) durch **Nichterreichbarkeit** und **Teilnahmeverweigerung** („unit non-response") ab. Die Relation zwischen Netto- und Brutto-Stichprobe wird als **Ausschöpfungsrate** („response rate") bezeichnet. Als Faustregel wird davon ausgegangen, dass geringere Ausschöpfungsraten mit einer geringeren Repräsentativität der Stichprobe für die Grundgesamtheit einhergehen. Ausschöpfungsraten sollten durch wiederholte Kontaktversuche (bei Nichterreichbarkeit) sowie durch Motivierung zur Teilnahme (bei Teilnahmeverweigerung) erhöht werden, was mit Aufwand verbunden ist.

In der Umfragepraxis wird relativ viel Aufwand betrieben, um Nonresponse möglichst gering bzw. die Ausschöpfungsrate möglichst hoch zu halten. Bei **Nichterreichbarkeit** können neben **wiederholten Kontaktversuchen** notfalls auch sog. **Proxy-Interviews** zum Einsatz kommen (eine dritte Person wird zu Angaben über die Zielperson gebeten; z. B. ein im Haushalt angetroffenes Elternteil macht Angaben zum Freizeitverhalten des Kindes). Dies ist nur dann sinnvoll, wenn begründet werden kann, dass Fehler durch den Interviewausfall die Studie stärker beeinträchtigen würden als mögliche Fehlereffekte durch ein Proxy-Interview. Eine weitere Strategie, um schwer erreichbare Respondenten zu kontaktieren, besteht im **Wechsel des Interview-Modus** (wer für ein persönliches Interview nicht zur Verfügung steht wird telefonisch kontaktiert) oder gar des Befragungs-Modus (wer für ein Interview nicht erreichbar wird, erhält eine Einladung zu einer schriftlichen Fragebogenerhebung).

Soll Nonresponse aufgrund von **Antwortverweigerung** reduziert werden, müssen zunächst die Gründe für die mangelnde Teilnahmebereitschaft eruiert und auf dieser Basis Gegenmaßnahmen eingeleitet werden. Generell wird eine nachlassende Bereitschaft zur Teilnahme an wissenschaftlichen Umfragen in der Bevölkerung beobachtet. Es existieren eine Reihe von mehr oder minder gut empirisch gesicherten **Umfrage-Theorien**, die erklären, was Menschen zur Teilnahme bzw. Nicht-Teilnahme an einer wissenschaftlichen Umfrage bewegt (Häder, 2006, S. 192ff.): So kann die Entscheidung für oder gegen die Umfrageteilnahme Ergebnis einer **rationalen Abwägung** von Kosten (z. B. Zeitaufwand gemäß Dauer des Inter-

views; Angst vor Verletzung der Privatsphäre, wenn der Interviewer die Wohnung betritt etc.) und Nutzen (z. B. Hoffnung auf ein angenehmes Gespräch, Zeitvertreib, Möglichkeit der Meinungsäußerung, finanzielle Aufwandsentschädigung) sein. Es können aber auch **sozialpsychologische Austauschprozesse** in der Weise eine Rolle spielen, dass auf eine Vorleistung von Seiten der Forschenden (z. B. vorab gezahlte Aufwandsentschädigung/Incentivierung, Eingehen auf individuelle Terminwünsche) mit größerer Antwortbereitschaft reagiert wird. Gemäß der **Theorie geplanten Handelns** („theory of planned behavior") spielen a) die Einstellung gegenüber Befragungen (z. B. Überzeugung, dass Umfragen dem Erkenntnisfortschritt dienen), b) die subjektive Norm (z. B. Überzeugung, dass Bezugsgruppen die Umfrageteilnahme erwarten) und c) die wahrgenommene Kontrolle in einer Befragungssituation (z. B. Erwartung eines geringen Schwierigkeitsgrades der gestellten Fragen) eine zentrale Rolle für die Teilnahmebereitschaft.

Zu beachten ist, dass eine Steigerung der Ausschöpfungsrate nicht nur durch Maßnahmen bei den potenziellen Befragungspersonen, sondern auch durch **Maßnahmen bei den Interviewenden** erreicht werden kann (z. B. umfassende Behandlung des Problems der Interviewverweigerung bei der Schulung der Interviewenden; Art und Umfang der Bezahlung und Incentivierung der Interviewenden; Menge der durchzuführenden Interviews; Zeitdruck und Arbeitslast der Interviewenden etc. (zur Nonresponse-Problematik s. Proner, 2011).

Durch Nichterreichbarkeit und Teilnahmeverweigerung **reduziert sich der Netto-Stichprobenumfang**, was ggf. durch die Befragung zusätzlicher Personen kompensiert werden kann. Methodisch problematischer sind Nichterreichbarkeit und Teilnahmeverweigerung, wenn die betreffenden Personen systematisch in für das Umfragethema relevanten Merkmalen von den übrigen Befragungspersonen abweichen. Wenn die Nichterreichbaren und Antwortverweigerer Eigenschaften, Einstellungen oder Gewohnheiten haben, die sie von den sonstigen Befragten abheben, so resultiert eine **verzerrte Stichprobe** („biased sample"). Allein der Hinweis, bei einer Studie sei eine hohe Ausschöpfungsrate erzielt worden, spricht also nicht automatisch für eine gute **Repräsentativität der Stichprobe** (► Abschn. 9.1.3). Ein pures Hochschrauben der Antwortrate durch Befragung weiterer erreichbarer und befragungswilliger Respondenten fügt der Stichprobe keinerlei Informationen über die Gruppe der Nichterreichbaren und Antwortverweigerer hinzu. Ebenso kann eine **Gewichtung der Stichprobe** einen Bias nicht immer ausgleichen (► Abschn. 11.4.4 „Gewichtung von Stichproben"). Bei größeren Umfragen sollte neben der Angabe der Ausschöpfungsrate immer auch diskutiert und ggf. in einer begleitenden Methodenstudie erhoben werden, welche Merkmale Nichterreichbare und Antwortverweigerer haben und inwiefern mit einer Stichprobenverzerrung hinsichtlich bestimmter Merkmale zu rechnen ist.

Weiterhin ist zu beachten, dass auch befragungswillige Personen oft einzelne Fragen unbeantwortet lassen („item nonresponse"), so dass sog. **Missings** (einzelne fehlende Werte bei einer Befragungsperson) zustande kommen. Problematisch sind hier vor allem Situationen, in denen viele fehlende Werte vorliegen und diese einem systematischen Muster folgen (► Abschn. 11.4.4 „Behandlung fehlender Werte"). Die Frage nach dem Grad der Verzerrung der Stichprobendaten durch Antwortverweigerung ist also nicht pauschal, sondern differenziert für die einzelnen abgefragten Variablen zu betrachten.

3. **Stichprobenfehler** („sampling error"): Wenn Aussagen über Populationen getroffen werden sollen, aber nur Stichproben untersucht werden, so können Stichprobenkennwerte (z. B. Mittelwert einer Variable in der Stichprobe) – je nach **Größe und Art der Stichprobe** – mehr oder minder gute Schätzer für die interessierenden Populationsparameter sein (z. B. Stichprobenmittelwert als Schätzer des Populationsmittelwertes). Bei Zufallsstichproben lässt sich der Stichprobenfehler als **Standardfehler** („margin of error") berechnen; er hängt maßgeblich vom Stichprobenumfang ab: Je größer der Umfang einer Zufallsstichprobe, desto genauer können Populationsparameter mithilfe von Stichprobenkennwerten geschätzt werden. Die Berechnung des Standardfehlers gehört routinemäßig zu jeder inferenzstatistischen Auswertung einer quantitativen Umfrage (z. B. Berechnung von Konfidenzintervallen, Durchführung von statistischen Signifikanztests, ► Kap. 12).
4. **Messfehler** („measurement error"): Während sich Abdeckungs-, Nonresponse- und Stichprobenfehler auf Qualitätsmängel bei der Auswahl der Befragungspersonen einer quantitativen Umfrage beziehen, adressiert der Messfehler Qualitätsmängel in der Aussagekraft der Antworten. Messfehler entstehen u. a. durch die bereits im Zusammenhang mit Interviewfehlern (► Abschn. 10.2.1) behandelten Probleme auf Seiten a) der Interviewenden (z. B. äußerliche Interviewermerkmale; zu schnelle Interviewdurchführung), b) des Fragebogeninstruments (z. B. missverständliche Fragen, unklare Antwortvorgaben, ungünstige Fragenreihenfolge) oder c) der Befragungspersonen (z. B. mangelnde Aufmerksamkeit, sozial erwünschtes Antworten). Auch die Problematik des

Tabelle 10.8 Verteilung der standardisierten Befragungsformen (Prozentangaben) bei den Mitgliedsinstituten des Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute ADM e.V. (Mod. nach ADM, 2013)

	1990	2000	2010	2013
	[%]			
Persönliches Interview	65	34	21	22
Telefon-Interview	22	41	35	36
Paper-Pencil-Fragebogen	13	22	6	6
Online-Fragebogen	–	3	38	36

Item-Nonresponse (Antwortverweigerung bei einzelnen Fragen) kann zu fehlerhaften Messwerten führen. Abgesehen von den zur Vermeidung von Interviewfehlern vorgesehenen Maßnahmen (z. B. Interviewerschulung) trägt eine sorgfältige Entwicklung und Überprüfung des Interview-Fragebogens (inklusive einer Überprüfung von dessen Gütekriterien) zur Reduktion von Messfehlern bei (zu quantitativen Fragebögen ▶ Abschn. 10.3.3).

5. **Datenverarbeitungsfehler** („postsurvey error"): Diese Fehler betreffen Defizite in der Archivierung, Codierung, Aufbereitung oder Auswertung von Umfragedaten. Sie kommen z. B. zustande, wenn eine Umpolung einzelner negativer Items vergessen wurde oder wenn sich im Datensatz Dubletten befinden (zur Aufbereitung quantitativer Daten ▶ Abschn. 11.4).

Vergleich der Interview-Modi

Betrachtet man die historische Entwicklung in der empirischen Markt- und Sozialforschung, so zeigt sich, dass das persönliche Interview seine Vormachtstellung unter den vollstrukturierten Befragungsmethoden in den letzten Dekaden verloren hat. War 1990 das strukturierte persönliche Interview mit Abstand am populärsten, so steht heute das Telefoninterview an erster Stelle, gefolgt vom Online-Fragebogen (Tab. 10.8). Dieser Wandel wird durch die **technische Entwicklung** ermöglicht (z. B. wachsende Internetverbreitung in der Bevölkerung) und durch die Möglichkeit zur **Kostensenkung** bei medienvermittelten Datenerhungen getrieben.

Im Bereich der **quantitativen Umfrageforschung** („survey research") werden die unterschiedlichen Formen der Administration standardisierter Fragebögen sowie die unterschiedlichen Kommunikationskanäle (persönlich, telefonisch, online) als **Befragungs-Modi** („survey modes") bezeichnet. Beim Interview handelt es sich um eine **intervieweradministrierte Befragung**, weil der Interviewer den Interview-Fragebogen präsentiert und die Antworten der Befragungsperson registriert. Demgegenüber handelt es sich bei der Fragebogentechnik (▶ Abschn. 10.3) um eine **selbstadministrierte Befragung**, weil die Respondenten den Fragebogen selbst lesen und eigenständig ausfüllen.

Welchen Einfluss es auf die erzielten Antworten hat, ob man dieselben geschlossenen Fragen in einem persönlichen Interview, einem telefonischen Interview, einem postalisch versendeten Fragebogen oder einem Online-Fragebogen stellt, wird als **Methoden- oder Modus-Effekt** („mode effect") bezeichnet und empirisch untersucht. Interview-Modi unterscheiden sich typischerweise hinsichtlich Abdeckungs- und Nonresponsefehlern sowie Messfehlern. So ist z. B. in einem persönlichen Interview die Interaktion zwischen den Befragten und den Interviewern sehr hoch, was zu einem starken Einfluss der Interviewenden auf das Antwortverhalten führen kann. Aus diesem Grund eignet sich diese Interviewmethode weniger für heikle Themen (Möhring & Schlütz, 2010). Demgegenüber kann der Einsatz von Befragungsmethoden ohne Face-to-Face-Kontakt (Telefoninterviews, postalisch versendete Fragebögen oder Online-Fragebögen) die Einflüsse sozialer Erwünschtheit reduzieren (Chan, Orlando, Hosh-Dastidar, Duan, & Sherbourne, 2004). Im Fall von postalischen und Online-Befragungen fehlen Informationen über die näheren Umstände der Befragungssituation (z. B. ob die Befragung ungestört durchgeführt wird oder evtl. Dritte anwesend sind). So konnten Häder & Kühne (2009) nachweisen, dass bei telefonischen Interviews in Anwesenheit dritter Personen tendenziell ein höheres Haushaltsnettoeinkommen angegeben wird. Ein weiterer Vorteil von persönlichen Interviews (neben der Einschätzung der Interviewsituation) besteht darin, dass visuelle Hilfsmittel (z. B. eine grafische Darstellung der abgefragten Skalenabstufungen) eingesetzt werden können, was teilweise zur Erhöhung der Reliabilität der Befragung beiträgt (Möhring & Schlütz, 2010). Demgegenüber entstehen erhöhte Personalkosten bei persönlichen Interviews, was häufig zu einer geringeren Reichweite bei der Rekrutierung der Interviewten führt. Eine große Reichweite ist stattdessen bei telefonischen Interviews und schriftlichen Befragungen gegeben.

Bei der **Entscheidung für einen Interview-Modus** sind folgende vier Aspekte zu berücksichtigen:

1. **Zielgruppe**: Über welchen Kommunikationskanal ist die Zielgruppe gut bzw. schlecht erreichbar? So wird man Hochbetagte oft am liebsten persönlich befragen. Da der Anteil der Offliner unter den Hochbetagten relativ groß ist, scheiden Online-Interviews aus. Auch

Telefon-Interviews scheinen problematisch, denn eingeschränkte Hörfähigkeit ist am Telefon schlechter kompensierbar als im direkten Gespräch, in dem bei Bedarf visuelle Hilfen einsetzbar sind.

2. **Thema und Umfang der Befragung**: Umfangreiche Fragebögen sind am besten im persönlichen Gespräch abzufragen, weil hier die Kopräsenz des Interviewers motivierend auf die Befragungspersonen wirkt und Interviewabbrüche verhindert. In der anonymeren Situation beim telefonischen oder Online-Interview kommt es bei langen Fragebögen häufiger zu Interviewabbrüchen. Bei Interview-Fragebögen, die heikle, umstrittene oder intime Themen behandeln, kann man anstelle des persönlichen Interviews lieber auf ein Telefon- oder Online-Interview zurückgreifen, um für die Befragungspersonen eine anonymere und weniger konfrontative Situation zu schaffen.
3. **Ressourcen**: Persönliche Interviews sind aufgrund der notwendigen Fahrten und Reisen am zeit-, personal- und kostenintensivsten. Bei begrenzten Ressourcen wird man deswegen eher Telefon-Interviews durchführen: Hier entfallen die Reisekosten zugunsten von Telefonkosten. Zudem kann bei Nutzung eines Telefonlabors die Zahl der Interviewer reduziert werden. Das Online-Interview ist für viele Forschungseinrichtungen und einzelne Forschende bislang meist der ökonomischste Interview-Modus, da weder Reise- noch Telefonkosten anfallen und die Internetanbindung ohnehin besteht.
4. **Auswahlrahmen**: Insbesondere wenn man eine probabilistische Stichprobe anstrebt, die eine möglichst hohe globale Repräsentativität hinsichtlich der Population aufweist, ist die Wahl des Interview-Modus an die Verfügbarkeit eines entsprechenden Auswahlrahmens gebunden.

Eine Umfrage kann auch als **Mixed-Mode-Befragung** („mixed-mode survey“, „multi-mode survey“) mehrere dieser Kommunikationswege kombiniert einsetzen, um Befragungspersonen auf dem jeweils am besten akzeptierten Kommunikationsweg zu erreichen. Auch bei den in einer globalisierten Welt an Bedeutung gewinnenden internationalen Studien handelt es sich häufig um Mixed-Mode-Umfragen, wobei die jeweilige Kombination der Kommunikationskanäle nationalen Besonderheiten Rechnung trägt (z. B. Verbreitung von Festnetzanschlüssen und Handys in der Bevölkerung). Mixed-Mode-Befragungen können Abdeckungs- und Nonresponsefehler reduzieren, sind aber ihrerseits fehlerbehaftet. Die modusspezifischen Verzerrungen in den Daten können bei einer Mixed-Mode-Befragung z. B. fälschlich als inhaltliche Effekte gedeutet werden. Aktuell existieren zwei unterschiedliche Strategien zur Gestaltung von Mixed-Mode-Befragungen:

- **Möglichst einheitliches Vorgehen über alle Befragungs-Modi hinweg**. Dieser als **Total Design Method** (TDM) in den 1970er-Jahre von dem amerikanischen Soziologen Don Dillman (1978) an der Universität Washington entwickelte Ansatz wurde inzwischen zur **Tailored Design Method** (Dillman, Smyth, & Christian, 2008) ausgebaut. Grundgedanke ist, dass bei der Kombination unterschiedlicher mündlicher wie schriftlicher Befragungs-Modi (z. B. Telefon-Interview, postalischer Fragebogen, Online-Fragebogen) möglichst derselbe vollstandardisierte Fragebogen verwendet wird. Dies bedeutet, dass auf umfangreiche oder bildliche Antwortskalen verzichtet werden sollte, wenn ein Teil der Interviews am Telefon durchgeführt wird, wo nur kurze Antwortskalen nutzbar sind. Ebenso muss bei persönlichen und telefonischen Interviews auf komplexe Filterführungen verzichtet werden, wenn ein Teil der Studie per postalischer Fragebogenmethode durchgeführt werden soll. Mit einem möglichst einheitlichen Fragebogen und Feldgeschehen sollen die modusspezifischen Fehler und Verzerrungen aneinander angeglichen werden.
- **Möglichst optimiertes Vorgehen für jeden einzelnen Befragungs-Modus**. Bei dieser Strategie sollen Feldgeschehen und Fragebogen möglichst für jeden einzelnen Befragungs-Modus separat optimiert werden. Dementsprechend wird man beim persönlichen Interview z. B. visuelle Hilfen verwenden und beim Telefoninterview die Antwortskalen kürzen. Der Vorteil dieses Vorgehens besteht darin, dass man Messfehler und Nonresponsefehler modusspezifisch reduzieren kann. Gleichzeitig vergrößern sich dadurch jedoch die Unterschiede zwischen den Teilstudien.

Vollstrukturiertes persönliches Interview

Im Folgenden werden Besonderheiten vollstrukturierter persönlicher Interviews hinsichtlich der Stichprobenverfahren sowie des Interviewablaufs behandelt.

Stichprobenverfahren für persönliche Umfragen

Für eine aussagekräftige persönliche Interviewstudie ist eine Stichprobe von Personen oder Haushalten zusammenzustellen, die von den Interviewenden persönlich zu erreichen sind. **Nicht-probablistische Stichproben** (► Abschn. 9.3.1) wie **Gelegenheitsstichproben** oder **Quotenstichproben** lassen sich z. B. ziehen, indem man Personen willkürlich bzw. systematisch gemäß einem Quotenplan auf der Straße oder bei bestimmten Institutionen anspricht. Bei schwer zugänglichen Populationen

kommt auch die Methode des **„Targeted Sampling“** infrage, indem gezielt Szene-Treffpunkte idenfiziert und auf dieser Basis ein Stichprobenplan zusammengestellt wird. Relativ häufig wird für persönliche Interviewstudien auch auf **Access-Panels** zurückgegriffen, d. h. auf selbst-selektierte Probandenpools, deren Adressdaten und soziodemografischen Kennwert in einer Datenbank vorgehalten werden und die grundsätzlich zur Teilnahme an Studien bereit sind.

Wenn im Rahmen vollstrukturierter persönlicher Interviewstudien **probabilistische Stichproben** realisiert werden sollen (► Abschn. 9.3.2), sind drei Ansätze verbreitet:

1. **Register-Stichprobe**: Die von den Einwohnermeldeämtern der Gemeinden geführten Daten (Name, akademischer Titel, Geburtstag, Wohnadresse etc.) dürfen für wissenschaftliche Untersuchungen genutzt werden, wenn diese inhaltlich von öffentlichem Interesse sind. Eine Zufallsstichprobe aus dem **Einwohnermelderegister** ergibt eine Adressliste privater Haushalte, die von den Interviewenden aufzusuchen sind. Auf diese Weise werden Ausschöpfungsquoten von ca. 50 % erreicht (Diekmann, 2004, S. 26). Register-Stichproben können auch für postalische Umfragen eingesetzt werden (► Abschn. 10.3.3), dann liegen die Ausschöpfungs- bzw. Rücklaufquoten jedoch deutlich geringer (ca. 10 %).
2. **Random-Route-/Random-Walk-Verfahren**: Ausgehend von einem Startpunkt (Straße, Hausnummer) wird eine Begehungsregel (z. B. „Wähle jedes vierte Haus und biege bei Kreuzungen immer abwechselnd nach rechts und nach links ab.“) befolgt, um zufällig Haushalte in einem Areal auszuwählen. Die Begehungsregel muss dabei den Umgang mit allen Besonderheiten (z. B. Ein- und Mehrfamilienhäuser, Baustellen etc.) berücksichtigen. Beim Random-Route-Verfahren erfolgt die Adressermittlung oft durch die Interviewer, die gleich die Interviews durchführen (**Standard-Random-Verfahren**). Eine Problematik dieser Vorgehensweise kann darin bestehen, dass die Interviewer im Sinne der Arbeitsökonomie geneigt sind, von der festgelegten Begehungsregel abzuweichen, um sofort viele Interviews zu bekommen. Wenn an der Zieladresse niemand die Tür öffnet und der Interviewer somit zu einem anderen Zeitpunkt erneut den Haushalt aufsuchen müsste, besteht die Versuchung, bei einer Nachbarwohnung zu klingeln. Das **Adress-Random-Verfahren** soll dieser Verfälschung entgegenwirken: Hier erfolgt im ersten Schritt nur die Adressermittlung strikt nach den Begehungsregeln. Die generierte Adressliste wird dann den Interviewern vorgegeben. Die Studienleitung kann somit beim Adress-Random-Verfahren nachprüfen (z. B. durch eine schriftliche Anfrage), ob die Interviewer tatsächlich die Zielhaushalte aufgesucht haben.
3. **ADM-Mastersample** (ADM-Stichprobensystem F2F: „face-to-face“): Der Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute ADM hat einen Auswahlrahmen für persönliche Interviewstudien entwickelt, der eine **dreistufige Zufallsauswahl** vorsieht. Dieser Auswahlrahmen erlaubt es, repräsentative Stichproben von Privathaushalten und den darin wohnenden Personen zu erstellen. Das ADM-Stichprobensystem basiert auf einer Flächenstichprobe, die das gesamte bewohnte Gebiet der Bundesrepublik Deutschland umfasst. Diese Stichprobe berücksichtigt die Gemeindegliederung der Bundesrepublik, und darauf aufbauend die intrakommunalen Gebietsgliederungen und zusätzlich die für Navigationssysteme erstellten elektronischen Regionaleinteilungen. Auf Grundlage dieser Daten werden in Deutschland rund 53 000 Flächen elektronisch abgegrenzt.
   Auf der ersten Ziehungsstufe erfolgt eine **Flächenauswahl**. Dazu werden die abgegrenzten Flächen erst regional nach Kreisen und BIK-Typen (einer Einteilung nach Grad der Verstädterung von Regionalzonen) geschichtet, so dass sie insgesamt in rund 1 500 Regionalschichten eingeteilt werden. Anschließend werden 128 sog. Netze proportional zur Verteilung der Privathaushalte gezogen – bestehend aus 210 Auswahlflächen in den alten Bundesländern und 48 in den neuen. Da eine Fläche jeweils nur für ein Netz gezogen wird, sind die 128 Netze überschneidungsfrei. Die in der Arbeitsgemeinschaft versammelten Institute erhalten je nach Wunsch und Intensität ihrer Feldarbeit mehr oder weniger Netze aus der Gesamtmenge zur eigenen Verwendung. Die zweite und dritte Auswahlstufe wird von den forschenden Instituten vorgenommen. Auf der zweiten Ziehungsstufe erfolgt in systematischer Zufallsauswahl die **Haushaltsauswahl**, z. B. mittels Random-Route-Verfahren oder Adressenauflistung. Auf der dritten Ziehungsstufe erfolgt die **Zielpersonenauswahl**: Aus allen Haushalten wird jeweils eine Befragungsperson zufällig ausgewählt, z. B. gemäß Schwedenschlüssel (► Auswahl von Zielpersonen aus Mehrpersonen-Haushalten). Das ADM-Mastersample kann von allen ADM-Mitgliedern genutzt werden, wobei mehr oder minder umfangreiche Stichproben gezogen werden können (ADM, 2008).

■ Interview-Ablauf bei persönlichen Umfragen

Üblicherweise erfolgt der **Erstkontakt** zu den interviewten Personen auch bei persönlichen Umfragen schriftlich oder telefonisch, um einen Termin zu vereinbaren. Gelegentlich wird versucht, ohne Voranmeldung durch direktes Aufsuchen der ausgewählten Wohnung zu einem

Auswahl von Zielpersonen aus Mehrpersonen-Haushalten

Bei Stichprobenverfahren, bei denen über die Wohnadressen oder die Festnetz-Telefonnummern jeweils komplette Haushalte ausgewählt werden, muss vor der Durchführung des persönlichen oder telefonischen Interviews per Zufall eine Befragungsperson **(Zielperson)** ausgewählt werden. Denn die Personen, die man spontan an der Tür oder am Festnetztelefon erreicht **(Kontaktperson)** stellen eine systematisch verzerrte Auswahl der Haushaltsmitglieder dar (z. B. sind hier kontaktfreudige und nichtberufstätige Personen sowie Frauen überrepräsentiert). Die Kontaktperson wird also über das Anliegen der Umfrage informiert und um Angaben zum Haushalt gebeten, auf deren Basis – im Falle eines Mehrpersonenhaushaltes – dann die Auswahl der Zielperson erfolgt. Für die Auswahl von Zielpersonen aus Mehrpersonenhaushalten werden in der Literatur mindestens 14 verschiedene Verfahren vorgeschlagen (für einen kritischen Vergleich dieser Verfahren s. Gaziano, 2005). In der Praxis haben sich vor allem zwei Verfahren etabliert: Die Geburtstagsmethode und der Schwedenschlüssel.

Geburtstagsmethode: Beim Kontakt mit dem Haushalt wird die Kontaktperson nach demjenigen volljährigen Haushaltsmitglied gefragt, das zuletzt Geburtstag hatte (**Last-Birthday-Methode**). Manchmal wird auch nach der Person gefragt, die als nächste Geburtstag hat (**Next-Birthday-Methode**). Die Last- oder Next-Birthday-Methode (entwickelt von Salmon & Nichols, 1983) hat den Vorteil, dass sie relativ schnell und einfach auszuführen ist. Sie hat jedoch den Nachteil, dass mit der Festlegung des Untersuchungszeitraums etliche Personen der Grundgesamtheit von vorne herein keine positive Auswahlwahrscheinlichkeit mehr haben. Dies lässt sich vermeiden, wenn man für jeden Haushalt ein zufälliges Datum generiert und dann die Person aus dem Haushalt auswählt, die vor oder nach dem Zufallsdatum Geburtstag hat (**randomisierte Geburtstagsauswahl**; Schneiderat & Schlinzig, 2009). Dieses methodisch stringentere Verfahren ist deutlich aufwändiger für die Kontaktperson und weniger praxistauglich. Die Last- bzw. Next-Birthday-Methode kommt häufig auch bei Telefonumfragen zum Einsatz.

Schwedenschlüssel („Kish-selection grid"): Dieses von Leslie Kish (1949) entwickelte und auf die heterosexuelle Kernfamilie zugeschnittene Verfahren verlangt es, dass die Auskunftsperson zu allen Haushaltsmitgliedern deren **Verhältnis zum männlichen Haushaltsvorstand** sowie ihr **Geschlecht** angibt. Die Angaben werden in eine Tabelle eingetragen. Alle Haushaltsmitglieder werden durchnummeriert (Erwachsenen-Nr.), vom ältesten bis zum jüngsten männlichen und dann vom ältesten bis zum jüngsten weiblichen Haushaltsmitglied. Sofern sich die Altersreihenfolge nicht aus den familiären Beziehungen ergibt (z. B. Tochter ist jünger als Mutter etc.), muss das **Alter** ausdrücklich erfragt werden (Tab. 10.9). Gemäß dem Beispiel in Tab. 10.9 (nach Kish, 1965) handelt es sich um einen Sechs-Personen-Haushalt. Nachdem der Interviewer alle 6 Haushaltsmitglieder nach dem Kish-Raster in eine feste Reihenfolge gebracht hat (Erwachsenen-Nr.), entnimmt er nun einer **speziellen Auswahltabelle** (Kish, 1965), welches Haushaltsmitglied aus einem Sechs-Personen-Haushalt zu befragen ist:

Anzahl der Erwachsenen im Haushalt	Wähle Nummer
1	1
2	2
3	2
4	3
5	4
6 oder höher	**4**

In diesem Fall ist gemäß Auswahltabelle in der ersten Spalte bei Anzahl der Erwachsenen im Haushalt „6 oder höher" (Sechs-Personen-Haushalt) zu wählen, so dass sich laut der zweiten Spalte für die Befragung Person Nr. 4 ergibt, also die im Haushalt lebende Tante der Ehefrau, für welche die Erwachsenen-Nr. 4 bestimmt worden war (Tab. 10.9). Bei der Kish-Grid-Methode wird systematisch mit acht verschiedenen Auswahltabellen gearbeitet, die jeweils bei einem Teil der Stichprobe einzusetzen sind, um gleiche Auswahlwahrscheinlichkeiten für alle Haushaltsmitglieder trotz unterschiedlicher Haushaltsgrößen zu erzielen. In Deutschland bekannt wurde das Verfahren als „Schwedenschlüssel" durch einen Artikel von Elmo Wilson (1950), in dem die Anwendung des Kish-Verfahrens u. a. am Beispiel einer schwedischen Studie erklärt wurde.

Interview zu gelangen. Diese Vorgehensweise führt in der Regel zu einer erhöhten Verweigerungsquote. Die erste Kontaktaufnahme entscheidet weitgehend darüber, ob ein Interview zustande kommt oder nicht. Sie sollte deswegen gründlich vorbereitet sein. Bei allen Anwerbungen sollte eine einheitliche Textvorlage verwendet werden. Diese sollte eine kurze Beschreibung der Grundsätze der Vertraulichkeit der Befragtenangaben, der allgemeinen Forschungsziele, der Identität der Forschungseinrichtung bzw. der verantwortlichen Stelle und einen Hinweis auf die Freiwilligkeit der Teilnahme enthalten. Idealerweise sollten für das angefragte Interview mehrere Auswahltermine angeboten werden.

Findet ein Interview statt, stellen sich die Interviewenden an der Haustür vor und weisen sich durch einen **Interviewer-Ausweis** aus. Diese werden von der forschenden Institution zur Verfügung gestellt. Neben Lichtbild und Namen des Interviewers, Ausstellungs- und Ablaufdatum, sind der Name und die Kontaktangaben der Institution auf dem Ausweis aufzuführen. Das Interview selbst sollte in der Wohnung der Befragten oder zumindest in einer ihnen vertrauten Umgebung stattfinden. Nach der Begrüßung erläutern die Interviewer nochmals ihr Anliegen und bedanken sich für die Gesprächsbereitschaft der Befragten. Anschließend werden die Interviewten erneut auf die Vertraulichkeit ihrer Angaben und die Freiwilligkeit der Teilnahme hingewiesen. Dabei sind den Befragten dokumentierte Zusicherungen anzubieten.

Bevor das eigentliche Interview beginnt, prüfen die Interviewenden Möglichkeiten, die situativen Bedingungen zu standardisieren (einheitliche Sitzordnung, gute Beleuchtung, keine Ablenkung durch andere Personen, abgeschaltete Rundfunk- und Fernsehapparate, keine ablenkenden Nebentätigkeiten während des Interviews etc.).

Tabelle 10.9 Anwendung des Schwedenschlüssels auf einen Beispielhaushalt mit sechs Personen – das Verhältnis der Haushaltsmitglieder zum männlichen Haushaltsvorstand ergibt laut Kish-Verfahren die Erwachsenen-Nr. von 1 bis 6, die zur Auswahl genutzte Erwachsenen-Nr. ergibt sich aus den vorgefertigten Auswahltabellen (die bei einem Haushalt mit 6 Personen z. B. Person 4 vorgeben)

Beziehung	Geschlecht	Alter	Erwachsenen-Nr. laut Kish-System	Auswahl laut Auswahltabelle
Mann („Haushaltsvorstand“)	M		2	
Frau	W	40	5	
Vater des Mannes	M		1	
Sohn	M		3	
Tochter	W		6	
Tante der Frau	W	44	4	X

Es ist selbstverständlich, dass evtl. erforderliche Korrekturen an den situativen Bedingungen nur mit Einverständnis der Befragten vorgenommen und zudem begründet werden. Während des Interviews unerwartet auftretende Störungen oder Beeinträchtigungen sind später in einem Interviewprotokoll festzuhalten.

Die **Durchführung des Interviews** erfolgt anhand des **standardisierten Interview-Fragebogens**. Dies kann in der Interviewpraxis durchaus eine Herausforderung darstellen. Interviewende können versucht sein, die Fragen nicht wortwörtlich abzulesen, sondern zu variieren, z. B. wenn ihnen die zu stellende Frage selbst unangenehm ist. Schon kleine Veränderungen können jedoch zu Abweichungen in den erhobenen Antworten führen und sind in vollstrukturierten Interviews deswegen nicht zulässig (Schuman & Presser, 1981). Die Antworten der Befragungspersonen sind in geeigneter Weise festzuhalten. Dies geschieht meist durch schriftliche Notizen in vorbereiteten Formularen oder durch direkte Eingabe in einen portablen Computer, z. B. Notebook, Netbook, Tablet PC mit entsprechender Umfrage-Software (CAPI-Interview; ▶ Studienbeispiel „Persönliches computergestütztes Interview“).

Am Ende des Interviews wird der Befragungsperson für ihre Mitwirkung gedankt. Eventuelle Versprechungen, nähere Erläuterungen zum Interview nach Abschluss des Gespräches zu geben, müssen jetzt eingelöst werden.

Der dargestellte Ablauf eines vollstrukturierten persönlichen Interviews zeigt, dass die Interviewendeneinen großen Einfluss auf dessen Erfolg haben. Aus diesem Grund ist eine gründliche **Interviewerschulung** (z. B. hinsichtlich Kontaktaufnahme, Herstellen einer guten Interviewbeziehung, Fertigkeiten der Interviewführung inklusive wortgetreuer Wiedergabe des Leitfadens und exakter Dokumentation der gegebenen Antworten, Bedienung von Laptop und entsprechender Interview-Software) unerlässlich und wird z. B. in der Interview-Software 20252 (Deutsches Institut für Normung, 2006) zu Markt-, Meinungs- und Sozialforschung explizit thematisiert (s. Bryman, 2008).

Studienbeispiel

Persönliches computergestütztes Interview

Der **Mikrozensus** (kleine Volkszählung) ist die **amtliche Repräsentativstatistik** über die Bevölkerung und den Arbeitsmarkt in Deutschland, an der jährlich **1 % aller Haushalte** beteiligt sind (laufende Haushaltsstichprobe). Insgesamt nehmen rund 390 000 Haushalte mit 830 000 Personen am Mikrozensus teil.

Das Mikrozensus-Frageprogramm umfasst rund **170 Fragen**, u. a. zu Erwerbstätigkeit, Alter, Geschlecht, Wohnsitz, Staatsangehörigkeit, Familienstand, Größe des Haushaltes, beruflicher Stellung, Einkommen, Bildung etc.

Der Mikrozensus ist eine **Zufallsstichprobe**, bei der alle Haushalte die gleiche Auswahlwahrscheinlichkeit haben. Dazu werden aus dem Bundesgebiet Flächen (Auswahlbezirke) ausgewählt, in denen alle Haushalte und Personen befragt werden (**einstufige Klumpenstichprobe**). Ausgewählte Haushalte bzw. Personen sind **gesetzlich zur Teilnahme verpflichtet**, da anderenfalls durch erhöhte Antwortverweigerung in bestimmten Bevölkerungsgruppen verzerrte Ergebnisse entstehen würden.

Die Datenerhebung erfolgt mit **persönlichen computergestützten Interviews** („computer-assisted personal interviewing“, **CAPI**). Die geschulten Interviewer suchen im Auftrag der statistischen Landesämter die Haushalte persönlich auf, weisen sich entsprechend aus und erfassen die Antworten auf Laptops. Auf Wunsch können die Befragten anstelle des mündlichen Interviews auch einen **schriftlichen Fragebogen** ausfüllen und postalisch zurücksenden, wovon rund 20 % Gebrauch machen. Die Möglichkeit zur **telefonischen Durchgabe der Daten** an das zuständige statistische Landesamt wird dagegen kaum genutzt. Die Datenerhebung beim Mikrozensus ist also **multimodal** angelegt (persönliches Interview, schriftlicher Fragebogen, Telefon), wobei die persönlichen Interviews dominieren (für ausführliche Informationen zum Mikrozensus s. die Websites des Statistischen Bundesamtes: http://www.destatis.de).

Vollstrukturiertes Telefon-Interview

Im Folgenden werden Besonderheiten vollstrukturierter Telefoninterviews hinsichtlich der Stichprobenverfahren sowie des Interviewablaufs behandelt.

Stichprobenverfahren für Telefonumfragen

Für eine aussagekräftige telefonische Interview-Studie muss man eine Stichprobe von Personen zusammenstellen, deren Telefonnummern man kennt (Personenstich-

probe), oder Haushalte telefonisch kontaktieren (Haushaltsstichprobe) und dann gezielt jeweils eine Befragungsperson auswählen.

Vollstrukturierte Telefon-Interview-Studien werden zuweilen mit **nicht-probabilistischen Stichproben** durchgeführt. Möglich sind z. B. Quotenstichproben oder Gelegenheitsstichproben (z. B. Umfrageeinladungen mit kostenloser Rufnummer zum Einwählen werden unter gut erreichbaren potenziellen Befragungspersonen verteilt). Auch Access-Panels können genutzt werden.

Typischer sind bei Telefonumfragen **probabilistische Stichproben**, die z. B. auf folgenden Wegen zusammengestellt werden:

- **Zufallsauswahl aus Telefonlisten von Personen**: Handelt es sich bei der interessierenden Population um die Mitglieder einer bestimmten Organisation (z. B. Firmenangestellte, Vereinsmitglieder), so existiert zuweilen ein vollständiges Personenverzeichnis mit individuellen beruflichen und/oder privaten Festnetz- und/oder Handynummern, das für eine Vollerhebung oder Stichprobenziehung genutzt werden kann, sofern die Liste für Forschungszwecke zur Verfügung gestellt wird (Berücksichtigung des Datenschutzes). Liegen keine geeigneten Telefonlisten vor, so lassen sie sich teilweise unter Mithilfe der Zielgruppe für ein Forschungsprojekt erstellen.
- **Auswahl aus dem Telefonbuch**: Bis 1992 war es in Westdeutschland Pflicht, die eigene Festnetz-Telefonnummer im Telefonbuch eintragen zu lassen. Seit dies nicht mehr der Fall ist, tauchen mehr als ein Viertel aller Festnetzanschlüsse nicht mehr in den öffentlichen Telefonbüchern bzw. auf den Telefonbuch-CDs auf. Diese **Nichteingetragenen** („Nonpubs“) unterscheiden sich systematisch von der Gesamtbevölkerung (z. B. leben sie häufiger in den neuen Bundesländern und sind jünger). Stichprobenauswahlen auf der Basis von Telefonbüchern sind somit nicht mehr bevölkerungsrepräsentativ (Häder & Glemser, 2004, S. 154). Zudem sind Festnetz-Telefonnummern nicht auf Einzelpersonen, sondern auf Haushalte bezogen. Sofern man **Mehrpersonen-Haushalte** erreicht, muss z. B. mit der Geburtstagsmethode (▶ Auswahl von Zielpersonen aus Mehrpersonen-Haushalten) noch die Zielperson ausgewählt werden.
- **Random-Digit-Dialing-Verfahren** (RDD, Zufallsziffernanwahl): Bei diesem Verfahren werden Festnetzrufnummern nach dem Zufallsprinzip erzeugt, so dass man vom Telefonbuch unabhängig ist. Will man etwa eine Stichprobe der Einwohner der Gemeinde Neustadt an der Weinstraße ziehen, so weiß man, dass entsprechende Festnetznummern mit der Vorwahl 06321- für den Ortsnetzbereich beginnen müssen, die niedrigste Nummer für einen privaten Haushaltsanschluss ist die 06321-2001, die höchste die 06321-6790151. Theoretisch könnte man nun nach dem Zufallsprinzip eine Stichprobe von Telefonnummern in dieser Spanne erzeugen lassen. Doch das Nummernsystem im Festnetzbereich ist in Deutschland sehr kompliziert. Es basiert nicht auf fortlaufenden Nummern, sondern einer blockweisen Nummernvergabe und enthält sehr viele Lücken. Somit werden per Zufallsziffernanwahl fast nur ungültige Nummern generiert. Für **Deutschland** ist das Verfahren deswegen **unbrauchbar** (Häder & Glemser, 2004, S. 154f.). In den USA ist es dagegen weit verbreitet.
- **Randomize(d)-Last-Digit-Verfahren** (RLD, zufällige letzte Ziffer). Im Unterschied zum RDD-Verfahren, bei dem die komplette Telefonnummer künstlich generiert wird, werden beim RLD-Verfahren Telefonnummern aus dem Telefonbuch gezogen und dann die letzte Ziffer oder die letzten beiden Ziffern durch Zufallszahlen ersetzt, um auch nicht eingetragene Anschlüsse erreichen zu können. Dieses Verfahren ist aufgrund des spezifischen Nummernsystems für **Deutschland untauglich**, in den USA jedoch etabliert.
- **Gabler-Häder-Design**: Angepasst an das spezifische, blockweise organisierte deutsche Festnetznummernsystem wurde Mitte der 1990er-Jahre am Zentrum für Umfragen, Methoden und Analysen (ZUMA; mittlerweile integriert in GESIS: Leibniz-Institut für Sozialwissenschaften) von Siegfried Gabler und Sabine Häder das sog. Gabler-Häder-Design entwickelt. Bei diesem Auswahlrahmen wird mit dem **Randomized-Last-Digit-Verfahren** gearbeitet, allerdings abgestimmt auf das deutsche Nummernsystem, so dass höhere Trefferquoten erzielt werden. Dennoch entsteht ein nennenswerter Anteil ungültiger Telefonnummern. Das Gabler-Häder-Design ist deswegen nur sinnvoll einsetzbar, wenn ein professionelles Telefonumfrage-Studio zur Verfügung steht, bei dem die Anwahl der Nummern nicht per Hand, sondern automatisiert erfolgt. Das Gabler-Häder-Design wird der **akademischen Sozialforschung** kostenfrei zur Verfügung gestellt, d. h. von der Abteilung „Survey Design and Methodology“ SDMder GESIS (Kontakt über http://www.gesis.org; ehemals ZUMA) kann eine repräsentative Stichprobe von Telefonnummern angefordert werden. Wie **Mobilfunknummern** geeignet in das bislang rein festnetzbasierte Gabler-Häder-Design einzubeziehen sind, wird aktuell erforscht. Dabei ist zu beachten, dass es sich bei Festnetznummern um Kontakte zu Haushalten, bei Mobilfunknummern dagegen um Kontakte zu Personen handelt, die es im Sinne eines **Dual-Frame-Ansatzes** (Kombination von zwei Auswahlrahmen) konsistent zu verknüpfen gilt (Hunsicker & Schroth, 2007).

- **ADM-Design** (ADM-Stichprobensystem für Telefonbefragungen): Aufbauend auf dem Gabler-Häder-Design wurde vom Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM) ein noch anspruchsvollerer, weil differenzierterer Auswahlrahmen entwickelt, der teilweise bereits Mobilfunknummern einbezieht. Das ADM-Design steht nur ADM-Mitgliedern zur Verfügung.

Interview-Ablauf bei Telefonumfragen

Professionelle Telefonumfragen werden in einem **Telefonlabor** mit einer **CATI-Anlage** (Computer-Assisted Telephone Interview) unter Betreuung durch einen Supervisor durchgeführt. Der Interview-Fragebogen muss in das System eingegeben sowie einem Pretest unterzogen werden. Die Interviewenden müssen auf dem System geschult und während der Feldphase betreut, z. B. kontrolliert und ggf. nachgeschult werden. Das System bietet den Interviewenden eine klare Filterführung, führt Plausibilitäts-Checks bei den Antworten durch und dokumentiert alle Eingaben. Da bei telefonischen Umfragen visuelle Vorlagen nicht verwendet werden, muss die Zahl der vorgegebenen Antwortalternativen einfach und überschaubar sein. Die Dauer von Telefon-Interviews sollte rund 20 Minuten nicht übersteigen. Für die Dokumentation der Studie sollten Anzahl und Anteil folgender Kennwerte ausgewiesen werden: angewählte Telefonnummern, ungültige Nummern, nicht erreichte Nummern, keine Teilnahmebereitschaft, abgebrochenes Interview, vollständiges Interview (AAPOR, 2008). Je nachdem, welche Kennwerte zugrunde gelegt werden, können **unterschiedliche Varianten von Ausschöpfungsraten** berechnet werden.

Bei Telefon-Interviews erfolgt der **Erstkontakt** üblicherweise telefonisch. Hier besteht die Herausforderung darin, die potenziellen Befragungspersonen für eine Teilnahme zu gewinnen. Angesichts der wachsenden Zahl von Werbeanrufen und Direktmarketing muss ein wissenschaftliches Anliegen als solches deutlich gemacht werden. Ansonsten befürchten Respondenten schnell, dass ihnen etwas verkauft werden soll. Dies ist noch aus einem anderen Grund wichtig: Telefonanrufe zu Werbe- und Verkaufszwecken in Privathaushalten sind durch gesetzliche Regelungen untersagt, sofern keine Vertragsbeziehung besteht oder ausdrückliche Einwilligung vorliegt (ADM, 2008).

Wissenschaftliche Telefonumfragen sind im **Zeitfenster zwischen 9 und 21 Uhr** zulässig (ADM, 2008). Üblicherweise wählt man für Telefonumfragen Zeiten zwischen 17 und 21 Uhr. Frühere Anrufe würden z. B. Erwerbstätige kaum erreichen, spätere Anrufe wären störend. Wird niemand erreicht, so erfolgt zu einem späteren Zeitpunkt eine erneute Kontaktaufnahme. Kommt der Anruf ungelegen, so wird ein Termin für einen Rückruf vereinbart. Bei schwer erreichbaren Teilnehmern werden mindestens **10 Kontaktversuche** an unterschiedlichen Wochentagen zu verschiedenen Tageszeiten empfohlen. Die Kontaktversuche werden durch das CATI-System gesteuert und dokumentiert.

Lehnen es angerufene Personen ab, im Rahmen der jeweiligen Umfrage Auskunft zu geben, ist es laut der Richtlinie für telefonische Befragungen des Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute (ADM) nicht zulässig, nochmals einen Kontaktversuch zur selben Studie zu unternehmen. Von der umfragespezifischen ist die generelle **Teilnahmeverweigerung** bei Telefonumfragen zu unterscheiden: Wird von einer angerufenen Person generell abgelehnt, an telefonischen Befragungen teilzunehmen, muss das Forschungsinstitut geeignete Maßnahmen treffen, damit die betreffende Telefonnummer nicht mehr in Stichproben für telefonische Befragungen gelangt (ADM, 2008).

Kontaktversuche bei telefonischen Befragungen sind frühestens nach 20 Sekunden und spätestens nach 40 Sekunden zu beenden, wenn der **Anruf nicht angenommen** wird. Damit die angerufenen Personen auch ohne eine Annahme des Anrufs die Identität der Forschungseinrichtung feststellen können, muss diese dafür sorgen, dass im Display des Telefons eine Rufnummer angezeigt werden kann, die einen Kontakt mit der Einrichtung ermöglicht. Diese Telefonnummer muss entweder zu Mitarbeitern der Einrichtung führen oder zu einer Voicebox, die durch Tastendruck folgende Optionen bietet: a) Informationen über die Forschungseinrichtung, b) Informationen über den Zweck des Anrufs einschließlich der Herkunft der Rufnummer und Informationen über die Berufsgrundsätze und Standesregeln der deutschen Markt- und Sozialforschung (z. B. Freiwilligkeit der Teilnahme, Wahrung der Anonymität) sowie c) direkten Kontakt zu einem Mitarbeiter, einer Mitarbeiterin der Forschungseinrichtung (ADM, 2008).

Alle **angenommenen Anrufe** wiederum sollten grundsätzlich vom System innerhalb einer Sekunde mit einem Interviewer verbunden werden, so denn ein solcher verfügbar ist. Erfolgt eine solche Verbindung nicht innerhalb von einer Sekunde, wird der Anruf abgebrochen und als stiller Anruf dokumentiert.

Bei erfolgtem Kontakt mit den angerufenen Personen müssen den Befragungspersonen folgende Informationen mitgeteilt werden (ADM, 2008):

- Anrufende Forschungseinrichtung,
- Zweck des Anrufs,
- Freiwilligkeit der Teilnahme und
- Herkunft der Telefonnummer sowie Widerspruchsrecht bzgl. ihrer Übermittlung und/oder Nutzung (► Kontaktaufnahme bei Telefon-Interviews).

Kontaktaufnahme bei Telefon-Interviews

Mustertext (ADM, 2008, S. 5):
„Guten Tag. Mein Name ist _____vom _____(Institut) _____in _____.
Wir führen eine wissenschaftliche Befragung zu _____(allgemeine Thematik) _____durch. Ihre Telefonnummer wurde durch ein wissenschaftliches (Zufalls-)Verfahren ausgewählt. Die Teilnahme ist freiwillig, aber es ist sehr wichtig, dass möglichst alle ausgewählten Personen teilnehmen, damit die Umfrage ein aussagekräftiges Ergebnis liefert.
Die Auswertung erfolgt anonym, also nicht in Verbindung mit Ihrem Namen, Ihrer Anschrift oder Telefonnummer.

Einwilligung in die Tonaufzeichnung

(bei Bedarf):
Wenn Sie damit einverstanden sind, würden wir das Interview zu wissenschaftlichen Zwecken und um Fehler zu vermeiden gerne aufzeichnen. Aber auch dadurch wird Ihre Anonymität nicht gefährdet."

Die **Durchführung des Telefon-Interviews** erfolgt vergleichbar zu persönlichen Interviews anhand eines vollstandardisierten Interview-Fragebogens. Die Antworten der Befragungspersonen werden durch die Interviewenden direkt im CATI-System dokumentiert. Nach dem Interview werden die Befragten freundlich verabschiedet. Dazu gehört ebenfalls, dass ihnen für ihre Teilnahme und ihre Zeit gedankt wird.

Als nachteilig wirkt sich bei Telefon-Interviews die Tatsache aus, dass die **situativen Umstände** des telefonischen Interviews wenig standardisierbar sind; die Begleitumstände des Interviews (ablenkende Reize, Lärmbelästigungen, Ermüdung etc.) bleiben teilweise unbemerkt Dies ist insbesondere bei Telefon-Interviews per Mobiltelefon zu beachten, bei denen Befragte z. B. während des Autofahrens erreicht werden können und somit abgelenkt sind bzw. auch die Verkehrssicherheit gefährden könnten (Döring, 2009). Weitere Hinweise zu Telefon-Interviews findet man in der einschlägigen Fachliteratur (z. B. Bourque & Fielder, 2002; Buchwald, 2006; Groves et al., 2001; Gwartney, 2007; Martens & Ritter, 2008; Schnell, Hill, & Esser, 2008).

Vollstrukturiertes Online-Interview

Bei Online-Interviews handelt es sich typischerweise um qualitative Leitfaden-Interviews (▶ Abschn. 10.2.2). Wenn Online-Befragungen auf der Basis eines vollstandardisierten Interviewinstruments durchgeführt werden, dann meist als selbstadministrierte Fragebogenerhebungen im Web, kaum als interviewerad ministrierte Online-Chat- oder Online-Videokonferenz-Interviews. Es kann jedoch im Rahmen einer **multimodalen Interviewstudie** zuweilen sinnvoll sein, mündliche und fernmündliche Interviews um standardisierte Chat- oder Videokonferenz-Interviews zu ergänzen, etwa wenn potenzielle Befragungspersonen für telefonische oder persönliche Interviews nicht zur Verfügung stehen. Die generelle Problematik von Online-Interviews für bevölkerungsrepräsentative Studien besteht darin, dass noch immer ein nennenswerter Teil der Bevölkerung zu den „Offlinern" gehört (v. a. in der älteren Generation) und keine Auswahlrahmen für die Internetpopulation zur Verfügung stehen (z. B. sind bislang keine Methoden bekannt, um E-Mail-Adressen per Zufallsverfahren zu generieren).

Lernquiz

1. Nennen Sie drei Techniken der unstrukturierten mündlichen Befragung!
2. Was stimmt?
   a) Bei einer offenen Frage werden keine Antwortalternativen vorgegeben.
   b) Qualitative Interviews arbeiten fast ausschließlich mit offenen Fragen.
   c) In quantitativen Interviews werden vereinzelt auch offene Fragen eingesetzt.
3. Grenzen Sie die Datenerhebung mittels Interview von der Fragebogenmethode ab und zwar hinsichtlich ihrer zentralen Merkmale sowie ihrer Vor- und Nachteile!
4. Worin bestehen die besonderen Herausforderungen bei Experteninterviews?
5. Was ist zu beachten, wenn ein standardisierter Interview-Fragebogen entwickelt wird, der sowohl in mündlichen als auch telefonischen Interviews zum Einsatz kommen soll?
6. Welche Unterschiede und Gemeinsamkeiten haben das fokussierte Interview und die Fokusgruppen-Diskussion?
7. Wie gehen Sie vor, um ein bevölkerungsrepräsentatives Sample von $n = 500$ Personen Ihres Bundeslandes telefonisch zu befragen (Interview-Fragebogen liegt vor)? Erläutern Sie Stichprobenziehung, Rekrutierung und Voraussetzungen der Interview-Durchführung!
8. Wie gehen Sie hinsichtlich Stichprobenkonstruktion vor, um in einer bestimmten Stadt eine bevölkerungsrepräsentative persönlich-mündliche Interviewstudie mit $n = 300$ Personen durchzuführen?
9. Welche Fehler beeinträchtigen gemäß Total-Survey-Error-Modell die Aussagekraft von quantitativen Interviewstudien?
10. Wie kann man Befragungsfehler durch die Interviewenden verhindern bzw. verringern?

Übungsaufgaben

1. Machen Sie sich mit dem Interview-Leitfaden zum Kernreaktorunfall von Tschernobyl im Jahr 1986 vertraut (▶ Abschn. 10.2.2). Suchen Sie in Ihrem Umfeld einen Zeitzeugen des Ereignisses und führen Sie das Leitfaden-Interview samt Audioaufzeichnung durch. Hören Sie sich die Aufzeichnung an und reflektieren Sie Ihre Erfahrung als Interviewer/in.
2. Laden Sie sich den Interview-Fragebogen zum Zensus 2011 aus dem Internet herunter. Suchen Sie in Ihrem Umfeld eine Befragungsperson und führen Sie das vollstrukturierte Interview samt Audioaufzeichnung mündlich durch. Hören Sie sich die Aufzeichnung an und reflektieren Sie Ihre Erfahrung als Interviewer/in.
3. Bestimmen Sie für einen Mehrpersonen-Haushalt aus Ihrem Umfeld gemäß dem Schwedenschlüssel („Kish selection grid") diejenige Person, die im Rahmen einer Zufallsziehung zu interviewen wäre.

Weiterführende Literatur

- Froschauer, U. & Lueger, M. (2003). *Das qualitative Interview: Zur Praxis interpretativer Analyse sozialer Systeme.* Stuttgart: UTB.
- Helfferich, C. (2005). *Die Qualität qualitativer Daten: Manual für die Durchführung qualitativer Interviews* (2. Aufl.). Wiesbaden: VS.
- Küsters, I. (2009). *Narrative Interviews: Grundlagen und Anwendungen* (2. Aufl.). Wiesbaden: VS.
- Mayer, H. O. (2009). *Interview und schriftliche Befragung: Entwicklung, Durchführung und Auswertung* (5. Aufl.). München: Oldenbourg.
- Nohl, A.-M. (2008). *Interview und dokumentarische Methode: Anleitungen für die Forschungspraxis* (2. Aufl.). Wiesbaden: VS.
- Renner, K.-H. (2011). *Das Interview.* Wiesbaden: VS.
- Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York: Guilford.
- Sapsford, R. (2007). *Survey research* (2. Aufl.). Thousand Oaks: Sage
- Weichbold, M. (2009). *Umfrageforschung: Herausforderungen und Grenzen.* Wiesbaden: VS.

Literatur

AAPOR. (2008). *Standard definitions. Final dispositions of case codes and outcome rates for surveys.* Retrieved February 24, 2012, from http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions&Template=/CM/ContentDisplay.cfm&ContentID=1273

Adams, J. & Neville, S. (2009). Men who have sex with men account for nonuse of condoms. *Qualitative Health Research, 19*(12), 1669–1677.

ADM. (2008). Richtlinie für telefonische Befragungen. Abgerufen 31. August, 2015, unter https://www.adm-ev.de/richtlinien

ADM. (2013). Jahresbericht 2013. Abgerufen 21. September, 2015, unter https://www.adm-ev.de/jahresberichte

Atteslander, P. (2010). *Methoden der empirischen Sozialforschung*. Berlin: Schmidt.

Ayling, R. & Mewse, A. J. (2009). Evaluating internet interviews with gay men. *Qualitative Health Research, 19*(4), 566–576.

Baar, R. (2010). *Allein unter Frauen: Der berufliche Habitus männlicher Grundschullehrer* (1. Aufl.). Wiesbaden: VS.

Bachleitner, R., Weichbold, M., & Aschauer, W. (2010). *Die Befragung im Kontext von Raum, Zeit und Befindlichkeit: Beiträge zu einer prozessorientierten Theorie der Umfrageforschung* (1. Aufl.). Wiesbaden: VS.

Balch, G. I. & Mertens, D. M. (1999). Focus group design and group dynamics: Lessons from deaf and hard of hearing participants. *American Journal of Evaluation, 20*(2), 265–277.

Barrios, E. X., Bayarri, S., Carbonell, I., Izquierdo, L., & Costell, E. (2008). Consumer attitudes and opinions towards functional foods: A focus group study. *Journal of Sensory Studies, 23*(4), 514–525.

Bennett, L. & McAvity, K. (1994). *Family Research: A Case for Interviewing Couples*. New York: de Gryuter.

Bernart, Y. & Krapp, S. (1997). *Das narrative Interview. Ein Leitfaden zur rekonstruktiven Interpretation*. Landau: VEP.

Blank, R. (2007). *Gruppendiskussionsverfahren*. Wiesbaden: Gabler.

Bock, M. (1992). Das halbstrukturierte-leitfadenorientierte Tiefeninterview: Theorie und Praxis der Methode am Beispiel von Paarinterviews. In J. H. P. Hoffmeyer-Zlotnik (Hrsg.), *Analyse verbaler Daten: Über den Umgang mit qualitativen Daten* (S. 90–109). Opladen: Westdeutscher Verlag.

Bogner, A., Littig, B., & Menz, W. (2005). *Das Experteninterview. Theorie, Methode, Anwendung*. Wiesbaden: VS.

Bohnsack, R., Przyborski, A., & Schäffer, B. (2010). *Das Gruppendiskussionsverfahren in der Forschungspraxis* (2. Aufl.). Leverkusen: Budrich.

Bourque, L. B. & Fielder, E. P. (2002). *How to conduct telephone surveys (survey kit)* (2nd ed.). Thousand Oaks: Sage.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

Bryman, A. (2008). *Social research methods* (3rd ed.). Oxford: Oxford University Press.

Buchwald, C. (2006). *Das Telefoninterview. Instrument der Zukunft?* Halle: Druckerei der Martin-Luther-Universität Halle-Wittenberg.

Chan, K. S., Orlando, M., Hosh–Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the center for epidemiological studies depression (CES-D) scale: An item response theory analysis. *Med Care, 42*(3), 281–289.

Charlton, M. & Neumann, K. (1990). *Medienrezeption und Identitätsbildung. Kulturpsychologische und kultursoziologische Befunde zum Gebrauch von Massenmedien im Vorschulalter*. Tübingen: Narr.

Clark, L. (2009). Focus group research with children and youth. *Journal for Specialists in Pediatric Nursing, 14*(2), 152–154.

Dammer, I. & Szymkowiak, F. (2008). *Gruppendiskussionen in der Marktforschung*. Köln: Rheingold.

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H., & Resnicow, K. (2009). Interviewer effects in public health surveys. *Health Education Research, 25*(1), 14–26.

Deutsches Institut für Normung. (2006). Markt-, Meinungs- und Sozialforschung – Begriffe und Dienstleistungsanforderungen (ISO 20252:2006). Berlin: Beuth.

Diekmann, A. (Hrsg.). (2004). *Methoden der Sozialforschung. Sonderheft 44/2004 der Kölner Zeitschrift für Soziologie und Sozialpsychologie.* Wiesbaden: VS.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method*. New York: Wiley.

Dixon, R., Maddison, R., Mhurchu, C. N., Jull, A., Meagher-Lundberg, P., & Widdowson, D. (2010). Parents' and children's perceptions of active video games: A focus group study. *Journal of Child Health Care, 14*(2), 189–199.

Döring, N. (2009). Psychological aspects of interviewing by cellular telephone. In E. Maxl, N. Döring & A. Wallisch (Hrsg.), *Mobile market research* (pp. 40–80). Köln: Herbert von Halem.

Dorussen, H., Lenz, H., & Blavoukos, S. (2005). Assessing the reliability and validity of expert interviews. *European Union Politics, 6*(3), 315–337.

Dresing, T. & Pehl, T. (2010). Transkription. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 723–733). Wiesbaden: Gabler.

Ericsson, K. & Simon, H. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251.

Ericsson, K. & Simon, H. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Boston: MIT Press.

Farnsworth, J. & Boon, B. (2010). Analysing group dynamics within the focus group. *Qualitative Research, 10*(5), 605–624.

Flick, U. (1995). Stationen des qualitativen Forschungsprozesses. In U. Flick, E. v. Kardorff, H. Keupp, L. von Rosenstiel & S. Wolff (Hrsg.), *Handbuch Qualitative Sozialforschung* (S. 148–176). München: PVU.

Flick, U. (2002). *Qualitative Sozialfoschung. Eine Einführung*. Reinbek: Rowohlt.

Garcia, A. C., Standlee, A. I., Bechkoff, J., & Cui, Y. (2009). Ethnographic approaches to the internet and computer-mediated communication. *Journal of Contemporary Ethnography, 38*(1), 52–84.

Garcia, C. A. & Lane, J. (2009). What a girl wants, what a girl needs: Findings form a gender-specific focus group study. *Crime & Delinquency*. Retrieved December 6, 2013, from http://cad.sagepub.com/content/early/2009/04/03/0011128709331790.abstract

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly, 69*(1), 124–157.

Gläser, J. & Laudel, G. (2004). *Experteninterviews und qualitative Inhaltsanalyse*. Wiesbaden: VS.

Goudy, W. J. & Potter, H. R. (1975). Interview rapport: Demise of a concept. *Public Opinion Quarterly, 39*(4), 529–543.

Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nicholls, W. L., & Waksberg, J. (Eds.). (2001). *Telephone survey methodology* (1st ed.). New York: Wiley.

Gwartney, P. A. (2007). *The telephone interviewer's handbook: How to conduct standardized conversations* (2nd ed.). New York: Wiley.

Häder, M. (2006). *Empirische Sozialforschung*. Wiesbaden: VS.

Häder, M. & Kühne, M. (2009). Die Prägung des Antwortverhaltens durch die soziale Erwünschtheit. In M. Häder & S. Häder (Hrsg.), *Telefonbefragungen über das Mobilfunknetz* (S. 175–186). Wiesbaden: VS.

Häder, S. & Glemser, A. (2004). Stichprobenziehung für Telefonumfragen in Deutschland. In A. Diekmann (Hrsg.), *Methoden der Sozialforschung. Sonderheft 44/2004 der Kölner Zeitschrift für Soziologie und Sozialpsychologie.* (Vol. 44, S. 148–171). Wiesbaden: VS.

Halcomb, E. J. & Davidson, P. M. (2006). Is verbatim transcription of interview data always necessary? *Applied Nursing Research, 19*(1), 38–42.

Henderson, S. & Gilding, M. (2004). „I've never clicked this much with anyone in my life": Trust and hyperpersonal communication in online friendships. *New Media & Society, 6*(4), 487–506.

Hoffmann, D., Krauß, F., & Gäbel, M. (2005). Erotische Körperinszenierungen. Lesearten von 16- bis 18-jährigen Jugendlichen. *tv diskurs – Verantwortung in audiovisuellen Medien, 34*(4), 26–32.

Hunsicker, S. & Schroth, Y. (2007). Die Kombination von Mobilfunk- und Festnetzstichproben. *Methoden – Daten – Analysen, 1*(2), 161–182.

Irvine, A., Drew, P., & Sainsbury, R. (2010). Mode effects in qualitative interviews: A comparison of semi-structured face–to–face and telephone interviews using conversation analysis. Retrieved December 6, 2013, from http://www.york.ac.uk/inst/spru/pubs/rworks/2010-03July.pdf.

Kincaid, H. V. & Bright, M. (1957). The tandem interview: A trial of the two-interviewer team. *Public Opinion Quarterly, 21*, 304–312.

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association, 44*(247), 380–387.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Krueger, R. & Casey, M. A. (2008). *Focus groups: A practical guide for applied research* (4th ed.). Thousand Oaks: Pine Forge Press.

Kuckartz, U. (2009). *Einführung in die computergestützte Analyse qualitativer Daten* (3. Aufl.). Wiesbaden: VS.

Küsters, I. (2009). *Narrative Interviews: Grundlagen und Anwendungen* (2. Aufl.). Wiesbaden: VS.

Lamnek, S. (2005). *Gruppendiskussion: Theorie und Praxis.* (2. Aufl.). München: UTB.

Legewie, H., Böhm, A., Boehnke, K., Faas, A., Gross, B., & Jaeggi, E. (1990). *Längerfristige psychische Folgen von Umweltbelastungen: Das Beispiel Tschernobyl. Abschlußbericht des Forschungsinitiativprojekts FIP 2/17 TU Berlin* (No. FIP 2/17). Berlin: Institut für Psychologie der TU Berlin.

Leithäuser, T. & Volmerg, B. (1988). *Psychoanalyse in der Sozialforschung. Eine Einführung am Beispiel einer Sozialpsychologie der Arbeit*. Opladen: Westdeutscher Verlag.

Littig, B. (2008). Interviews with the elite and with experts: Are there any differences? *FQS Forum Qualitative Social Research, 9*(3), Art. 16. Retrieved December 6, 2013, from http://nbn-resolving.de/urn:nbn:de:0114-fqs0803161.

Lucius-Hoene, G. & Deppermann, A. (2002). *Rekonstruktion narrativer Identität. Ein Arbeitsbuch zur Analyse narrativer Interviews*. Opladen: Leske & Budrich.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah: Erlbaum.

Martens, B. & Ritter, T. (Hrsg.). (2008). *Eliten am Telefon. Neue Formen von Experteninterviews in der Praxis*. Baden-Baden: NOMOS.

Medjedovic, I. & Witzel, A. (2010). *Wiederverwendung qualitativer Daten: Archivierung und Sekundärnutzung qualitativer Interviewtranskripte*. Wiesbaden: VS.

Merton, R. K., Fiske, M., & Kendall, P. L. (1990). *The focused interview. A manual of problems and procedures* (2nd ed.). New York: The Free Press.

Meuser, M. & Nagel, U. (1991). ExpertInneninterviews. Vielfach erprobt, wenig bedacht. Ein Beitrag zur qualitativen Methodendiskussion. In D. Garz & K. Kraimer (Hrsg.), *Qualitativ-empirische Sozialforschung : Konzepte, Methoden, Analysen.* (S. 441–471). Opladen: Westdeutscher Verlag.

Mey, G. (2000). Erzählungen in qualitativen Interviews: Konzepte, Probleme, soziale Konstruktion. *Sozialer Sinn, 1*(1), 135–151.

Mey, G. (2005). Das (Wieder-)Erfinden von Interviewverfahren. Kommentar zu „Das existenzielle Interview". *Journal für Psychologie, 12*(3), 273–282.

Mey, G. & Mruck, K. (2007). *Qualitative Interviews*. Wiesbaden: Gabler.

Mey, G. & Mruck, K. (Hrsg.). (2010). *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS.

Möhring, W. & Schlütz, D. (2010). *Die Befragung in der Medien- und Kommunikationswissenschaft* (2. Aufl.). Wiesbaden: VS.

o. A. (1989). *Goldmanns großer Cartoonband. Schweine mit Igeln*. München: Goldmann.

Oldfield, R. C. (1951). *The psychology of the interview*. London: Methuen.

Olson, K. & Peytchev, A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly, 71*(2), 273–286.

Opdenacker, R. (2006). Advantages and Disadvantages of Four Interview Techniques in Qualitative Research. *FQS Forum Qualitative Social Research, 7*(4), Art. 11. Retrieved December 6, 2013, from http://nbn-resolving.de/urn:nbn:de:0114-fqs0604118.

Paus-Haase, I. (2000). *Medienrezeptionsforschung mit Kindern*. München: Kopäd.

Perleberg, K., Schütze, F., & Heine, V. (2006). Sozialwissenschaftliche Biographieanalyse von chronisch kranken Patientinnen auf der empirischen Grundlage des autobiographisch-narrativen Interviews. Exemplifiziert an der Lebensgeschichte einer jungen Patientin mit Morbus Crohn. *Psychotherapie und Sozialwissenschaft, 8*(1), 95–145.

Petersen, T. (2002). *Das Feldexperiment in der Umfrageforschung.* Frankfurt a. M.: Campus.

Proner, H. (2011). *Ist keine Antwort auch eine Antwort? Die Teilnahme an politischen Umfragen*. Wiesbaden: VS.

Raczka, R. (2005). A focus group enquiry into stress experienced by staff working with people with challenging behaviours. *Journal of Intellectual Disabilities, 9*(2), 167–177.

Rehbein, J., Schmidt, T., Meyer, B., Watzke, F., & Herkenrath, A. (2004). Handbuch für das computergestützte Transkribieren nach HIAT. Arbeiten zur Mehrsprachigkeit. (56). Abgerufen 6. Dezember, 2013, unter http://www1.unihamburg.de/exmaralda/files/azm_56.pdf

Reinders, H. (2005). *Qualitative Interviews mit Jugendlichen führen. Ein Leitfaden*. München: Oldenbourg.

Rezabek, R. J. (2000). Online focus groups: Electronic discussions for research. *Forum Qualitative Sozialforschung, 1*(1). Retrieved December 6, 2013, from http://www.qualitative-research.net/index.php/fqs/article/view/1128.

Rochlen, A. B., Paterniti, D. A., Epstein, R. M., Duberstein, P., Willeford, L., & Kravitz, R. L. (2010). Barriers in diagnosing and treating men with depression: A focus group report. *American Journal of Men's Health, 4*(2), 167–175.

Rosenthal, G. (2002). Biographisch-narrative Gesprächsführung: Zu den Bedingungen heilsamen Erzählens im Forschungs- und Beratungskontext. *Psychotherapie und Sozialwissenschaften, 4*, 204–227.

Salmon, C. T. & Nichols, J. S. (1983). The next-birthday method of respondent selection. *Public Opinion Quarterly, 47*(2), 270–276.

Salmons, J. (2009). *Online interviews in real time*. Thousand Oaks: Sage.

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2009). *Research methods of business students* (5th ed.). Harlow: Pearson.

Scheibelhofer, E. (2008). combining narration-based interviews with topical interviews: Methodological reflections on research practices. *International Journal of Social Research Methodology, 11*(5), 403–416.

Schneider, S. & Margraf, J. (2008). *DIPS. Diagnostisches Interview bei psychischen Störungen: Handbuch, Interview-Leitfaden, Protokollbogen* (3. Aufl.). Berlin: Springer.

Schneiderat, G. & Schlinzig, T. (2009). Teilnahmebereitschaft und Teinahmeverhalten bei Telefonumfragen in der Allgemeinbevölkerung über das Mobilfunknetz. In M. Häder & S. Häder (Hrsg.), *Telefonumfragen über das Mobilfunknetz* (S. 83–98). Wiesbaden: VS.

Schnell, R. (1991). Wer ist das Volk? Zur faktischen Grundgesamtheit bei „ allgemeinen Bevölkerungsumfragen". Undercoverage, Schwererreichbare und Nichtbefragbare. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43*(1), 106–137.

Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der empirischen Sozialforschung*. München: Oldenbourg.

Schorn, A. (2000). Das „themenzentrierte Interview". Ein Verfahren zur Entschlüsselung manifester und latenter subjektiver Wirklichkeit. *FQS Forum Qualitative Social Research, 1*(2), Art. 23. Abgerufen 6. Dezember, 2013, unter http://www.qualitativeresearch.net/index.php/fqs/article/view/1092/2393.

Schuman, H. & Presser, S. (1981). *Questions & answers in attitude surveys: Experiments on question form, wording, and context.* San Diego: Academic Press.

Schütze, F. (1976a). Zur Hervorlockung und Analyse thematisch relevanter Geschichten im Rahmen soziologischer Feldforschung. In Arbeitsgruppe Bielefelder Soziologen (Hrsg.), *Kommunikative Sozialforschung* (S. 159–260). München: Fink.

Schütze, F. (1976b). *Zur soziologischen und linguistischen Analyse von Erzählungen.* Wiesbaden: Westdeutscher Verlag.

Schütze, F. (1977a). *Die Technik des narrativen Interviews in Interaktionsfeldstudien.* Unpublished Unveröffentlichtes Manuskript. Fakultät für Soziologie, Universität Bielefeld.

Schütze, F. (1977b). *Die Technik des narrativen Interviews in Interaktionsfeldstudien – dargestellt an einem Projekt zur Erforschung von kommunalen Machtstrukturen (MS)* (Arbeitsberichte und Forschungsmaterialien). Bielefeld: Fakultät für Soziologie.

Schütze, F. (1983). Biographieforschung und narratives Interview. *Neue Praxis, 3*, 283–293.

Schütze, F. (1984). Kognitive Figuren des autobiographischen Stegreiferzählens. In M. Kohli & G. Robert (Hrsg.), *Biographie und soziale Wirklichkeit* (S. 78–117). Stuttgart: Enke.

Selting, M. (2001). Gesprächsanalytisches Transkritionssystem (GAT). *Linguistische Berichte, 173*, 91–122.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., et al. (2009). Gesprächsanalytisches Transkriptionssystem 2. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion, 10*, 353–402. Abgerufen 6. Dezember, 2013, unter http://www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2009/px-gat2.pdf

Silver, H. J. (2004). The lived experience of home total parenteral nutrition: An online qualitative inquiry with adults, children, and mothers. *Nutrition in Clinical Practice, 19*(3), 297–304.

Stassen, P. & Seefeldt, S. (1991). HIV als Grenzsituation. Unveröffentlichte Diplomarbeit. Technische Universität Berlin, Institut für Psychologie.

Stock, D. (1994). Biografische Sinnfindung in einem sozialistischen Land. Unveröffentlichte Diplomarbeit. Technische Universität Berlin, Institut für Psychologie.

Straus, S. G., Miles, J. A., & Levesque, L. L. (2001). The effects of videosconference, Telephone, and face–to–face media on interviewer and applicant judgments in employment interviews. *Journal of Management, 27*(3), 363–381.

Sylow, M. & Holm, L. (2009). Building groups and independence: The role of food in the lives of young people in danish sports centres. *Childhood, 16*(2), 213–228.

Trautmann, T. (2009). *Interviews mit Kindern: Grundlagen, Techniken, Besonderheiten, Beispiele*. Wiesbaden: VS.

Verhoeven, P., Kirkland, D., Rheims, B., Mark , M. E., & Close, M. (1995). *Showgirls. Portrait of a Film*. New York: Newmarket Press.

Völzke, R. (2005). Erzählen. Brückenschlag zwischen Leben und Lernen. Angeleitete biografisch-narrative Kommunikation in Ausbildung und Praxis sozialer Arbeit. *Sozialextra. Zeitschrift für Soziale Arbeit und Sozialpolitik, 29*, 12–15.

Walther, E., Preckel, F., & Mecklenbräuker, S. (2010). *Befragung von Kindern und Jugendlichen: Grundlagen, Methoden und Anwendungsfelder*. Göttingen: Hogrefe.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press.

Wengraf, T. (2001). *Qualitative research interviewing. Biographic narrative and semistructured method*. Thousand Oaks: Sage.

Wilson, E. C. (1950). Adapting probability sampling to western europe. *Public Opinion Quarterly, 14*(2), 215–223.

Witzel, A. (1982). *Verfahren der qualitativen Sozialforschung. Überblick und Alternativen*. Frankfurt a. M.: Campus.

Witzel, A. (1989). Das problemzentrierte Interview. In G. Jüttemann (Hrsg.), *Qualitative Forschung in der Psychologie: Grundfragen, Verfahrensweisen, Anwendungsfelder* (S. 227–256). Heidelberg: Asanger.

Witzel, A. (1996). Auswertung problemzentrierter Interviews. Grundlagen und Erfahrungen. In R. Strobl & A. Böttger (Hrsg.), *Wahre Geschichten? Zur Theorie und Praxis qualitativer Interviews* (S. 49–76). Baden-Baden: Nomos.

Witzel, A. (2000). Das problemzentrierte Interview. *FQS Forum Qualitative Social Research, 1*(1), Art. 22. Abgerufen 6. Dezember, 2013, unter http://www.qualitative-research.net/index.php/fqs/article/view/1132/2519

Woelfer, C. (2000). Das personenzentrierte Interview als qualitative Forschungsmethode. *Journal für Psychologie, 8*(1), 3–13.

Woolsey, C. & Bracy, K. (2010). Emergency response and the psychological needs of school-age children. *Traumatology, 16*(2), 1–6.

10.3 Fragebogen

Lernziele

- Wissen, was die Fragebogenmethode als wissenschaftliche Datenerhebungsmethode auszeichnet.
- Verschiedene Typen wissenschaftlicher Fragebögen im Hinblick auf Herangehensweise und Anwendungsfelder voneinander abgrenzen können.
- Potenzielle Fragebogenfehler und Gegenmaßnahmen kennen.
- Eine qualitative sowie eine quantitative Fragebogenstudie planen und durchführen können.
- Die jeweiligen Vor- und Nachteile qualitativer und quantitativer schriftlicher Befragungen abwägen können.

Dieses Kapitel erklärt, wodurch sich eine wissenschaftliche schriftliche Befragung bzw. die wissenschaftliche Fragebogenmethode als empirische Datenerhebungsmethode auszeichnet (▶ Abschn. 10.3.1). Daraufhin werden jeweils verschiedene Varianten der qualitativen (▶ Abschn. 10.3.2) sowie der quantitativen Fragebogenmethode (▶ Abschn. 10.3.3) erläutert und an Studienbeispielen veranschaulicht.

10.3.1 Die wissenschaftliche Fragebogenmethode

Wie ist die Fragebogentechnik definiert und für welche Zwecke ist sie in der empirischen Sozialforschung geeignet bzw. ungeeignet? Wie unterscheidet sich der wissenschaftliche Fragebogen von nicht-wissenschaftlichen schriftlichen Befragungen im Alltag? Anhand welcher Klassifikationskriterien lassen sich unterschiedliche Typen von Fragebögen voneinander abgrenzen? Welche typischen Fehler sind mit Fragebogenstudien verbunden und welche Gegenmaßnahmen sollten ergriffen werden? Die folgenden Abschnitte geben hierauf Antwort.

Definition und Indikation der wissenschaftlichen Fragebogenmethode

Die Befragungsmethode ist in der empirischen Sozialforschung generell dominierend. Besonders populär sind dabei in der qualitativen Forschung die nicht- oder halbstrukturierten mündlichen Befragungentechniken (v. a. Leitfaden-Interiews; ▶ Abschn. 10.2.2). In der quantitativen Forschung dagegen wird am häufigsten nicht mit mündlichen, sondern mit vollstrukturierten schriftlichen Befragungen (auf der Basis standardisierter Fragebögen) gearbeitet. Es existieren jedoch auch qualitative schriftliche Befragungen.

Wissenschaftliche Fragebogenmethode – Unter der wissenschaftlichen Fragebogenmethode („questionnaire method“) verstehen wir die zielgerichtete, systematische und regelgeleitete Generierung und Erfassung von verbalen und numerischen Selbstauskünften von Befragungspersonen zu ausgewählten Aspekten ihres Erlebens und Verhaltens in schriftlicher Form. Als Erhebungsinstrument fungiert ein **wissenschaftlicher Fragebogen** („research questionnaire“, „scientific questionnaire“), der von den Befragungspersonen eigenständig bzw. selbstadministriert ausgefüllt wird („self-administered questionnaire“). Der Fragebogen kann in einer Face-to-Face-Situation, postalisch, per Internet oder mobilem Endgerät zur Verfügung gestellt werden. Die Antworten werden dokumentiert und systematisch analysiert. Die drei zentralen Elemente der schriftlichen Befragung sind a) die Befragungspersonen, b) der Fragebogen sowie c) die Situation, in der der Fragebogen beantwortet wird.

Unter welchen Umständen ist die Fragebogentechnik als selbstadministrierte schriftliche Befragungsmethode besonders gut geeignet **(Indikation)?**

- Als Befragungstechnik ist die Fragebogenmethode – ebenso wie die Interviewmethode – in der Lage, Aspekte des **subjektiven Erlebens** sowie des **vergangenen oder privaten Verhaltens** zu erfassen, die nicht direkt beobachtbar und auch nicht in Verhaltensspuren oder Dokumenten manifestiert sind.
- Im Vergleich zur Interviewtechnik ist die Fragebogenmethode **durch Selbstadministration viel effizienter**: In kurzer Zeit können Fragebogenantworten von vielen Befragungspersonen zu sehr vielen Merkmalen gesammelt werden. Es müssen keine Interviewerinnen und Interviewer rekrutiert, geschult und ins Feld geschickt werden. Auch sind viele Menschen eher bereit, einen Fragebogen auszufüllen als einen Interviewtermin zu verabreden und einzuhalten.
- Das Ausfüllen eines Fragebogens ist aus Sicht der Befragungspersonen **diskreter und anonymer** als eine Interviewsituation. Dementsprechend können mittels Fragebogen heikle und intime Themen besser erhoben werden als im Interview.

Die Fragebogenmethode ist jedoch auch mit **Nachteilen** verbunden:

- Das selbstständige Ausfüllen eines Fragebogens erfordert **Lese- und Schreibkompetenzen** und schließt eine Reihe von Zielgruppen aus (z. B. Kinder, Menschen mit geringer formaler Bildung). Es sei denn, für diese Zielgruppen werden spezielle Fragebögen entwickelt (etwa sehr einfache Fragebögen für Kinder, die z. B. auf symbolischen Ratingskalen mit Smileys zu beantworten sind, zu Ratingskalen ▶ Abschn. 8.4.4).
- Sehr umfangreiche und komplexe Antworten sind schriftlich nicht zu erwarten, vielmehr muss man sich auf **wenige** und **klar umschriebene Befragungsinhalte** beschränken, während im mündlichen Interview längere Erzählungen möglich sind.

- Die Fragebogentechnik ist wie die Interviewtechnik eine **reaktive Methode**, denn die Befragungspersonen sind sich dessen bewusst, dass sie an einer wissenschaftlichen Erhebung teilnehmen. Die mithilfe des Fragebogens **forschungsgenerierten Daten** würden normalerweise nicht entstehen und sind maßgeblich von den Eigenschaften des Fragebogens sowie der Situation, in der der Fragebogen bearbeitet wird, abhängig. Diese Faktoren können die Aussagekraft der gewonnenen Daten einschränken.
- Da die Befragungspersonen den Fragebogen eigenständig und meist in Abwesenheit der Forschenden bearbeiten, besteht im Unterschied zum Live-Kontakt beim Interview nicht die Möglichkeit, a) unmittelbar Rückfragen zu beantworten, b) individuell auf die Befragungsperson einzugehen oder c) sich einen Eindruck von der Befragungsperson und der Ausfüllsituation zu verschaffen. Die **Umstände der Datenerhebung sind somit weniger transparent und gestaltbar** als im Interview.

Abgrenzung der wissenschaftlichen Fragebogenmethode von nichtwissenschaftlichen schriftlichen Befragungen

Fragebögen sind im Alltag weit verbreitet. Egal, ob wir einen Wandschrank bestellen, in einen Sportverein oder eine Partei eintreten, eine Operation durchführen lassen, heiraten, Transferleistungen beantragen, einen Diebstahl melden, ein Visum benötigen, ein Hotel bewerten oder eine Online-Kontaktanzeige schalten – in jedem Fall müssen mehr oder minder umfangreiche **Formulare** bzw. **Fragebögen** ausgefüllt werden. Der Einsatz dieser Fragebögen erfolgt, genau wie bei der wissenschaftlichen schriftlichen Befragung, zielgerichtet und systematisch.

Was die wissenschaftliche schriftliche Befragung von anderen Fragebogenanwendungen grundlegend unterscheidet, ist ihre konsequente Orientierung an **Regeln der Wissenschaftlichkeit** mit dem Ziel, zu **generalisierbaren Aussagen** über die Erfahrungswirklichkeit zu gelangen. Demgegenüber werden die nicht-wissenschaftlich erfassten Fragebogendaten meist für die Beurteilung einzelner Personen bzw. die Abwicklung bestimmter Prozesse verwendet. Auch beschränken sich Formulare im Alltag meist auf wenige einfache Variablen, während wissenschaftliche Fragebögen oft auf die Erfassung **komplexer** und **latenter Merkmale** von Personen (z. B. Erziehungsstil, Religiosität, Mediennutzungsmotive, Selbstwertgefühl) hinauslaufen. Nicht zuletzt orientiert sich die wissenschaftliche Datenerhebung mittels Fragebogenmethode an forschungsethischen Kriterien und garantiert z. B. Freiwilligkeit und Anonymität.

Eine wissenschaftliche schriftliche Befragungsstudie zeichnet sich dadurch aus, dass

- der gesamte Forschngsprozess auf die Lösung eines bestimmten **Forschungsproblems** ausgerichtet ist und daraus auch das **Untersuchungsdesign** abgeleitet wurde (z. B. Entscheidung für eine qualitative oder quantitative Fragebogenstudie),
- die **Konstruktion des Fragebogeninstruments** wissenschaftlichen Kriterien entspricht und diese überprüft wurden,
- die **Auswahl der Befragungspersonen** wissenschaftlichen Kriterien der Stichprobenziehung entspricht (sofern keine Vollerhebung stattfindet),
- **Verbreitung, Bearbeitung** und **Rücklauf der Fragebögen** systematisch und dokumentiert ablaufen,
- die Aufbereitung und **Auswertung der Fragebogendaten** wissenschaftlichen qualitativen bzw. quantitativen Analysemethoden folgt und
- die gesamte Studie **forschungsethischen Kriterien** genügt.

Klassifikationskriterien für wissenschaftliche schriftliche Befragungen

Zur Abgrenzung unterschiedlicher schriftlicher Befragungstechniken bzw. Fragebogenformen lassen sich vor allem **sechs Klassifikationskriterien** heranziehen.

1. **Grad der Strukturierung der Befragung** bzw. **Grad der Standardisierung des Fragebogens** mit den drei Stufen nicht-standardisiert, teil-/halbstandardisiert und vollstandardisiert. Bei einem **nichtstandardisierten Fragebogen** äußern sich die Befragungspersonen schriftlich in eigenen Worten zu einem Thema, ohne dass ein detaillierter Fragenkatalog vorgelegt wird. In einem **teilstandardisierten Fragebogen** wird – analog dem Interviewleitfaden – eine Liste von offenen Fragen schriftlich vorgelegt. Nicht- und teilstandardisierte Fragebögen werden als **qualitative Fragebögen** („qualitative questionnaires“) bezeichnet. Ein **vollstandardisierter Fragebogen** besteht aus geschlossenen Fragen (d. h. Fragen oder Items mit Antwortvorgaben), die in der Regel anzukreuzen oder mit Zahlenwerten zu beantworten sind. **Vollstandardisierte Fragebögen** („standardized questionnaires“) werden auch als **quantitative Fragebögen** („ quantitative questionnaires“) bezeichnet und stellen die Grundlage der quantitativen Umfrageforschung dar („survey research“). Bei einer vollstrukturierten Fragebogenstudie füllen die Befragten den vollstandardisierten Fragebogen selbst aus, bei einer vollstrukturierten Interviewstudie lesen die Interviewer/innen den vollstandardisierten Fragebogen vor und

die Befragten beantworten alle Fragen nacheinander mündlich („interviewer-administered questionnaire“, ► Abschn. 10.2.3).

2. **Modus der schriftlichen Befragung** („questionnaire mode“) in den Varianten Paper-Pencil-Fragebogen und elektronischer Fragebogen. Der **Paper-Pencil-Fragebogen** („paper-pencil questionnaire“) liegt ausgedruckt auf Papier vor und wird mit einem Stift ausgefüllt. Die Daten müssen anschließend elektronisch erfasst werden, entweder per Hand oder – bei Einsatz eines maschinenlesbaren Fragebogens – automatisch. Der **elektronische Fragebogen** („electronic questionnaire“) liegt digital vor und wird auf einem Computermedium ausgefüllt, typischerweise am Desktop-PC bzw. Laptop oder auf einem mobilen Endgerät (z. B. Smartphone, Tablet PC). Bei elektronischen Fragebögen werden die Antworten gleich digital erfasst. Elektronische Fragebögen sollten auf das genutzte Endgerät abgestimmt sein, so können Fragebögen, die auf dem Desktop-PC auszufüllen sind („computer questionnaire“) umfangreicher und komplexer gestaltet sein als Fragebögen, die auf dem Smartphone zu bearbeiten sind („mobile questionnaire“).
3. **Verbreitungsweg des Fragebogens** („questionnaire administration“) in den Varianten Austeilen und Einsammeln, postalisch, online und mobil. Paper-Pencil-Fragebögen werden häufig an anwesende Personen ausgeteilt und direkt wieder eingesammelt („delivery and collection questionnaire“). Am effizientesten ist diese Vorgehensweise im Gruppenkontext (z. B. Klassenraumbefragung). Ein anderer Verbreitungsweg ist die **postalische Umfrage.** Hier werden die Paper-Pencil-Fragebögen (sehr selten auch elektronische Fragebögen auf Datenträgern) samt Rückumschlag mit der Post verschickt oder persönlich ausgeteilt und per Post zurückgeschickt („mail/postal questionnaire“). **Online-Umfragen** gewinnen an Beliebtheit, hier wird der elektronische Fragebogen im Internet oder innerhalb von Organisationen im Intranet zur Verfügung gestellt („internet-mediated questionnaire“, „intranet-mediated questionnaire“). Im Zuge der Veralltäglichung der Mobilkommunikation werden inzwischen auch **mobile Umfragen** genutzt, bei denen die Befragungspersonen den elektronischen Fragebogen auf ihrem mobilen Endgerät bearbeiten („mobile questionnaire“). Nicht zuletzt lassen sich Fragebögen auch selbstadministriert über **automatisierte Telefonbefragungen** erheben („interactive voice response“, IVR; „telephone audio computer assisted self-interviewing“, T-ACASI). Mit der zunehmenden Mediatisierung des Alltags wächst die Zahl der mediengestützten Verbreitungswege für Fragebögen (Couper, 2005, 2011).

Vom Verbreitungsweg abzugrenzen ist die Art der Stichprobe bzw. die Rekrutierung, bei der ein aktives und ein passives Vorgehen zu differenzieren ist. Bei der **aktiven Stichprobenziehung** wird der Fragebogen gezielt genau den für die Stichprobe ausgewählten Personen bereitgestellt (z. B. persönliche E-Mail-Einladungen zu einer Online-Umfrage), bei der **passiven Stichprobenziehung** wird der Fragebogen mehr oder minder breit gestreut, so dass sich per Selbstselektion eine Stichpobe bildet (z. B. Link zur Online-Umfrage wird in mehreren Foren und auf einer Website publiziert).

4. **Art der Befragungspersonen** in den Varianten Betroffene versus Experten. Die meisten Fragebögen zielen darauf ab, Selbstauskünfte von **Betroffenen** zu erhalten. Zuweilen werden jedoch analog dem Experteninterview (► Abschn. 10.2.2) schriftliche **Expertenbefragungen** durchgeführt.
5. **Befragung zu Einzelpersonen oder zu Gruppen.** Üblicherweise werden Fragebögen unabhängig voneinander von einzelnen Personen ausgefüllt und beziehen sich auf das Verhalten und Erleben des jeweiligen Individuums **(Befragung zu Einzelpersonen)**. Zusätzlich existieren zwei verschiedene Varianten von **Befragungen zu Gruppen**.

   Bei der **Delphi-Methode** („delphi survey“, „delphi poll“, „delphi method“), geht es darum, das fachliche Urteil einer Expertengruppe zu einem (meist zukunftsbezogenen) Sachverhalt systematisch zu bündeln und zu schärfen, indem mindestens zwei Beurteilungsrunden durchgeführt und den Experten die Urteile der Mitexperten vorgelegt werden.

   Daneben existiert ein weiterer Typus der Gruppenbefragung. Dieser hat **natürliche soziale Gebilde** wie Paare, Kleingruppen, soziale Netzwerke oder Organisationen zum Gegenstand. Diese Befragungstechniken beziehen sich inhaltlich auf die Erfassung der Verhältnisse zwischen den befragten Personen. Zudem verlangen bzw. erlauben es diese auf soziale Gebilde bezogenen Befragungstechniken teilweise, dass alle Mitglieder des jeweiligen sozialen Gebildes an der Studie teilnehmen und somit nicht unabhängige, sondern miteinander bekannte Personen zu rekrutieren und zu befragen sind. Dabei werden die Antworten der jeweils miteinander bekannten Befragungspersonen direkt untereinander in Beziehung gesetzt, um das interessierende soziale Gebilde aus Sicht aller Beteiligten zu beleuchten: z. B. werden alle Mitglieder einer natürlichen Kleingruppe zu ihren Beziehungen zu allen anderen Gruppenmitgliedern mit speziellen soziometrischen Fragebögen befragt und die Ergebnisse zu einer Darstellung der Gruppenstruktur (dem Soziogramm) zusammengefasst.

Tabelle 10.10 Verbreitete Varianten der schriftlichen Befragung und ihre Einordnung

Schriftliche Befragung				
	Qualitative schriftliche Befragung		Quantitative schriftliche Befragung	
Grad der Strukturierung	Unstrukturierte schriftliche Befragung	Halbstrukturierte schriftliche Befragung	Vollstrukturierte schriftliche Befragung	
Befragung zu Einzelpersonen oder Gruppen	Einzelpersonen	Einzelpersonen	Einzelpersonen	Gruppen
Varianten schriftlicher Befragung (Art der Befragungspersonen, Verbreitungsweg des Fragebogens, spezielles Abfrageformat)	– Aufforderung zu schriftlichen autobiografischen Erzählungen – Aufforderung zu schriftlichen Erklärungen – Schriftliche projektive Methoden	– Befragung von Experten – Befragung von Betroffenen – Halbstandardisierte Tagebuchmethode	– Befragung per Austeilen und Einsammeln – Postalische Befragung – Online-Befragung – Mobile Befragung – Automatische Telefonbefragung – Vollstandardisierte Tagebuchmethode (ggf. kombiniert mit Ambulantem Assessment)	– Delphi-Befragung – Befragung zu Paaren – Befragung zu Kleingruppen – Befragung zu sozialen Netzwerken – Befragung zu Organisationen

6. **Spezielles Abfrageformat** mit einer Fülle von Varianten, etwa der Tagebuchmethode als Instrument zur fortlaufenden (mindestens täglichen) detaillierten Protokollierung alltäglichen Verhaltens und Erlebens oder dem semantischen Differenzial als Skalierungstechnik zur Erfassung des emotionalen Bedeutungsgehaltes von Objekten (▶ Abschn. 8.6.4).

Eine Auswahl verbreiteter Varianten der Fragebogenmethode wird im vorliegenden Kapitel behandelt (zur Übersicht ◘ Tab. 10.10).

10.3.2 Qualitative Fragebogenmethode

Qualitative Fragebögen werden in der Forschungspraxis viel seltener eingesetzt als qualitative Interviews. Im Folgenden wird die Datenerhebung mittels nicht-strukturierter und halbstrukturierter schriftlicher Befragung anhand von Beispielen behandelt.

Qualitative Fragebogenmethode – Die qualitative Fragebogenmethode beinhaltet unstrukturierte und halbstrukturierte schriftliche Befragungen. Dabei kommen als Erhebungsintrumente qualitative Fragebögen („qualitative questionnaires“) zum Einsatz, die offene Fragen beinhalten, so dass sich die Befragten schriftlich in eigenen Worten äußern. Man unterscheidet nicht-standardisierte und teilstandardisierte qualitative Fragebögen.

Unstrukturierte schriftliche Befragung

Im Kern basiert die unstrukturierte schriftliche Befragungstechnik darauf, dass die Befragten die Aufforderung erhalten, einen Aufsatz („essay“) zu schreiben oder sich anderweitig ausführlich schriftlich zu äußern. Der nicht-standardisierte Fragebogen enthält die entsprechende Arbeitsaufforderung sowie viel Platz, da sich die Befragten in eigenen Worten artikulieren sollen. Ergänzend können am Ende eines nicht-standardisierten Fragebogens einige Fragen nach sozialstatistischen oder sonstigen relevanten Hintergrundinformationen stehen.

Nicht-standardisierter Fragebogen – Bei der nicht-strukturierten bzw. unstrukturierten schriftlichen Befragung wird ein nicht-standardisierter Fragebogen („nonstandardized questionnaire“) genutzt. In diesem werden die Befragungspersonen mittels einer einzelnen Frage oder Aufgabe zu einem bestimmten Thema um ausführliche schriftliche Beschreibungen, Erklärungen oder Erzählungen in Aufsatzform gebeten.

Nicht-strukturierte schriftliche Befragungen und die zugehörigen nicht-standardisierten Fragebogeninstrumente werden in der Methodenliteratur kaum behandelt. Das Spektrum der Möglichkeiten sei hier anhand von Beispielen verdeutlicht.

- **Schriftliche autobiografische Erzählungen**: Befragungspersonen können in einem nicht-standardisierten Fragebogen gebeten werden, autobiografische Erlebnisse aufzuschreiben, analog dem narrativen Interview (▶ Abschn. 10.2.2). Dabei kann es sich etwa um berufliche, Medien- oder Krankheitsbiografien handeln. In einer unstrukturierten qualitativen Fragebogenstudie wurden von 276 Jugendlichen autobiografische Erzählungen zum Rauchen der ersten Zigarette ihres Lebens gesammelt (Delorme, Kreshel, & Reid, 2003). Die Auswertung dieser Aufsätze zeigte, dass die erste Zigarette für die Jugendlichen ein symbolisch wichtiges Lebensereignis war, das

sie mit den entsprechenden Begleitumständen detailliert schilderten. Erkenntnisse über die symbolische Bedeutung des Rauchens der ersten Zigarette können z. B. dazu dienen, die Aufklärung über die Risiken des Rauchens besser auf die Zielgruppe abzustimmen.

- **Schriftliche Erklärungen**: Um die mentalen Modelle zu Ursache-Wirkungs-Relationen von Befragungspersonen zu explorieren, kann man sie statt um Erzählungen auch schriftlich um Erklärungen bitten. So wurden in einer unstrukturierten Fragebogenstudie zur Friedensforschung 74 protestantische und katholische Jugendliche aus Nordirland gebeten, in Aufsätzen darzustellen, aus welchen Gründen sich ihrer Ansicht nach Menschen paramilitärischen Organisationen anschließen (Muldoon, Mclaughlin, Rougier, & Trew, 2008). Ergebnisse der Studie könnten z. B. in der Friedenserziehung fruchtbar gemacht werden.
- **Schriftliche projektive Methoden:** Bei projektiven Methoden werden u. a. uneindeutige Reizvorlagen (z. B. Fotos, Geschichten- oder Satzanfänge) präsentiert, die von den Befragungspersonen zu beschreiben oder zu ergänzen sind. Indem man nicht direkt nach den eigenen Erfahrungen oder Einstellungen fragt, sondern die Untersuchungspersonen ihre innerpsychischen Vorgänge auf Reizvorlagen übertragen (projizieren) lässt, sollen stereotype und sozial erwünschte Antworten vermieden und unbewusste Einstellungen oder Motive hervorgelockt werden (zu projektiven psychologischen Testverfahren ▶ Abschn. 10.4.2). Solche indirekten und projektiven qualitativen Verfahren werden als Befragungstechniken u. a. in der **Marktforschung** eingesetzt (Kirchmaier, 2007): So kann man Befragungspersonen z. B. den Satzanfang ergänzen lassen „Menschen, die Champagner trinken, wollen damit zeigen, dass . . .“ (**Satzergänzungsmethode**) oder man kann ihnen eine bildliche Dialogsituation vorlegen, in der sie die Sprechblase einer Person ausfüllen sollen (**Sprechblasentest**). In der Markenforschung kann mit der Methode der **Personifizierung** gearbeitet werden, indem man die Befragungspersonen bittet, sich vorzustellen, ein Unternehmen (z. B. eine bestimmte deutsche Großbank) sei eine Person: „Stellen Sie sich einmal vor, die Bank wäre kein Unternehmen, sondern ein lebendiger Mensch: Wie sieht dieser aus? Wo wohnt er? Wie lebt er? Ist er verheiratet? Wie sieht sein Familienleben aus?“ Die qualitative Auswertung entsprechender schriftlicher Beschreibungen soll Hinweise auf das Image des jeweiligen Unternehmens liefern. Projektive schriftliche Verfahren lassen sich z. B. auch in der **Geschlechterforschung** einsetzen, etwa indem man Situationsbeschreibungen (sog. **Vignetten**) vorgibt (▶ Studienbeispiel „Anne-Experiment“).

Studienbeispiel

Anne-Experiment

Eine 1972 an der Universität Harvard durchgeführte Studie, das sog. Anne-Experiment, arbeitete mit einer projektiven schriftlichen Befragungstrechnik (Horner, 1972). Die studentischen Untersuchungspersonen erhielten in Form einer Vignette den Anfang einer Geschichte und sollten diese weitererzählen. Laut Vignette hatte in der einen Versuchsbedingung eine Studentin „Anne“, in der anderen Versuchsbedingung ein Student „John“ bei der Zwischenprüfung im Medizinstudium Bestnoten erreicht. Aufgabe der Untersuchungspersonen war es, den weiteren Lebensweg von „Anne“ bzw. „John“ frei in eigenen Worten schriftlich zu schildern. Es zeigt sich, (dass 65 % der befragten Studentinnen der im Studium so erfolgreichen Studentin „Anne“ im nicht-standardisierten Fragebogen einen negativen weiteren Lebensweg vorzeichneten, im Extremfall sogar inklusive Depression, Krankheit und Tod, während 90 % der befragten Studenten in ihren Aufsätzen für den erfolgreichen Studenten „John“ eine Zukunft voll Glück und Wohlstand ausmalten. Die Autorin interpretierte diesen Befund als Indikator für Erfolgsangst („fear of success“) bei Frauen: Da beruflicher Erfolg traditionell mit der männlichen und nicht mit der weiblichen Geschlechterrolle assoziiert wird, sei er für Frauen ambivalenter, nicht zuletzt müssten sie real negative Reaktionen im Umfeld befürchten, wenn sie als „Karrierefrauen“ gelten. Deswegen projizierten, so die Annahme der projektiver Methode, die Studentinnen ihre eigenen Zukunftsängste in „Anne“ hinein. Es ist zu beachten, dass es sich bei dem ursprünglichen Anne-Experiment nicht um einen vollständigen Versuchsplan handelte, da „Anne“ nur den weiblichen und „John“ nur den männlichen Studierenden vorgelegt wurde.

Das Anne-Experiment wurde in den folgenden Dekaden mehrfach repliziert und kritisch diskutiert (z. B. Engle, 2003). Kritik bezog sich auf das aus der Geschlechterrollen-Theorie abgeleitete Konzept der Erfolgsangst, auf die Aussagekraft der projektiven Befragungsmethode (als Alternative wurden psychometrische Skalen zur Messung von Erfolgsangst entwickelt) und nicht zuletzt auf die mangelnde Konsistenz der Befunde (eine Reihe von Studien zeigten größere Erfolgsangst bei Männern als bei Frauen).

Der größte Vorteil des nicht-strukturierten Vorgehens bei der schriftlichen Befragung besteht darin, dass die Äußerungen der Befragungspersonen ihren eigenen subjektiven Strukturen und Schwerpunktsetzungen entsprechen und nicht durch einen Fragenkatalog der Forschenden begrenzt bzw. vorstrukturiert werden. Das Aufsatzschreiben mag jedoch vor allem bei formal besser gebildeten Untersuchungspersonen auf Akzeptanz stoßen. Zudem ist darauf zu achten, dass Schreibhemmungen vermieden werden und die Befragten wissen, dass es um keine schulähnliche Leistungsüberprüfung geht und dass orthografische oder grammatikalische Fehler überhaupt keine Rolle spielen. Je nach Schreibgewohnheiten der Zielgruppe sollte der nicht-standardisierte Fragebogen handschriftlich (Paper-Pencil-Fragebogen) oder maschinen-

schriftlich (elektronischer Fragebogen) angeboten werden. Die Auswertung der auf diese Weise generierten Textdokumente erfolgt mit einer der Methoden der qualitativen Daten- bzw. Dokumentenanalyse (► Abschn. 12.1, ► Abschn. 10.6.2).

Halbstrukturierte schriftliche Befragung

Im Unterschied zur unstrukturierten schriftlichen Befragung, bei der die Befragten nur einen Anstoß für ausführlichere schriftliche Beschreibungen oder Erklärungen erhalten, wird ihnen bei der halbstrukturierten schriftlichen Befragung eine Fragenliste vorgelegt. Allerdings gibt es keinerlei Antwortvorgaben, damit die Befragten ihre Antworten in eigenen Worten formulieren. Die Datenerhebung mittels halbstandardisierter qualitativer Fragebögen wird auch als qualitative Umfrageforschung („qualitative survey research") bezeichnet.

Teil-/halbstandardisierter Fragebogen – Bei der teilstrukturierten bzw. halbstrukturierten schriftlichen Befragung kommt ein teil-/halbstandardisierter Fragebogen („semi-standardized questionnaire") zum Einsatz. Er besteht aus einer Liste offener Fragen, die von den Untersuchungspersonen in eigenen Worten zu beantworten sind. Der halbstandardisierte Fragebogen für halbstrukturierte schriftliche Befragungen ist das Pendant zum Interviewleitfaden bei halbstrukturierten mündlichen Befragungen.

Teilstandardisierte Fragebögen können in unterschiedlichen Varianten auftreten, auf Papier und elektronisch, für Betroffene und Experten, als einmalige Befragung oder Tagebuchinstrument zum regelmäßigen Ausfüllen über mehrere Tage hinweg (zur Logik der qualitativen Umfrageforschung s. Jansen, 2010). Während im Bereich der vollstandardisierten Fragebögen Hunderte von etablierten und überprüften Instrumenten bzw. psychometrischen Skalen (► Abschn. 8.6) vorliegen, die man für eigene Studien ganz oder teilweise übernehmen kann, werden qualitative Fragebögen in der Regel individuell für jede einzelne Studie neu erstellt. Grundlage für die **qualitative Fragebogenerstellung –** d. h. die Auswahl und Formulierung der offenen Fragen und die Zusammenstellung des Fragebogens (Fragenanzahl, Fragenreihenfolge, Layout) – sind a) das eigene Forschungsanliegen, b) relevante Theorien, c) frühere Studien zum Themenfeld sowie d) eigene Vorstudien (► Regeln zur Formulierung von Fragen für qualitative Befragungen).

Vor der Hauptuntersuchung sollte der qualitative Fragebogen erprobt werden, indem ausgewählte Mitglieder der Zielgruppe den Bogen ausfüllen, die dafür notwendige Zeit registriert wird (ggf. ist eine Kürzung des Fragenkatalogs notwendig), mögliche Verständnisprobleme bei einzelnen Fragen identifiziert werden (ggf. ist eine Umformulierung einzelner Fragen notwendig) und Qualität und Umfang der Antworten betrachtet werden (ggf. ist es notwendig, mehr Platz in den Antwortboxen einzuräumen oder ausdrücklicher zum ausführlichen Antworten aufzufordern). Auch die Fragebogenform (auf Papier und/oder als elektronisches Formular) ist mit Blick auf die bei der Erprobung des Fragebogens artikulierten Vorlieben der Zielgruppe auszuwählen.

Beispiele für drei Varianten von halbstrukturierten Fragebogenstudien sollen die Anwendungsmöglichkeiten verdeutlichen:

- **Halbstrukturierte schriftliche Befragung von Experten**: Was verstehen professionelle Business-Coaches unter „Selbstmanagement", und wie fördern sie dieses bei ihrer Klientel? Um dieser Forschungsfrage nachzugehen wurde ein halbstandardisierter Online-Fragebogen entwickelt. Qualifizierte professionelle Coaches (identifiziert über die Websites von Coa-

Regeln zur Formulierung von Fragen für qualitative Befragungen)

Diese Regeln gelten für schriftliche und mündliche Befragungen:

1. Formuliere **jede Frage einfach und direkt** in einer der Zielgruppe angepassten Sprache (pro Frage nur ein Sachverhalt, keine doppelten Verneinungen, keine Fremdwörter und Fachbegriffe).
2. Beschränke **Faktenfragen** (z. B. Alter, berufliche Position, Anzahl der Kinder, Jahre Berufserfahrung etc.) auf die für die Studie notwendigen Hintergrundinformationen. Faktenfragen stehen meist am Anfang (Vorteil beim selbstadministrierten Fragebogen: Einstieg mit einfachen Fragen; Vorteil beim intervieweradministrierten Leitfaden: Vorinformation des Interviewers über das Gegenüber) oder am Ende (Vorteil: der qualitative Charakter der Befragung ist von Anfang an klar, eine strukturierte Abfragesituation entsteht erst am Schluss).
3. Verwende vor allem **offene Fragen**, die sich auf das Verhalten und Erleben beziehen (statt auf einfache Fakten) und dabei zum **Beschreiben** („Wie haben Sie sich dabei gefühlt?") und **Erzählen** („Was ist dann passiert?") anregen. Um zu knappe Antworten zu verhindern, kann vertiefend nachgefragt werden („Was fällt Ihnen noch dazu ein, alle Details sind wichtig!").
4. Vermeide zu viele **Warum-Fragen** („Warum haben Sie nicht studiert?"), denn Begründungen und Erklärungen werden oft in stereotyper Form abgegeben und treiben die Befragten schnell in die Enge.
5. Vermeide **Suggestivfragen**, die eine bestimmte Antwortrichtung vorgeben und nicht alle Möglichkeiten einbeziehen („Was haben Sie getan, um wieder gesund zu werden?"; besser: „Wie sind Sie mit der Krankheit umgegangen?").
6. Nutze **geschlossene Fragen** (Ja-/Nein-Fragen) bei Bedarf für eine Filterführung („Sind Sie schon einmal gekündigt worden? Falls ja: Wie ist Ihre letzte Kündigung verlaufen?").

Studienbeispiel

Halbstrukturierte schriftliche Befragung von Betroffenen

Körperveränderungen („body modification“) wie Piercings und Tattoos sowie **Körpergestaltung** („body styling“) durch Kosmetik, Solarium oder Sport sind für viele sozialwissenschaftliche Forschungsfelder relevante Themen – etwa in der Psychologie (Körperselbstbild sowie damit verbundene Störungen), in der Soziologie (soziale und kulturelle Schönheitsnormen), in der Marktforschung (Vermarktung entsprechender Produkte und Dienstleistungen), in der Medizin (gesundheitliche Risiken) oder in der Kommunikationswissenschaft (mediale Vorbildwirkung). Seit Beginn des 21. Jhd. wird **Körperhaarentfernung** („body depilation“) inklusive Schamhaarentfernung als neues Massenphänomen unter Jugendlichen und jungen Erwachsenen verstärkt öffentlich diskutiert und empirisch untersucht. Die Soziologin Merran Toerien führte zusammen mit der Psychologin Sue Wilkinson eine teilstrukturierte qualitative Fragebogenstudie unter britischen Frauen zum Thema Körperhaarentfernung durch (Toerien & Wilkinson, 2004), wobei der Fragebogen neben einem quantitativen einen umfangreichen qualitativen Teil mit neun offenen Fragen enthielt.

Wahl der Datenerhebungsmethode: Eine qualitative schriftliche Datenerhebungsmethode sollte genutzt werden, um die subjektiven Erfahrungen der Befragten herauszuarbeiten. Dabei wurde ein relativ großes Sample angestrebt, um eine möglichst große Bandbreite an Sichtweisen zu erfassen. Bei großen Stichproben sind mündliche Leitfadeninterviews zu zeit- und kostenaufwändig, so dass auf die ökonomischere schriftliche Fragebogentechnik zurückgegriffen wurde.

Fragebogenform, Verbreitung des Fragebogens und Stichprobe: Der Paper-Pencil-Fragebogen wurde breit gestreut an Studierende, Erwerbstätige, Nichterwerbstätige, sexuelle und ethnische Minoritäten und Majoritäten verteilt (teils über Multiplikatorinnen, teils postalisch). Insgesamt konnte eine Gelegenheitsstichprobe von 678 Umfrageteilnehmerinnen zwischen 16 und 70 Jahren zusammengestellt werden.

Qualitativer Fragebogen: Der qualitative Teilfragebogen bestand aus neun offenen Fragen, in denen bewusst nicht Gründe für oder gegen, sondern individuelle Erfahrungen mit Körperhaarentfernung erfasst werden sollten. Dabei wurden positive wie negative Erfahrungen und Reaktionen des Umfeldes auf praktizierte sowie nicht-praktizierte Körperhaarentfernung völlig symmetrisch offen abgefragt:

1. „If you DO remove your body hair, what, for you, are the BEST things about doing so?“
2. „If you DO NOT remove your body hair, what, for you, are the BEST things about not doing so?“
3. „If you DO remove your body hair, what, for you, are the WORST things about doing so?“
4. „If you DO NOT remove your body hair, what, for you, are the WORST things about not doing so?“
5. „What would make you change your current practices to do with your body hair?“
6. „Have you had any NEGATIVE reactions from other people to your body hair and/or your lack of body hair? If so, please give details!“
7. „Have you had any POSITIVE reactions from other people to your body hair and/or lack of body hair? If so, please give details!“
8. „Has anyone ever suggested that you SHOULD remove your body hair? If so, please give details!“
9. „Has anyone ever suggested that you SHOULD NOT remove your body hair? If so, please give details!“

An jede offene Frage schloss sich eine mehrzeilige leere Antwortbox an, um zum Ausdruck zu bringen, dass ausführliche schriftliche Antworten gewünscht waren.

Datenanalyse und Ergebnisse: Das verbale Datenmaterial wurde im Sinne einer qualitativen Inhaltsanlyse ausgewertet und zu Kategorien verdichtet. Dabei wurden zur Bedeutung der Haarentfernung vier dichotome Themenfelder identifiziert, wobei der positive Pol jeweils mit Haarentfernung, der negative mit fehlender Haarentfernung verbunden war: 1. Attraktivität (attraktiv/nicht attraktiv), 2. Glattheit (glatt/stopplig), 3. Sauberkeit (sauber/unsauber) und 4. Weiblichkeit (weiblich/nicht weiblich). Die negativen Reaktionen des Umfeldes (Verwandte, Freunde, Partner, Fremde) auf fehlende Haarentfernung ließen sich in vier Gruppen einteilen: 1. Ratschläge, 2. Kritik, 3. Witze, 4. Blicke.

Interpretation: Die Autorinnen argumentieren, dass angesichts der negativen Bewertung und Sanktionierung bei Behaarung die Körperhaarentfernung nicht als Frage des persönlichen Geschmacks, sondern bei Frauen als notwendige Orientierung an einer sozialen Norm zu verstehen ist.

Einordnung der Studie: Da die Studie hinsichtlich der verwendeten Literatur sowie der erhobenen Daten die westliche Welt adressiert, wären Studien, die z. B. asiatische, arabische oder afrikanische Kulturen einbeziehen, eine sinnvolle Ergänzung. Weiterhin wären die Perspektiven von Frauen unterschiedlicher Generationen und sexueller Orientierungen zu differenzieren. Zudem ist unter Geschlechterperspektive auch die Körperenthaarungsnorm für Männer zu berücksichtigen, die sich in der jüngeren Generation ebenfalls zu etablieren scheint (Boroughs, Cafri, & Thompson, 2005: vollstrukturierte Fragebogenstudie) und von jungen Frauen in den USA und Neuseeland als attraktiv eingestuft wird (Dixson, Dixson, Bishop, & Parish, 2010: vollstrukturierte Fragebogenstudie).

10

ching-Berufsverbänden) wurden zur Befragung eingeladen und 20 Expertinnen und Experten füllten den teilstandardisierten Fragebogen aus (Buhl, Roth, & Düx, 2007). Sie sollten dabei in eigenen Worten z. B. ihre Definition von „Selbst-Management“ angeben und Coaching-Beispiele aus ihrer Praxis schriftlich beschreiben.

- **Halbstrukturierte schriftliche Befragung von Betroffenen**: Welche Erfahrungen machen Nicht-Heterosexuelle mit chronischen Erkrankungen einerseits im Gesundheitssystem, andererseits in ihren Netzwerken und Szenen? Eine halbstrukturierte Online-Befragung mit 190 Nicht-Heterosexuellen aus acht Ländern zeigte, dass die Betroffenen über mangelnde Unterstützung in beiden Bereichen klagten (Jowett & Peel, 2009). Zur Illustration der Methode dient auch das ► Studienbeispiel „Halbstrukturierte schriftliche Befragung von Betroffenen“.

- **Halbstrukturierte Tagebuchmethode**: Die positive Psychologie (für Hintergrundinformationen s. das Positive Psychology Center: http://www.ppc.sas.upenn.edu) empfiehlt als Intervention zur Steigerung der Lebenszufriedenheit das Führen eines Dankbarkeits-Tagebuchs, in dem man täglich drei Dinge vermerkt, die an diesem Tag positiv verlaufen sind und welche Ursache es dafür jeweils gab („gratitude diary", „three good things"; Seligman, Steen, Park, & Petersen, 2005). Ein teilstandardisiertes Tagebuchinstrument, in dem über eine bestimmte Zeitspanne hinweg täglich ausführliche Freitext-Eintragungen zu ausgewählten Themen vorgenommen werden, kann in vielen Forschungsbereichen als Datenerhebungsmethode genutzt werden. Dabei wird oft ein Paper-Pencil-Instrument eingesetzt (z. B. ein Schreibheft). Teilstandardisierte Tagebücher lassen sich aber auch mediengestützt führen, etwa mittels mobiler Endgeräte. So können Befragungspersonen aufgefordert werden, ihre Reaktionen zu beschreiben, wenn sie im Alltag mit einer bestimmten Marke konfrontiert werden. Neben verbalen Tagebuchdaten können hierbei visuelle Daten (z. B. Fotografien mit der Handykamera) erfasst werden (zur qualitativen mobilen Tagebuchmethode s. Maxl & Döring, 2010). Die halbstrukturierte Tagebuchmethode kommt seltener zum Einsatz als die vollstrukturierte Tagebuchmethode (► Abschn. 10.3.3).

10.3.3 Quantitative Fragebogenmethode

Wenn Daten mittels schriftlicher Befragung erhoben werden, dann überwiegend im quantitativen Forschungsansatz mithilfe vollstandardisierter Fragebögen. In der Meinungsforschung, in der Soziologie, in der Kommunikations-, Gesundheits- und Politikwissenschaft werden zuweilen Umfragen mit bevölkerungsrepräsentative Stichproben durchgeführt mit dem Ziel der Populationsbeschreibung durch Parameterschätzung (zu bevölkerungsrepräsentativen vollstrukturierten Befragungen s. Schnell, 2011). Darüber hinaus sind explorative sowie explanative Fragebogenerhebungen in vielen sozialwissenschaftlichen Forschungsfeldern weit verbreitet, wobei meist mit nicht-repräsentativen bzw. nicht-probabilistischen Stichproben gearbeitet wird (zu Stichprobenarten ► Kap. 9).

(Voll)standardisierter Fragebogen – Der vollstrukturierten schriftlichen Befragung liegt ein quantitativer bzw. (voll)standardisierter Fragebogen („quantitative questionnaire", „standardized questionnaire") zugrunde. Er besteht überwiegend aus geschlossenen Fragen bzw. Aussagen mit Antwortvorgaben, so dass die Befragten die jeweils passenden Antwortalternativen auswählen können.

Im Folgenden wird der Ablauf einer vollstrukturierten schriftlichen Befragung bzw. Umfrage („written survey") geschildert. Die Konstruktion eines vollstandardisierten Fragebogens steht dabei im Zentrum. Anschließend gehen wir auf die Besonderheiten unterschiedlicher Arten und Verbreitungswege von quantitativen Fragebögen ein: Auf die klassische Erhebung eines Paper-Pencil-Fragebogens in Gruppensituationen durch Austeilen und Einsammeln, auf die postalische Befragung, die Online-Befragung, die mobile Befragung und die automatische Telefonbefragung. Schließlich widmen wir uns besonderen Fragebogentechniken, nämlich der Tagebuchmethode (inklusive Ambulantem Assessment), der Delphi-Methode sowie Befragungen, die auf Aussagen über größere soziale Gebilde wie Paare, Kleingruppen, soziale Netzwerke und Organisationen hinauslaufen.

Ablauf einer vollstrukturierten Fragebogenstudie

Die **Konstruktion des standardisierten Fragebogeninstruments** erfolgt in zwei Schritten, zunächst als Grob-, dann als Feinkonzeption. Der konstruierte Fragebogen wird schließlich einem **Fragebogen-Pretest** unterzogen und bei Bedarf überarbeitet (Fragebogenrevision).

Erst danach erfolgt die Datenerhebung im Zuge einer nach wissenschaftlichen Kriterien geplanten **Stichprobenziehung** (► Abschn. 9.3). Die Stichprobenart ist bei vollstrukturierten schriftlichen Umfragen mit dem avisierten Distributionskanal des Fragebogens abzustimmen. So können per postalischer Befragung andere Zielgruppen erreicht werden als per Online-, mobiler oder automatisch-telefonischer Befragung. Entsprechende Besonderheiten werden in den jeweiligen Abschnitten zu den Befragungsvarianten erläutert.

Nicht zuletzt ist bei schriftlichen Umfragen mit aktiver Rekrutierung der **Fragebogenrücklauf** zu dokumentieren und zu analysieren, um Anhaltspunkte für die Aussagekraft der Studie abzuleiten. Die Auswertung quantitativer Fragebogenstudien läuft auf statistische Datenanalysen hinaus (► Kap. 12). Zuvor ist eine sorgfältige Datenaufbereitung nötig, etwa eine Gewichtung des Datensatzes oder eine Kompensation für fehlende Werte (► Abschn. 11.4).

Grobkonzept des Fragebogens

Die Konstruktion eines vollstandardisierten Fragebogens ist ein aufwändiger Prozess, wenn ein Instrument entstehen soll, das wissenschaftlichen Gütekriterien der Objektivität, Reliabilität und Validität genügt (zu diesen Gütekriterien ► Abschn. 8.6, ► Abschn. 10.4.1).

Standardisierte Fragebögen sind in ihrer Grundstruktur immer gleich aufgebaut und bestehen aus sechs Elementen (◘ Tab. 10.11).

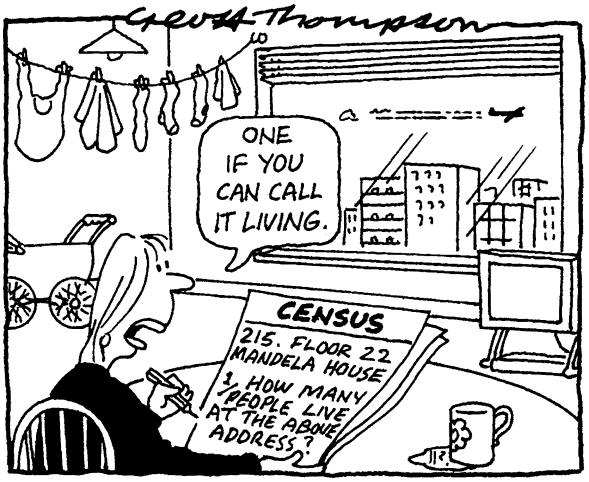
Tabelle 10.11 Aufbau eines standardisierten Fragebogens

Fragebogenelement mit Erläuterung	Beispiel
1. Fragebogentitel Der Fragebogentitel gibt einen Anhaltspunkt zu Thema und Zielgruppe der Befragung und soll die gesamte Zielgruppe zur Teilnahme motivieren. Heikle Themen werden in der Regel begrifflich entschärft. Man verwendet allgemeine Oberbegriffe, um die eigentliche Zielsetzung oder Hypothesenrichtung nicht vorab preiszugeben. Personengruppen, die befragt werden sollen, sich aber möglicherweise weniger angesprochen fühlen, können explizit genannt werden.	„Fragebogen für Frauen und Männer zu Sexualität im Internet“ Weniger gut: „Befragung zu Internetpornos“
2. Fragebogeninstruktion Die Instruktion (Ausfüllanleitung) erklärt den Befragungspersonen Zielsetzung und Ablauf der Fragebogenerhebung und gibt eine Kontaktmöglichkeit zur verantwortlichen Person oder Institution an (Postadresse, Telefonnummer oder E-Mail-Adresse). Zudem kann auf die Incentivierung (Gewinnspiel, Bonuspunkte für Mitglieder von Access-Panels) hingewiesen oder inhaltlich zur Teilnahme motiviert werden (z. B. Anwendungsrelevanz des Themas hervorheben; Methode der Stichprobenziehung erwähnen und die Bedeutung der Teilnahme jeder einzelnen Person für eine unverzerrte Stichprobe betonen). Auch Verweise auf die Forschungsethik (Freiwilligkeit, Anonymität) sind notwendig.	Diese Studie untersucht, wie Menschen mit den vielfältigen sexuellen Informations-, Unterhaltungs-, Kontakt- und Konsumangeboten im Internet umgehen. Der Fragebogen richtet sich an Frauen und Männer aller Altersgruppen und Lebenssituationen. Für die Studie sind Ihre ganz persönlichen Erfahrungen wichtig. Es gibt keine richtigen oder falschen Antworten. Alle Angaben erfolgen anonym und dienen rein wissenschaftlichen Zwecken. Das Ausfüllen des Fragebogens dauert etwa zehn Minuten. Vielen Dank, dass Sie an unserer Studie teilnehmen!
3. Inhaltliche Fragenblöcke Die inhaltlichen Fragenblöcke sollten so gestaltet sein, dass die Befragten sich gut in die Thematik hineindenken und nicht zu oft gedanklich springen müssen. Dies gilt ebenso für die formale Gestaltung der Fragen, so sollten Antwortformate nicht zu oft im Verlauf des Fragebogens wechseln. Im ersten Fragenblock stehen möglichst unproblematische, leicht und schnell zu beantwortende Fragen, um den Befragten den Einstieg in die Befragung zu erleichtern. Die Fragenblöcke können durch Zwischenüberschriften voneinander abgesetzt werden. Fragen, die sich nur an Teilgruppen richten, können von den anderen Befragten übersprungen werden (bei Paper-Pencil-Fragebögen erfolgt die Filterführung durch ausdrücklichen Hinweis, zu welcher Frage gesprungen werden soll; bei elektronischen Fragebögen erfolgt die Filterführung automatisch). Vereinzelt können offene Fragen in einen standardisierten Fragebogen integriert werden. Hier ist jedoch mit vielen Antwortverweigerungen (Item Nonresponse) zu rechnen.	A. Sexuelle Information im Internet 1. Wie oft nutzen Sie das Internet, um sexuelle Informationen abzurufen? [nie – selten – manchmal – oft – sehr oft] 2. Haben Sie schon einmal in einem sexualbezogenen Onlineforum mitdiskutiert? nein → bitte weiter mit Frage 5 ja 3. Denken Sie an die letzte sexualbezogene Onlinediskussion, an der Sie teilgenommen haben. Um welches Thema ging es? _____ B. Sexuelle Unterhaltung im Internet [. . .]
4. Statistische Angaben Statistische Angaben dienen vor allem der Beschreibung der befragten Stichprobe anhand allgemeiner soziodemografischer Merkmale. Sie sollten sparsam und mit etablierten Items erfasst werden (► Abschn. 8.5.1) Zudem können hier weitere einfache Hintergrundvariablen erhoben werden, die im Zusammenhang mit dem Studienthema relevant sind, aber sich nicht sinnvoll in die inhaltlichen Fragenblöcke integrieren lassen.	Allgemeine soziodemografische Merkmale: Geschlecht, Alter, Bildung, Tätigkeit, Einkommen, Familienstand, sexuelle Orientierung, Nationalität, Religion etc. Themenspezifische einfache Hintergrundvariablen: z. B. Intensität der Internetnutzung allgemein, Nutzung anderer Medien, Anzahl der bisherigen Sexualpartner
5. Fragebogen-Feedback Um den Befragungspersonen Gelegenheit zu bieten, bei Bedarf ein Feedback zum Fragebogen zu äußern, kann zum Schluss ein offenes Antwortfeld angeboten werden.	Haben Sie Anmerkungen zu diesem Fragebogen oder zu dieser Studie? _____ _____
6. Verabschiedung Meist endet der Fragebogen mit einem Wort des Dankes. Sofern eine Incentivierung in Form eines Gewinnspiels erfolgt, werden die Befragungspersonen, die daran teilnehmen möchten, um freiwillige Angaben zu ihren Kontaktdaten gebeten. (Diese Angaben müssen getrennt vom Fragebogen verwaltet werden.)	Vielen Dank für Ihre Mithilfe!

Die inhaltlichen Fragenblöcke stellen das „Herz“ des Instrumentes dar. Hier finden sich die Items, mit denen die laut Forschungsfragen und Forschungshypothesen (▶ Abschn. 5.2) interessierenden Variablen operationalisiert (▶ Abschn. 8.3) werden. Man unterscheidet **zwei Operationalisierungsvarianten** im Kontext von standardisierten Fragebögen:

1. **Einzelitem** („single item“): Ein einzelnes Item bzw. ein einzelner Indikator dient zur Messung eines Merkmals (▶ Abschn. 8.5). Dies ist bei soziodemografischen Variablen (z. B. Alter), einfachen Verhaltensvariablen (z. B. Anzahl der täglichen Telefonate), aber teilweise auch bei komplexeren Konstrukten der Fall (z. B. Selbstwertgefühl: „I have high self-esteem“, Antwortskala: 1 = „very true of me“ bis 5 = „not very true of me“; Robins, Hendin, & Trzesniewski, 2002). Die Operationalisierung mittels Einzelitem hat den Vorteil maximaler Ökonomie: Der Aufwand für die Befragten ist gering, der Fragebogen bleibt kurz. Nachteilig ist, dass die Messgenauigkeit (Reliabilität) bei einem Einzelitem anhand des vorliegenden Datensatzes nicht überprüft werden kann. Generell gelten Messungen mit Einzelitems bzw. Einzelindikatoren als weniger messgenau (◘ Abb. 10.7).
2. **Psychometrische Skala** („psychomeric scale“): Mehrere ähnliche Items bzw. Indikatoren werden verwendet, die gemeinsam ein Merkmal messen (Zusammenfassung der inhaltsähnlichen Items zu einem Gesamtwert; ▶ Abschn. 8.6). Insbesondere komplexere Konstrukte, die für die jeweilige Studie eine besonders große inhaltliche Bedeutung haben, sollten nicht mit Einzelitems, sondern lieber mit Skalen erfasst werden. So lässt sich das Selbstwertgefühl mit der aus 10 Items bestehenden „Rosenberg self esteem scale“ (RSE) zuverlässig erfassen: 1.) „On the whole I am satisfied with myself.“, 2.) „At times I think that I am no good at all.“, 3.) „I feel that I have a number of good qualities.“ etc. Antwortskala: 1 = „strongly agree“ bis 4 = „strongly disagree“ (Rosenberg, 1965). Die Verwendung von psychometrischen Skalen ist zwar aufwändiger, liefert aber reliablere und validere Daten.

Grundsätzlich sollten bei einer Fragebogenkonstruktion möglichst erprobte Einzelitems und Skalen aus der Literatur übernommen werden. Vielen Publikationen ist im Anhang das verwendete Instrument beigefügt, es kann in der Regel auch bei den Autorinnen und Autoren angefragt werden. Zudem lassen sich Test- und Fragebogendatenbanken nutzen, um Instrumente zu finden (z. B. http://www.zpid.de). Aus umfangreichen Skalen sind bei Platzproblemen auch Kurzskalen oder Einzelitems zu entnehmen. Englischsprachige Skalen lassen sich ins Deutsche



◘ **Abb. 10.7** Soziodemografische Angaben werden meist mit Einzelitems erfasst. (© Geoff Thompson/Search ID: gth0196, Rights Available from CartoonStock.com)

übertragen. Der **Rückgriff auf etablierte Items und Skalen** hat den Vorteil, dass deren Gütemerkmale bereits überprüft worden sind und dass die eigenen Ergebnisse dann mit anderen Studien vergleichbar sind.

▪ Feinkonzept des Fragebogens

Wenn das Grobkonzept des Fragebogens steht, ist am Feinkonzept zu arbeiten. Dabei sollten folgende Aspekte betrachtet, angepasst und bei Bedarf überarbeitet werden:

1. **Art der Items und Antwortformate**: Ein guter Fragebogen enthält einfache, leicht verständliche und schnell zu beantwortende Items. Günstig ist es, einen Fragebogen so zu gestalten, dass nicht zu viele verschiedene Arten von Items und Antwortformaten vorkommen, da dies ein ständiges Umdenken von den Befragungspersonen verlangt (▶ Unterschiedliche Typen von Items in standardisierten Fragebögen). Die Methodenforschung befasst sich u. a. damit, wie Befragungspersonen bestimmte Fragentypen wahrnehmen, verarbeiten und beantworten und welche Rückschlüsse daraus auf die optimale Gestaltung von Fragebogenitems zu ziehen sind (für einen Forschungsüberblick s. Schaeffer & Dykema, 2011).
2. **Reihenfolge der Items:** Die inhaltlichen Fragen sollten in thematisch zusammenhängende Blöcke gegliedert werden. Die Reihenfolge der Blöcke sowie der einzelnen Items innerhalb der Blöcke sollte mit Blick auf die kognitiven Verarbeitungsprozesse der Befragungspersonen hin optimiert werden. Eine logische Abfolge, die von allgemeinen schrittweise zu spezielleren Fragen führt, erleichtert die Beantwortung. Zu beachten sind auch unerwünschte Reihenfolgen-

Unterschiedliche Typen von Items in standardisierten Fragebögen

Fragen, Aussagen und Aufforderungen als Fragebogenitems

Die einzelnen Bausteine eines Fragebogens sind die Fragebogenfragen bzw. allgemeiner: die Fragebogenitems („questionnaire items"). Ein Fragebogenitem kann bestehen aus

- einer **Frage** („question") plus Antwortvorgaben („Wie oft gehen Sie ins Kino?" nie – selten – manchmal – oft – sehr oft) oder
- einer **Aussage** („statement") plus Antwortvorgaben („Ich gehe oft ins Kino." stimmt gar nicht – stimmt wenig – stimmt teils-teils – stimmt ziemlich – stimmt völlig) oder
- einer **Aufforderung** plus Antwortvorgaben („Bewerten Sie die Zufriedenheit mit Ihrem letzten Kinobesuch auf einer Schulnotenskala!" 1 – 2 – 3 – 4 – 5 – 6).

Antwortformate

Die Wahl des Antwortformats muss inhaltlich zum Item passen. Das Antwortformat ist entscheidend dafür, auf welchem Skalenniveau die interessierenden Variablen gemessen werden. Besonders einschlägig sind Ratingskalen als Antwortformat, da sie sich schnell ankreuzen lassen und in der Regel intervallskalierte Daten liefern, die sich statistisch gut auswerten lassen (zu Ratingskalen ► Abschn. 8.4.4). Bei elektronischen bzw. Online-Fragebögen sind teilweise noch andere Antwortformate möglich als im Paper-Pencil-Fragebogen (z. B. Pull-Down-Liste, Drag & Drop-Fragen).

Geschlossene, quasi-geschlossene und offene Items

Ein standardisierter Fragebogen besteht definitionsgemäß primär aus **geschlossenen Fragen** bzw. Items mit vorgegebenen Antwortalternativen, aus denen die Respondenten zu wählen haben (z. B. Ratingskalen). Es werden jedoch auch **quasi-geschlossene Faktenfragen** zu Daten, Anzahlen, Zeitdauern gestellt: „Seit wann sind Sie verheiratet? Jahr: ____ Monat: ___"; „Wie viele E-Mails haben Sie gestern versendet? ___ versendete E-Mails"). Diese Fragen sind insofern quasi-geschlossen, als der Antwortraum vorgegeben ist. Es würde allerdings zu viel Platz im Paper-Pencil-Bogen beanspruchen, alle Antwortmöglichkeiten aufzulisten, um diese dann ankreuzen zu lassen (in elektronischen Fragebögen können indessen durch Pull-Down-Listen beliebig lange Listen von Anzahlen, Jahreszahlen, Monaten etc. vorgegeben werden). **Offene qualitative Fragen**, die in eigenen Worten eine Beschreibung oder Erzählung zum Sachverhalt verlangen (► Abschn. 10.3.2), werden innerhalb standardisierter Fragebögen kaum eingesetzt. Allenfalls wird mit offenen qualitativen Fragen nach Beispielen für interessierende Sachverhalte gefragt („Nennen Sie die drei von Ihren Kindern am häufigsten genutzten Computerprogramme: _____"; „Haben Sie in Ihrem Studium schon einmal geschlechtsbezogene Benachteiligung durch Mitstudierende erlebt? Wenn ja, schildern Sie bitte die Situation(en): _____"). Ergänzende Angaben in eigenen Worten werden bei polytomen Items bei der Antwortkategorie „sonstiges/anderes" erbeten („Ihr Beziehungsstatus: verheiratet/feste Partnerschaft – Single – sonstiges, und zwar: _____").

Inhaltsitems und Funktionsitems

Die zur Erfassung der interessierenden Sachverhalte genutzten Fragen bzw. Items sind die **Inhaltsitems**. Sie werden zuweilen um **Funktionsitems** ergänzt, die den Befragungsablauf steuern, aber inhaltlich nicht von Belang sind. So werden manchmal Fragebögen mit einer **Eisbrecherfrage** eingeleitet, die leicht zu beantworten ist und einen positiven Einstieg in die Befragung erlauben soll. Ebenso werden zuweilen **Pufferfragen** eingesetzt, die unterschiedliche Fragenblöcke voneinander abgrenzen und Ausstrahlungseffekte vermeiden sollen (z. B. können nach Fragen zu negativen Emotionen inhaltlich positive oder neutrale Pufferfragen eingesetzt werden, um zu verhindern, dass die folgenden Items aus negativem Blickwinkel betrachtet werden). **Distraktoritems** werden eingestreut, wenn man Befragungspersonen vom eigentlichen Untersuchungsthema ablenken möchte. Eisbrecher-, Puffer- und Distraktoritems stellen sog. **Wegwerf-Items** dar, da sie nicht ausgewertet werden. Im Sinne der Ökonomie sind Wegwerffragen sparsam einzusetzen. Gegebenenfalls können zusätzliche Funktionsitems überflüssig werden, wenn man für die inhaltlichen Items eine optimale Reihenfolge und Darstellungsform findet. **Filterfragen**, die dafür sorgen, dass Befragungspersonen jeweils nur für sie relevante Fragen vorgelegt bekommen und nicht passende Fragenblöcke übersprungen werden, lassen sich ebenfalls den Funktionsitems zuordnen („Haben Sie Kinder? ja/nein. Wenn Sie keine Kinder haben, überspringen Sie bitte die folgende Fragen und machen Sie weiter mit Frage 13 auf Seite 5"). Komplizierte Filterführungen, die ein Hin- und Herblättern der Befragten im Fragebogen erfordern, eignen sich nicht für Paper-Pencil-Fragebögen; sie sollten Online-Fragebögen vorbehalten bleiben, bei denen der Filterprozess automatisch erfolgt.
Kontrollfragen dienen der Überprüfung der Antwortqualität (z. B. Konsistenz-Checks), sie werden jedoch selten eingesetzt. Wichtige Konstrukte werden über psychometrische Skalen (mehrere inhaltsähnliche Items, die zu einem Gesamtwert zusammengefasst werden; ► Abschn. 8.6) erfasst, wodurch Messfehler reduziert und Kontrollfragen obsolet werden. Kontrollfragen bergen die Gefahr, die Befragungspersonen zu verärgern, wenn sie sich überprüft fühlen oder den Eindruck haben, redundante Fragen beantworten zu müssen.

Selbstauskunftsitems und Testitems

Bei mündlichen und schriftlichen Befragungen geht es um die Erhebung von Selbstauskünften, nicht um das Testen von Leistungen. Items, die darauf abzielen, Leistungen zu testen (z. B. Wissen, Intelligenz, Konzentration etc.) werden als Testitems in ► Abschn. 10.4.3 zu psychologischen Testverfahren behandelt. Bei diesen Testitems existieren objektiv richtige und falsche Lösungen, ein Abschreiben oder Raten ist möglich, während es bei Selbstauskünften um möglichst ehrliche subjektive Auskünfte zum eigenen Erleben und Verhalten geht.

Inhaltliche Varianten von Selbstauskunftsitems

Oft werden Inhaltsitems danach differenziert, ob sie Meinungen (d. h. Bewertungen von Sachverhalten, z. B. „Zwischen Ehe und eingetragener gleichgeschlechtlicher Partnerschaft sollten keinerlei Unterschiede gemacht werden."), Überzeugungen (d. h. Glaube an bestimmte Fakten oder Werte, z. B. „Computerspiele schaden Kindern."), Gedanken („Ich denke oft an die Vergangenheit."), Gefühle („Ich fühle mich oft traurig."), aktuelle Verhaltensweisen („Wie oft treiben Sie Sport?"), Verhaltensabsichten („Haben Sie vor, das Rauchen aufzugeben?"), körperliche Merkmale („Haben Sie Tattoos?") sowie sozialstatistische Eigenschaften (z. B. „Welches ist Ihr höchster Schulabschluss?") abfragen. Je nach Untersuchungsthema kombiniert ein standardisierter wissenschaftlicher Fragebogen in der Regel verschiedene inhaltliche Varianten von Selbstauskunftsitems, etwa Fragen nach Verhaltensweisen, Einstellungen und Gefühlen. Unverzichtbar in jedem Fragebogen ist ein Fragenblock zu sozialstatistischen bzw. soziodemografischen Angaben (► Abschn. 8.5.1). Sie dienen der Stichprobenbeschreibung, damit nachvollziehbar ist, wie die Stichprobe zusammengesetzt ist, auf deren Basis die inhaltlichen Ergebnisse gewonnen wurden.

effekte in der Weise, dass Fragen zu bestimmten Themen die Befragten in eine negative Stimmung versetzen können, aus der heraus die nachfolgenden Fragen negativer beantwortet werden. Zur Vermeidung derartiger **Ausstrahlungseffekte** ist die Itemreihenfolge zu überdenken. Zudem können bei Bedarf neutralisierende Pufferfragen verwendet werden (► Unterschiedliche Typen von Items in standardisierten Fragebögen). Welche kognitiven Prozesse im Einzelnen beim Lesen, Verarbeiten und Beantworten von Fragebogenitems und ganzen Fragebögen ablaufen, ist Gegenstand von **Theorien des Antwortverhaltens** und entsprechenden Methodenstudien (Tourangeau, Rips, & Rasinski, 2000).

3. **Filterführung im Fragebogen**: Enthält ein Fragebogen einzelne Items oder ganze Fragenblöcke, die nicht von allen Befragungspersonen zu beantworten sind, so werden Filterführungen eingesetzt. Während bei elektronischen Fragebögen die Filterführungen automatisiert werden können, muss beim Paper-Pencil-Fragebogen hin- und hergeblättert werden. Bei der Feinkonzeption des Fragebogens ist also eine möglichst übersichtliche und ökonomische Einbindung der Filterführung zu entwickeln.
4. **Layout des Fragebogens**: Es lohnt sich, bei jeder Fragebogenform (Paper-Pencil-, elektronischer, mobiler Fragebogen), auf ein **ansprechendes Layout** zu achten, da der Fragebogen dadurch professioneller wirkt, stärker akzeptiert wird und auch Antwortverzerrungen vermieden werden können. Muss bei einem Paper-Pencil-Fragebogen eine Abwägung zwischen einer gewünschten Reduktion von Seitenzahl und Druckkosten einerseits und einer großzügigen, gut lesbaren Formatierung andererseits getroffen werden, so sollte letztere den Ausschlag geben. In der Methodenforschung werden u. a. Eyetracking-Verfahren eingesetzt, um das Fragebogenlayout zu optimieren (für Datenerhebung mittels Blickbewegungskamera ► Abschn. 10.5.6).
5. **Fragebögen für unterschiedliche Distributionswege:** Je nachdem, ob ein vollstandardisierter Fragebogen auf Papier, am Computermonitor, auf dem Handy oder automatisiert am Telefon beantwortet werden soll, muss das Instrument in Aufbau, Länge und Layout gegebenenfalls entsprechend angepasst werden. So sind mobile Fragebögen meist deutlich kürzer als Online-Fragebögen. Bei größeren Umfragen mit Paper-Pencil-Fragebögen lohnt es sich, maschinenlesbare Fragebögen einzusetzen, sofern ein entsprechendes System zur Verfügung steht (viele Hochschulen nutzen derartige Systeme für die Lehrveranstaltungsevaluation; z. B. „Education Survey Automation Suite", EvaSys).
6. **Fragebögen für interkulturelle Studien:** Bei interkulturellen Studien muss derselbe Fragebogen in mehreren Sprachen vorliegen. Dafür reicht eine wortwörtliche Übersetzung nicht aus. Um sicherzugehen, dass die Fragebogenitems auch wirklich sinngemäß in allen Ländern gleich verstanden werden, muss der Fragebogen einer eingehenden Kontrolle durch Muttersprachler unterzogen werden.

■ Typische Fehler bei der Konstruktion standardisierter Fragebögen

Es gibt eine Reihe von typischen Fehlern, die bei der Fragebogenkonstruktion immer wieder gemacht werden. Diese werden im Folgenden kurz vorgestellt.

Die ersten Fehler werden häufig bereits beim Rückbezug auf den theoretischen Hintergrund gemacht. So weisen viele Fragebögen einen **mangelnden direkten Bezug der Fragebogenitems zu den Forschungsfragen** bzw. Forschungshypothesen auf. Dies führt dazu, dass die Forschungsfragen nach der Datenauswertung nicht wirklich beantwortet werden können (z. B. wichtige inhaltliche Bereiche des Forschungsproblems wurden im Fragebogen nicht oder unzureichend abgebildet).

Des Weiteren müssen bei der Formulierung der Items zwingend **die Merkmale der zu untersuchenden Ziel-**

gruppe beachtet werden. Das betrifft Wortwahl, Umfang und Anspruch des Fragebogens, die zum Bildungsniveau passen sollten, aber auch verwendete Beispiele und Erläuterungen, die Bezüge zur alltäglichen Lebenswelt der Befragten haben sollten.

Eine weitere Fehlerquelle besteht darin, ad hoc eigene Items zu formulieren, obwohl **etablierte Items und psychometrische Skalen** vorliegen, die entweder direkt übernommen oder adaptiert werden können. Sollen eigene Items und Skalen entwickelt werden, etwa, weil für das interessierende Phänomen noch keine eingeführten Instrumente existieren, so muss dies im Rahmen einer systematischen Skalenkonstruktion geschehen (▶ Abschn. 8.6).

Generell sollte auf eine **klare und verständliche Sprache** geachtet werden. Dies beginnt bereits beim Fragebogentitel und setzt sich über die Instruktion bis zu den einzelnen inhaltlichen Items fort. So ist z. B. eine Verwendung von Fremdwörtern, komplizierten Ausdrücken oder auch doppelten Verneinungen unbedingt zu vermeiden (▶ Empfehlungen zur sprachlichen Gestaltung von Items für standardisierte Fragebögen; für Hinweise zur Konstruktion von Ratingskalen ▶ Abschn. 8.4.4).

Zusätzlich liegen typische Fehlerquellen in der Festlegung des **Skalenniveaus** begründet. Je höher das Skalenniveau der erhobenen Variablen, desto mehr Verfahren können bei der statistischen Datenanalyse sinnvoll berechnet werden. So sollten Häufigkeiten und Anzahlen wenn möglich direkt („Wie viele E-Mails haben Sie gestern versendet? ___ E-Mails“) und nicht über von den Forschenden vorgegebene Abstufungen erhoben werden (0 E-Mails, 1–5 E-Mails, 6–10 E-Mails, > 10 E-Mails), da hierdurch die Auswertungsmöglichkeiten eingeschränkt sind. Wichtig ist bei direkter Abfrage von Häufigkeiten oder Anzahlen, dass eine klare Definition des zu erhebenden Verhaltens und ein klarer Referenzzeitraum, in dem das Verhalten gezeigt worden sein soll (z. B. gestern, in den letzten 14 Tagen etc.) vorgegeben werden. Derartige offene Fragen nach numerischen Angaben erzeugen deutlichen Zusatzaufwand bei der Datenaufbereitung, da es hier zu unplausiblen Antworten und Ausreißerwerten (z. B. „999 E-Mails“) kommt. Nicht selten müssen die numerischen Antworten dann im Nachhinein für eine handhabbare Auswertung mittels Skalentransformation doch wieder zu plausiblen Kategorien zusammengefasst werden (▶ Abschn. 8.4.6).

Ein zu **häufiger Wechsel des Antwortformats** (Wechsel zwischen 4- und 6-stufigen Intensitäts- und Häufigkeitsskalen etc.) führt zu einer hohen kognitiven Beanspruchung der Befragungspersonen. Dies sollte vermieden werden. Ein möglichst einheitliches Antwortformat ist nicht nur angenehmer für die Befragten, sondern führt auch zu einer besseren Datenqualität. Mit dem Einsatz negativ gepolter Items, die ebenfalls ein künstliches Umdenken verlangen, ist deswegen sparsam umzugehen.

Empfehlungen zur sprachlichen Gestaltung von Items für standardisierte Fragebögen

(mod. nach Bryman, 2008; Bühner, 2011; Hussy, Schreier, & Echterhoff, 2009)

- Achten Sie auf eine klare und informative Instruktion.
- Formulieren Sie konkret, anstatt Verallgemeinerungen zu verwenden.
- Vermeiden Sie Begriffe mit mehreren Bedeutungen.
- Verwenden Sie keine ungebräuchlichen Begriffe und Ausdrücke.
- Verwenden Sie nach Möglichkeit keine Fremdwörter und Fachbegriffe. Sollte sich das nicht vermeiden lassen, dann müssen diese kurz erklärt werden.
- Vermeiden Sie doppelte Verneinungen.
- Verwenden Sie keine Abkürzungen.
- Formulieren Sie in einfachen und möglichst kurzen Sätzen. Vermeiden Sie Verschachtelungen und ungewöhnliche Satzkonstruktionen.
- Formulieren Sie die Items zielgruppengerecht.
- Erfragen Sie Häufigkeiten – sofern inhaltlich notwendig und forschungsökonomisch machbar – konkret und benennen Sie immer die Einheit (z. B. Stunden pro Woche).
- Definieren Sie Zeitspannen genau (z. B. „in den letzten 14 Tagen“; „in den vergangenen drei Monaten“).
- Wichtige Wörter können sparsam hervorgehoben werden (z. B. Kursivsetzung, Unterstreichung).

Insgesamt sollten standardisierte Fragebögen eine gewisse **Länge** nicht überschreiten, da die Teilnahmemotivation mit zunehmender Ausfüllzeit absinkt. Eine tolerierbare Zeitgrenze kann im Pretest ermittelt werden, da diese stark von der Zielgruppe, dem Thema und dem Beantwortungssetting abhängig ist. Kürzungspotenzial besteht bei den meisten Fragebögen z. B. bei den sozialstatistischen Angaben. Hier werden häufig **zu viele Detailabfragen** vorgenommen (z. B. sehr differenzierte Abfragen zu Beruf, Bildung, Einkommen), die am Ende gar nicht sinnvoll inhaltlich ausgewertet werden. Das Kürzungspotenzial in anderen Fragebogenteilen sollte ebenfalls geprüft werden. Meist ist es ratsamer, bei näherer Betrachtung überflüssige Items zu streichen, anstatt den Fragebogen möglicherweise unübersichtlich zu formatieren, um Platz zu sparen und ihn kürzer erscheinen zu lassen. Notfalls muss angesichts der Obergrenze der Fragebogenlänge auch auf die Beantwortung einer der Forschungsfragen verzichtet werden.

■ Fragebogen-Pretest

Je mehr selbst konstruierte Elemente ein standardisierter Fragebogen enthält (im Unterschied zu Items und Skalen, die aus der Literatur direkt übernommen werden), umso

wichtiger sind gründliche empirische Pretests, die in mehreren Stufen stattfinden sollten. Ziel von Fragebogen-Pretests ist es, mögliche Probleme der Befragungspersonen beim Beantworten zu identifizieren, um das Instrument dann entsprechend zu verbessern. Es existieren eine Reihe unterschiedlicher Pretesttechniken, die oft kombiniert eingesetzt werden (Häder, 2006, S. 385ff.). Drei besonders praxisrelevante Strategien seien hier vorgestellt:

- **Qualitativer Pretest**: Empfehlenswert ist meist zunächst ein qualitativer Pretest. Dafür legt man den Fragebogen einzeln oder in Kleingruppen etwa fünf bis zehn ausgewählten Personen der Zielpopulation vor, die später nicht an der Hauptuntersuchung teilnehmen. Die Befragten bearbeiten den Bogen sorgfältig und geben anschließend mündlich im Rahmen eines **qualitativen Interviews** in eigenen Worten Feedback: Welche Fragen ließen sich gut beantworten, welche waren schwer verständlich oder wirkten unsinnig und welche Aspekte wurden vermisst? Alternativ kann man die Fragebogen-Pretester bitten, beim Ausfüllen des Fragebogens die **Methode des lauten Denkens** anzuwenden (▶ Abschn. 10.2.2). Auf diese Weise erhält man Rückmeldungen über technische Probleme (z. B. fehlerhafte Filterführung), mangelnde Verständlichkeit (z. B. ungünstiger Satzbau, unverständliche Fremdwörter) oder die Eignung verschiedener Antwortformate (Bühner, 2006). Auch andere **kognitive Pretesttechniken** sind einsetzbar: Bei der **Nachfragetechnik** („probing“) werden die Pretest-Teilnehmenden gebeten zu begründen, warum sie eine bestimmte Antwortalternative gewählt haben oder wie sie das Item im Einzelnen interpretieren. Bei der **Paraphrasierungstechnik** („paraphrasing“) bittet man die Pretester, ein gerade vorgelegtes Item in eigenen Worten zu wiederholen. Derartig aufwändige kognitive Techniken sind angemessen, wenn es sich um groß angelegte populationsbeschreibende Studien handelt oder wenn der Fragebogen in mehreren Studien eingesetzt werden soll. Im Pretest wird auch eruiert, wie lange das Beantworten einzelner Items bzw. das Ausfüllen des gesamten Fragebogens dauert. Die im qualitativen Pretest erhaltenen Rückmeldungen werden zunächst alle zusammengetragen und anschließend im Forschungsteam sortiert, hinsichtlich Relevanz und Brauchbarkeit bewertet und im Zuge der sog. **Fragebogenrevision** selektiv umgesetzt.
- **Fragebogenkonferenz:** Bei der Fragebogenkonferenz geben nicht potenzielle Befragungspersonen, sondern **Fachkollegen bzw. Experten** Feedback zum Instrument. Dazu setzt man sich mehreren Kolleginnen und Kollegen, die über Expertise hinsichtlich des Befragungsthemas und/oder der Fragebogenmethode verfügen, zu einer sog. Fragebogenkonferenz zusammen und geht den Fragebogen Item für Item durch, um Unstimmigkeiten zu identifizieren und zu beheben. Auch hier stellen Verbesserungsvorschläge die Grundlage für eine **Fragebogenrevision** dar. Zudem kann die positive Bewertung des Fragebogens durch Fachexperten auch als Methode der Überprüfung der **Inhaltsvalidität** („content validity“) des Instruments aufgefasst werden (▶ Abschn. 10.4.1).
- **Quantitativer Pretest:** Die letzte Stufe der Fragebogen-Überprüfung ist der quantitative Pretest. Hier füllt eine kleine Teilmenge (z. B. $n = 40$) der Zielpopulation, die später nicht an der Hauptuntersuchung teilnimmt, die Endfassung des Fragebogens unter Realbedingungen aus und gibt die ausgefüllten Bögen ab. Die Daten werden entsprechend elektronisch erfasst, aufbereitet und statistisch ausgewertet (z. B. Betrachtung der Häufigkeitsverteilungen). Auf diese Weise können noch einmal der **reibungslose Ablauf der gesamten Datenerhebung** geprüft (z. B. auch korrektes Einlesen der Daten aus Online-Fragebögen) sowie eine hohe Datenqualität sichergestellt werden. Dabei besteht die Möglichkeit, Items zu identifizieren, die in der Zielgruppe zu wenig Varianz erzeugen oder bei denen es zu einem Deckeneffekt („ceiling effect“: nur die oberen Skalenwerte werden angekreuzt) oder Bodeneffekt („floor effect“: nur die unteren Skalenwerte werden ausgewählt) oder sonstigen **extremen Verteilungsformen** kommt. Ein letztes Nachsteuern durch eine Umformulierung oder Eliminierung einzelner Items im Zuge einer erneuten **Fragebogenrevision** ist möglich. Enthält der Fragebogen Itembatterien, die zu Skalengesamtwerten zusammengefasst werden sollen, so können psychometrische Kennwerte wie z. B. die Item-Trennschärfen, die Item-Schwierigkeiten oder die faktorielle Struktur der Skala anhand der quantitativen Pretestdaten geprüft werden (zu Skalen- und Testkonstruktion und Gütekriterien ▶ Abschn. 8.6; ▶ Abschn. 10.4.1; ▶ Abschn. 10.4.4).

Fragebogenrücklauf

Bei strukturierten schriftlichen Umfragen erfolgt die Rekrutierung der Befragten entweder aktiv oder passiv.

- Bei der **aktiven Rekrutierung** wird zunächst eine Stichprobe gezogen (z. B. Zufallsstichprobe aus der Zielpopulation, Quotenstichprobe mit spezifischem Quotenplan) und dann wird allen Mitgliedern der Stichprobe der Fragebogen vorgelegt. Möglich ist bei kleinen Populationen auch eine Vollerhebung (z. B. alle Teilnehmenden einer Bildungsmaßnahme erhalten den Evaluationsfragebogen).
- Man spricht dagegen von **passiver Rekrutierung**, wenn der Fragebogen öffentlich gestreut wird und die Befragten durch Selbstselektion eine Stichpro-

be bilden. Bei passiver Rekrutierung entstehen reine Selbstselektions- bzw. Gelegenheitsstichproben, deren Inferenzpopulationen schwer zu rekonstruieren sind (► Abschn. 9.3.1).

Das erste Kriterium zur Bewertung der Stichprobe einer strukturierten schriftlichen Umfrage ist also die Stichprobenart, die an die Rekrutierung gekoppelt ist. Das zweite Kriterium ist der **Fragebogenrücklauf**. Er sollte bei einer wissenschaftlichen Studie stets dokumentiert, mehr oder minder umfassend ausgewertet und in Publikationen auch berichtet werden in Form von Rücklaufkurve, Rücklaufquote und Rücklaufstatistik.

Rücklaufkurve Die Rücklaufkurve beschreibt die Anzahl der eingehenden ausgefüllten Fragebögen im Zeitverlauf. Während bei der Distributionsmethode des Austeilens-und-Einsammelns die ausgefüllten Fragebögen sofort zur Verfügung stehen, ist bei allen anderen Verteilungsvarianten eine mehr oder minder große zeitliche Verzögerung einzukalkulieren. Dabei werden mobile Fragebögen am schnellsten beantwortet und Online-Fragebögen am zweitschnellsten, während der Rücklauf postalischer Fragebögen einige Tage beansprucht. Wenn die Rücklaufkurve z. B. nach ca. einer Woche deutlich abflacht, kann eine **Nachfassaktion** durchgeführt werden. Bei dieser werden alle Personen, die bei aktiver Rekrutierung zur Umfrage eingeladen worden waren, noch einmal kontaktiert. Ihnen wird für die Kooperation gedankt und bisherige Nicht-Antworter werden erneut zur Teilnahme motiviert. In seltenen Fällen wird sogar eine zweite Nachfassaktion nach ca. drei Wochen durchgeführt, um die Rücklaufquote weiter zu steigern und auch die Gruppe der Spät-Antworter noch zu erreichen.

Rücklaufquote Als Rücklaufquote bezeichnet man die Anzahl der ausgefüllten Fragebögen nach Abschluss der Erhebungsphase, relativiert an der Anzahl aller ausgeteilten Fragebögen. Die Rücklaufquote liegt oft zwischen 5 % und 40 % und hängt u. a. vom Distributionsweg des Fragebogens, von der Zielpopulation, von der Thematik der Studie, von der Länge des Fragebogens, vom Zeitpunkt der Fragebogenversendung, von der Incentivierung, von Art und Anzahl der Nachfassaktionen ab. Postalische Umfragen haben meist die geringsten Rücklaufquoten. Die Rücklaufquote sollte angegeben werden, damit man eine strukturierte Fragebogenstudie und deren Stichprobe beurteilen kann. Es ist jedoch zu beachten, dass die Rücklaufquote kein Maß für die Repräsentativität der befragten Stichprobe im Hinblick auf die Zielpopulation darstellt (► Abschn. 9.1.3). So kann ein hoher Rücklauf aus einer besonders themeninteressierten Teilmenge der Population zu starken Verzerrungen der Ergebnisse führen. Bei aktiver Rekrutierung ist die Zahl der verteilten Fragebögen klar. Bei passiver Rekrutierung im Internet zieht man als Anzahl der „verteilten" Fragebögen z. B. die Anzahl der Abrufe der Umfrage heran, an der dann die Zahl der teilweise oder vollständig beantworteten Fragebögen relativiert wird. Welche Bezugsdaten zur Berechnung der Rücklaufquote herangezogen werden, verändert deren Größe beträchtlich. Dasselbe gilt für die Berechnung von Response-Raten bei strukturierten mündlichen Befragungen (z. B. Telefoninterviews; ► Abschn. 10.2.3). Hinweise und Tools zur Berechnung von Rücklaufquoten stellt z. B. die Amercian Association for Public Opion Research (AAPOR) zur Verfügung (http://www.aapor.org; zur Berechung von Rücklaufquoten bei aktiv rekrutierten bzw. personalisierten postalischen und Online-Umfragen s. AAPOR, 2011). Ein geringer Fragebogenrücklauf bzw. ein **hoher Nonresponse** („nonresponce error") stellt neben Abdeckungsfehlern („coverage error") und Messfehlern („measurement error") eine Hauptfehlerquelle bei Umfragen dar (zum **Modell des Total Survey Error** ► Abschn. 10.2.3; Groves, Dillman, Eltinge, & Little, 2002).

Rücklaufstatistik Merkmale der Fragebogen-Respondenten im Vergleich zu den Nicht-Respondenten werden in der Rücklaufstatistik erfasst. Inwiefern eine Stichprobe von ausgefüllten Fragebögen Rückschlüsse auf die Population zulässt, hängt nicht nur von der Art der Stichprobenziehung (► Abschn. 9.3) und der Anzahl der Antworten (Rücklaufquote), sondern vor allem auch von der Art der Personen ab, die antworten bzw. nicht antworten (Rücklaufstatistik). So ist bekannt, dass es sich bei den **motivierten, freiwilligen Untersuchungsteilnehmenden** häufiger um Personen mit höherer Bildung, weiblichem Geschlecht, größerem Themeninteresse, engerem Bezug zur Forschungseinrichtung etc. handelt, während andere Bevölkerungsgruppen eher dazu neigen, die Antwort zu verweigern bzw. auf eine Teilnahme zu verzichten. Eine Möglichkeit, etwas über die Merkmale von Nicht-Respondeten zu erfahren, ist der **Vergleich von Sofort- und Spät-Antwortern**. Dabei betrachtet man die soziodemografischen Merkmale derjenigen Respondenten, deren Antworten als letztes – oder erst nach einer Nachfassaktion – eingingen und vergleicht diese mit den Sofort-Antwortern. Unterscheiden sich beide Gruppen nicht oder geringfügig, so ist eine Verzerrung der Ergebnisse durch Nicht-Beantworter unwahrscheinlich. Statt anhand der Merkmale von Spät-Antwortern Rückschlüsse auf die Nicht-Antworter zu ziehen, kann man auch versuchen, die **Merkmale der Nicht-Respondeten** direkt zu eruieren. Dies ist z. B. möglich, wenn die Befragten aus einem Access-Panel rekrutiert wurden und somit die soziodemografischen Merkmale der Nicht-Antworter be-

kannt sind. Es besteht auch die Möglichkeit, im Rahmen einer Nachfassaktion die Nicht-Respondenden nur um Auskunft über einige soziodemografische Merkmale zu bitten, um dann einschätzen zu können, inwiefern sie sich von den Respondenten unterscheiden. Sofern verlässliche Informationen darüber vorliegen, welche Teilgruppen in der Stichprobe durch Antwortverweigerung unterrepräsentiert sind, können entsprechende Verzerrungen durch statistische Gewichtungsverfahren im Rahmen der Datenaufbereitung teilweise kompensiert werden (► Abschn. 11.4.4). Ansonsten sind mögliche Differenzen zwischen Respondeten und Nicht-Respondeten zumindest in der Diskussion der Ergebnisse zu berücksichtigen.

Befragung per Austeilen und Einsammeln

Die einfachste und ökonomischste Form der Datenerhebung mittels Paper-Pencil-Fragebogen stellt das **Austeilen und Einsammeln** in einer Face-to-Face-**Gruppensituation** dar („delivery and collection questionnaire"; z. B. Schulklassen, Hochschulveranstaltungen, Konferenzen, Tagungen, Elternabende, Betriebsversammlungen etc.). Hier kann das Anliegen der Studie durch ein Mitglied des Forschungsteams noch einmal mündlich erläutert werden, allerdings sollte dies in standardisierter Form erfolgen, um Versuchsleitereffekte zu vermeiden. Solange die anderen anwesenden Befragungspersonen den Fragebogen noch bearbeiten, fällt es den Befragten leichter, sich ebenfalls auf diese Tätigkeit zu konzentrieren (Effekt der sozialen Erleichterung). Demgegenüber wird beim solitären Ausfüllen allein zu Hause eher abgebrochen. Indem die ausgefüllten Bögen vor Ort sofort wieder eingesammelt werden, kann eine Rücklaufquote nahe 100 % erzielt werden. Zu beachten ist jedoch der Datenschutz, der eingeschränkt sein kann, wenn Befragte fürchten, dass der Sitznachbar auf ihr Blatt schaut und deswegen bestimmte Fragen gar nicht oder sozial erwünscht beantworten.

Befragung per Austeilen und Einsammeln – Bei einer Befragung per Austeilen und Einsammeln („delivery and collection survey") wird ein Paper-Pencil-Fragebogen an die anwesenden Befragungspersonen (meist in einer Gruppe) ausgeteilt und nach dem Ausfüllen vor Ort direkt wieder eingesammelt.

Wird in einer willkürlich herausgegriffenen Gruppe (z. B. einer Schulklasse) per Austeilen und Einsammeln ein Fragebogen erhoben, so handelt es sich um eine nicht-probabilistische **Gelegenheitsstichprobe** (► Abschn. 9.3.1). Wird dagegen aus einer Liste aller Gruppen, aus denen sich die Zielpopulation zusammensetzt (z. B. alle Schulklassen aller Schulen einer Stadt), eine Zufallsauswahl von Gruppen gezogen, die dann vollständig untersucht werden, so handelt es sich um eine probabilistische **Klumpenstichprobe** (► Abschn. 9.3.2). Per Austeilen und Einsammeln können auch **Einzelpersonen** rekrutiert und **Vollerhebungen** realisiert werden. So kann z. B. im Rahmen einer Evaluationsstudie jeder Person, die an einer bestimmten Maßnahme oder Veranstaltung teilnimmt, vor Ort ein Evaluationsfragebogen ausgehändigt werden, der unmittelbar wieder eingesammelt wird.

Postalische Befragung

Bei postalischen Umfragen („postal questionnaire", „postal survey") wird ein Paper-Pencil-Fragebogen (seltener ein auf einem Datenträger gespeicherter elektronischer Fragebogen) per Post an die Untersuchungspersonen versendet. Eine passive Rekrutierung – Fragebögen mit frankierten Rückumschlägen werden breit gestreut in Briefkästen eingeworfen – ist bei postalischen Befragungen untypisch. Stattdessen erfolgt eine **aktive Rekrutierung**, die jedoch daran gebunden ist, dass man die Adressen der in die Stichprobe gezogenen Zielpersonen kennt. Dies ist z. B. der Fall, wenn Stichproben aus **Mitgliederverzeichnissen von Organisationen** oder Institutionen gezogen werden. Zudem können bevölkerungsrepräsentative postalische Befragungen durchgeführt werden, indem man auf das **Einwohnermelderegister** zurückgreift. Eine Auskunft über Namen und Anschriften lokaler Einwohner (Gruppenauskunft) aus dem Melderegister einer Kommune ist in Deutschland nach dem Melderechtsrahmengesetz des Bundes nach § 2 möglich. Die Auskunft wird jedoch nur gegeben, wenn das Anliegen im öffentlichen Interesse liegt, was bei vielen Forschungsprojekten der Fall sein dürfte. Wenn für eine wissenschaftliche postalische Befragung ein Auszug aus dem Melderegister benötigt wird, so muss dies schriftlich beim zuständigen Einwohnermeldeamt beantragt werden, und die Daten dürfen nur für den angegebenen Zweck verwendet werden. Hierbei ist zu beachten, dass die Regelung des Meldewesens eine Aufgabe der Länder ist. Im konkreten Fall muss das Meldegesetz des jeweiligen deutschen Bundeslandes zugrunde gelegt werden.

Postalische Befragung – Bei einer postalischen Befragung („postal survey") wird ein Paper-Pencil-Fragebogen (oder ein auf einem Datenträger gespeicherter elektronischer Fragebogen) per Post verbreitet und auch auf dem Postweg wieder eingesammelt (beiliegender frankierter Rückumschlag).

Bei einer postalischen Befragung wird dem Fragebogen meist ein **ausführliches Anschreiben** beigelegt (ggf. mit Visitenkarte, Informationsflyer etc.), in dem sich die Forschungseinrichtung vorstellt, das Anliegen der Studie deutlich gemacht und zur Teilnahme motiviert wird. Da viele Privathaushalte mit Werbe- und Informationsmaterialien überhäuft werden, sollten die wissenschaftlichen Anschreiben sich schon äußerlich deutlich von Werbesendungen abheben, um Beachtung zu finden. Auch müssen die meisten Menschen zur Teilnahme motiviert werden



Abb. 10.8 Zielgruppen postalischer Umfragen müssen oft aktiv zur Teilnahme motiviert werden, denn das Ausfüllen wird selten begeistert in Angriff genommen. (© Chris Madden/Search ID: cman23, Rights Available from CartoonStock.com)

und sind nicht von sich aus vom Ausfüllen begeistert (Abb. 10.8). Nicht zuletzt stellt das Zurücksenden eine Hürde dar. Denn selbst wenn ein frankierter Rückumschlag beiliegt (und auch gefunden wird) bzw. die Rücksendung portofrei erfolgen kann, müssen die Befragten noch die Mühe auf sich nehmen, den Brief in einen Briefkasten einzuwerfen.

Insgesamt gehen postalische Umfragen meist mit einer **sehr geringen Rücklaufquote** einher (oft um 5 %; Gratton & Jones, 2004). Dies ist bei der Aussendung zu beachten. Soll etwa postalisch ein Netto-Sample von $n = 500$ erzielt werden, so müssen bei einer angenommenen Rücklaufquote von 5 % immerhin 10 000 Briefe versendet werden. Damit gehen entsprechend hohe Druck-/Kopier- und Portokosten einher, ganz abgesehen vom Aufwand für das Eintüten der versendeten Briefe, das Öffnen der eingehenden Briefe etc.

Die **geringe Effizienz** der postalischen Umfrage ist dafür verantwortlich, dass diese mittlerweile zu großen Teilen durch die sehr viel ökonomischere Online-Befragung ersetzt wird. Ausnahmen bilden Fragebogenstudien, in denen auch Personen angesprochen werden sollen, die nicht das Internet nutzen. Zudem greift man bei der **Befragung von Organisationen, Experten oder Persönlichkeiten des öffentlichen Lebens**, deren Büroadressen im Internet jeweils leicht recherchierbar sind, teilweise lieber auf postalische Umfragen zurück, um dem Anliegen größeres Gewicht zu verleihen. Hier wird vermutet, dass ein aufwändigeres postalisches Schreiben stärker beachtet wird als eine E-Mail-Anfrage, die im überlaufenden Posteingang eher untergeht.

Online-Befragung

Wenn wissenschaftliche Umfragen durchgeführt werden, dann heute zunehmend als Online-Befragung. Neben dem mündlichen Telefon-Tnterview (CATI: „computer-assisted telephone interview“; ▶ Abschn. 10.2.3) ist der schriftliche Online-Fragebogen (**CAWI**: „computer-assisted web interview“, „online questionnaire“) mittlerweile die wichtigste vollstrukturierte Befragungstechnik, die sowohl in der privatwirtschaftlichen Markt- und Sozialforschung als auch in der akademischen Grundlagenforschung eingesetzt wird (Couper, 2008; Jackob, Schoen, & Zerback, 2009; Kuckartz, Ebert, Rädiker, & Stefer, 2009). Der Hauptvorzug liegt in der großen **Effizienz** des Verfahrens.

Online-Befragung – Bei einer Online-Befragung („online survey“, „web survey“) wird ein elektronischer Fragebogen per Internet (vor allem Web) verbreitet und meist an stationären Computern bearbeitet. Die gewählten Antworten der Befragungspersonen werden auf einem Befragungsserver gespeichert und stehen als digitale Datensätze zur Verfügung.

Online-Umfragen werden üblicherweise über entsprechende **Befragungsserver** im Web durchgeführt (z. B. Enterprise Feedback Suite [EFS]: http://www.globalpark.de/efs-uebersicht/, Lime Survey: http://www.limesurvey.org, Survey Monkey: de.surveymonkey.com, eine Übersicht zu „Software für Onlinebefragungen“ bietet GESIS: http://www.gesis.org). Ein Versenden von elektronischen Fragebögen per E-Mail ist im Vergleich zur Web-Umfrage wesentlich unkomfortabler und schwieriger zu anonymisieren. Soll eine Online-Umfrage realisiert werden, so muss zunächst der Zugang zu einem Befragungsserver sichergestellt werden (viele Hochschulinstitute haben Lizenzverträge abgeschlossen). Zudem muss man sich in die Bedienung der Plattform einarbeiten, um den Fragebogen dort entsprechend umzusetzen. Gründliche technische Pretests sind empfehlenswert, um sicherzustellen, dass die Daten korrekt eingelesen werden.

Umfrageserver ermöglichen ein **Teilnehmer-Management** und eine geordnete **Präsentation des Fragebogens** sowie die **Dokumentation der Antworten**. Im Unterschied zum Paper-Pencil-Fragebogen bietet der elektronische Fragebogen den Vorteil, dass z. B. Bilder, Audio- und Videoelemente sowie neue Itemformate (z. B. Drag-and-Drop-Aufgaben, Pull-Down-Menüs, Analogskalen mit Schiebereglern, ▶ Abschn. 8.4.4) einbezogen

werden können, eine Filterführung automatisch vorgenommen wird, ein Zurückblättern oder Überspringen von Items unterbunden, Ausfülldatum und Ausfüllzeit automatisch registriert und zwischendurch Vollständigkeits- und Plausibilitäts-Checks durchgeführt werden können. In der laufenden Umfrage lassen sich auch Zwischenergebnisse anzeigen. Ebenso kann die Befragungsplattform experimentelle Designs realisieren, indem die Befragten randomisiert bestimmten Bedingungen (z. B. Reizvorlagen) zugeordnet werden, und danach den Fragebogen beantworten.

Ein guter Online-Fragebogen muss auf das Medium abgestimmt (z. B. Lesen am Bildschirm) und hinsichtlich seiner **Usability** überprüft worden sein (zu Layout und Usability von Webfragebögen s. z. B. Kaczmirek, 2009; Lumsden & Morgan, 2005). Zur optimalen Gestaltung eines Online-Fragebogens gibt es unterschiedliche Empfehlungen, etwa dahingehend, ob jede Frage einzeln auf dem Bildschirm angezeigt werden sollte (Vorteil: Übersichtlichkeit, verringerte Ausstrahlungseffekte der Items; Nachteil: Fragebogen wird sehr lang, Befragte müssen sich durch sehr viele Seiten „durchklicken"; Abb. 10.9), oder ob pro Bildschirmseite eine größere Zahl von Items zu präsentieren ist. Ein **Fortschrittsbalken** sollte in jedem Fall genutzt werden, um den Befragten eine Orientierung über den noch anstehenden Aufwand zu geben. Als Faustregel geht man davon aus, dass Online-Befragungen nicht länger als **10–15 Minuten** dauern sollten (Bosnjak, 2002) – in der Praxis sind jedoch zahlreiche längere Fragebögen anzutreffen. Mit wachsender Länge steigt bei Online-Fragebögen die Antwortverweigerung, zudem sinkt im hinteren Teil des Fragebogens die Datenqualität (z. B. weil Fragen zunehmend schneller und oberflächlicher beantwortet oder übersprungen werden, vgl. Galesic & Bosnjak, 2009).



"There´s never an option that reflects exactly what I want to say."

Abb. 10.9 Ein Online-Fragebogen muss nicht nur technisch auf das Verbreitungsmedium zugeschnitten, sondern wie jeder andere Fragebogen auch gut verständlich sein. (© Rob Murray/Search ID: rmun48, Rights Available from CartoonStock.com)

Bei der Online-Befragung sind einige methodische Besonderheiten und Limitationen zu beachten.

- **Repräsentativität von Online-Umfragen**: Ein Hauptkritikpunkt an Online-Umfragen war und ist der Vorwurf, diese seien nicht repräsentativ. Tatsächlich können auf dem Onlineverbreitungsweg nur **Personen mit Internetzugang** erreicht werden, die Bevölkerungsgruppe der Offlinerinnen und Offliner bleibt außen vor. Sollen Personen ohne oder mit äußerst sporadischer Onlinenutzung befragt werden, so sind andere Kanäle (z. B. postalische Befragung) zu wählen. Abgesehen von dieser Einschränkung lassen sich per Online-Umfrage bei **aktiver Rekrutierung** alle möglichen Stichprobenarten realisieren, auch probabilistische Stichproben. Wird die Online-Befragung mit **passiver Rekrutierung** durchgeführt (Veröffentlichung des Umfragelinks über Websites, Online-Foren, soziale Netzwerkplattformen), so resultiert indessen eine Selbstselektionsstichprobe, von der unklar ist, welche Inferenzpopulation sie repräsentiert (z. B. Population aller Webnutzenden, die in Kalenderwoche 32 die Website xy und/oder das Onlineforum xy besucht, dort den Link auf die Umfrage bemerkt und angeklickt und schließlich den Fragebogen ausgefüllt haben). Auch bei der in der Online-Forschung verbreiteten Nutzung von **Online-Access-Panels** ist zur Beurteilung der Repräsentativität zu beachten, wie das Panel rekrutiert und zusammengesetzt ist. Ein aktiv nach einem probabilistischen Stichprobenverfahren rekrutiertes Panel hat hierbei eine höhere Aussagekraft als ein passiv rekrutiertes Panel, zu dem sich willkürlich Freiwillige anmelden.
- **Rücklaufquote von Online-Umfragen**: Wie bei jeder selbstadministrierten Befragung ist auch bei Online-Befragungen mit höheren Verweigerungs- und Abbruchraten als z. B. bei mündlichen Interviews zu rechnen, da hier die motivierende Wirkung der Kommunikation mit dem Interviewer entfällt und die Befragten auf sich gestellt sind. Die stark wachsende Zahl an Online-Befragungen kann zusätzlich zu Übersättigung und Antwortverweigerung führen. Auch ein thematisch uninteressant wirkender, schlecht lesbarer, überlanger oder umständlich bedienbarer Online-Fragebogen reduziert den Rücklauf. Abgesehen davon, dass derartige Gestaltungsfehler wenn möglich zu vermeiden sind, können Incentives (z. B. Verlosungen) dazu beitragen, die Teilnahmebereitschaft zu steigern, wenn diese durch einen langen Fragebogen und/oder ein uninteressantes Thema beeinträchtigt ist (Marcus,

Bosnjak, Lindner, Pilischenko, & Schütz, 2007). Als Anreiz kann auch eine Rückmeldung über die Studienergebnisse angeboten werden. Die Rücklaufquote hängt bei jedem Umfrage-Modus von vielen Faktoren ab und ist themen- und zielgruppenspezifisch zu diskutieren (Bowling, 2005). So ist sie bei bevölkerungsrepräsentativen Umfragen im postalischen Modus teilweise höher als im Online-Modus (Bech & Kristensen, 2009). Bei internetaffinen Zielgruppen sollte der Rücklauf bei Online-Umfragen hingegen deutlich über dem Rücklauf postalischer Umfragen liegen. Bei groß angelegten Umfragen ist zuweilen eine Kombination unterschiedlicher Befragungs-Modi empfehlenswert, um den Rücklauf zu steigern, etwa indem man Befragten die Wahl lässt, ob sie den Fragebogen lieber Online, über eine Smartphone-App oder klassisch auf Papier beantworten möchten („mixed-mode survey"; Couper, 2011; Dillman et al., 2009). Bei einer vollstrukturierten **Mixed-Mode-Befragungsstudie** kann man nicht nur unterschiedliche schriftliche Befragungs-Modi kombinieren (z. B. Online-Befragung plus postalische Befragung), sondern bei Bedarf zusätzlich auch unterschiedliche mündliche Befragungs-Modi einbeziehen (z. B. persönliches und telefonisches Interview). Der mit einer Mixed-Mode-Befragung verbundene Aufwand bei der Datenerhebung und Datenanalyse ist jedoch nur gerechtfertigt, wenn die verbesserte Stichprobenqualität zur Lösung des Forschungsproblems notwendig ist. Andernfalls beschränkt man sich auf einen einzigen Befragungs-Modus (z. B. Online-Befragung) und diskutiert die Grenzen der Aussagekraft der Studie angesichts mangelnder Berücksichtigung von internetabstinenten Personen. Zudem ist zu beachten, dass jeder Befragungs-Modus wiederum spezifische Verzerrungen und Fehler im Sinne des Total-Survey-Error-Modells mit sich bringt (Bowling, 2005) und ein Hochschrauben der Rücklaufquote somit nicht pauschal zu einer Verbesserung von Stichproben- und Datenqualität führt.

- **Forschungsethik bei Online-Umfragen**: Aus forschungsethischen Gründen (▶ Abschn. 4.1) sind empirische Daten grundsätzlich vertraulich und anonym zu behandeln. Bei einer Online-Umfrage wäre eine direkte oder indirekte Identifizierbarkeit gegeben, wenn im ausgefüllten Fragebogen z. B. E-Mail- oder IP-Adressen gespeichert und weiterverarbeitet würden. Obwohl eine Personalisierung für die Forschung keinerlei Erkenntnisgewinn bietet, von den Forschenden nicht intentiert und forschungsethisch und datenschutzrechtlich gar nicht zulässig ist, mehren sich angesichts öffentlicher Datenschutzdebatten entsprechende Sorgen auf Seiten der Befragungspersonen. Hier gilt es, darüber aufzuklären, dass die Umfrageteilnahme stets freiwillig und anonym erfolgt und dass dies selbstverständlich auch gewährleistet ist, wenn ein Umfrageserver individualisierte Umfrageeinladungen und Erinnerungsmails verschickt. Für jede Online-Umfrage ist abzuwägen, ob bestimmte Optionen des Teilnehmer-Managements (z. B. Vergabe eines einmal verwendbaren Umfragezugangslinks, um ein die Stichprobe verzerrendes Mehrfachbeantworten oder die Weitergabe des Links auszuschließen) umgesetzt werden sollen, wenn diese möglicherweise unter den Befragten unnötige Datenschutzsorgen hervorrufen.
- **Messfehler bei Online-Umfragen**: Gemäß dem **Total-Survey-Error-Modell** (▶ Abschn. 10.2.3) entstehen Fehler bei Umfragen, wenn die Zielpopulation nicht vollständig im Auswahlrahmen abgebildet ist (Coverage-Problem, z. B. Unterabdeckung von älteren Menschen bei Online-Umfragen), wenn die Stichprobe durch Antwortverweigerungen verzerrt ist (Nonresponse-Problem, z. B. verringerte Teilnahmebereitschaft bei manchen Befragungspersonen aufgrund Datenschutzbedenken) und wenn der Befragungs- bzw. Fragebogen-Modus das Antwortverhalten im Sinne eines Messfehlers beeinflusst (Measurement-Problem). Diese modusspezifischen Messfehler können sich bei Online-Umfragen z. B. darauf beziehen, wie aufmerksam die Fragebogenitems am Bildschirm gelesen und kognitiv verarbeitet werden und inwiefern beim Antworten am Computer soziale Kontextinformationen mitberücksichtigt werden. Sozialpsychologische Effekte, die aus der computervermittelten Kommunikation bekannt sind (z. B. Enthemmung, De-Individuation; Döring, 2003) könnten sich auch bei der Beantwortung von Online-Fragebögen niederschlagen, so dass dieselben Personen dieselben Items anders beantworten, wenn sie diese Online oder auf Papier vorgelegt bekommen (z. B. offener, weniger sozial erwünscht, provokativer etc.). Bislang wird oft von geringen modusspezifischen Messfehlern ausgegangen; zudem werden diese durch bestimmte Präventions- und Korrekturverfahren teilweise kompensiert, insbesondere wenn in einer Mixed-Mode-Umfrage die Daten verschiedener Befragungs-Modi gemeinsam ausgewertet werden sollen (für einen Forschungsüberblick s. Couper, 2011).

Mobile Befragung

Seit der stärkeren Verbreitung von tragbaren, drahtlos vernetzten Computermedien (Smartphone, Tablet PC etc.) wird auch diese Kommunikationsplattform für Forschungszwecke eingesetzt. Erste mobile Umfragen mit

standardisierten Fragebögen (mobile selbstadministrierter Non-Voice-Verfahren; Maxl & Döring, 2010) arbeiteten Anfang der 1990er-Jahre mit dem SMS- und MMS-Dienst, um einzelne Fragen und Antworten mit Respondenten auszutauschen. Mit dem Aufkommen von Smartphones, die internetfähig sind und bei denen durch Datenflatrates auch keine zusätzlichen Kosten entstehen, werden **mobile Online-Umfragen** und **mobile Apps** als Fragebogenvarianten attraktiver.

Mobile Befragung – Bei einer mobilen Befragung („mobile survey") wird ein elektronischer Fragebogen über drahtlose Netzwerke bzw. Mobilfunknetze und mobile (portable) Endgeräte verbreitet. Die gewählten Antworten werden auf einem Befragungsserver gespeichert und stehen als digitale Datensätze zur Verfügung. Mobile Fragebögen müssen für die angezielten Endgeräte und Umfragesituationen (z. B. das Ausfüllen unterwegs in Nischenzeiten oder im Freien) gestaltet werden und unterscheiden sich somit von herkömmlichen Online-Fragebögen.

Insbesondere um Jugendliche oder sehr mobile, kaum per Festnetz oder Internet erreichbare Zielgruppen anzusprechen, ist die mobile Befragung prädestiniert. Ebenso dienen mobile Endgeräte für **kontextgebundene Befragungen**, etwa indem ortsbezogen ein Fragebogenlink zugespielt wird, wenn sich eine Person gerade in einer bestimmten Einkaufspassage, im Fussballstadion oder bei einer Touristenattraktion aufhält (zur mobilen Marktforschung s. Maxl, Döring, & Wallisch, 2009).

Automatische Telefonbefragung

Die Fragebogenmethode ist eine schriftliche, selbstadministrierte Form der Datenerhebung. Es besteht aber auch die Möglichkeit einer **selbstadministrierten mündlichen Befragung**, sozusagen eines **Selbstinterviews**. Im Unterschied zum Telefoninterview, bei dem ein Livekontakt zwischen Befragungspersonen und Interviewenden zustande kommt (► Abschn. 10.2.3), handelt es sich bei der automatischen Telefonbefragung um ein selbstadministriertes Verfahren mithilfe eines computergesteuerten Telefonsystems (Janda, Janda, & Tedford, 2001).

Automatische Telefonbefragung – Bei einer automatischen Telefonbefragung („interactive voice response", **IVR**; „telephone audio computer assisted self-interviewing", **T-ACASI**) wird der Fragebogen über ein telefonisch erreichbares Audiosystem vorgelesen. Bei dieser selbstadministrierten akustischen Fragebogenpräsentation geben die Respondenten ihre Antworten per Tastendruck („touch tone input") oder per Spracheingabe („voice input").

Die selbstadministrierte automatische Telefonbefragung wird als Alternative zum vollstrukturiertenTelefoninterview empfohlen, wenn es um die Erhebung sensibler Themen geht (Cooley, Miller, Gribble, & Turner, 2000). Zudem wird das vollstrukturierte mündliche Selbstinterview dem selbstausgefüllten schriftlichen Fragebogen vorgezogen, wenn man den Befragten ein mündliches Beantworten der Fragen erlauben will. Die Datenerhebung per automatischer Telefonbefragung kann im Alltag der Befragten zwischendurch oder unterwegs erfolgen, wenn keine feste Unterlage zum Schreiben zur Verfügung steht. Die Kundenzufriedenheit mit einem Restaurant oder Freizeitpark kann z. B. über eine automatisierte telefonische Kurzbefragung erhoben werden, wenn auf der Rechnung die entsprechende kostenfreie Telefonnummer des Systems angegeben ist und mit der Umfrageteilnahme ggf. eine Verlosung oder ein Gutschein verbunden ist.

Die **telefonische Selbstbedienung** ist vielen Menschen heute von Unternehmens-Hotlines bekannt, in denen sie sich per Spracheingabe oder Tastendruck durch Menüs bewegen. Dort erzeugt sie jedoch oft Frustration und es wird als Mangel empfunden, keinen menschlichen Ansprechpartner zu erreichen (Dean, 2008). Die Akzeptanz der automatischen Telefonbefragung muss deswegen zielgruppenspezifisch geprüft werden. Es ist im Vergleich zum Telefoninterview mit erhöhten Abbruchquoten zu rechnen (Dillman et al., 2009). Die besten Erfahrungen mit automatischen Telefonbefragungen wurden bislang im Rahmen von Tagebuchstudien gesammelt.

Tagebuchmethode und Ambulantes Assessment

Wie viel Radio hören Sie? Wie oft lügen Sie? Wie viel Zeit verbringen Sie mit Hausarbeit? Wann sind Sie im Tagesverlauf gut oder schlecht gelaunt? Wie viel Fleisch essen Sie? Wann und wie lange leiden Sie unter Kopfschmerzen? Es fällt uns normalerweise schwer, derartige Fragen exakt zu beantworten. Denn um hier zu genauen Einschätzungen zu gelangen, müssen wir längere zurückliegende Zeiträume einbeziehen („retrospective recall") und uns dabei auf typische Abläufe konzentrieren, Ausnahmefälle dagegen vernachlässigen. Abgesehen davon, dass es schon kognitiv schwer fällt, alle notwendigen Informationen vollständig abzurufen und zu einer passenden Antwort zu verarbeiten (z. B. Fernsehkonsum in der letzten Woche), spielen auch soziale Verzerrungen eine Rolle, etwa weil wir vor uns selbst und anderen gut dastehen möchten: So werden wir z. B. die für Arbeiten und Lernen aufgewendete Zeit mehr oder minder unbewusst durch großzügiges Aufrunden eher über-, die Fernsehzeit eher unterschätzen („overreporting", „underreporting"). Einen Lösungsansatz für diese Problematik stellen die vollstrukturierte Tagebuchmethode („diary method") und das Ambulante Assessment („ambulatory assessment") dar (zu Methodenvergleichen zwischen retrospektiven Fragebogendaten und fortlaufend erhobenen Tagebuchdaten zum Arbeitszeitbudget s. Otterbach & Sousa-Poza,

2010; zum Sexualverhalten s. Horvath, Beadnell, & Bowen, 2007; zum Wasserverbrauch s. Wutich, 2009).

Tagebuchmethode – Die vollstrukturierte Tagebuchmethode („diary method") ist eine genuin sozialwissenschaftliche Befragungsmethode. Das Tagebuch besteht aus einer chronologisch geordneten Serie vollstandardisierter Fragebogenformulare, die von den Befragten fortlaufend (mindestens einmal pro Tag) über längere Zeitperioden (mehrere Tage bis Wochen) hinweg ausgefüllt werden. Ziel ist es, möglichst unverzerrte und detaillierte Angaben über alltägliche Erlebens- und Verhaltensweisen zu erhalten, was jedoch mit großem Aufwand für die Befragten verbunden ist.

Anstelle einer einmaligen summarischen Gesamtabschätzung (z. B. des TV-Konsums) werden die Befragten bei der Tagebuchmethode angeleitet, sich selbst im Alltag genau zu beobachten und diese Selbstbeobachtungen über mehrere Tage oder Wochen hinweg schriftlich fortlaufend in einem **Paper-Pencil-Tagebuch** (als Heft oder in Form einzelner Formulare) detailliert zu notieren (für Zeittagebücher Belli, Stafford, & Alwin, 2008; Pentland, Powell Lawton, Harvey, & McColl, 1999). Neuere Varianten verwenden auch **elektronische Tagebücher** („eletronic diaries"), die auf dem Computer oder einem mobilen Endgerät der Befragten auszufüllen sind. Elektronische Tagebücher haben den Vorteil, dass sie z. B. Datum und Uhrzeit automatisch eintragen, durch Signaltöne oder Erinnerungs-Nachrichten an den fälligen Eintrag erinnern und – im Falle von vernetzten Endgeräten – die Daten auch gleich auf einem zentralen Forschungsserver sammeln (für einen Methodenvergleich zwischen Papier-, Online- und SMS-Tagebuch s. Lim, Sacks-Davis, Aitken, Hocking, & Hellard, 2010). Nicht zuletzt wird auch die **automatische Telefonbefragung** für Tagebuchstudien erfolgreich eingesetzt, etwa indem Teilnehmerinnen und Teilnehmer über einen Zeitraum mehrerer Tage oder Wochen hinweg täglich vom Telefonsystem angerufen werden, um mündlich oder per Tastendruck Informationen zum aktuellen Gesundheitsverhalten zu geben (Perrine, Mundt, Searles, & Lester, 1995; Stritzke, Dandy, Durkin, & Houghton, 2005).

Im Unterschied zur Tagebuchmethode als sozialwissenschaftlicher Befragungstechnik stammt das **Ambulante Assessment**, bei dem es primär um objektive Messungen von Verhalten oder körperlichen Prozessen im Alltag geht, aus der klinischen Praxis. Es wird aber auch sozialwissenschaftlich eingesetzt und kann mit der Tagebuchmethode sinnnvoll kombiniert werden, weshalb es an dieser Stelle behandelt wird (zu physiologischen Messungen ▶ Abschn. 10.5). Während die Tagebuchmethode auch rein papierbasiert umgesetzt werden kann, sind für das Ambulante Assessment entsprechende technische Geräte notwendig. In Forschungsprojekten stellt man die Geräte üblicherweise leihweise zur Verfügung, damit alle Teilnehmenden mit derselben Plattform arbeiten (Fahrenberg, Myrtek, Pawlik, & Perrez, 2007; Wilhelm & Perrez, 2008).

Methode des Ambulanten Assessments – Die Methode des Ambulanten Assessments (auch: Ambulantes Monitoring; „ambulatory assessment", „ambulatory monitoring", „ecological momentary assessment", EMA) stammt aus dem klinischen Anwendungsfeld. Beim Ambulanten Assessment werden physiologische Messungen (z. B. Blutdruck, Herzfrequenz, Atmung), Verhaltensmessungen (z. B. Schrittzähler) sowie Messungen von Umweltvariablen (z. B. Lärm, Temperatur, Helligkeit) mithilfe entsprechender mobiler Messgeräte über längere Zeitperioden fortlaufend im Alltag der Untersuchungspersonen (anstatt z. B. in der Klinik oder im Forschungslabor) durchgeführt. Die objektiven Messungen können durch eine fortlaufende Erhebung von Selbstauskunftsdaten im Sinne der vollstrukturierten Tagebuchmethode sinnvoll ergänzt werden.

Ein wichtiges Anwendungsfeld für alltagsnahe fortlaufende Erhebungen von Selbstauskünften ist die Verkehrsforschung. Hier geht es darum, die alltägliche Mobilität von Menschen zu erfassen, etwa für die Verkehrsplanung. Reisetageücher („travel diaries") wurden dazu zunächst in Papierform, später auch elektronisch eingesetzt. Heute werden kombinierte Verfahren erprobt, bei denen die Datenerfassung mittels GPS-fähigen Smartphones erfolgt: Die Studienteilnehmenden nutzen dafür eine entsprechende Applikation auf dem Handy, geben für jede Fahrtstrecke Ausgangspunkt und Zielort ein, wobei anhand der GPS-Daten die genauen Routen nachvollziehbar sind. Abends werden die per Handy im Tagesverlauf mitprotokollierten Selbstauskunfts- und GPS-Daten automatisch in ein Web-Tagebuch übertragen und können dort von den Studienteilnehmenden komfortabel überprüft und ergänzt werden (Itsubo & Hato, 2006).

Ob eher eine klassische papierbasierte Tagebuchform oder eine computergestützte mobile Assessmentvariante einzusetzen ist, hängt a) von der Fragestellung (z. B. werden auch physiologische Messwerte benötigt?), b) vom Budget des Projektes (z. B. können Leihgeräte bereitgetellt werden?) sowie c) von den Technikeinstellungen der Zielgruppe ab (so kann die Verwendung innovativer mobiler Endgeräte bei manchen Zielgruppen die Teilnahmemotivation steigern, bei anderen eher auf Akzeptanzprobleme stoßen).

Bei der Datenerfassung mit der strukturierten Tagebuchmethode sind drei Vorgehensweisen zu differenzieren:

1. **Durchgängiges Protokoll über den ganzen Tag hinweg**: Vorbereitete Formulare decken die gesamten 24 Stunden des Tages ab (Robinson, 2002). Die für die Studie relevanten Aktivitäten oder Ereignisse werden fortlaufend (typischerweise alle 15 Minuten mit Ausnahme der Schlafenszeit) eingetragen (◘ Abb. 10.10). Diese Methode erlaubt sehr genaue Analysen der Zeitdauer unterschiedlicher Aktivitäten und ist für

Uhrzeit	Was haben Sie gemacht? Bitte dokumentieren Sie ihre Aktivitäten alle 15 Minuten. Eine Aktivität pro Zeile.	Was haben Sie noch gemacht? Schreiben Sie eine weitere Aktivität in die Zeilen, welcher Sie neben der ersten zur gleichen Zeit nachgegangen sind. z.B. telefonieren, aufräumen, Musik hören oder etwas trinken.	Wo waren Sie? z.B. zu Hause, am Arbeitsplatz, im Auto, im Zug, auf dem Fahrrad, beim Sport, bei Verwandten	Mit wem waren Sie zusammen? Bitte kreuzen Sie an. Siehe dazu das Beispiel auf Seite 1. Alleine oder mit einem Unbekannten	Kinder bis 9 Jahre, die in Ihrem Haushalt leben	Kinder zwischen 10 und 14 Jahren, die in Ihrem Haushalt leben	Andere Mitglieder aus Ihrem Haushalt	Andere Personen, die Sie kennen
18:00 – 18:15	Von der Arbeit nach Hause gefahren	Zeitung lesen	Im Zug	☒				
18:15 – 18:30	↓	↓	↓					
18:30 – 18:45	Vom Bahnhof nach Hause gelaufen		auf der Straße	☒				
18:45 – 19:00	Abendbrot gemacht	Mit Partner geredet	zu Hause				☒	
19:00 – 19:15	Essen	↓						
19:15 – 19:30	abwaschen							
19:30 – 19:45	Im Internet surfen	Telefonieren						
19:45 – 20:00	Fernsehen	Wäsche bügeln	↓				☒	

In diesem Bereich müssen Sie keine Angaben machen für die Zeit auf Arbeit, in der Schule oder im Schlaf. Bitte kontrollieren Sie, dass für alle anderen Aktivitäten/Zeiten Angaben gemacht wurden.

Bitte führen Sie dieses Tagebuch während des gesamten Tages mit sich. Machen Sie zeitnah – am besten sofort – die entsprechenden Einträge. Vielen Dank für Ihre Unterstützung.

Abb. 10.10 Auszug aus einem Tagebuch-Formular für ein vollständiges Tagesprotokoll in Anlehnung an UK 2000 Time Use Survey, durchgeführt durch das Office for National Statistics (ONS)

Zeitbudgetstudien einschlägig (z. B. Zeitaufwand für Hausarbeit, Erwerbsarbeit, Mediennutzung etc.).

2. **Protokollierung zu ausgewählten Zeitpunkten im Laufe des Tages** („time sampling diary", TSD): Weniger aufwändig für die Befragungspersonen ist eine stichprobenartige Protokollierung zu vorgegebenen festen oder flexiblen Zeitpunkten.
   - Beispiel für **Protokollierung zu festen Zeitpunkten**: In den USA lebende Latinas und Latinos (n = 91) füllten 13 Tage lang **jeden Abend** auf einer Internetplattform ein standardisiertes Online-Tagebuch zu ihren subjektiven Diskriminierungserfahrungen in unterschiedlichen Lebensbereichen (z. B. „Today, because I am Latina/o, I was passed up for a promotion and/or benefits at work." Yes/No) sowie zu ihrer Depressivität bezogen auf den aktuellen Tag aus (Torres & Ong, 2010). Die Befragungspersonen wurden online rekrutiert und erhielten eine Erinnerungs-E-Mail, wenn sie ihr tägliches Tagebuchformular nicht ausgefüllt hatten. Es zeigte sich, dass die Depressivität jeweils am Tag nach einer erlebten Diskriminierung signifikant anstieg.
   - Beispiel für **Protokollierung zu unterschiedlichen Zeitpunkten**: Studierende in den USA (n = 124) wurden im Zuge einer Studie zum Alkoholkonsum mit Handhelds ausgestattet, auf denen ein elektronisches Tagebuch für den Zeitraum von zwei Wochen installiert war (Piasecki, Slutske, Wood, & Hunt-Carter, 2010). Nach einem individuellen Zufallsprinzip wurden die Teilnehmenden per Signalton **4-mal am Tag zu unterschiedlichen Zeitpunkten** aufgefordert, Einträge im Tagebuch vorzunehmen, die sich auf subjektive Symptome eines „Katers" („hangover") in den jeweils letzten 15 Minuten vor dem Signalton bezogen (z. B. extreme Müdigkeit, Kopfschmerzen, Schwächegefühl, extremer Durst). Das mehrfach am Tag per Tagebuch erfasste „Kater"-Erleben wurde mit der jeweils einmal pro Tag erfragten Menge des Alkoholkonsums am Vorabend in Verbindung gesetzt. Alle drei bis vier Tage sollten sich die Befragungspersonen persönlich beim Forschungsteam melden. Die auf dem Handheld gespeicherten Daten wurden dann ausgelesen und die Teilnehmenden zum weiteren sorgfältigen Ausfüllen des Tagebuchs animiert. Die Studie lieferte Hinweise darauf, welche Alkoholmenge zu wie starken „Kater"-Symptomen führten.

3. **Protokollierung nur der relevanten Ereignisse** („experience sampling method", ESM): Bei diesem Verfahren ist die Protokollierung nicht von Zeitfaktoren, sondern lediglich vom Auftreten der relevanten Ereignisse gesteuert, etwa wenn die Befragten die Inhalte ihrer Telefonate oder ihrer Einkäufe festhalten. Tagebuchstudien, in denen Untersuchungspersonen alle Situationen protokollierten, in denen sie im Alltag lügen – inklusive Merkmale der Lüge und Adressaten der Lüge – zeigten z. B., dass Lügen ein sehr verbreitetes Phänomen ist und vor allem in Form von „weißen Lügen" dazu dient, soziale Konflikte zu vermeiden (DePaulo & Kashy, 1998). Indem in örtlicher und zeitlicher Nähe zum Ereignis sofort die Daten protokolliert werden, sollen Messfehler reduziert werden. Sofortiges Protokollieren kann stichpunktartig in der jeweiligen Situation mithilfe eines mitgeführten kleinen Notizbuchs erfolgen, wobei die über den Tag gesammelten ereignisbezogenen Notizen dann abends gebündelt in das Tagebuchformular übertragen werden (DePaulo & Kashy, 1998). Im Rahmen einer Tagebuchstudie mithilfe automatischer Telefonbefragung sollte die Handynutzung im Alltag erfasst werden: Die Teilnehmenden wurden instruiert, nach jedem Handytelefonat die Telefonnummer des Umfrageservers zu wählen und dann im Telefonsystem mündlich anzugeben, wo sie sich gerade befanden und mit wem sie das Handytelefonat geführt hatten (Cohen & Lemish, 2003). Diese Tagebucherhebung zeigte, dass Menschen sehr viel häufiger zu Hause mit dem Handy telefonieren, als sie es bei einer summarischen Abfrage in einem Fragebogen angeben.

Bei der Datenerhebung mittels standardisiertem Tagebuch (ggf. flankiert durch Ambulantes Assessment) sind folgende Besonderheiten im Unterschied zur Nutzung einfacher standardisierter Fragebögen zu beachten:

- **Wahl eines geeigneten Erhebungszeitraums**: Sozialwissenschaftliche Tagebuchstudien erfassen typischerweise einen Zeitraum von mehreren Tagen bis Wochen. Relativ üblich ist eine einwöchige Erhebungszeit. Neben der Dauer ist auch die Platzierung der Erhebungsperiode entscheidend (z. B. Vermeidung von Ferienzeiten oder extremen Wetterlagen).
- **Schulung aller Studienteilnehmenden**: Alle Befragungspersonen sind gründlich zu schulen, damit sie genau wissen, wie die Formulare auszufüllen oder die technischen Erfassungsgeräte zu bedienen sind.
- **Fortlaufende Motivierung und Kontrolle der Teilnehmenden während des Erhebungszeitraums**: Erfahrungsgemäß müssen Teilnehmende von Tagebuch-Studien regelmäßig motiviert und erinnert werden, die Protokollierung über Tage hinweg sorgfältig aufrecht zu erhalten. Neben einer Incentivierung sind hier regelmäßige Erinnerungsanrufe einschlägig. Auch kann man sich die ausgefüllten Bögen täglich elektronisch zusenden lassen (Sicherstellung der Anonymität durch Vergabe eines Identifizierungscodes, der die Zuordnung der Daten einer Person erlaubt, nicht jedoch ihre Identität preisgibt), so dass bei fehlendem Rücklauf sofort nachgehakt werden kann.

Delphi-Methode

Bei der Delphi-Methode handelt es sich um eine strukturierte schriftlichen Expertenbefragung zu einem zukunftsbezogenen Sachverhalt, die in mindestens zwei Runden durchzuführen ist und bei der alle Befragten die Antworten der anderen Befragten erfahren, um ihr eigenes Urteil zu überdenken (► Studienbeispiel „Delphi-Methode").

Delphi-Methode – Die Delphi-Methode („delphi technique", „delphi poll", „delphi survey") ist eine nach dem berühmten griechischen Orakel benannte spezielle Form der vollstrukturierten schriftlichen Befragung, die auf die Suche nach Problemlösungen und die Prognose zukünftiger Entwicklungen abzielt. Kennzeichnend für die Delphi-Methode ist, dass eine Gruppe von Expertinnen und Experten („delphi panel") in mehreren (mindestens zwei) Runden befragt und ihnen dabei ab Runde zwei jeweils die aggregierten Antworten der Mitexperten zugänglich gemacht werden. Die Befragten können somit ihre eigenen Antworten im Kontext der Antworten des gesamten Delphi-Panels reflektieren und ggf. modifizieren und somit im Idealfall zu besseren Einschätzungen gelangen.

Der Ablauf einer Delphi-Befragung ist folgendermaßen zu beschreiben (Häder & Häder, 2009; Linstone & Turoff, 1975):

1. Ein Forschungsteam entwickelt zunächst für eine anstehende Problematik oder Prognosefrage einen ausführlichen vollstandardisierten Fragebogen (ggf. ergänzt um offene Fragen), der für die **erste Befragungsrunde** an eine größere Expertengruppe unterschiedlicher Fachrichtungen postalisch bzw. per Internet verschickt wird. Die **Onlineform der Delphi-Methode** („electronic/internet/online Delphy study") beschleunigt das Verfahren und erleichtert internationale Studien. Ebenso wie bei Interview- oder Fragebogenstudien mit einzelnen Experten steht und fällt die Qualität einer Delphi-Studie mit der tatsächlichen **Expertise des rekrutierten Delphi-Panels** (z. B. definiert über Ausbildung und Berufserfahrung der Sachverständigen).
2. Das Forschungsteam wertet die ausgefüllten Fragebögen der ersten Runde aus und fertigt auf der Basis der Resultate einen neuen bzw. überarbeiteten Fragenkatalog an, der demselben Expertenkreis erneut vorgelegt wird. Diese **zweite Befragungsrunde** informiert zusätzlich über die Standpunkte und Lö-

Studienbeispiel

Delphi-Methode

Wie wird sich die Psychotherapie in den kommenden zehn Jahren entwickeln? Zur Beantwortung dieser Forschungsfrage wurde 2001 in den USA eine zweistufige postalische Delphi-Studie durchgeführt (Norcross, Hedges, & Prochaska, 2002). Dem Delphi-Panel gehörten $n = 62$ US-amerikanische Expertinnen und Experten für Psychotherapie mit durchschnittlich 30 Jahren postdoktoraler Berufserfahrung an. Der standardisierte Delphi-Fragebogen enthielt rund 100 Items (Tab. 10.12).

Es zeigte sich inhaltlich, dass das befragte Delphi-Panel bei verhaltenstherapeutischen und technologiegestützten Interventionstechniken einen verstärkten (Mittelwerte > 5.00), bei psychoanalytischen Techniken einen verminderten (Mittelwerte < 3.00) Einsatz in der kommenden Dekade prognostizierte.

Vergleicht man nun die Mittelwerte und Standardabweichungen jeweils zwischen der ersten und zweiten Delphi-Runde, so bemerkt man bei den beiden verhaltenstherapeutischen und den beiden psychoanalytischen Techniken jeweils eine Reduktion der Standardabweichungen auf unter (oder nahe) $SD = 1.00$, d. h. die Experten näherten sich in ihren Einschätzungen einander an. Bei den technologiegestützten Verfahren trat dies nicht ein. Hinsichtlich der vermuteten Zunahme von Virtual-Reality-Therapie waren sich die Experten in der zweiten Runde uneiniger ($SD = 1.20$) als in der ersten Runde ($SD = 1.07$). Die Autoren interpretieren die Befunde insgesamt als Hinweis darauf, dass Effizienz eine wachsende Rolle im Gesundheitssystem spielt und deswegen Langzeittherapien wie die Psychoanalyse an Bedeutung verlieren, während Kurzzeittherapien sowie Interventionsansätze für zu Hause (Hausaufgaben, computergestützte Therapie) laut Experteneinschätzung mehr Zulauf bekommen werden.

sungsbeiträge aller anderen beteiligten Expertinnen und Experten, so dass jedes einzelne Gruppenmitglied Gelegenheit erhält, die eigenen Beiträge nach Kenntnisnahme der Antworten des gesamten Delphi-Panels gewissermaßen aus einer höheren Warte zu überdenken und ggf. zu korrigieren, so dass Missverständnisse geklärt und widersprechende Aussagen vereinheitlicht werden können.

3. An diese zweite Runde („two-round delphi-survey“) kann sich bei Bedarf eine **dritte Runde** („three-round delphi-survey“) anschließen.
4. Auf der Basis aller Informationen erarbeitet das Forschungsteam schließlich einen umfassenden Lösungsvorschlag für das behandelte Problem. Das mithilfe der Delphi-Methode erarbeitete **Gruppenergebnis** soll einen höheren prognostischen Wert haben als die Einzelurteile der Experten und zudem effizienter sein als andere Prognosemethoden.

Die psychologische Literaturdatenbank PsycInfo verzeichnet für 2010 bis 2013 knapp 150 Delphi-Studien in referierten Fachzeitschriften („delphi“ als Keyword). Delphi-Studien werden jedoch nicht nur in der Psychologie, sondern auch in der Medizin, Pädagogik, in der Wissenschafts-, Wirtschafts-, Technik- und Zukunftsforschung eingesetzt. Sie dienen z. B. dazu, um aus Expertensicht herauszufinden, welche Forschungsfragen in der Palliativmedizin vordringlich sind (Malcolm, Knighting, Forbat, & Kearney, 2009), welches die wichtigsten nicht-akademischen Eigenschaften guter Ärztinnen und Ärzte sind (Lambe & Bristow, 2010), mit welchen Risiken IT-Projekte verbunden sind (Liu, Zhang, Keil, & Chen, 2010) oder welche Typen von Mobbing am Arbeitsplatz sich differenzieren lassen (Rodríguez-Carballeira, Escartín Solanelles, Visauta Vinacua, Porrúa García, & Martín-Peña, 2010). In der **Evaluationsforschung** können Delphi-Studien ein Hilfsmittel sein, um Konsens zwischen Stakeholdern hinsichtlich der Zielsetzungen einer Evaluationsstudie oder der aus ihren Ergebnissen abzuleitenden Maßnahmen herzustellen (Geist, 2010). Mit dem wachsenden Einsatz von Delphi-Methoden mehren sich auch die **Methodenstudien** zu dieser speziellen Befragungstechnik (Akins, Tolson, & Cole, 2005; Häder & Häder, 2009).

Befragungstechniken für größere soziale Gebilde

Normalerweise zielt die Datenerhebung per standardisiertem Fragebogen auf Informationen über die einzelnen Befragungspersonen und ihr individuelles Erleben und Verhalten ab. Zuweilen wird jedoch ein Erkenntnisgewinn über größere soziale Gebilde angestrebt, etwa über **Paare, Kleingruppen, soziale Netzwerke** und **Organisationen**. Hierfür haben sich spezielle standardisierte schriftliche Befragungstechniken etabliert, die es teilweise erfordern bzw. ermöglichen, dass alle Mitglieder desselben sozialen Gebildes befragt und die Antworten der miteinander bekannten Personen dann direkt in Beziehung zueinander gesetzt werden.

Befragung zu Paaren

Interessiert man sich z. B. für die Beziehungsqualität bei Liebespaaren, das Konkurrenzverhältnis bei Geschwisterpaaren oder die Übereinstimmung in Erziehungsfragen bei Elternpaaren, so kann man einerseits die Sichtweisen der Individuen erfassen (z. B. Ehezufriedenheit von Männern und Frauen allgemein) und andererseits im Rahmen einer **Paarbefragung** jeweils Daten mithilfe desselben Fragebogens von beiden Beziehungspartnern erheben und bei der Auswertung paarweise miteinander in Verbindung setzen (z. B. Ehezufriedenheit von miteinander verheirateten Männern und Frauen; für eine paarzentrierte fragebogenbasierte Datenerhebung und Datenanalyse siehe Luo & Klohnen, 2005).

Tabelle 10.12 Vorhergesagte Veränderungen der eingesetzten Psychotherapiemethoden in den kommenden zehn Jahren. Als Antwortskala wurde eine 7-stufige Ratingskala verwendet von 1 (starker Rückgang des Einsatzes der Technik) bis 7 (starke Zunahme des Einsatzes der Technik). Ein Wert von 4 (Mittelkategorie der Skala) bedeutet keine Veränderung im Einsatz der Technik. *M* = Mittelwert, *SD* = Standardabweichung. (Mod. nach Norcross, Hedges, & Prochaska, 2002. Copyright © 2002 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

	Erste Delphi-Runde		Zweite Delphi-Runde	
	M	*SD*	*M*	*SD*
Verhaltenstherapeutische Techniken				
Hausaufgaben	5.25	1.18	5.52	0.94
Rückfallprophylaxe	5.25	1.11	5.44	0.92
Technologiebasierte Techniken				
Einsatz von Virtual Reality	5.42	1.07	5.32	1.20
Computergestützte Therapie	5.50	1.21	5.28	1.16
Psychoanalytische Techniken				
Traumdeutung	2.78	1.18	2.69	0.99
Freies Assoziieren	2.68	1.37	2.54	1.06

Der praktische und wissenschaftliche Nutzen der Paarbefragung ist besonders offensichtlich im Kontext der **Paar- bzw. Eheberatung**. Beispiel: Im Rahmen der an der Universität Göttingen entwickelten Online-Paarberatung theratalk (http://www.theratalk.de) können Paare unabhängig voneinander in standardisierten Online-Fragebögen u. a. ihre sexuellen Wünsche eintragen und auch sehr detailliert angeben, zu welchen sexuellen Praktiken sie selbst bereit oder nicht bereit sind. Das System stellt dann automatisch einen Match der Antworten der Partner her in dem Sinne, dass nur die Wünsche an den eigenen Partner weitergeleitet werden, die dieser laut seines Fragebogens auch bereit wäre zu erfüllen. Die Methode der Online-Paarbefragung mit paarweisem Matching der Antworten und selektiver Weiterleitung nur der erfüllbaren Wünsche umgeht die von vielen Paaren gefürchtete Situation, Wünsche zu äußern, die vom Partner abgelehnt werden und für Konfliktstoff sorgen könnten. Eine interne Evaluationsstudie dieser sog. Ressourcenaktivierungsmethode zeigte, dass sowohl bei Frauen als auch bei Männern mehr als die Hälfte der bislang unerfüllten und verschwiegenen sexuellen Wünsche beim eigenen Partner de facto auf positive Resonanz stoßen (zu Hintergrundinformationen s. http://www.theratalk.de).

Befragung zu Kleingruppen

Eine standardisierte Methode zur Untersuchung der Beziehungen zwischen den Mitgliedern von Kleingruppen stellt die auf den österreichischen Psychiater und Soziologen Jacob Levy Moreno zurückgehende **Soziometrie** („sociometrics“) dar (Moreno, 1934). Im Rahmen einer soziometrischen Befragung (auch soziometrischer Test genannt) **beantworten alle Mitglieder einer Kleingruppe** (z. B. Schulklasse, Arbeitsteam, Kollegium, Sportmannschaft) schriftlich, wie gern oder ungern sie mit allen anderen Gruppenmitgliedern zusammenarbeiten bzw. bestimmte fachliche Aufgaben lösen. Um der Verknüpfung von leistungs- und sachbezogenen (instrumentellen) mit sozio-emotionalen Faktoren in der Gruppendynamik Rechnung zu tragen, können zusätzlich auch Präferenzen hinsichtlich sozialer Aktivitäten abgefragt werden („Mit welchen Mitgliedern des Teams besprechen Sie persönliche Themen besonders gern oder ungern“). Die Auswertung läuft auf eine **Soziomatrix** (Adjazenzmatrix), **soziometrische Kennwerte** sowie ein **grafisches Soziogramm** hinaus, das die Gruppenkonfiguration darstellt (Schwehm, 2008; Terry, 2000). Für statistische und grafische soziometrische Analysen liegen entsprechende Software-Tools vor (z. B. Soziogramm-Editoren; für ein soziometrisches Analysewerkzeug für Schulklassen s. Hrabel, 2010).

Soziometrische Daten sind zur Beantwortung vielfältiger Forschungsfragen geeignet (zum Zusammenhang zwischen soziometrischem Status und Aggression s. Spörrle & Strobel, 2007), haben aber auch einen hohen praktischen Anwendungswert. Soziometrische Befragungen werden im Rahmen der Pädagogik, Organisationsentwicklung, Teamdiagnostik und Teamentwicklung eingesetzt, um den Gruppenmitgliedern die Struktur ihrer Gruppe und ihre Position darin zu verdeut-

„Mit welchen Mitgliedern des Teams arbeiten Sie besonders gern oder ungern zusammen?"

Zusammenarbeit mit A	❒ *ungern*	❒ *neutral*	❒ *gern*
Zusammenarbeit mit B	❒ *ungern*	❒ *neutral*	❒ *gern*
...			
Zusammenarbeit mit F	❒ *ungern*	❒ *neutral*	❒ *gern*

Abb. 10.11 Beispiel eines Fragebogens für eine soziometrische Gruppenbefragung

Studienbeispiel

Soziometrische Gruppenbefragung

Zur Ermittlung des Soziogramms einer 6-köpfigen Arbeitsgruppe wird jedes Mitglied zu allen anderen Mitgliedern schriftlich befragt (Abb. 10.11):

Die wechselseitigen negativen (−1), neutralen (0) und positiven (+1) Präferenzwahlen der $n = 6$ Personen werden in eine **Soziomatrix** eingetragen (Abb. 10.12). Da Personen sich nicht selbst wählen, bleibt die Diagonale leer. Insgesamt enthält die Soziomatrix $\frac{n(n-1)}{2} = 15$ wechselseitige Wahlen aller Mitglieder. Jedes Mitglied gibt $n - 1$ Wahlen ab und kann $n - 1$mal selbst gewählt werden.

Die Auswertung der wechselseitigen Präferenzwahlen erfolgt statistisch über unterschiedliche **soziometrische Kennwerte.** Entsprechende Kennwerte lassen sich für Individuen sowie für die gesamte Gruppe berechnen:

- **Soziometrische Kennwerte für Individuen**. Der **soziometrische Status** einer Person ist definiert als Anzahl der positiven Wahlen durch andere Gruppenmitglieder (positive Spaltensumme). Dementsprechend ist Person A (+4) als Mitglied mit dem höchsten soziometrischen Status anzusehen, gefolgt von D und E (jeweils +3). Person B (−3) dagegen ist mit drei Negativwahlen das Mitglied mit dem geringsten soziometrischen Status. Person F (0) ist als Außenseiter zu betrachten, da hier weder positive noch negative Wahlen erfolgt sind, sondern die anderen Gruppenmitglieder alle gleichgültig bzw. neutral votiert haben.
- **Soziometrische Kennwerte für die Gruppe**. Die **Kohäsion** einer Gruppe ist definiert als Anteil positiver wechselseitiger Wahlen an allen möglichen Wahlen zwischen den Gruppenmitgliedern. Insgesamt sind in einer 6-köpfigen Gruppe 15 wechselseitige Wahlen möglich. Ist keine davon beidseitig positiv so beträgt die Kohäsion 0 %, sind alle beidseitig positiv so beträgt sie 100 %. Im vorliegenden Fall sind 3 von 15 Wahlen beidseitig positiv, dementsprechend beträgt die Kohäsion $3/15 = 20\,\%$ und ist als eher niedrig zu bewerten.

Die Ergebnisdarstellung einer soziometrischen Befragung erfolgt schließlich auch grafisch als **Soziogramm** (Abbildung des Gesamtnetzwerkes der Kleingruppe; Abb. 10.13), wobei jede Person als Kreis dargestellt wird, die Beziehungen der Personen durch ein- oder zweiseitige positive oder negative Verbindungslinien zwischen den Kreisen.

	A wird gewählt	**B**	**C**	**D**	**E**	**F**
A wählt		−1	**1**	**1**	0	0
B	1		−1	0	0	0
C	**1**	−1		1	1	0
D	**1**	−1	0		**1**	0
E	1	0	0	**1**		0
F	0	0	0	0	1	
Zeilensumme	+4	−3	+1/−1	+3	+3	0

Abb. 10.12 Soziomatrix einer 6-köpfigen Gruppe. *fett* drei wechselseitige positive Wahlen

lichen und dann bei Bedarf an einer Verbesserung zu arbeiten (Spörrle, Landes, Otrebski, & Schwehm, 2010). Entsprechende Verfahren sollten von Profis eingesetzt werden, um der Gefahr vorzubeugen, dass sich Gruppenkonflikte oder Außenseiterpositionen durch die Methode eher verschärfen als abbauen. Ohnehin sind soziometrische Ergebnisse immer als Momentaufnahmen zu verstehen, da Gruppen einem permanenten Wandel unterliegen (▶ Studienbeispiel „Soziometrische Gruppenbefragung").

▪ Befragung zu sozialen Netzwerken

Auf der Basis der Soziometrie wurde die **Netzwerkanalyse** („social network analysis", SNA; Analyse sozialer Netzwerke) als Forschungsmethode entwickelt, die darauf abzielt, die Beziehungen zwischen einer festgelegten Zahl von Personen zu untersuchen. Formal werden dabei die Personen als Knoten, ihre paarweisen Beziehungen als Kanten im Netzwerk aufgefasst, so dass zur Analyse z. B. die mathematische Graphentheorie einsetzbar ist. Während die Soziometrie nur Kleingruppen abdeckt,

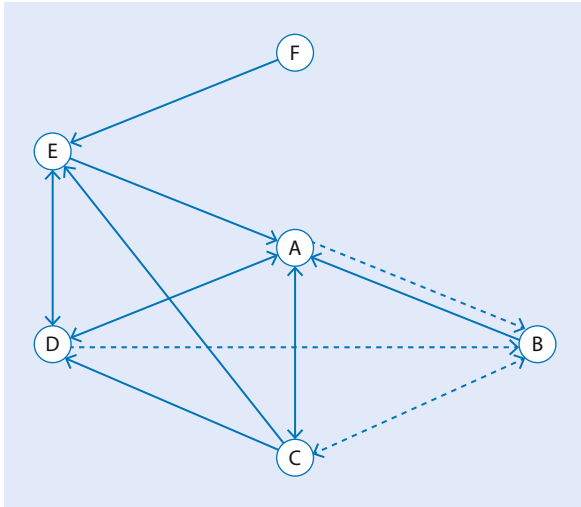


Abb. 10.13 Soziogramm einer 6-köpfigen Gruppe; die *gestrichelten Pfeile* stellen negative Wahlen dar

kann die Netzwerkanalyse prinzipiell auch große bis sehr große Mengen von Personen und deren Beziehungen untereinander einbeziehen. Die Netzwerkanalyse hat seit den 1990er-Jahren einen Boom erlebt und zwar sowohl als innovativer **methodischer Ansatz der Datenerhebung und Datenanalyse** als auch als neues **theoretisches Erklärungsmodell** (Netzwerktheorie) für unsere auch im Alltagsverständnis zunehmend „vernetzte" Welt (zur Einführung in die Netzwerkanalyse siehe Jansen, 2011; Stegbauer, 2010; Trappmann, Hummell, & Sodeur, 2010). Aktuelle Forschung auf diesem Gebiet präsentieren Zeitschriften wie *Cyberpsychology, Behavior and Social Networks; Social Networks; Journal of Social Structure* sowie das *International Network for Social Network Analysis* (INSNA; http://www.insna.org).

Grundsätzlich sind bei der Erhebung und Analyse sozialer Netzwerke zwei Herangehensweisen zu unterscheiden:

1. **Gesamtnetzwerk** („whole network"): Ein Gesamtnetzwerk besteht aus allen Beziehungen zwischen einer abgegrenzten Anzahl von Akteuren. Zur Erhebung eines Gesamtnetzwerkes werden (genau wie in der kleingruppenbezogenen Soziometrie) Daten von allen Akteuren des betreffenden Netzwerkes erhoben. Daten für Gesamtnetzwerkanalysen können nicht nur mithilfe standardisierter Fragebögen gewonnen werden, sondern z. B. auch anhand von Verhaltensspuren wie z. B. **Protokolldateien** (z. B. E-Mail-Verkehr zwischen den Mitgliedern eines Unternehmens; Profilverlinkungen auf einer Social-Networking-Plattform wie Facebook).
2. **Egozentriertes/persönliches Netzwerk** („egocentric network", „personal network"): Ein egozentriertes Netzwerk besteht aus den wichtigen Beziehungen einer Person (Ego, fokale Person) zu anderen Personen (Alteri) sowie deren Beziehungen untereinander. Zur Erhebung von egozentrierten Netzwerken werden nur Daten der jeweiligen Fokalpersonen erfasst und nicht von allen Netzwerkmitgliedern. Für die Erhebung persönlicher Netzwerke sind bestimmte Fragenformate (sog. **Namensgeneratoren** und **Namensinterpretatoren**) notwendig (Wolf, 2010; ▶ Studienbeispiel „Erhebung persönlicher Netzwerke mithilfe standardisierter Fragebögen"). Bei der Konstruktion eines eigenen Fragebogens sollte auf etablierte Instrumente (z. B. das Burt-, Fischer-, oder Wellman-Instrument) zurückgegriffen werden (Hennig, 2006). Über persönliche Netzwerke werden Informationen, praktische Hilfe, emotionale Unterstützung etc. vermittelt, aber auch negative Vorbildwirkungen (z. B. gesundheitsschädigendes Verhalten). Ihre Untersuchung ist zudem interessant im Hinblick auf soziale Schlüsselfiguren wie Meinungsführer und Angehörige von Eliten.

Studienbeispiel

Erhebung persönlicher Netzwerke mithilfe standardisierter Fragebögen

Fragestellung: Angenommen man interessiert sich dafür, wie sich die **persönlichen Netzwerke** von Personen verschiedener Altersgruppen unterscheiden und zwar hinsichtlich ihrer **Größe** (Anzahl der Personen im Netzwerk), ihrer **Zusammensetzung** (z. B. Anteil von Verwandten im Netzwerk, Anteil von Personen mit ähnlichen oder abweichenden politischen Meinungen) und ihrer **Dichte** (Intensität der Beziehungen zwischen den Netzwerkmitgliedern).

Aufbau des Fragebogens: Ein entsprechender standardisierter Fragebogen muss neben soziodemografischen Angaben (vor allem Alter, Parteipräferenz etc.) der befragten Fokalpersonen insbesondere drei Fragenblöcke zur Beschreibung des persönlichen Netzwerkes enthalten (Wolf, 2010): 1.) Fragen zu den konkreten Kontaktpersonen (Namensgeneratoren), 2.) Fragen nach den Eigenschaften der Kontaktpersonen (Namensinterpretatoren), 3.) Fragen nach den Beziehungen zwischen den Kontaktpersonen (Abb. 10.14).

Ergebnisse: Es zeigte sich, dass in höherem Alter die sozialen Netzwerke weniger Kontaktpersonen enthalten, der Verwandtenanteil im Netzwerk bei knapp 50 % liegt und dementsprechend auch die Netzwerkdichte höher ist als bei jüngeren Menschen mit größeren und überwiegend aus Nicht-Verwandten bestehenden Netzwerken (Tab. 10.13).

▪ Befragung zu Organisationen

In der Grundlagen- wie auch der Anwendungsforschung (z. B. Evaluationsforschung) interessiert man sich oft für Struktur- und Prozessmerkmale von Organisationen, et-

1. Fragenblock: Anzahl der Kontaktpersonen

Die meisten Leute besprechen hin und wieder wichtige Angelegenheiten mit anderen Personen. Wenn Sie an die letzten 6 Monate zurückdenken: An wen haben Sie sich gewandt, um Dinge zu besprechen, die Ihnen wichtig waren? Es kann sich dabei sowohl um Verwandte als auch um nicht-verwandte Freunde oder Bekannte handeln, nur nicht um Personen, die mit Ihnen im selben Haushalt wohnen.

Bitte schreiben Sie nun alle Personen auf, an die Sie sich in den letzten 6 Monaten zur Besprechung wichtiger Dinge gewandt haben (Verwenden Sie dabei einfach die Vornamen oder die Anfangsbuchstaben der Namen):

Tanja, Irene, Udo

Anmerkung zur Fragebogenkonstruktion:
***Namengeneratoren** können sich auf unterschiedliche Teile des persönlichen Netzwerkes beziehen (z.B. Personen, mit denen man wichtige Dinge bespricht, Personen, von denen man praktische Hilfe bekommt, Personen, mit denen man Freizeitaktivitäten nachgeht). Interessiert man sich nicht für die Größe, sondern nur für die Zusammensetzung und Dichte des Netzwerkes, so kann die Zahl der genannten Personen festgelegt werden (z.B. "Wir haben jetzt einige Fragen zu den Personen, mit denen Sie häufig privat zusammen sind. Denken Sie bitte einmal an die drei Personen, mit denen Sie am häufigsten privat zusammen sind.")*

2. Fragenblock: Merkmale der Kontaktpersonen

Tragen Sie nun die erste Person aus Ihrer Liste hier ein und beantworten Sie zu dieser Person (wir nennen Sie A) einige Fragen:

Person Tanja (Vorname oder Initialen der ersten Person aus der Liste oben)	
Welches Geschlecht hat A?	❒ männlich ☒ weiblich
Wie alt ist A? (wenn Sie es nicht genau wissen, bitte schätzen)	56 Jahre
Ist A mit Ihnen verwandt?	☒ Mit mir nicht verwandt ❒ Eigenes Kind (Sohn, Tochter) ❒ Bruder, Schwester ❒ Vater, Mutter ❒ Andere verwandte Person (z.B. Onkel, Tante, Nichte, Neffe, Schwager, Schwägerin etc.)
Welche Partei hat A bei der letzten Bundestagswahl gewählt?	❒ CDU/CSU ❒ SPD ❒ FDP ❒ Die Linke ☒ Bündnis 90/Die Grünen ❒ Andere Partei ❒ Noch nicht wahlberechtigt ❒ Hat nicht gewählt ❒ Weiß nicht

Abb. 10.14 Beispielfragebogen für die Erhebung persönlicher Netzwerke (eingetragen sind fiktive Daten)

Tragen Sie nun die zweite Person aus Ihrer Liste hier ein und beantworten Sie zu dieser Person (wir nennen Sie B) einige Fragen:

Person Irene (Vorname oder Initialen der zweiten Person aus der Liste oben)	
Welches Geschlecht hat B?	❐ männlich ☒ weiblich
Wie alt ist B? (wenn Sie es nicht genau wissen, bitte schätzen)	61 Jahre
Ist B mit Ihnen verwandt?	...

Anmerkung zur Fragebogenkonstruktion:
*Zur Kennzeichnung der Kontaktpersonen (**Namensinterpretatoren**) werden unterschiedliche Merkmale herangezogen, neben soziodemografischen Variablen (Geschlecht, Alter, Bildung, Tätigkeit, Nationalität) auch Angaben zu politischen Einstellungen oder zum Lebensstil (z.B. politische Meinungen, Freizeitaktivitäten, Gesundheitsverhalten). Um die Befragungspersonen zu entlasten, kann man die Abfrage der Merkmale auf maximal drei bis fünf der genannten Kontaktpersonen beschränken.*

3. Fragenblock: Beziehungen zwischen den Kontaktpersonen

Abschließend geht es jetzt darum, ob sich die eben beschriebenen Personen, mit denen Sie sich über wichtige Dinge austauschen, untereinander kennen.

A und B ☒ kennen sich nicht ❐ kennen sich wenig ❐ kennen sich gut

A und C ❐ kennen sich nicht ❐ kennen sich wenig ❐ kennen sich gut

B und C ❐ kennen sich nicht ❐ kennen sich wenig ❐ kennen sich gut

Anmerkung zur Fragebogenkonstruktion:
*Der **Grad des Kennens** kann auch grober (kennen, nicht kennen) oder differenzierter (stehen sich sehr nahe, stehen sich nahe, stehen sich nicht nahe, kennen sich überhaupt nicht) abgefragt werden. Um die Befragungspersonen zu entlasten, kann man die Abfrage der Beziehungen zwischen den Kontaktpersonen ebenfalls beschränken. Bei fünf Kontaktpersonen sind bereits 10 wechselseitige Beziehungen zu bewerten.*

Abb. 10.14 (Fortsetzung)

Tabelle 10.13 Ausgewählte Eigenschaften sozialer Netzwerke von Personen unterschiedlicher Altersgruppen (Kecskes & Wolf, 1996, S. 130; zitiert nach Wolf, 2010, S. 479)

Alter der Fokalpersonen	Netzwerkgröße	Netzwerkzusammensetzung	Netzwerkdichte
Jahre	Anzahl der genannten Kontaktpersonen	Anteil der Verwandten im Netzwerk in %	Anteil der Kontakte, die sich kennen, in %
18–34	10.89	31	37
35–49	9.90	40	50
50–64	8.84	47	57
65 und älter	8.02	49	56
Varianzaufklärung ($^* p < .05$)	.20*	.28*	.26*

wa für die Computerausstattung und den PC-Einsatz an Grundschulen, für die Personalfluktuation in Krankenhäusern, den Männer- und Frauenanteil in unterschiedlichen Positionen an Hochschulen oder die Anzahl und soziodemografische Zusammensetzung der Mitglieder von Bürgerinitiativen. Entsprechende Daten sind mittels standardisierter Fragebögen von **zentralen Auskunftspersonen** in den entsprechenden Organisationen zu gewinnen (z. B. Schulleitung, Geschäftsführung).

Inwiefern basiert die Arbeit britischer Nichtregierungsorganisationen („non-governmental organizations" NGO), die sich für landwirtschaftliche Projekte in Entwicklungsländern einsetzen, auf aktuellen Ergebnissen der Landwirtschaftsforschung? Um dieser Fragestellung zum Wissenstransfer zwischen Wissenschaft und Nichtregierungsorganisationen nachzugehen, wurden alle einschlägigen britischen NGO im Internet recherchiert ($N = 78$), ein standardisierter Fragebogen wurde postalisch verschickt und erbrachte einen Rücklauf von $n = 46$ (Morton, Quan, Nelson, & Albright, 2002). Wenn Informationen zu ganzen Organisationen über einzelne Schlüsselpersonen erfragt werden, ist es sinnvoll, die Stellung der Auskunftspersonen in der Organisation (z. B. Mitgliedschaftsdauer, Position) zu erfassen, um mögliche Grenzen ihrer Auskunftsfähigkeit abschätzen zu können. Vorsicht ist auch geboten, wenn man von Auskunftspersonen nicht nur Fakteninformationen für die Organisation abfragt, sondern auch Bewertungen vornehmen lässt, die ggf. stark subjektiv gefärbt sind.

Lernquiz

1. Welche Fehler enthält dieses Item? „Intensive Computernutzung beeinträchtigt Kinder nicht pauschal in ihrer Intelligenzentwicklung, vielmehr kommt es auf die Förderung ihrer Interessen im Elternhaus an." stimmt gar nicht – stimmt wenig – stimmt teils-teils – stimmt ziemlich – stimmt völlig
2. Nennen Sie zwei Vor- und zwei Nachteile selbstadministrierter Fragebögen im Vergleich zu intervieweradministrierten Befragungen!
3. Nennen Sie wichtige Klassifikationskriterien zur Differenzierung wissenschaftlicher Fragebogenstudien!
4. Was versteht man unter T-ACASI?
5. Grenzen Sie die strukturierte Tagebuch-Methode und das Ambulante Assessment voneinander ab!
6. Mit welchen Maßnahmen lässt sich ein neu konstruierter standardisierter Fragebogen vor der Haupterhebung überprüfen und verbessern?
7. Nennen Sie fünf Typen von Funktionsitems! Welche davon sind Wegwerf-Fragen bzw. Wegwerf-Items?

	A wird gewählt	B wird gewählt	C wird gewählt	D wird gewählt
A wählt		**1**	1	0
B wählt	**1**		**–1**	**1**
C wählt	–1	**–1**		**1**
D wählt	0	**1**	**1**	

Abb. 10.15 Soziomatrix für Übungsaufgabe 2

8. Was versteht man unter projektiven Fragetechniken?
9. Differenzieren Sie zwischen Rücklaufquote, Rücklaufkurve und Rücklaufstatistik!
10. Grenzen Sie eine Mixed-Mode-Umfrage von einer Mixed-Methods-Studie ab!

Übungsaufgaben

1. Angenommen, Sie möchten bei einer Schülerbefragung in Köln bei den soziodemografischen Angaben auch „Religion" miterfassen. Wie lautet Ihr Fragebogenitem? Begründen Sie Ihre Wahl!
2. Interpretieren Sie die folgende Soziomatrix (Abb. 10.15) und erstellen Sie das Soziogramm.
3. Angenommen, Sie möchten das Ernährungsverhalten von Schichtarbeiterinnen und Schichtarbeitern erforschen. Entwickeln Sie ein geeignetes standardisiertes Tagebuchinstrument für die Datenerhebung.

Weiterführende Literatur

- Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research*. London: Kogan Page.
- Groves, R. M., Fowler Jr., F. J., Couper, M. P., & Lepkowski, J. M. (2009). *Survey methodology*. Hoboken: Wiley
- Kuckartz, U., Ebert, T., Rädiker, S., & Stefer, C. (2009). *Evaluation Online. Internetgestützte Befragung in der Praxis*. Wiesbaden: VS.
- Porst, R. (2011). *Fragebogen: Ein Arbeitsbuch* (3. Aufl.). Wiesbaden: VS.
- Raab-Steiner, E. & Benesch, M. (2010). *Der Fragebogen: Von der Forschungsidee zur SPSS/PASW- Auswertung* (2. Aufl.). Wiesbaden: VS.
- Schnell, R. (2011). *Survey-Interviews: Methoden standardisierter Befragungen*. Wiesbaden: VS.

Literatur

AAPOR. (2011). *Standard Definitions 7th Edition, 2011 Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* Retrieved February 6, 2013, from http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156

Akins, R. B., Tolson, H., & Cole, B. (2005). Stability of response characteristics of a Delphi panel: Application of bootstrap data expansion. *BMC Medical Research Methodology, 5*(1), 1–12.

Bech, M. & Kristensen, M. B. (2009). Differential response rates in postal and Web-based surveys among older respondents. *Survey Research Methods, 3*(1), 1–6.

Belli, R. F., Stafford, F. P., & Alwin, D. F. (2008). Calendar and time diary methods in life course research. Thousand Oaks: Sage.

Boroughs, M., Cafri, G., & Thompson, J. K. (2005). Male body depilation: Prevalence and associated features of body hair removal. Sex Roles, 52(9–10), 637–644.

Bosnjak, M. (2002). (Non)Response bei Web-Befragungen. Aachen: Shaker.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27,* 281–291.

Bryman, A. (2008). *Social research methods* (3rd ed.). Oxford: Oxford University Press.

Buhl, C., Roth, W. L., & Düx, B. (2007). Selbstmanagement-Entwicklung durch Coaching? Eine Experimentbefragung über die Förderung von Selbstmanagement-Kompetenzen *Organisationsberatung, Supervision, Coaching, 14*(3), 243–255.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2. Aufl.). München: Pearson Studium.

Bühner, M. (2011). *Einführung in die Test– und Fragebogenkonstruktion* (3. Aufl.). München: Pearson Studium.

Cohen, A. A. & Lemish, D. (2003). Real time and recall measures of mobile phone use: Some methodological concerns and empirical applications. *New Media and Society, 5*(2), 167–183.

Cooley, P. C., Miller, H. G., Gribble, J. N., & Turner, C. F. (2000). Automating telephone surveys: Using T-ACASI to obtain data on sensitive topics. *Computers in Human Behavior, 16*(1), 1–11.

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review, 23*(4), 486–501.

Couper, M. P. (2008). *Designing effective web surveys.* Cambridge: Cambridge University Press.

Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quaterly, 75*(5), 889–908.

Dean, D. H. (2008). What's wrong with IVR self-service. *Managing Service Quality, 18*(6), 594–609.

Delorme, D. E., Kreshel, P. J., & Reid, L. N. (2003). Lighting up: Young adults' autobiographical accounts of their first smoking experiences. *Youth and Society, 34*(4), 468–496.

DePaulo, B. M. & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology, 74*(1), 63–79.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Social Science Research, 38*(1), 1–18.

Dixson, B. J., Dixson, A. F., Bishop, P. J., & Parish, A. (2010). Human physique and sexual attractiveness in men and women: A New Zealand-U.S. comparative study. *Archives of Sexual Behaviour, 39*(3), 798–806.

Döring, N. (2003). *Sozialpsychologie des Internet. Die Bedeutung des Internet für Kommunikationsprozesse, Identitäten, soziale Beziehungen und Gruppen.* (2. Aufl.). Göttingen: Hogrefe.

Engle, J. (2003). 'Fear of success revisited': A replication of Matina Horner's study 30 years later. Paper presented at the Annual Meeting of the American Education Association, April 21–25, 2003, Chicago..

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulantes Assessment – Verhalten im Alltagskontext erfassen. Eine verhaltenswissenschaftliche Herausforderung an die Psychologie. *Psychologische Rundschau, 58*(1), 12–23.

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of quality of answers in a web survey. *Public Opinion Quarterly, 73*(2), 349–360.

Geist, M. R. (2010). Using the Delphi method to engage stakeholders: a comparison of two studies. *Evaluation and Program Planning, 33*(2), 147–154.

Gratton, C. & Jones, I. (2004). *Research methods for sport studies.* Milton Park: Routledge.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. A. J. (Eds.). (2002). *Survey nonresponse.* Hoboken: Wiley.

Häder, M. (2006). *Empirische Sozialforschung.* Wiesbaden: VS.

Häder, M. & Häder, S. (Hrsg.). (2009). *Telefonbefragung über das Mobilfunknetz: Konzept, Design und Umsetzung einer Strategie zur Datenerhebung.* Wiesbaden: VS.

Hennig, M. (2006). *Individuen und ihre sozialen Beziehungen.* Wiesbaden: VS.

Horner, M. S. (1972). Toward an understanding of achievement-related conflicts in women. *Journal of Social Issues, 28*(2), 157–175.

Horvath, K. J., Beadnell, B., & Bowen, A. M. (2007). A daily web dairy of sexual experiences of men who have sex with men: Comparisons with a retrospective recall survey. *AIDS and Behavior, 11*(4), 537–548.

Hrabel, V. (2010). *SORAT-M Soziometrische Rating-Methode für die Diagnostik und Planung von Interventionsstrategien bei schwierigen Schulklassen und gefährdeten Schülern an Sekundarschulen. Version für Schulpsychologen.* Göttingen: Hogrefe.

Hussy, W., Schreier, M., & Echterhoff, G. (2009). *Forschungsmethoden in Psychologie und Sozialwissenschaften. Für Bachelor.* Berlin: Springer.

Itsubo, S. & Hato, E. (2006). Effectiveness of household travel survey using GPS-equipped cell phones and web diary: comparative study with paper-based travel survey. Paper presented at the Transportation Research Board Annual Meeting, January 22–26 2006, Washington DC.

Jackob, N., Schoen, H., & Zerback, T. (Hrsg.). (2009). *Sozialforschung im Internet: Methodologie und Praxis der Online-Befragung.* Wiesbaden: VS.

Janda, L. H., Janda, M., & Tedford, E. (2001). IVR test & survey: A computer program to collect data via computerized telephonic applications. *Behavior Research Methods, Instruments and Computers, 33*(4), 513–516.

Jansen, D. (2011). *Einführung in die Netzwerkanalyse: Grundlagen, Methoden, Forschungsbeispiele* (3. Aufl.). Wiesbaden: VS.

Jansen, H. (2010). The logic of qualitative survey research and its position in the field of social research methods. *Forum Qualitative Social Research, 11*(2). Retrieved December 10, 2013, from http://www.qualitative-research.net/index.php/fqs/article/view/1450

Jowett, A. & Peel, E. (2009). Chronic illness in non-heterosexual contexts: An online survey of experiences. *Feminism and Psychology, 19*(4), 454–474.

Kaczmirek, L. (2009). *Human-survey interaction. Usability and nonresponse in online surveys.* Köln: Halem.

Kecskes, R. & Wolf, C. (1996). *Konfession, Religion und soziale Netzwerke.* Wiesbaden: VS.

Kirchmaier, R. (2007). *Indirekte psychologische Methoden.* Wiesbaden: Gabler.

Kuckartz, U., Ebert, T., Rädiker, S., & Stefer, C. (2009). *Evaluation Online. Internetgestützte Befragung in der Praxis*. Wiesbaden: VS.

Lambe, P. & Bristow, D. (2010). What are the most important non-academic attributes of good doctors? A Delphi survey of clinicians. *Medical Teacher, 32*(8), e347–e354.

Lim, M. S. C., Sacks-Davis, R., Aitken, C. K., Hocking, J. S., & Hellard, M. E. (2010). Randomised controlled trial of paper, online and SMS diaries for collecting sexual behaviour information from young people. *Journal of Epidemiology and Community-Health, 64*(10), 885–889.

Linstone, H. A. & Turoff, M. (Eds.). (1975). *The Delphi method*. London: Addison-Wesley.

Liu, S., Zhang, J., Keil, M., & Chen, T. (2010). Comparing senior executive and project manager perceptions of IT project risk: A Chinese Delphi study. *Information System Journal, 20*(4), 319–355.

Lumsden, J. & Morgan, W. (2005). *Online-questionnaire design: Establishing guidelines and evaluating existing support.* Paper presented at the 16th International Information Resources Management Association Conference (IRMA 2005), May 15–18 2005, San Diego, US.

Luo, S. & Klohnen, E. C. (2005). Assortative mating and martial quality in newlyweds: A couple-centered approach. *Journal of Personality and Social Psychology, 88*(2), 304–326.

Malcolm, C., Knighting, K., Forbat, L., & Kearney, N. (2009). Prioritisation of future research topics for children's hospice care by its key stakeholders: A Delphi study. *Palliative Medicine, 23*(5), 398–405.

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys: A field experiment on nonresponse in Web surveys. *Social Science Computer Review, 25*(3), 372–383.

Maxl, E. & Döring, N. (2010). Selbstadministrierte mobile Non-Voice-Marktforschung: Methoden- und Forschungsüberblick. *Transfer – Werbeforschung und Praxis, 56*(1), 22–32.

Maxl, E., Döring, N., & Wallisch, A. (Hrsg.). (2009). *Mobile market research.* Köln: Halem.

Moreno, J. L. (1934). *Who shall survive?* Washington, DC: Nervous and Mental Disease.

Morton, J., Quan, J., Nelson, V., & Albright, K. (2002). Improving communication with UK agricultural and related scientific expertise: UK NGO perspectives. *Science Communication, 23*(4), 442–462.

Muldoon, O. T., Mclaughlin, K., Rougier, N., & Trew, K. (2008). Adolescents' explanations for paramilitary involvement. *Journal of Peace Research, 45*(5), 681–695.

Norcross, J. C., Hedges, M., & Prochaska, J. O. (2002). The face of 2010: A Delphi poll on the future of psychotherapy. *Professional Psychology: Research and Practice, 33*(3), 316–322.

Otterbach, S. & Sousa-Poza, A. (2010). How accurate are German *work-time data*? A comparison of time-diary reports and stylized estimates. *Social Indicators Research, 97*(3), 325–339.

Pentland, W. E., Powell Lawton, M., Harvey, A. S., & McColl, M. A. (1999). *Time use research in the social sciences.* Heidelberg: Springer.

Perrine, M. W., Mundt, J. C., Searles, J. S., & Lester, L. S. (1995). Validation of daily self-reported alcohol consumption using interactive voice response (IVR) technology. *Journal of Studies on Alcohol, 56*(5), 487–490.

Piasecki, T. M., Slutske, W. S., Wood, P. K., & Hunt-Carter, E. E. (2010). Frequency and correlates of diary-measured hangoverlike experiences in a college sample. *Psychology of Addictive Behaviors, 24*(1), 163–169.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2002). Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality and social Psychology Bulletin, 27*(2), 151–161.

Robinson, J. P. (2002). The time-diary method structure and uses. In W. E. Pentland, M. P. Lawton, A. S. Harvey & M. A. McColl (Eds.), *Time Use Research in the Social Sciences* (pp. 47–89). Heidelberg: Springer.

Rodríguez-Carballeira, A., Escartín Solanelles, J., Visauta Vinacua, B., Porrúa García, C., & Martín-Peña, J. (2010). Categorization and hierarchy of workplace bullying strategies: A Delphi survey. *The Spanish Journal of Psychology, 13*(1), 297–308.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton: Princeton University Press.

Schaeffer, N. & Dykema, J. (2011). Questions for surveys: Trends and future direction. *Public Opinion Quaterly 75*(5), 909–961.

Schnell, R. (2011). *Survey-Interviews: Methoden standardisierter Befragungen*. Wiesbaden: VS.

Schwehm, H. (2008). *Soziometrie – Die Methode der Wahl*. Wiesbaden: VS.

Seligman, M., Steen, T., Park, N., & Petersen, C. (2005). Positive psychology progress. *American Psychologist, 60*(5), 410–421.

Spörrle, M., Landes, M., Otrebski, D. A., & Schwehm, H. (2010). Zeig uns, wie wir zueinander stehen: Soziometrische Verfahren und assoziierte visuellskalierende Techniken im berufsbezogenen Anwendungskonzept. *Zeitschrift für Psychodrama und Soziometrie, 9*, 175–188.

Spörrle, M. & Strobel, M. (2007). Zum Zusammenhang zwischen soziometrischem Status und Aggression: Eine Bestandsaufnahme der aktuellen Forschungslage *Zeitschrift für Psychodrama und Soziometrie, 6*, 43–66.

Stegbauer, C. (2010). *Netzwerkanalyse und Netzwerktheorie: Ein neues Paradigma in den Sozialwissenschaften* (2. Aufl.). Wiesbaden: VS.

Stritzke, W. G. K., Dandy, J., Durkin, K., & Houghton, S. (2005). Use of interactive voice response (IVR) technology in health research with children. *Behavior Research Methods, 37*(1), 119–126.

Terry, R. (2000). Recent advances in measurement theory and the use of sociometric techniques. In A. H. N. Cillessen & W. M. Bukowski (Eds.), *Recent advances in the measurement of acceptance and rejection in the peer system*. San Francisco: Jossey-Bass.

Toerien, M. & Wilkinson, S. (2004). Exploring the depilation norm: A qualitative questionnaire study of women's body hair removal. *Qualitative Research in Psychology, 1*(1), 69–92.

Torres, L. & Ong, A. D. (2010). A daily diary investigation of latino ethnic identity, discrimination, and depression. *Cultural Diversity and Ethnic Minority Psychology, 16*(4), 561–568.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Trappmann, M., Hummell, H. J., & Sodeur, W. (2010). *Strukturanalyse sozialer Netzwerke: Konzepte, Modelle, Methoden* (2. Aufl.). Wiesbaden: VS.

Wilhelm, P. & Perrez, M. (2008). Ambulantes Assessment in der klinischen Psychologie und Psychiatrie. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 56*(3), 169–179.

Wolf, C. (2010). Egozentrierte Netzwerke: Datenerhebung und Datenanalyse. In C. Stegbauer & R. Häußling (Hrsg.), *Handbuch Netzwerkforschung* (S. 471–484). Wiesbaden: VS.

Wutich, A. (2009). Estimating household water use: A comparison of diary, prompted recall, and free recall methods. *Field Methods, 21*(1), 49–68.

10.4 Psychologischer Test

Lernziele

- Wissen, was wissenschaftliche psychologische Tests von nicht-wissenschaftlichen „Psychotests“ unterscheidet.
- Verschiedene Arten projektiver und psychometrischer psychologischer Tests im Hinblick auf Zielsetzung und Anwendungsfelder voneinander abgrenzen können.
- Die Testgütekriterien Objektivität, Reliabilität und Validität mit ihren jeweiligen Unterformen erklären können.
- Das Prinzip der Itemanalyse gemäß Klassischer Testtheorie verstehen.
- Die Klassische Testtheorie und die Probabilistische Testtheorie in ihren Grundannahmen skizzieren und miteinander vergleichen können.

Der **Intelligenztest** ist einer der bekanntesten psychologischen Tests. Er eignet sich deswegen gut, um die Besonderheiten psychologischer Testverfahren als Datenerhebungsmethode zu verdeutlichen: „Intelligenz“ – allgemein definiert als geistige Leistungsfähigkeit – ist ein Merkmal, dessen exakte Ausprägung man einer Person nicht direkt ansehen kann und das sich im Alltag auch schwer über Verhaltensbeobachtungen erschließen lässt. Ebenso sind Selbstauskünfte per Interview oder Fragebogen unbefriedigend, denn wir können nicht davon ausgehen, dass Menschen ihre Intelligenz allgemein sowie deren Subdimensionen (z. B. logisches Denken, räumliches Denken, numerisches Denken etc.) zutreffend und präzise in Relation zu anderen Personen einschätzen können. Ein **psychologischer Test** übernimmt diese Aufgabe: Er besteht aus einer Batterie von empirisch geprüften, meist vollstandardisierten Testaufgaben (d. h. Aufgaben und Antwortformate sind genau festgelegt), die von den Testpersonen – teilweise unter Zeitdruck – bearbeitet werden. Bei einem **eindimensionalen Test** werden alle Testaufgaben zu einem einzigen Testwert zusammengefasst, der die Ausprägung des interessierenden Konstruktes (z. B. allgemeine Intelligenz) erfasst. Ein **mehrdimensionaler Test** dagegen misst verschiedene Unterdimensionen des Konstruktes (z. B. sprachliches, numerisches und räumliches Denken) und läuft auf die Bestimmung mehrerer Subtest-Werte hinaus.

Während die Bezeichnung **psychologischer** Test auf die Eigenheit des gemessenen Merkmals (psychologische Eigenschaften von Personen) verweist, bezieht sich der ebenfalls verwendete Begriff **psychometrischer** Test auf die methodische Fundierung quantitativer bzw. strukturierter Testverfahren durch eine statistische Testtheorie (▶ Abschn. 10.4.4).

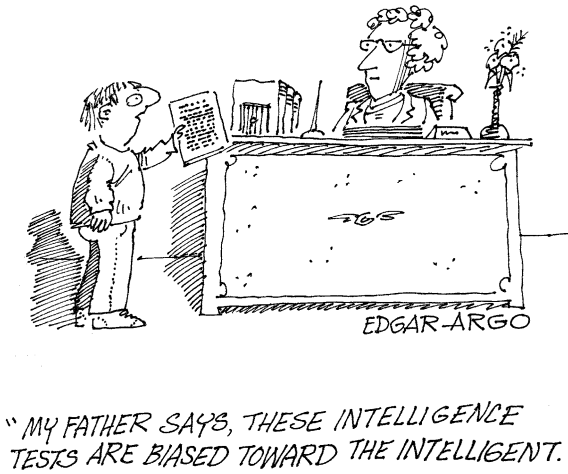


Abb. 10.16 Psychologische Leistungstests sollen die Leistungsfähigkeit der Testpersonen messen, wobei jedoch gemäß Testfairness einzelne Bevölkerungsgruppen nicht benachteiligt werden dürfen. (© Edgar Argo/Search ID: ear0868, Rights Available from CartoonStock.com)

Während standardisierte Fragebögen als Forschungsinstrumente dienen und meist nur anonym auf Aggregatebene ausgewertet werden (z. B. Gruppenmittelwerte), kommen psychologische Tests sehr oft im Rahmen der **Diagnostik** („psychological assessment“) zum Einsatz, wo es um die Beurteilung der Merkmale einzelner identifizierbarer Person geht: Psychologische Tests werden z. B. – zusammen mit anderen Verfahren – eingesetzt, um die Eignung einer konkreten Person für eine berufliche Position festzustellen, über die Einschulung eines Kindes zu entscheiden, die Notwendigkeit der Verlängerung einer Psychotherapie festzustellen, über die Rückfallgefahr bei einer möglichen Entlassung aus dem Strafvollzug zu befinden oder die Voraussetzungen zum Wiedererwerb einer wegen Alkohol- oder Drogenmissbrauchs verlorenen Fahrerlaubnis zu prüfen. Psychologische Tests, die im Rahmen der Diagnostik eingesetzt werden, können also sehr weitreichende Folgen für die Testpersonen haben (sog. **„high-stakes testing“**). Dementsprechend spielen auch Fragen der **Testethik, Testfairness** und **Testverfälschung** eine wichtige Rolle im Kontext psychologischer Testverfahren (Abb. 10.16).

In diesem Kapitel wird zunächst erläutert, was unter **wissenschaftlichen Tests** zu verstehen ist und welche Rolle diese im Rahmen der sozial- und humanwissenschaftlichen Forschung sowie der psychologischen Diagnostik spielen (▶ Abschn. 10.4.1). Anschließend werden ausgewählte **projektive Testverfahren** vorgestellt (▶ Abschn. 10.4.2) sowie v. a. **psychometrische Tests** präsentiert (▶ Abschn. 10.4.3). Diese Ausführungen sollen dabei helfen, vorhandene Studien, die auf Testwerten

basieren, besser zu verstehen. Zudem sollen sie Forschende dabei unterstützen, für eigene Studien passende Testverfahren auszuwählen und sachgerecht einzusetzen. Für all diejenigen, die selbst psychometrische Testverfahren entwickeln oder einfach nur die methodischen Grundlagen psychologischer Tests besser verstehen möchten, wird auch die Testtheorie behandelt (▶ Abschn. 10.4.4). Und zwar zunächst die sog. **Klassische Testtheorie**, die die Grundlage der meisten etablierten psychologischen Testverfahren darstellt. Abschließend widmen wir uns der **Probabilistischen Testtheorie** bzw. **Item-Response-Theorie**.

10.4.1 Wissenschaftliche psychologische Tests

Wie sind psychologische Tests definiert und für welche wissenschaftlichen und praktischen Problemstellungen eignen sie sich? Was unterscheidet psychologische Tests als wissenschaftliche Methoden der Datenerhebung von Fragebögen sowie von nicht-wissenschaftlichen „Psychotests"? Welche Typen von psychologischen Tests sind zu unterscheiden? Was ist zu beachten, damit der Einsatz psychologischer Tests ethischen Standards entspricht und zudem Verzerrungen und Verfälschungen von Testergebnissen im Zuge der Testdurchführung vermieden werden? Die folgenden Abschnitte geben Antworten auf diese Fragen.

Definition und Indikation psychologischer Tests

Der Begriff „Test" hat im alltäglichen und im wissenschaftlichen Sprachgebrauch mehrere Bedeutungen und bezeichnet (Lienert & Raatz, 1998, S. 1):

1. ein Verfahren zur Untersuchung eines Persönlichkeitsmerkmals,
2. den Vorgang der Durchführung einer Untersuchung,
3. die Gesamtheit der zur Durchführung notwendigen Requisiten,
4. jede Untersuchung, sofern sie Stichprobencharakter hat,
5. gewisse mathematisch-statistische Prüfverfahren (z. B. *t*-Test).

Mit einem **psychologischen Test** ist die erstgenannte Bedeutung angesprochen (Bühner, 2011; Lienert & Raatz, 1998; Rost, 2004).

Psychologischer Test – Ein psychologischer Test („psychological test") ist ein wissenschaftliches Datenerhebungsverfahren, das aus mehreren Testaufgaben **(Testbogen/Testmaterial)** sowie festgelegten Regeln zu deren Anwendung und Auswertung **(Testmanual)** besteht. Ziel eines psychologischen Tests ist es, ein latentes psychologisches Merkmal (Konstrukt) – typischerweise eine **Fähigkeit** oder **Persönlichkeitseigenschaft** – in seiner absoluten oder relativen Ausprägung zu Forschungszwecken oder für praktische Entscheidungen zu erfassen.

Theoretische und empirische Fundierung von Tests – Grundlage psychologischer Tests sind immer sowohl **inhaltsbezogene Theorien** (Theorien über das zu erfassende Konstrukt; z. B. Theorie der Intelligenz für einen Intelligenztest) als auch **methodenbezogene Theorien** (sog. Testtheorien, die Aussagen über Aufbau und Eigenschaften von psychologischen Tests und Testaufgaben treffen; z. B. sog. Klassische oder Probabilistische Testtheorie) sowie **empirische Belege**, die diese Theorien stützen.

Testpersonen und Testanwender – Im Kontext der Datenerhebung mittels psychologischer Tests sind die untersuchten Testpersonen („test taker") und die untersuchenden Testanwender („test user") zu unterscheiden.

Dass psychologische Tests theoretisch und empirisch fundiert sein sollen, bedeutet folgendes: Zunächst muss der Test als eine spezifische Form der Operationalisierung des interessierenden Merkmals auf einer **theoretisch fundierten Konzeptspezifikation** basieren (▶ Abschn. 8.2). Neben inhaltlichen Theorien und Befunden zu dem zu messenden Konstrukt, ist auf methodologischer Ebene eine Fundierung des Verfahrens durch eine **Testtheorie** sowie eine empirische Überprüfung der Testgütekriterien notwendig (▶ Abschn. 10.4.4). Psychologische Tests gehören zu den anspruchsvollsten und am gründlichsten überprüften Formen der Operationalisierung von Persönlichkeitseigenschaften und Fähigkeiten des Menschen. Sie kommen seit dem Ende des 19. Jahrhunderts zum Einsatz und erfreuen sich wachsender Beliebtheit. Die Frage, was im Einzelnen gute Testqualität ausmacht und wie man sie am besten sicherzustellt, ist seit Jahrzehnten Gegenstand methodologischer Überlegungen und wird bis heute kontrovers diskutiert (Newton & Shaw, 2013).

Psychologische Testverfahren werden häufig als **Paper-Pencil-Tests** durchgeführt, d. h., die Testaufgaben werden von den Testpersonen mit Stift auf dem Papier beantwortet in Anwesenheit der Testanwender. Alternativ werden Tests heute statt auf Papier verstärkt auch elektronisch (offline oder online) durchgeführt. **Computer- sowie Internet-basierte Tests** weisen spezifische Vor- und Nachteile für Testpersonen und Testanwender auf (Naglieri, et al., 2004). Neben Paper-Pencil- oder elektronischen Tests, bei denen Aufgaben schriftlich zu bearbeiten sind, existieren auch Testverfahren, die ergänzend auf andere Formen der Datenerhebung zurückgreifen und z. B. **persönliche Interviews** oder **physiologische Messungen** einbeziehen. Manche Testverfahren quantifizieren die zu erfassenden Leistungen oder Eigenschaftsausprägungen auch durch Reaktionszeiten. **Reaktionszeit-Tests** kommen z. B. in der allgemeinen Psychologie (Wahrnehmungsforschung), Neuropsychologie (z. B. Erfassung von exekutiver Funktionen) sowie in der Sozialpsychologie (z. B. Stereotypenforschung mithilfe des

Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) vor.

Psychometrische Persönlichkeitstests weisen auf den ersten Blick eine große **Ähnlichkeit mit vollstandardisierten Fragebögen** auf (▶ Abschn. 10.3.3). Tatsächlich werden standardisierte Paper-Pencil-Instrumente, die Persönlichkeitseigenschaften mittels Selbstauskunfts-Items erfassen (z. B. „Ich fühle mich oft traurig und niedergeschlagen" – stimmt gar nicht – wenig – ziemlich – völlig), manchmal synonym sowohl als **Persönlichkeitstest** als auch als **Persönlichkeitsfragebogen** oder **Persönlichkeitsinventar** bezeichnet. Nicht nur im Aufbau und Aussehen, sondern auch hinsichtlich der wissenschaftlichen Konstruktionsprinzipien ähneln sich beide Verfahren: Psychometrische Tests und Fragebögen werden beide in der Regel auf der Basis der Klassischen Testtheorie entwickelt und stellen meist Likert-Skalen dar (▶ Abschn. 8.6.1). Folgende **Unterschiede** zwischen psychometrischen Tests und Fragebögen sind jedoch zu beachten:

- Im Unterschied zum standardisierten Forschungsfragebogen (▶ Abschn. 10.3.3) zeichnen sich Persönlichkeitstests v. a. durch ihre **Normierung** aus, die es ermöglicht, individuelle Messwerte differenziert zu beurteilen und im Rahmen der **Diagnostik** zur Beurteilung einer Einzelperson einzusetzen. Psychometrische Fragebögen – für die keine Normwerte vorliegen – sind dagegen nur für Gruppenaussagen zu Forschungszwecken geeignet, nicht für diagnostische Zwecke.
- Generell werden psychologische Persönlichkeitstests als etablierte Verfahren einer **sehr viel gründlicheren Überprüfung der Testgütekriterien** (insbesondere auch der Validität) unterzogen als Fragebögen, die im Zuge von Forschungsprojekten viel häufiger selbst konstruiert und nur einmalig eingesetzt werden.
- Des Weiteren zählen zu den psychologischen Tests nicht nur die Persönlichkeitstests auf der Basis von Selbstauskünften, sondern auch die **Leistungstests**, bei denen die Testpersonen anhand von Testaufgaben (z. B. Denk- oder Rechenaufgaben) ihre Leistungsfähigkeit durch entsprechende Performanz direkt unter Beweis stellen müssen. Standardisierte Verfahren zur Messung von Fähigkeiten, die aus **richtig oder falsch zu lösenden Testaufgaben** bestehen, werden nie als Fragebögen, sondern immer als Tests bezeichnet.
- Psychologische Testverfahren konzentrieren sich inhaltlich v. a. auf Fähigkeiten und Persönlichkeitseigenschaften – das sind die Merkmalsbereiche, die in der Individualdiagnostik die wichtigste Rolle spielen. Forschungsfragebögen widmen sich dagegen so gut wie allen **vorstellbaren Themen**. Neben latenten Merkmalen wie Eigenschaften oder Fähigkeiten thematisieren Fragebögen oftmals auch Lebensereignisse (z. B. Daten einer Berufskarriere), Verhaltensweisen (z. B. Fernsehgewohnheiten, Freizeitaktivitäten, Energienutzung) oder andere Sachverhalte (z. B. politische Einstellungen, Urlaubspräferenzen), die dann jedoch oft nicht über Skalen, sondern über Einzel-Indikatoren bzw. Single-Item-Skalen erfragt werden (▶ Abschn. 8.5).
- Nicht zuletzt ist zu beachten, dass psychologische Tests – im Unterschied zu Forschungsfragebögen – aufgrund ihres Einsatzes in der Diagnostik sehr viel strengeren **ethischen Richtlinien** unterliegen und aus Gründen des **Testschutzes** oft nur von autorisierten Personen (Studium der Psychologie oder Medizin) genutzt und auch nicht frei veröffentlicht werden dürfen.

Die große Bedeutung psychologischer Tests im Rahmen der **psychologischen Diagnostik** haben wir bereits betont. Unter psychologischer Diagnostik (Jäger & Petermann, 1995; S. 11) versteht man das systematische Sammeln und Aufbereiten von Informationen mit dem Ziel, individuumsbezogene Entscheidungen und daraus resultierende Handlungen zu begründen, zu kontrollieren und zu optimieren. Inhaltlich geht es dabei im Zusammenhang mit Fähigkeitsdiagnostik oft um **Personalselektion** im Bildungsbereich, beim Militär oder im Erwerbsleben sowie im Zusammenhang mit Persönlichkeitseigenschaften oft um **psychologische Beratung** und **Psychotherapie**, **Psychiatrie** oder **Forensik**. Dabei wird im Prozess des Diagnostizierens auf Regeln, Anleitungen und Algorithmen zurückgegriffen. Die so gewonnenen Daten über die Merkmalsträger werden zu einem Urteil integriert, das meist in einem **Gutachten** festgehalten wird und der Diagnose und Prognose dient. Als Merkmalsträger kommen je nach Anwendungsbereich der Diagnostik neben Einzelpersonen zuweilen auch Personengruppen, Institutionen, Situationen und Gegenstände in Betracht (Amelang & Schmidt-Atzert, 2006). Die Anwendungsbereiche der Diagnostik sind vielfältig: Sie reichen von der Feststellung der Glaubwürdigkeit von Zeugen in der forensischen Psychologie, über Schulempfehlungen für Kinder bis hin zur Diagnose von Angststörungen in der klinischen Psychologie. Die psychologische Diagnostik bedient sich dabei verschiedenster Methoden, wobei Paper-Pencil-Testverfahren den größten Raum einnehmen (Kubinger, 1995), Internet-basierte Tests aber an Zulauf gewinnen.

Normierung von Tests – Kennzeichnend für einen psychometrischen Test – im Unterschied zum standardisierten Forschungsfragebogen – ist insbesondere seine **Normierung** bzw. **Eichung** („standardization"), d. h. die Bereitstellung von Normwerten (Testergebnisse repräsentativer Bevölkerungsgruppen). Erst durch die Normierung sind Testwerte nicht nur für wissenschaftliche Aussagen auf Aggregatebene, sondern auch für die

Hilfen bei der Auswahl passender psychologischer Tests

Eine Reihe von Testinformationssystemen stehen zur Verfügung, um einen inhaltlich passenden Test mit hoher Testgüte für eine eigene Studie zu finden (mod. nach Bühner, 2011, S. 78):

- Wie bei der Literatursuche (► Abschn. 6.1) kann auch bei der Recherche nach psychologischen Tests auf **elektronische Datenbanken** zurückgegriffen werden. Das Zentrum für Psychologische Information und Dokumentation **ZPID** ist die überregionale Dokumentations- und Informationseinrichtung für das Fach Psychologie in den deutschsprachigen Ländern. Das ZPID liefert über sein elektronisches Testarchiv umfangreiche Informationen zu Testverfahren, darunter auch frei verfügbare Persönlichkeitstests (http://www.zpid.de). Die APA als international führende amerikanische Fachgesellschaft für Psychologie bietet seit 2012 die APA-Testdatenbank PsycTEST an, die kostenpflichtig online zugänglich ist (http://www.apa.org/pubs/databases).
- Hochschulbibliotheken betreiben teilweise eigene **Testotheken** bzw. **Testbibliotheken**, d. h. Sammlungen von Tests und Testmaterialien, die von berechtigten Personen eingesehen oder entliehen werden können (eine Liste entsprechender Einrichtungen im deutschsprachigen Raum bietet das ZPID unter http://www.zpid.de).
- **Psychologische Fachverlage**, die Testverfahren vermarkten (z. B. Testbögen, Manuale, Auswertungssoftware etc.), bieten Testbeschreibungen in ihren Katalogen und auf ihren Websites. Im deutschsprachigen Raum führend ist der Hogrefe-Verlag mit seiner **Online-Testzentrale** (http://www.testzentrale.de). Die Testdarstellung folgt hier jedoch auch Marketing-Gesichtspunkten.
- Testverfahren werden in Buchform als **Testkompendien** dargestellt. Das Testkompendium „Brickenkamp" – benannt nach seinem ursprünglichen Herausgeber Rolf Brickenkamp – bietet eine umfassende Sammlung von Beschreibungen psychologischer und pädagogischer Tests (Brähler, Holling, Leutner, & Petermann, 2002a, 2002b). Zu wirtschaftspsychologischen Testverfahren liegen ebenfalls einschlägige Kompendien vor (Sarges, Wottawa, & Roos, 2010), wobei hier nicht nur Merkmale von Individuen, sondern auch von Teams oder Organisationen gemessen werden. Kompendien zu neuro-psychologischen, klinisch-psychologischen oder auch – in der Sozialforschung breit anwendbaren – gemeindepsychologischen Skalen und Testverfahren (Johnson, 2009) sind verfügbar. Die Buchreihe „Tests unter der Lupe" (aktueller Band: Fay, 2005) bietet ausführliche Testrezensionen unterschiedlicher, mehr oder minder bekannter psychologischer Testverfahren.
- Das **Testkuratorium** der Föderation deutscher Psychologenvereinigungen bemüht sich als Fachgremium um die Qualitätssicherung psychologischer Testverfahren (http://www.zpid.de/Testkuratorium). Das Testkuratorium (TK) hat ein eigenes Testbeurteilungssystem (TBS) entwickelt, das seit einigen Jahren eingesetzt wird. Dieses standardisierte Testbeurteilungssystem des Testkuratoriums (**TBS-TK**) umfasst drei Stufen (Testkuratorium, 2010):
  - **Stufe 1:** Prüfung der Informationsgrundlage gemäß den Anforderungen nach DIN 33430 (DIN-Screen Checkliste 1; Kersting, 2008; Westhoff, et al., 2010).
  - **Stufe 2:** Testkategorisierung nach dem ZPID-System und Teilen des EFPA (European Federation of Psychologists' Associations)-Systems (http://www.efpa.eu/professional-development/assessment)
  - **Stufe 3:** Testbeurteilung und Testrezension durch zwei unabhängige Fachleute auf Grundlage der standardisierten Beurteilungsrichtlinie.

Bekannte Testverfahren wie z. B. das Beck Depressionsinventar, der Intelligenz-Struktur-Test oder das NEO-Persönlichkeitsinventar wurden bereits einer entsprechenden systematischen Testbeurteilung unterzogen und die Ergebnisse publiziert. Informationen sind der Website des Testkuratoriums zu entnehmen (http://www.zpid.de/Testkuratorium).

Beurteilung von Einzelpersonen im Rahmen der psychologischen Diagnostik („psychological assessment") nutzbar.

Normorientiertes und kriteriumsorientiertes Testen – Wird ein individueller Testwert mit empirischen Normwerten verglichen und daraus das Testergebnis abgeleitet (z. B. über- oder unterdurchschnittliche Intelligenz einer Person in Relation zu ihrer Bezugsgruppe), so spricht man von **normorientiertem Testen** („norm-oriented testing"), hierbei geht es um interindividuelle Vergleiche. Wird ein individueller Testwert mit einem vorab definierten Kriterium oder Schwellenwert verglichen (z. B. Festlegung von Mindestpunktzahl 50 zum Bestehen eines Schulleistungstests), so handelt es sich um **kriteriumsorientiertes Testen** („criteria-oriented testing").

Während in der Diagnostik die psychologischen Tests einer praktischen Anwendung unterliegen und die betrachteten Einzelfälle im Fokus stehen, dienen psychologische Tests im Rahmen der empirischen Sozial- und Humanforschung dem allgemeinen Erkenntnisgewinn anhand von stichprobenbezogenen Kennwerten (z. B. Mittelwerte oder Korrelationen). Empirische Intelligenz-, Persönlichkeits-, Entwicklungs-, Bildungs- und Aggressionsforschung wären ohne entsprechende Testverfahren kaum möglich (für Empfehlungen zur Recherche nach inhaltlich passenden Tests mit hoher Testgüte ► Hilfen bei der Auswahl passender psychologischer Tests).

Ein psychologischer Test ist immer dann für Forschungszwecke besonders geeignet (**Indikation**) wenn

- zentrale und etablierte psychologische Merkmale (z. B. Intelligenz, Extraversion, Depression) mit einem besonders **gut geprüften Verfahren** erfasst werden sollen,
- Ergebnisse einer Studie mit Ergebnissen anderer Studien oder einer Referenzstichprobe **vergleichbar** sein sollen,
- **leistungsbezogene Merkmale** objektiv – d. h. unabhängig von subjektiven Selbsteinschätzungen der

Testpersonen – erfasst werden sollen (z. B. Intelligenz, Berufseignung, Schuleignung).

Der Einsatz psychologischer Tests ist jedoch auch mit **Nachteilen und Einschränkungen** verbunden:

- Um ein Merkmal mit einem psychologischen Test zu messen, müssen von den Testpersonen in der Regel mehr oder minder **umfangreiche Item-Batterien** für jedes einzelne zu messende Konstrukt abgearbeitet werden. Dies verlängert Untersuchungsinstrumente und nimmt viel Zeit in Anspruch – mangelnde Akzeptanz auf Seiten der Untersuchungsteilnehmenden kann die Folge sein.
- Insbesondere Leistungstests müssen **unter Aufsicht** durchgeführt werden (z. B. Kontrolle des Zeitlimits), was die Einsatzmöglichkeiten einschränkt (z. B. keine Möglichkeit einer Leistungstest-Durchführung im Rahmen einer postalischen Befragung, während eine Online-Administration möglich ist).
- Psychologische Testverfahren unterliegen dem **Testschutz**, d. h. Intelligenztestaufgaben, die z. B. zu Zwecken der Personalauswahl verwendet werden, sollen nicht öffentlich in Umlauf kommen.
- Der Einsatz vieler psychologischer Testverfahren ist nicht zuletzt **kostenpflichtig**, d. h. es fallen Kosten für jede einzelne Testperson an (z. B. Bezahlung der einzelnen Testbögen, die nicht frei kopiert werden dürfen), was rasch das Budget eines Forschungsprojekts übersteigen kann.

Abgrenzung zwischen wissenschaftlichen psychologischen Tests und nicht-wissenschaftlichen Tests

„Wie kompromissbereit sind Sie?“ (Elle), „Wie temperamentvoll bist du?“ (Bravo), „Sind Sie eine gute Freundin?“ (Amica), „Wie zufrieden sind Sie mit ihrem Körper?“ (Glamour), „Welcher Wohn-Typ sind Sie?“ (Men’s Health). In Zeitschriften und auf Websites existieren Hunderte von sog. **Psychotests**, die anhand sehr weniger, szenarisch ausgeschmückter Fragen eine Einschätzung von Persönlichkeitsmerkmalen bzw. eine Zuordnung zu einem bestimmten Persönlichkeitstyp versprechen. Der genannte Test zur Kompromissbereitschaft bietet im Ergebnis z. B. die Einteilung in einen von vier qualitativen Typen: die Ausgeglichene, die Diplomatische, die Kritikerin und die Ängstliche (◘ Tab. 10.14). Psychotests in Zeitschriften werden nicht nach wissenschaftlichen Kriterien, sondern auf der Basis des **gesunden Menschenverstandes** konstruiert und dienen in der Praxis in erster Linie der **Unterhaltung**. Klischeehafte und konfrontative Fragen und Ergebnisse können deswegen sogar gewünscht sein, um dem Publikum Diskussionsstoff zu bieten.

Neben den sog. Psychotests als nicht-wissenschaftlicher Variante von Persönlichkeitstests kennen wir alle auch nicht-wissenschaftliche Varianten von Leistungstests, etwa **Klausuren**, **Tests** und **Prüfungen im Bildungsbereich**, die von Lehrenden ad hoc erstellt werden und keinen wissenschaftlichen Konstruktionsprozess durchlaufen. Hier handelt es sich meist um ein kriteriumsorientiertes Testen, d. h. es wird vorab anhand theoretischer Kriterien festgelegt, wie viele Punkte für ein Bestehen des Tests bzw. für eine bestimmte Note erzielt werden müssen.

Klassifikationskriterien für wissenschaftliche Tests

Ein einheitliches Klassifikationssystem psychologischer Testverfahren hat sich bislang nicht etabliert. Allerdings lassen sich vier Beurteilungsdimensionen angeben, die in der Literatur oft verwendet werden (Brähler et al., 2002a, S. XIIff.).

1. **Grad der Standardisierung des Testverfahrens**: Nicht- oder teilstrukturierte (sozusagen qualitative) projektive Tests („projective tests“) stehen den vollstrukturierten (sozusagen quantitativen) psychometrischen Tests („psychometric tests“) gegenüber. Bei den meisten psychologischen Tests handelt es sich um **psychometrische Tests**, bei denen die Aufgaben und Antwortmöglichkeiten genau vorgegeben sind und die Auswertung im Sinne einer Punktvergabe bzw. Messung (▶ Abschn. 8.4) exakt festgelegt ist. **Projektive Tests** dagegen arbeiten mit unstrukturiertem Stimulusmaterial und offenen Aufgaben und werden oft interpretativ ausgewertet.
2. **Merkmale der Testaufgaben**: Bei psychometrischen Tests kommen definitionsgemäß immer standardisierte Items (Frage/Aussage/Aufgabe plus festes Antwortformat) zum Einsatz; für eine Subdifferenzierung psychometrischer Tests ist die Art der Testaufgaben deswegen nicht geeignet. Demgegenüber werden bei projektiven Tests sehr unterschiedliche Aufgaben und Stimulus-Materialien genutzt. **Innerhalb der Gruppe der projektiven Tests** ist deswegen eine Klassifikation der Verfahren nach Merkmalen der Testaufgaben sinnvoll, wobei man v. a. drei Varianten von Testaufgaben unterscheidet: **Formdeuteverfahren**, **verbal-thematische Verfahren** sowie **zeichnerische und Gestaltungsverfahren** (Brähler et al., 2002a, S. XII).
3. **Art des zu messenden Merkmals**: Für Subklassifikationen **innerhalb der sehr großen Gruppe der psychometrischen Tests** ist die Art des zu messenden Merkmals ausschlaggebend. Dabei unterscheidet man Tests, bei denen die Testpersonen ihre Leistungsfähigkeit unter Beweis stellen müssen (**Leistungs-**

Tabelle 10.14 Exemplarische Gegenüberstellung von wissenschaftlichem psychologischen Test und nicht-wissenschaftlichem „Psychotest“ (Beispiele aus NEO-FFI und „Wie kompromissbereit sind Sie?“ aus der Zeitschrift „Elle“; http://www.elle.de/Typentest/Wie-kompromissbereit-sind-Sie-_1202972.html [26.11.2013])

	Wissenschaftlicher psychologischer Test	Nicht-wissenschaftlicher „Psychotest“
Testentwicklung	Erfolgt theoretisch und empirisch fundiert durch Psychologinnen und Psychologen	Erfolgt intuitiv außerhalb psychologischer Forschungsprozesse durch Testautoren
Testgütekriterien	Sind bekannt und publiziert	Sind nicht bekannt
Durchführung, Auswertung	Erfolgt durch geschultes psychologisches Personal	Erfolgt eigenständig durch Testpersonen (Selbstadministration)
Einsatzgebiete	Sozial- und humanwissenschaftliche Forschung, Diagnostik	Selbsterkenntnis, Unterhaltung
Anzahl der Items pro Test	Ca. 10–300	Ca. 5–10
Itemaussagen	Einfache Aussagen Beispiel: *„Ich ärgere mich oft darüber, wie Leute mich behandeln.“*	Komplexe Szenarien Beispiel: *„Sie kommen Sonntagabend müde und hungrig von einer Geschäftsreise nach Hause und müssen feststellen: Ihr Partner ist nicht da, der Kühlschrank ist leer und sie haben außer trockenem Brot und Reis nichts im Haus, um etwas zu essen zu kochen. Ihr erster Gedanke?“*
Item-Antwortformate	Ratingskala: – *starke Ablehnung* – *Ablehnung* – *neutral* – *Zustimmung* – *starke Zustimmung*	Single-Choice-Auswahl aus ca. 3–5 sehr spezifischen Antwortoptionen, die konkrete Denk- und Verhaltensweisen ansprechen Beispiel: – „Wir hatten doch vereinbart, dass er mit Einkaufen dran ist. Es ist nicht fair, dass er sich nicht darum gekümmert hat.“ – „Hat er völlig vergessen, dass ich heute zurückkomme? Er hätte mir wirklich einen angenehmeren Empfang bereiten können. Er muss mich ja nicht gerade vermisst haben …“ – „Gut, dass unser Lieblingsitaliener auch nach Hause liefert. Ich hätte sowieso keine Lust gehabt, noch etwas zu kochen.“ – „Typisch! Er hat die letzten Tage wahrscheinlich von Tiefkühlpizza gelebt und gar nicht gemerkt, dass der Kühlschrank leer ist. Wieso ist er immer so nachlässig?“
Testergebnis	Quantitativer Testwert (Wertebereich von 0–4 für das Persönlichkeitsmerkmal Neurotizismus) Beispiel: *Testwert: 2*	Qualitative Typenbeschreibung, oft kombiniert mit Verhaltenstipps Beispiel: **Die Diplomatin:** *„Kommunikation ist von essenzieller Wichtigkeit für eine funktionierende Beziehung. Das wissen Sie und aus diesem Grund sprechen Sie mit Ihrem Partner über alles: Wie Sie sich in einer bestimmten Situation gefühlt haben, was Sie besorgt, was Ihr Partner an sich ändern sollte und was Sie an sich ändern möchten. Was Sie sich für die gemeinsame Zukunft wünschen, was Sie von Ihrem Partner erwarten, wer von Ihnen mit dem Einkauf dran ist, wer im Restaurant bezahlen soll, wie genau Sie das kommende Wochenende verbringen möchten, wie den nächsten Urlaub und wie die Weihnachtsfeiertage bei der Familie. Haben Sie sich jemals gefragt, ob Sie vielleicht zu viel diskutieren? […]“*
Normierung	Bewertung des individuellen Testergebnisses anhand empirischer Normwerte (z. B. $M = 1.63$ in der Vergleichsstichprobe; Körner, et al., 2008)	Keine Normwerte

NEO-FFI NEO (Neurotizismus, Extraversion, Offenheit)-Fünf-Faktoren-Inventar

bzw. Fähigkeitstests), indem sie – meist unter Zeitdruck – Aufgaben bearbeiten, für die es objektiv richtige oder falsche Lösungen gibt (z. B. Rechen-, Denk-, Gedächtnisaufgaben). Auf der anderen Seite stehen Tests, bei denen Personen Selbstauskünfte über ihre Eigenschaften, Vorlieben und Verhaltensweisen geben **(Persönlichkeitstests).** Sowohl bei den Leistungs- als auch bei den Persönlichkeitstests werden diverse Unterformen nach inhaltlichen Aspekten der gemessenen Merkmale unterschieden (Brähler et al., 2002a, S. XIII). Von den Persönlichkeitstests abzugrenzen sind **klinische Tests**, mit deren Hilfe die Ausprägung von Krankheitsbildern gemessen werden soll, teils mit Selbstauskünften (z. B. Alkoholmissbrauch), teils mit Leistungsaufgaben (z. B. Demenz). Bei jedem zu messenden Merkmal ist festzulegen, ob es global über einen einzelnen Testwert **(eindimensionaler Test)** oder nach Unterdimensionen differenziert über mehrere Subtest-Werte erfasst **(mehrdimensionaler Test)** erfasst werden soll.

4. **Testtheoretische Grundlagen**: Psychometrische Tests auf der Basis der Klassischen Testtheorie und auf der Basis der Probabilistischen Testtheorie. Die Mehrzahl der etablierten psychologischen Testverfahren basiert auf der **Klassischen Testtheorie** (KTT) und ihren Erweiterungen. Es existieren jedoch auch Tests, die auf der Basis der aktuelleren **Probabilistischen Testtheorie** (PTT) bzw. **Item-Response-Theorie** (IRT) entwickelt wurden. Die Schulleistungstests der international vergleichenden PISA-Studie (http://www.oecd.org/de/pisa) basieren z. B. auf der IRT.

Projektive und psychometrische Tests – Man unterscheidet nicht- oder teilstrukturierte **projektive Tests** („projective tests") mit offenen Aufgaben (z. B. eine Zeichnung erstellen, ein Bild interpretieren) von vollstrukturierten **psychometrischen Tests** („psychometric tests"), die aus einer Batterie vollstandardisierter Testitems (z. B. Rechen- oder Logik-Aufgaben mit genau einer korrekten Lösung; Selbstauskunftsfragen mit Ratingskalen als Antwortformat) bestehen. In Forschung und Praxis dominieren die psychometrischen Tests.

Eindimensionale und mehrdimensionale Tests – Bei einem **eindimensionalen Test** („one dimensional test") messen alle Testfragen bzw. Testaufgaben („test items") dasselbe Konstrukt. Die Beantwortung aller Items wird zu einem einzelnen (Gesamt-)Testwert („test score", „total test score") zusammengefasst, der global über die Ausprägung des gemessenen Merkmals informiert (z. B. Ausprägung der Intelligenz).

Ein **mehrdimensionaler Test** („multidimensional test") dagegen besteht aus zwei oder mehr **Untertests** (Teiltests, Subtests, „subtests"), die verschiedene Dimensionen des gemessenen Konstruktes separat erfassen. Alle Items, die zu einem Subtest gehören, werden jeweils zu einem Subtest-Wert („subtest score") zusammengefasst. Im Ergebnis liegen für jede Testperson mehrere Subtest-Werte vor, die über die Ausprägungen verschiedener Teilaspekte des interessierenden Merkmals informieren (z. B. Ausprägungen verschiedener Dimensionen von Intelligenz).

Fehler beim Einsatz psychologischer Tests und Gegenmaßnahmen

Werden psychologische Testverfahren zur Datenerhebung eingesetzt, so kann es aus drei unterschiedlichen Gründen zu **fehlerhaften Daten** kommen: 1.) Fehler durch den Test, 2.) Fehler durch die Testanwendung und 3.) Fehler durch die Testpersonen.

1. **Fehler durch den Test**: Fehler durch den Test selbst entstehen, wenn das verwendete Testverfahren nicht (mehr) den gängigen wissenschaftlichen **Gütekriterien entspricht**, es z. B. durch mittlerweile ungültige Normwerte veraltet ist oder wenn bestimmte Bevölkerungsgruppen – z. B. Personen mit geringerem Bildungsgrad, höherem Lebensalter oder einem anderen kulturellen Hintergrund – benachteiligt werden (mangelnde Testfairness). Um derartige Fehler zu vermeiden, sollten grundsätzlich nur Tests eingesetzt werden, die alle Gütekriterien in mindestens zufriedenstellendem Maße erfüllen.
2. **Fehler durch die Testanwendung**: Fehler bei der Testanwendung entstehen, wenn das verwendete Testverfahren zwar den Testgütekriterien entspricht, aber nicht sachgerecht eingesetzt wird: Dies geschieht z. B., wenn **Durchführungsvorschriften laut Testmanual** nicht genau eingehalten werden, oder der Test auf Personengruppen angewendet wird, für die er gar nicht konzipiert wurde. Um derartige Fehler zu vermeiden, muss sichergestellt werden, dass die Vorgaben des Testmanuals genau befolgt werden und der Test nur von **geschultem Personal** durchgeführt, ausgewertet und interpretiert wird.
3. **Fehler durch die Testpersonen**: Fehler durch die Testpersonen entstehen, wenn diese ganz bewusst versuchen, das Testergebnis zu beeinflussen bzw. zu verfälschen, wobei sie im Testergebnis entweder eine übermäßig starke Merkmalsausprägung (Simulation) oder eine besonders schwache Merkmalsausprägung (Dissimulation) anstreben können. Auch weniger bewusste Antwortverzerrungen sind möglich (z. B. sozial erwünschtes Antworten). Dabei stellen sich bei Persönlichkeitstests andere Probleme der Testverfälschung als bei Leistungstests.

Testverfälschung bei Persönlichkeitstests

Simulation und Dissimulation: Die gezielte Testverfälschung durch taktisches (anstatt ehrliches) Beantworten der Items eines Persönlichkeitstests ist v. a. dann ein Problem, wenn das Testergebnis unmittelbare individuelle Konsequenzen hat, wie das beim Einsatz von Tests in der Eignungs-, klinischen oder forensischen **Diagnostik** der Fall ist („high stakes testing"): Wer eine berufliche Anstellung anstrebt, wird

die eigene Leistungsmotivation in einem Test möglicherweise ebenso übertreiben wie eine Person, die auf Berentung spekuliert, ihre psychischen Beschwerden dramatisiert (Simulation). Wer eine vorzeitige Entlassung aus dem Strafvollzug erreichten möchte, wird in einem klinischen Test psychische Auffälligkeiten eher verleugnen (Dissimulation). In der Forschungspraxis, in der – im Unterschied zur Diagnostik – Daten anonym erhoben und nur auf Gruppenebene ausgewertet werden (entsprechende Richtlinien der **Forschungsethik** ▶ Abschn. 4.1) sollten Simulations- oder Dissimulations-Effekte meist vernachlässigbar gering ausgeprägt sein.

- **Selbsttäuschung**: Tests und Befragungen bedeuten für die Untersuchungspersonen eine Selbstkonfrontation. Sie sind gezwungen, über die im Erhebungsinstrument angesprochenen Themen nachzudenken und sehen in ihren Antworten einen Spiegel ihrer Verfassung. Eigene Erlebens- und Verhaltensweisen als widersprüchlich, unvernünftig oder unakzeptabel wahrzunehmen, ist unangenehm. Die eigenen Äußerungen zu „glätten" und mit **Selbstkonzept** und Gruppenidentität in Übereinstimmung zu bringen, ist häufig intrapersonal motiviert und dient somit eher der Selbst- als der Fremdtäuschung. Um diese Effekte abzufangen, sollten Rahmenbedingungen geschaffen werden, die den Untersuchungspersonen eine Auseinandersetzung mit problematischen Selbstaspekten erleichtern. Negative Aspekte können z. B. leichter zugelassen werden, wenn die Testpersonen damit einen Lernerfolg (Selbsterkenntnis) verbinden. Schließlich sei noch erwähnt, dass positiv gefärbte Selbstdarstellungen (z. B. stellt man sich im Persönlichkeitstest sehr durchsetzungsfähig dar, obwohl man – wie Freunde und Angehörige bestätigen könnten – im Alltag nur sehr wenig Durchsetzungsfähigkeit zeigt), nicht nur als Selbst- oder Fremdtäuschungen, sondern auch als eine Art **Zukunftsprognose** aufzufassen sind: Wenn man sich darstellt, wie man gerne wäre, kommt darin auch zum Ausdruck, wie man sich in Zukunft vielleicht entwickelt. Markus und Nurusius (1986) sprechen in diesem Zusammenhang von „possible selfs". Selbstdarstellungseffekte sind also nicht nur als Fehler, sondern auch als Informationsquellen nutzbar (Mummendey, 1990, 1999).
- **Adressatenspezifische Selbstdarstellung**: Der Begriff „Testverfälschung" hat sich zwar eingebürgert, er ist jedoch reflektiert zu verwenden. Abgesehen von zuweilen auftretender bewusster taktischer Simulation oder Dissimulation sollte Untersuchungsteilnehmenden nicht pauschal unterstellt werden, Testergebnisse zu „fälschen", „unehrliche Antworten" zu geben oder gar zu „lügen". Wenn sich Testpersonen dafür entscheiden, bewusst in einer bestimmten Weise zu antworten (sog. Antwortverfälschung) oder auf die Teilnahme an einer Untersuchung zu verzichten (sog. Teilnahmeverweigerung), mögen sie dafür ihre guten Gründe haben. Aus Sicht der Untersuchungspersonen wird das Ausfüllen von Tests oder Fragebögen als **soziale Kommunikation** erlebt. Testpersonen wissen, dass sie anderen Menschen durch den Test etwas über sich mitteilen und machen sich Gedanken darüber, was sie mitteilen wollen und was nicht, bei wem die Informationen ankommen, wie der Empfänger auf sie reagieren könnte und was mit ihnen geschieht. Diese Form der Informationskontrolle nennt man (etwas missverständlich) **Selbstdarstellung** („self presentation") oder **Eindruckssteuerung** („impression management"). Die Art der Selbstdarstellung ist dabei grundsätzlich adressatenabhängig. So konnte Mummendey (1990) zeigen, dass dieselben männlichen Studenten Persönlichkeitsfragebögen anders ausfüllten, wenn sie angeblich von einer Forschungsgruppe „Auswirkungen der Frauenbewegung" oder einer Forschungsgruppe „Selbstkonzept" untersucht wurden. Für die Forschungspraxis lässt sich die Forderung ableiten, die eigene Selbstdarstellung – d. h. die **Vorstellung des Forschungsprojektes** bzw. der Forschungsgruppe – gut zu überdenken. Auch empfiehlt es sich, grundlagenwissenschaftliche Untersuchungen explizit als solche zu kennzeichnen, da Testpersonen bei psychologischen Untersuchungen meist automatisch einen „Psychotherapeuten" oder gar „Psychiater" als Adressaten vermuten und somit zu Unrecht eine Individualdiagnose befürchten.
- **Soziale Erwünschtheit**: Sozial erwünschtes Antworten kann als Sonderform der Selbstdarstellung in Persönlichkeitstests aufgefasst werden: Motiviert durch die Furcht vor sozialer Verurteilung neigt man zu konformem Verhalten und orientiert sich in seinen Verhaltensäußerungen an verbreiteten gesellschaftlichen Normen und Erwartungen (Edwards, 1957, 1970). Wie stark ein Test durch die Tendenz zum sozial erwünschten Antworten „verfälscht" werden kann, wird mit einer einfachen Technik empirisch ermittelt: Eine Gruppe von Testpersonen beantwortet den fraglichen Test unter normalen Bedingungen. Anschließend erhalten dieselben Testpersonen die Instruktion, den Test im zweiten Durchgang so zu beantworten, dass ein maximal positiver, günstiger Eindruck entsteht, sog. **Faking-Good-Instruktion**. Je größer die Diskrepanzen zwischen beiden Testdurchgängen, umso fälschungsanfälliger ist der Test.

Soziale Erwünschtheit – Soziale Erwünschtheit („social desirability") meint die Tendenz, Selbstauskunftsfragen – sei es in Interviews, Frage-

Fünf Maßnahmen zur Kontrolle sozial erwünschten Antwortens bei Persönlichkeitstests

Die gängigste Maßnahme zur Verhinderung sozial erwünschten Antwortens in der Forschung (in der Diagnostik ist das nicht möglich) ist die konsequente Anonymisierung kombiniert mit dem Hinweise an die Befragungspersonen, dass ihre Angaben anonym sind und ehrliche, spontane Antworten erbeten werden.

Im Folgenden werden darüber hinaus **fünf spezielle Techniken** vorgestellt, die dazu dienen sollen, sozial erwünschtes Antworten zu verhindern oder zumindest zu kontrollieren. Diese Verfahrensweisen stellen jedoch keine „Patentrezepte" dar und sind ihrerseits nicht unproblematisch. In vielen Fällen sind sie aus ethischen und/oder forschungsökonomischen Gründen gar nicht umsetzbar.

- **Die Bogus-Pipeline-Methode** („bogus pipeline method", „BPL method"). Bei dieser von den Sozialpsychologen Edward Jones und Harold Sigall (1971) entwickelten Methode werden die Testpersonen, während sie den Test bearbeiten, an einen angeblichen Lügendetektor angeschlossen (engl. „bogus pipeline": gefälschter Anschluss). Sofern die Testpersonen überzeugt sind, dass man auf diese Weise zweifelsfrei ihre möglichen Falschangaben im Test aufdecken kann, antworten sie offenbar deutlich ehrlicher als unter herkömmlicher Anonymitätszusicherung. Studien zeigen z. B., dass unter der Bogus-Pipeline-Bedingung die bekannten geschlechtsrollenkonformen Unterschiede im selbstberichteten Sexualverhalten von Frauen und Männern verschwinden (Alexander & Fisher, 2003), Sexualstraftäter mehr Straftaten einräumen (Gannon, Keown, & Polaschek, 2007) und Jugendliche häufiger Zigarettenkonsum zugeben (wobei hier anstelle eines fingierten Lügendetektortests auch ein angebliches Atemanalysegerät eingesetzt werden kann; Adams, Parkinson, Sanson-Fisher, & Walsh, 2008)
  Die Wirksamkeit der Methode ist – u. a. auch durch Metaanalysen – gut gesichert (Aguinis, Pierce, & Quigley, 1993; Roese & Jamieson, 1993). Kontrovers diskutiert wird die **ethische Vertretbarkeit des Vorgehens** (Aguinis & Handelsman, 1997), bei dem die Testpersonen anfangs bewusst getäuscht, am Ende jeder Studie dann jedoch aufgeklärt werden (Debriefing). Dabei sind forschungsethische Dilemmata einerseits philosophisch anhand ethischer Normen zu erörtern (► Abschn. 4.1), andererseits können auch empirisch die Sichtweisen von Testpersonen eingeholt werden. Eine entsprechende Befragungsstudie zeigte, dass Laien den wissenschaftlichen Nutzen der BPL-Methode als relativ hoch einschätzten und die Unannehmlichkeiten für die Testpersonen für vertretbar hielten (Aguinis & Henle, 2001). Unabhängig von den ethischen Fragen ist die BPL-Methode jedoch ohnehin kaum alltagstauglich, sondern nur für spezielle Fragestellungen geeignet. Denn sie erfordert nicht nur eine Testung vor Ort, sondern erzeugt für eine glaubwürdige Inszenierung des Lügendetektortests **nicht unerheblicher Zusatzaufwand**
- **Objektive Persönlichkeitstests** („objective personality tests"). Bei diesen Tests soll die Verfälschbarkeit von Testergebnissen dadurch reduziert werden, dass das Testziel durch eine geeignete Aufgabenwahl und Auswertungstechnik für die Testpersonen möglichst undurchschaubar **(geringe Augenscheinvalidität)** gemacht wird (Cattell & Warburton, 1967; Kubinger, 1997; Schmidt, 1975). Man beachte, dass „objektiv" in diesem Zusammenhang eine andere Bedeutung hat als das Testgütekriterium der Objektivität. Der Aufforderungscharakter, den Test zu verfälschen, soll bei objektiven Persönlichkeitstests dadurch reduziert werden, dass nicht nach personenbezogenen Inhalten, sondern scheinbar nach neutralen Sachverhalten gefragt wird. Ein objektiver Persönlichkeitstest, der das Merkmal „Risikobereitschaft in Verkehrssituationen" erfasst, fragt nicht nach Selbstauskünften zur Risikobereitschaft am Steuer, sondern zeigt computergestützt Videosequenzen von Verkehrssituationen (z. B. Möglichkeit zum Überholen) und lässt die Testpersonen per Tastendruck den Zeitpunkt bestimmen, ab dem sie nicht mehr überholen würden (WRBTV Wiener Risikobereitschaftstest Verkehr: Hergovich, Arendasy, Bognar, & Olbrich, 2004). Es existieren jedoch nur sehr **wenige objektive Persönlichkeitstests**.
- **Random-Response-Technik.** Die Random-Response-Technik (Warner, 1965) geht von der plausiblen Annahme aus, dass sich die Tendenz zu verfälschten Antworten reduzieren lässt, wenn die Testperson absolut sicher ist, dass sich ihr „wahres" Antwortverhalten nicht rekonstruieren lässt. Die auf Alternativantworten (z. B. ja/nein) bezogene Random-Response-Technik könnte etwa wie folgt aussehen: Die Testperson wird gebeten, vor jedem zu beantwortenden Item (z. B. „Ich rauche manchmal Haschisch") zu würfeln. Würfelt sie eine 1, 2, 3 oder 4, soll das Item ehrlich beantwortet werden. Bei einer 5 ist „ja" und bei einer 6 „nein" anzukreuzen. Da nun bei der Auswertung nicht mehr unterschieden werden kann, welche Antworten ehrlich bzw. erwürfelt sind (d. h., eine Individualauswertung ist nicht möglich), hat die Person keine Veranlassung, die Antworten zu verfälschen. Man vergleicht nun eine Stichprobe, die den Test nach der Random-Response-Technik bearbeitet hat, mit einer anderen, parallelen Stichprobe ohne Random-Response-Instruktion, von der man annimmt, dass sie den Test in üblicher Weise verfälscht. Unter Berücksichtigung des Anteils derjenigen Items, deren Antworten in der Random-Response-Stichprobe erwürfelt wurden, informiert ein Vergleich der Testdurchschnitte der (ehrlichen) Random-Response-Stichprobe und der (unehrlichen) Normalstichprobe, in welchem Ausmaß der Test verfälschbar ist. Nach einem Verfahren von Fidler und Kleinknecht (1977) lässt sich zudem ermitteln, welche Items statistisch bedeutsam verfälscht werden.
  Für die Random-Response-Technik wurden zahlreiche Varianten entwickelt (vgl. Fox & Tracey, 1986). Neben dem Anliegen, mit dieser Technik die Verfälschbarkeit von Tests zu ermitteln, geht es darum, Prävalenzraten für sensible Themenbereiche zu schätzen (sexueller Missbrauch, Serostatus HIV-positiv, Drogenkonsum etc.). Man kann davon ausgehen, dass sozial wenig erwünschte Verhaltensweisen bei Persönlichkeitstests und Befragungen, die mit der Random-Response-Technik operieren, eher zugegeben werden. Wie derartige Untersuchungen statistisch ausgewertet werden, wird z. B. bei Bierhoff (1996, S. 60ff.) bzw. Schnell, Hill und Esser (2008, S. 340ff.) beschrieben. Die Wirksamkeit der Technik ist jedoch umstritten (für eine entsprechende Methodenstudie ► Abschn. 7.3). Zudem besteht ein großer Nachteil der Random-Response-Technik in ihrem recht **großen Aufwand** für Anwender und Untersuchungsteilnehmer. Sie wird kaum eingesetzt, um ganze Tests zu absolvieren, sondern eher für Messungen mit **Einzelindikatoren** (► Abschn. 8.5).

- **Ausbalancierte Antwortvorgaben.** Einige Tests versuchen, das Problem der Verfälschbarkeit von Testergebnissen dadurch zu lösen, dass für die Test-Items Antwortalternativen vorgegeben werden, die bezüglich des Merkmals soziale Erwünschtheit ausbalanciert sind. Wenn die für ein Item zur Auswahl gestellten Antwortalternativen alle sozial gleich erwünscht (oder unerwünscht) sind, bleibt der Testperson keine Möglichkeit, durch ihre Antwort einen besonders guten oder schlechten Eindruck vorzutäuschen. Die Wahrscheinlichkeit, dass sie diejenige Antwortalternative wählt, die tatsächlich am besten auf sie zutrifft, wird damit erhöht.
  Verdeutlicht wird dieser Ansatz z. B. in einem von Edwards (1953) entwickelten Test zur Messung von Werten und Interessen (Edwards Personal Preference Schedule = EPPS; über weitere Tests, die diese Technik nutzen, berichtet Anastasi, 1963, S. 510ff.). Der Gehalt an sozialer Erwünschtheit der in diesem Test vorgegebenen Antwortalternativen (es werden pro Item zwei, hinsichtlich ihrer sozialen Erwünschtheit gleich attraktive Antwortalternativen angeboten) erwies sich nach mehreren Kontrolluntersuchungen gegenüber verschiedenen Alters-, Geschlechts-, Bildungs-, Einkommens- und Nationalitätsgruppen als relativ stabil (Edwards, 1957).
  Neben dem **Aufwand**, der mit der Konstruktion derartiger Testskalen verbunden ist, stellt eine **Reliabilitätsverringerung**, die mit der Vorgabe balancierter Antwortalternativen üblicherweise einhergeht, einen weiteren Nachteil dar. Offensichtlich erschwert oder verunsichert die Vorgabe von Antwortalternativen, die gleichermaßen sozial erwünscht sind, die Wahl einer geeignet erscheinenden Antwortalternative (Cronbach, 1960, S. 449ff.)
- **Kontrollskalen.** Während die zuvor genannten Techniken darauf abzielen, sozial erwünschtes Antworten zu verhindern, versucht man mit Kontrollskalen – also zusätzlichen Testskalen – das **Ausmaß der Verfälschung zu messen**. Kontrollskalen (sog. „Lügenskalen“) bestehen aus Items, die besonders sensibel auf Tendenzen zu sozial erwünschtem Antworten reagieren. Sie erfassen typischerweise Eigenschaften oder Verhaltensweisen, die allgemein negativ beurteilt werden, aber doch so oft vorkommen, dass eine ablehnende Antwort unglaubwürdig erscheint (z. B. „Manchmal benutze ich Notlügen“ – Antwort: „Nein“; oder umgekehrt: „Ich bin immer freundlich und hilfsbereit“ – Antwort: „Ja“). Ein sehr bekanntes Kontrollinstrument ist die **Social Desirability Scale** (SD-Skala) von Crowne und Marlowe (1964), die den Testpersonen zusammen mit dem eigentlich interessierenden Test vorgelegt wird. Hohe Korrelationen zwischen dem Punktwert der SD-Skala und dem interessierenden Testwert sprechen für eine Verzerrung des Testwertes in Richtung sozialer Erwünschtheit (deutschsprachige SD-Skalen stammen z. B. von Lück & Timaeus, 1969; oder Mummendey, 1999, S. 177f.). Die **Paulhus Deception Scale** (PDS) ist ein aus 40 Items bestehender Fragebogen, der die Tendenz zur Selbstdarstellung (Fremdtäuschung) sowie zur Selbstwerterhöhung (Selbsttäuschung) erfasst. Bei Personen, die hier hohe Werte erreichen, ist mit verzerrten Testergebnissen zu rechnen. Mithilfe statistischer Methoden (z. B. Kovarianzanalyse) lassen sich Testwerte um den mit einer Kontrollskala gemessenen Einfluss sozialer Erwünschtheit statistisch bereinigen.
  Ein inhaltliches Problem kann jedoch darin gesehen werden, dass es sich bei der Tendenz zu sozial erwünschtem Antworten um eine Persönlichkeitsdisposition handelt und somit eine **Bereinigung von Persönlichkeitstest-Ergebnissen um Persönlichkeitsaspekte** vorzunehmen wäre. Ein alternatives Verfahren schlägt deswegen vor, die sozial erwünschte Verfälschung von Persönlichkeitstests als Messfehler („spurious measurement errors“) aufzufassen, Daten in einem experimentellen Prä-Post-Kontrollgruppen-Design zu erheben und über Strukturgleichungsmodelle (▶ Kap. 17) auszuwerten (Bühner & Ziegler, 2009). Ein forschungsökonomischer Nachteil des Einsatzes von Kontrollskalen besteht zudem darin, dass sie zusätzlich zum eigentlichen Test erhoben werden müssen, also das **Untersuchungsinstrument verlängern** und die Untersuchungspersonen somit belasten.

bögen oder Persönlichkeitstests – in der Weise zu beantworten, dass die eigenen Aussagen weniger dem realen Erleben und Verhalten und dafür stärker sozialen Normen und Erwartungen entsprechen. Dies geschieht etwa, indem normverletzende Verhaltensweisen (z. B. Drogenkonsum) heruntergespielt („underreporting“) und sozial konforme Verhaltensweisen (z. B. Hilfsbereitschaft) übertrieben werden („overreporting“).

Problematisch am Konzept der sozialen Erwünschtheit ist die Tatsache, dass es in vielen Bereichen einer Gesellschaft gar **keine allgemeinverbindlichen Normen** über „gutes“ Verhalten oder „positive“ Eigenschaften gibt, sondern dass in Abhängigkeit von der Bezugsgruppe und der Situation unterschiedliche Erwartungen bestehen. So mögen sich im Persönlichkeitstest manche Testpersonen als besonders „dominant“ darstellen, weil sie dies für eine positive und in manchen gesellschaftlichen Positionen und Rollen erwartete Eigenschaft halten, während andere ihre Dominanz lieber untertreiben, um sympathischer zu wirken. Beide Gruppen haben somit den Test „verfälscht“. Wenn zwischen normaler Instruktion und Faking-Good-Instruktion keine Differenz im Gruppenmittelwert erscheint, ist dies nicht zwangsläufig ein Indiz für die Unverfälschbarkeit des Tests, sondern könnte das Resultat unterschiedlicher Vorstellungen über erstrebenswertes Verhalten sein, da sich divergierende Verfälschungstendenzen bei der Durchschnittsberechnung kompensieren (Gordon & Gross, 1978). Vorstellungen von sozial erwünschtem Verhalten sind nicht zuletzt auch stark (sub)kulturabhängig (vgl. Keillor, Owens, & Pettijohn, 2001; Middleton & Jones, 2000).

Die wichtigste und **gebräuchlichste Gegenmaßnahme gegen sozial erwünschtes Antworten** in Persönlichkeitstests sowie Fragebögen und Interviews besteht in

der Forschungspraxis darin, die Testpersonen bzw. Untersuchungsteilnehmenden in der Instruktion noch einmal ausdrücklich darauf hinzuweisen, dass alle ihre Angaben vollkommen freiwillig und anonym erfolgen, und sie **um ehrliche Antworten zu bitten**, da damit dem Forschungsprojekt am besten geholfen wird. Diese Maßnahme kann und sollte routinemäßig eingesetzt werden und normalerweise hinreichend unverfälschte Selbstauskünfte liefern. Darüber hinaus existierten **fünf weitere Maßnahmen** zur Verhinderung oder Kontrolle sozial erwünschten Antwortens (▶ Fünf Maßnahmen zur Kontrolle sozial erwünschten Antwortens bei Persönlichkeitstests).

Testverfälschung bei Leistungstests

Während Verfälschungen bei Persönlichkeitstests – so sie denn auftreten – v. a. darauf hinauslaufen, dass Testpersonen sich in einem günstigen Licht darstellen, geht es bei Leistungstests im Falle von Verfälschungsabsichten darum, ein möglichst gutes Ergebnis zu erzielen. Dafür wird v. a. auf drei unterschiedliche Verfälschungs-Methoden zurückgegriffen, die möglichst minimiert werden sollen: Abschreiben, Üben und Raten. Zudem kommt es zuweilen auch zu einer Verfälschung im Sinne bewusst schlechter Leistungen (Dissimulation).

- **Abschreiben**: Diesem Problem begegnet man durch eine entsprechende Sitzordnung (bei Gruppentestungen) sowie durch Paralleltests (Testbogen A und B).
- **Üben**: Eine Testverfälschung kann auch darin bestehen, dass Testpersonen die Testaufgaben (ggf. auch die Lösungen) im Vorfeld kennen und diese somit üben bzw. auswendig lernen können. Die Lösungen wären dann kein Ausdruck ihrer eigenen Leistungsfähigkeit mehr. Um dieses Problem zu vermeiden, unterliegen Tests – und insbesondere Leistungstests – dem sog. **Testschutz** und dürfen nicht frei veröffentlicht werden. So sind z. B. die Aufgaben der PISA-Studie nicht öffentlich verfügbar.
- **Raten**: Stellt man in Leistungstests standardisierte Wissensaufgaben z. B. in Form von **Multiple-Choice-Fragen**, so können Testpersonen ohne entsprechendes Wissen teilweise mit hoher Trefferquote die richtige Antwort erraten: Bei Aufgaben mit zwei Antwortmöglichkeiten beträgt die zufällige Trefferwahrscheinlichkeit immerhin 50 %. Dieser Nachteil könnte vernachlässigt werden, wenn die Verfälschung der Testergebnisse durch Raten bei allen Testpersonen konstant wäre. Dies ist jedoch nicht der Fall. Der prozentuale Anteil der durch Raten richtig beantworteten Aufgaben nimmt **mit abnehmender Fähigkeit** der Untersuchungsteilnehmer zu. Es ist deshalb erforderlich, die Ergebnisse von Tests mit Antwortvorgaben durch eine **Ratekorrektur** („correction for guessing“) zu bereinigen (Lienert & Raatz, 1994, S. 168f.; ▶ Maßnahmen zur Kontrolle und Korrektur des Ratens in Leistungstests)
- **Dissimulation:** Zuweilen kann bei Leistungstests neben Raten (als Methode zur Simulation guter Ergebnisse) auch **Dissimulation** auftreten, d. h. Testpersonen versuchen, eine besonders schlechte Leistung abzuliefern, indem sie absichtlich Fehler produzieren. So könnten Testpersonen motiviert sein, in einem Konzentrationstest sehr schlechte Leistungen zu erzeugen, um die Diagnose einer Aufmerksamkeitsstörung zu erhalten. Die Verfälschbarkeit eines Tests im Sinne von Dissimulation kann empirisch geprüft werden, indem man die Testpersonen ausdrücklich bittet, ein besonders schlechtes (Faking-Bad-Instruktion) Ergebnis plausibel zu fälschen. Eine entsprechende Überprüfung des Konzentrationstest d2 (▶ Abschn. 10.4.3) ergab, dass dieser kaum verfälschbar ist: Gute Testergebnisse lassen sich kaum durch Raten bzw. zufälliges Antworten generieren und die absichtlichen Fehler sind sehr auffällig, weil hier teilweise Ankreuzmuster vorkommen, die nicht einmal hirngeschädigte Personen zeigen (Schmidt-Atzert, Bühner, Rischen, & Warkentin, 2004). Hinweise zur Verfälschbarkeit (einem der sieben Nebengütekriterien von psychologischen Tests) sind hilfreich für die Bewertung und Auswahl sowohl von Persönlichkeits- sowie Leistungstests, wobei hier – wie bereits betont – in der diagnostischen Praxis mehr Brisanz liegt als in der anonymisierten Forschung.

Hauptgütekriterien psychologischer Tests

Psychologische Tests verfolgen das Ziel, die Ausprägung latenter – d. h. nicht direkt beobachtbarer – Merkmale wie Intelligenz, Gewissenhaftigkeit, Motivation oder Ängstlichkeit von Menschen messbar zu machen, um bestimmte wissenschaftliche Aussagen oder praktische Entscheidungen treffen zu können. Dabei müssen **Standards und Kriterien der Wissenschaftlichkeit** (▶ Abschn. 3.1) eingehalten werden. Der Testeinsatz muss also im Rahmen eines **wissenschaftlichen Forschungsprozesses** erfolgen und dem Kriterium der **methodischen Strenge** entsprechen, so dass sich die Frage nach der Testqualität stellt. **Testqualität** wird heute als breiter Oberbegriff für diverse Teilaspekte des Umgangs mit Tests verstanden, die letztlich dafür sorgen, dass aus Testergebnissen gezogene Schlussfolgerungen im konkreten Kontext als wissenschaftlich tragfähig gelten können (Newton & Shaw, 2013). Für eine konkrete Studie muss also nach heutigem Verständnis von Testqualität idealerweise differenziert anhand unterschiedlicher theoretischer Argumente und empirischer Befunde aufgezeigt werden, inwiefern die vorgenommene Verwendung und Interpretation der Testwerte Gültigkeit beanspruchen kann.

Maßnahmen zur Kontrolle und Korrektur des Ratens in Leistungstests

Ratekorrektur bei Testaufgaben mit zwei Antwortmöglichkeiten

Betrachten wir zunächst eine Testaufgabe aus einem Leistungstest mit zwei Antwortmöglichkeiten, von denen eine richtig und eine falsch ist. Wie sollte die Auswertung am besten erfolgen, wenn man das Raten einbezieht. Angenommen, man würde als Testergebnis die Anzahl aller richtig beantworteten Aufgaben gelten lassen. Eine völlig unfähige Person A würde bei diesem Verfahren ca. 50 % aller Aufgaben allein durch Raten richtig lösen und hätte damit das gleiche Ergebnis erzielt wie eine mittelmäßig befähigte Person B, die auf Raten verzichtet und 50 % der Aufgaben aufgrund ihres Wissens richtig löst und die übrigen Aufgaben unbearbeitet lässt. Die beiden Personen unterscheiden sich damit nicht in der Anzahl der richtig gelösten Aufgaben, sondern in der Anzahl der falsch gelösten Aufgaben. Zu einem angemesseneren Testergebnis käme man in diesem Fall, wenn als Testergebnis nicht die Anzahl der richtig gelösten Aufgaben, sondern die **Anzahl der richtig gelösten Aufgaben abzüglich der falsch gelösten Aufgaben** verwendet wird. Person A hätte dann ca. 0 Punkte und Person B die Hälfte der möglichen Punktzahl. Allgemein formuliert:

$$x_{\text{corr}} = N_{\text{R}} - N_{\text{F}}$$

mit

x_{corr}: korrigiertes Testergebnis,

N_{R}: Anzahl richtig gelöster Aufgaben,

N_{F}: Anzahl falsch gelöster Aufgaben.

Aiken und Williams (1978) haben sieben verschiedene Auswertungsstrategien mit unterschiedlichen Ratekorrekturen für Alternativaufgaben verglichen. Sie kommen zu dem zusammenfassenden Ergebnis, dass keine Auswertungstechnik generell zu bevorzugen sei. Sie empfehlen jedoch, die Testpersonen in der Testinstruktion **über die Art der Ratekorrektur bei der späteren Testauswertung aufzuklären**. Dadurch werden Benachteiligungen, die sich je nach Auswertungsart für ratende oder nicht ratende Personen ergeben, minimiert (Hsu, 1979; Ortmann, 1973; Rützel, 1972).

Ratekorrektur bei Testaufgaben mit mehr als zwei Antwortmöglichkeiten

Auch bei mehr als zwei Antwortalternativen können Testergebnisse durch Raten verfälscht werden. Stehen z. B. vier Antwortmöglichkeiten pro Test-Item zur Verfügung, von denen eine korrekt, die anderen drei inkorrekt (d. h. **Distraktoren**) sind, wird eine Person allein durch Raten ca. 25 % aller Aufgaben richtig lösen. An welcher Stelle unter den Distraktoren die richtige Antwort platziert wird, beeinflusst dabei den Rateerfolg nicht (Eheim, 1977). Allerdings wird das Raten sehr erleichtert, wenn die Distraktoren auf den ersten Blick sehr unplausibel wirken bzw. in inhaltlicher oder formaler Hinsicht offensichtlich von der richtigen Antwort abweichen. Eine bezüglich des Rate-Einflusses korrigierte Punktzahl resultiert, wenn man von der **Anzahl richtig gelöster Aufgaben die durch die Anzahl der Distraktoren dividierte Fehleranzahl** abzieht:

$$x_{\text{corr}} = N_{\text{R}} - \frac{N_{\text{F}}}{k-1}$$

mit k: Anzahl der Antwortalternativen.
Beispiel: Bei 100 Items mit jeweils $k = 4$ Antwortvorgaben wird eine ratende Testperson ca. $N_{\text{R}} = 25$ Items zufällig richtig und $N_{\text{F}} = 75$ Items zufällig falsch beantworten. Sie hätte damit eine korrigierte Punktzahl von Null:

$$x_{\text{corr}} = 25 - 75/(4-1) = 0$$

Wenn pro Item mehrere Antwortvorgaben richtig sind, können Rate-Effekte neutralisiert werden, indem man für jedes richtige Ankreuzen einen **Pluspunkt**, für jedes falsche Ankreuzen einen **Minuspunkt** und für jede nicht angekreuzte Antwortvorgabe **keinen Punkt** vergibt. Auch hier sollten jedoch die Testpersonen zuvor darüber informiert werden, in welcher Weise in der Auswertung Falschantworten berücksichtigt werden (so können richtige und falsche Antworten statt mit exakt einem Punkt auch unterschiedlich gewichtet werden). Weitere Ratekorrekturen bei Aufgaben mit vorgegebenen Antwortmöglichkeiten diskutiert Barth (1973). Die Art der Ratekorrektur scheint die Testreliabilität nicht zu beeinflussen (Jaradad & Tollefson, 1988).

Ratekorrekturen mittels Item-Response-Theorie

Die bislang dargestellten Korrekturmöglichkeiten beziehen sich auf Tests, die nach der Klassischen Testtheorie konstruiert wurden. Um den Einfluss des Ratens aus Testergebnissen zu eliminieren, wird bei Leistungstests mit dichotomen Items auf der Basis der Probabilistischen Testtheorie das sog. **3PL-Modell** (Drei-Parameter-Logistisches Modell) angewendet. Gemäß dem 3PL-Modell hängt die Lösungswahrscheinlichkeit eines Items nicht nur von der latenten Merkmalsausprägung einer Testperson ab (Personenparameter), sondern von drei weiteren Parametern, nämlich dem Itemschwierigkeits-, dem Trennschärfe- und dem **Rateparameter** (vgl. hierzu Baker & Kim, 2004, S. 18ff.; Bühner, 2011, S. 508ff.; Rost, 2004, S. 133).

Diesem aktuellen Verständnis von Testqualität steht eine lange Tradition anderer Denk- und Sprechweisen gegenüber (Newton & Shaw, 2013): Insbesondere hat es sich stark eingebürgert, Testqualität als Merkmal des Tests zu bezeichnen und in einen **Kanon von Testgütekriterien** aufzusplitten. Für die Testgütekriterien wird dann anhand von **Normwerten oder Faustregeln** betrachtet, ob sie als erfüllt oder nicht erfüllt einzustufen sind (z. B. Reliabilitätskoeffizient > .90: Messgenauigkeit erfüllt; Validitätskoeffizient > .40: Gültigkeit erfüllt), um auf dieser Basis dann den **Test pauschal für brauchbar oder unbrauchbar** zu erklären. Diese **schematische Vorgehensweise** erfreut sich bis heute großer Beliebtheit, da sie vergleichsweise gut umsetzbar und kommunizierbar ist. Sie hat jedoch den Nachteil, dass gemäß Standards der Wissenschaftlichkeit das konkrete Vorge-

hen beim Einsatz von Tests auf diese Weise nicht genau genug reflektiert wird und häufige Fehlschlüsse zu erwarten sind. Wenn z. B. mit einem Korrelationskoeffizienten von $r = .63$ eine „hohe Kriteriumsvalidität" beansprucht und somit der Test für gültig erklärt wird, dann können unbemerkt dennoch völlig falsche Schlussfolgerungen resultieren. Denn um gültige Schlussfolgerungen über die Testergebnisse zu erzielen, muss man viel grundlegender und systematischer theorie- und empiriebasiert aufzeigen, dass und warum die Testwerte für die jeweilige Zielgruppe und Fragestellung (auch im Vergleich zu anderen möglicherweise einsetzbaren Tests) ausreichende Aussagekraft haben. Die Korrelation mit einem einzelnen Außenkriterium zur „Kriteriumsvalidität des Tests" zu erklären, befördert eher falsche Gewissheit, während ein Verständnis von Testqualität, dem gemäß alle möglichen Argumente für und gegen die Aussagekraft des Tests und der daraus abgeleiteten Interpretationen abgewogen werden müssen, eben eine gründlichere und kritischere Reflexion des Vorgehens nahelegt.

Aktuell befinden wir uns in der paradoxen Situation, dass die international verbindlichen Richtlinien zum Einsatz psychologischer Tests es sowohl verbieten, den Tests Testgütekriterien zuzuschreiben (Validität gilt nicht für Tests, sondern nur für Interpretationen und Schlussfolgerungen auf der Basis von Testergebnissen) als auch die gängigen Untertypen von Testgütekriterien (z. B. Inhalts-, Kriteriums- und Konstruktvalidität) offiziell abgeschafft haben. Gleichzeitig werden aber genau diese Gütekriterien bis heute in der Fachliteratur – einschließlich der führenden Methodenzeitschriften – in traditioneller Sprechweise verwendet, teilweise sogar in den Titeln der Aufsätze (Newton & Shaw, 2013).

Da das vorliegende Lehrbuch dazu befähigen möchte, die Fachliteratur zu verstehen, kann auf die Einführung und Erklärung der in der Literatur bis heute gängigen Testgütekriterien (sowohl ihrer Differenzierung in Subtypen als auch ihrer Interpretation als Testmerkmale) nicht verzichtet werden, auch wenn dieses traditionelle Verständnis von Testgütekriterien seit den 1980er-Jahren einer grundlegenden Kritik unterzogen wird. Wir schließen uns im vorliegenden Lehrbuch der **Strategie der bewusst ungenauen Sprechweise** an (Newton & Shaw, 2013, S. 308): Das moderne Verständnis von Testqualität – insbesondere im Hinblick auf Validität -, das seit den 1980er-Jahren wesentlich von dem Psychologen Samuel Messick (1931–1998) entwickelt und von den internationalen Test-Standards aufgenommen wurde (Messick, 1980, 1988, 1995; AERA, APA, & NCME, 1995 [4. Aufl.], 1999 [5. Aufl.], 2014 [6. Aufl.]), halten wir für das angemessene Denkmodell. Gleichzeitig gehen wir in unpräziser Sprechweise in diesem und anderen Teilen des Buches auf die traditionellen Testgütekriterien ein.

Im Folgenden werden also die drei Hauptgütekriterien Objektivität, Reliabilität und Validität mit ihren jeweiligen Unterformen inhaltlich erläutert (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2012a). Kenntnisse über die Bedeutung der Testgütekriterien und ihrer Unterformen sind notwendig, um die Qualität vorliegender Testverfahren einschätzen und eigene Tests entwickeln zu können. Angesichts der gravierenden Konsequenzen, die Testergebnisse im Rahmen der Diagnostik für einzelne Menschen haben können, werden Qualitätskriterien für Tests nicht nur wissenschaftsintern behandelt, sondern sind mittlerweile auch in DIN-Normen festgeschrieben (z. B. **DIN 33430** zu „Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen).

Objektivität

Die **Objektivität** bzw. **Anwenderunabhängigkeit** gibt an, ob ein Test bzw. dessen Ergebnis von der Person des Testanwenders unabhängig ist. Ein Test ist objektiv, wenn verschiedene Testanwender bei denselben Testpersonen zu denselben Testergebnissen gelangen (Tab. 10.15). Umgekehrt ist ein Test nicht objektiv, wenn in die Durchführung oder Auswertung z. B. besonderes Expertenwissen oder individuelle Deutungen des Anwenders einfließen, die intersubjektiv nicht reproduzierbar sind. Entsprechende Einschränkungen der Objektivität können z. B. bei projektiven Testverfahren auftreten (▶ Abschn. 10.4.2).

Die Objektivität ist bei vollstandardisierten psychologischen Tests (ebenso wie bei vollstandardisierten Fragebögen) ein relativ unkritisches und eigentlich redundantes Testgütekriterium, denn durch die Standardisierung des Instruments bleibt den Testanwendenden eigentlich gar kein Raum für subjektive Abweichungen (entsprechend wird dieses Kriterium in der internationalen Literatur nicht gesondert aufgeführt). Objektivität kann in der Regel als erfüllt betrachtet werden, wenn das **Testmanual** zur Durchführung, Auswertung und Interpretation des Tests präzise Angaben macht sowie Hilfestellungen gibt (z. B. Auswertungsschablonen, Interpretationsbeispiele) und die **Testanwendenden** sich dann bei der Datenerhebung präzise daran halten (zur Objektivität von Entwicklungstests Macha & Petermann, 2013). Die Objektivität besteht aus den drei Elementen Durchführungs-, Auswertungs- und Interpretations-Objektivität (Tab. 10.15).

Reliabilität

Die **Reliabilität** bzw. **Zuverlässigkeit**, **Präzision** oder **Messgenauigkeit** („reliability") gibt an, wie gering oder stark ein Test durch Messfehler verzerrt ist: Ein reliabler Test bildet mit seinem Testwert („test score") gemäß

Tabelle 10.15 Das Testgütekriterium der Objektivität (Anwenderunabhängigkeit) und seine Unterformen

Testgütekriterium	Definition	Überprüfung
Objektivität	Unabhängigkeit des Testergebnisses von der Person des Testanwenders/Testleiters	Bei vollstandardisierten Tests ist durch die präzisen Vorschriften im Testmanual, an die sich alle Testanwendenden halten müssen, eine nahezu perfekte Objektivität erreichbar. Auf eine statistische Prüfung der Objektivität wird meist verzichtet. Entscheidend ist die Schulung und ggf. Kontrolle der Testanwendenden.
– **Durchführungsobjektivität**	Unabhängigkeit des Testergebnisses von der Person, die den Test mit der Testperson durchführt	Das Vorgehen bei der Testdurchführung (z. B. Bearbeitungszeit) ist im Testmanual genau vorgegeben und wird von den Testanwendern auch so umgesetzt.
– **Auswertungsobjektivität**	Unabhängigkeit des Testergebnisses von der Person, die den Test auswertet	Das Vorgehen bei der Testauswertung (z. B. Zuordnung von Punkten zu bestimmten Aufgabenlösungen) ist im Testmanual genau vorgegeben und wird von den Testanwendern auch so umgesetzt bzw. erfolgt semi-automatisch (anhand von Schablonen) oder vollautomatisch (computergestützt).
– **Interpretationsobjektivität**	Unabhängigkeit des Testergebnisses von der Person, die den Testwert interpretiert	Das Vorgehen bei der Testinterpretation (z. B. Einordnung eines Testwertes als über- oder unterdurchschnittliches Ergebnis anhand von Normwerten) ist im Testmanual genau vorgegeben und wird von den Testanwendern auch so umgesetzt.

dem Verständnis der Klassischen Testtheorie (▶ Abschn. 10.4.4) die wahre Merkmalsausprägung („true score“) der Person genau ab, er ist allenfalls geringfügig durch einen Messfehler („error score“) belastet. Die Objektivität ist eine Voraussetzung für die Reliabilität, denn mangelnde Objektivität führt zu Messfehlern und reduziert die Reliabilität. Doch auch wenn perfekte Objektivität auf Seiten der Testanwender gegeben ist, kann es auf Seiten der Testpersonen zu Messfehlern kommen, etwa wenn sie Aufgaben und Antwortmöglichkeiten nicht richtig verstehen. Die Messgenauigkeit eines Tests hängt also wesentlich von der Formulierung und Präsentation der Aufgaben und Antwortmöglichkeiten ab.

Die Reliabilität eines psychometrischen Tests (oder einer psychometrischen Skala; ▶ Abschn. 8.6) wird in der Regel empirisch überprüft, wofür auf der Basis der Klassischen Testtheorie **vier verschiedene Arten von Reliabilitätskoeffizienten** zur Verfügung stehen (Tab. 10.16).

Am häufigsten eingesetzt wird die Reliabilitätsbestimmung mit der **Methode der internen Konsistenz**, die auf die Berechnung von **Cronbachs Alpha-Koeffizienten** hinausläuft (Cronbach, 1951). In SPSS können Reliabilitätskoeffizienten über das Menü „Analysieren“ → „Skalierung“ → „Reliabilitätsanalyse“ (Modell: Parallel(test), Split-Half oder Alpha) berechnet werden. Näheres zur Berechnung der Reliabilitätskoeffizienten ist ▶ Abschn. 10.4.4 zu entnehmen.

Es ist zu beachten, dass es sich bei den Reliabilitätskoeffizienten um **Schätzungen der Reliabilität** auf der Basis der Annahmen der Klassischen Testtheorie handelt (Schermelleh-Engel & Werner, 2012): Die Reliabilität ist als theoretische Größe eindeutig definiert. In der Praxis kann sie jedoch nicht wirklich exakt berechnet werden, da wahre Werte und Messfehler nicht für jede einzelne Person bestimmbar sind. Bei Tests auf der Basis der Klassischen Testtheorie liefern die Reliabilitätskoeffizienten **pauschale Genauigkeitsschätzungen** der Testwerte. Demgegenüber ist bei Tests auf der Basis der Probabilistischen Testtheorie mittels der Informationsfunktion auch eine **testwertabhängige Genauigkeitsbeurteilung** der Testwerte möglich (Moosbrugger & Kelava, 2012a, S. 13ff.)

Der Wertebereich des Reliabilitätskoeffizienten *(Rel)* bewegt sich zwischen 0 (völlig unzuverlässige Messung, die nur aus zufälligen Messfehlern besteht) und 1 (perfekt zuverlässige Messung, gänzlich ohne Beeinträchtigung durch Messfehler): $\mathbf{0} \leq \mathbf{Rel} \leq \mathbf{1}$. Reliabilitätskoeffizienten **über .90** werden üblicherweise als hoch und Reliabilitätskoeffizienten **über .80** als ausreichend eingestuft (Fisseni, 1990; Bühner, 2011, S. 81). Wie in vielen anderen Bereichen (z. B. Bewertung von Effektgrößenmaßen ▶ Abschn. 14.2) ist jedoch auch bei Reliabilitätsmaßen vor einer mechanischen Interpretation statistischer Kennwerte zu warnen. Ob die Reliabilität eines Tests als ausreichend akzeptiert werden kann, muss in einem zeitgemäßen Verständnis von Testqualität unter Berücksichtigung der Art des gemessenen Merkmals sowie der methodischen Alternativen differenziert beurteilt werden. Schermelleh-Engel und Werner 2012 (S. 135ff.) schlagen u. a. folgende Anhaltspunkte für die **inhaltliche Beurteilung von Reliabilitätskoeffizienten** vor:

Tabelle 10.16 Das Testgütekriterium der Reliabilität (Zuverlässigkeit, Präzision, Messgenauigkeit) und seine Unterformen

Testgütekriterium	Definition	Überprüfung
Reliabilität („reliability“)	Keine oder geringe Verzerrung des Testwertes durch Messfehler, d. h. der Testwert bildet die wahre Merkmalsausprägung der Testperson sehr präzise ab	Die Reliabilität eines Tests wird standardmäßig bei der Testentwicklung (und oft auch bei der Testanwendung) mithilfe mindestens eines der vier möglichen Reliabilitätskoeffizienten statistisch geschätzt und inhaltlich interpretiert.
– **Testwiederholungs-Reliabilität** (Stabilität, Retest-Reliabilität, „test-retest reliability“, „stability“)	Hohe positive Korrelation des Testwertes mit einem zeitversetzt mit demselben Test an denselben Testpersonen gemessenen zweiten Testwert	Die Testwiederholungs-Methode ist nur für solche Tests ein sinnvolles Reliabilitätsmaß, die ein sehr stabiles Merkmal messen, so dass Unterschiede zwischen Erst- und Zweitmessung tatsächlich als Messfehler und nicht als Merkmalsveränderung zu interpretieren sind. Die Testwiederholungs-Methode ist durch zwei Messzeitpunkte relativ aufwändig für Testanwender und Testpersonen. Die Festlegung des Retest-Intervalls muss inhaltlich begründet werden (typisch sind 2–3 Wochen Zeitabstand zwischen den Messungen).
– **Paralleltest-Reliabilität** (Äquivalenz, „paralleltest reliability“)	Hohe positive Korrelation des Testwertes mit einem zum gleichen Zeitpunkt mit einem inhaltsähnlichen (parallelen) Test gemessenen Testwert	Die Paralleltest-Methode setzt die Existenz eines nahezu identischen Parallel-Tests (Testversion A, Testversion B) voraus. Sofern dieser nicht existiert, kann die Methode nicht angewendet werden.
– **Testhalbierungs-Reliabilität** (Split-Half-Reliabilität, „split half reliability“)	Hohe positive Korrelation der einen Testhälfte mit der zum gleichen Zeitpunkt erhobenen anderen Testhälfte (z. B. erste versus zweite Testhälfte, gerade versus ungerade Items)	Die Testhalbierungs-Methode ist sehr ökonomisch, da sie bei einem einzigen Testdurchgang angewendet werden kann. Sie stellt eine Spezialform der etablierten Methode der Internen Konsistenz dar.
– **Interne Konsistenz** („internal consistency“)	Hohe positive durchschnittliche Korrelation jedes einzelnen Testitems mit jedem anderen, zum selben Zeitpunkt gemessenen Item (korrigiert um die Testlänge)	Die Interne Konsistenz-Methode ist eine Verallgemeinerung der Testhalbierungs-Methode. Anstatt den hinsichtlich seiner Messgenauigkeit zu prüfenden Test willkürlich in zwei Hälften einzuteilen, wird er in seine einzelnen Items aufgeteilt und es werden alle bivariaten Korrelationen zwischen den Items berechnet und gemittelt. Die Interne Konsistenz wird typischerweise mit dem Cronbach Alpha-Koeffizienten berechnet. Er ist das mit Abstand gebräuchlichste Reliabilitätsmaß.

- **Art des zu erfassenden Merkmals und Vergleich mit konkurrierenden Verfahren**: Leistungsvariablen lassen sich häufig präziser messen als z. B. Einstellungen oder Persönlichkeitsmerkmale. Sollten keine besser geeigneten Testverfahren vorhanden sein, kann der Einsatz von Instrumenten mit niedriger Reliabilität aufschlussreicher sein als ein völliger Verzicht.
- **Individual- versus Kollektivdiagnostik**: In der Individualdiagnostik ist eine hohe Reliabilität unverzichtbar, um Fehlurteile zu vermeiden. In der Kollektivdiagnostik ist eine mangelnde Reliabilität zwar störend, da sie inferenzstatistische Nachweise von Gruppenunterschieden erschwert. Die Gruppenmittelwerte, die in der Forschung interessieren, würden aber auch bei individuell stärker messfehlerbehafteten Testwerten korrekt geschätzt.
- **Einsatzbedingungen des Testverfahrens**: Beim Screening, der groben Einschätzung von Merkmalsausprägungen mit geringem Aufwand, werden kürzere und damit ökonomischere Instrumente bevorzugt, dafür müssen Einschränkungen der Reliabilität in Kauf genommen werden (eine Alternative stellt teilweise das adaptive Testen dar; ▶ Abschn. 10.4.4).
- **Kosten-Nutzen-Abwägung**: Da hohe Reliabilität eine Voraussetzung hoher Validität darstellt, ist bei Reliabilitätseinschränkungen abzuwägen, ob die daraus folgenden Validitätseinbußen für den Einsatzzweck des Tests noch vertretbar sind. Andernfalls muss entsprechender Aufwand für die Entwicklung eines reliableren Tests betrieben werden.
- **Objektivität als Voraussetzung für Reliabilität**: Bei einem Test, der sich aus inhaltlichen Gründen nicht vollständig standardisieren lässt und somit geringere Objektivität aufweist, muss auch verringerte Reliabilität in Kauf genommen werden.
- **Passung von Homogenität/Heterogenität des Verfahrens und Methode der Reliabilitätsschätzung**: Enthalten Testverfahren bzw. Testskalen aus inhaltlichen Gründen heterogene Items, um ein heterogenes

Merkmal in seinen verschiedenen Aspekten zu erfassen, wird durch Konsistenzanalysen (Cronbachs Alpha) die tatsächliche Reliabilität unterschätzt. Dennoch kann die Retest-Reliabilität hoch sein, wenn es sich um ein zeitlich stabiles Merkmal handelt.

Werden im Zuge einer Testkonstruktion Reliabilitätsmängel festgestellt, so können diese im Rahmen der Klassischen Testtheorie auf der Basis einer **Itemanalyse** (▶ Abschn. 10.4.4) behoben werden: Items mit extrem hoher oder extrem geringer Schwierigkeit sowie mit sehr geringer Trennschärfe werden identifiziert und – sofern dies theoretisch vertretbar ist – umformuliert oder aus der Skala eliminiert. Es ist zu beachten, dass das Kriterium der Reliabilität, das sich auf den Testwert bezieht, inhaltlich voraussetzt, dass dieser Testwert aus einer sinnvollen Berechnungsvorschrift hervorgeht **(Nebengütekriterium der Skalierung).** Während man im Rahmen der Probabilistischen Testtheorie die Skalierbarkeit aller Items empirisch genau prüft und nur modellkonforme Items verwendet, die echte Messwerte der latenten Fähigkeiten darstellen, wird bei Tests auf der Basis der Klassischen Testtheorie nach der Itemanalyse routinemäßig aus den Antworten der ausgewählten Items ein additiver Summen- oder Durchschnittsscore gebildet, dem ohne weitere empirische Prüfung bestimmte messtheoretische Eigenschaften (z. B. Intervallskalenniveau) zugeschrieben werden (zur Kritik der Klassischen Testtheorie ▶ Abschn. 10.4.4).

Dass die Reliabilität von standardisierten Messinstrumenten im quantitativen Paradigma große Beachtung findet, begründet sich vor allem daraus, dass die Reliabilität eine Voraussetzung für das inhaltlich noch wichtigere Kriterium der **Validität des Instruments** bzw. der aus den Testergebnissen abgeleiteten Schlussfolgerungen ist. Darüber hinaus gefährden unrelieble bzw. messfehlerbehaftete Datenerhebungsinstrumente ganz grundlegend die Gültigkeit der Schlussfolgerungen, die aus empirischen Studien gezogen werden: Reliabilitätsmängel bei der Messung der abhängigen und/oder unabhängigen Variablen können die **externe Validität** (inklusive ihres Unteraspekts der Konstruktvalidität) sowie die **interne Validität** (inklusive ihres Unteraspekts der statistischen Validität) **einer Studie** beeinträchtigen (zu den vier Typen der Validität wissenschaftlicher Aussagen in der Campbell-Tradition ▶ Abschn. 3.2.2). Denn wenn verminderte Reliabilität eines Messinstruments dazu führt, dass auch seine Validität verringert ist, dann kann von den Messwerten definitionsgemäß nicht mehr ohne weiteres auf die angezielten Konstrukte geschlossen werden (verringerte Konstruktvalidität) und folglich ist auch eine Verallgemeinerung der Befunde auf andere Kontexte in ihrer Gültigkeit gefährdet (externe Validität). Reliabilitätsmängel erschweren es, Effekte im Signifikanztest nachzuweisen (statistische Validität), so dass man folglich auch Gefahr läuft, die Wirkung der unabhängigen Variablen auf die abhängigen Variablen verzerrt oder falsch einzuschätzen (interne Validität).

Validität

Die **Validität** bzw. **Gültigkeit** („validity“) als psychometrisches Hauptgütekriterium gibt im traditionellen Verständnis an, ob ein psychologischer Test tatsächlich das Merkmal misst, das er zu messen beansprucht (Hartig, Frey, & Jude, 2012). Die Validität ist das **wichtigste psychometrische Gütemerkmal psychologischer Tests**. Denn wenn ein Test gar nicht das Zielkonstrukt erfasst, ist er inhaltlich für die Diagnostik ebenso wie für die Forschung unbrauchbar. Eine hohe Reliabilität ist eine notwendige Voraussetzung für eine hohe Validität, denn ein mit Messfehlern belasteter Testwert kann das Zielkonstrukt auch nicht treffsicher erfassen. Doch selbst bei hoher Reliabilität kann ein Test trotzdem invalide sein, etwa wenn er mit großer Messgenauigkeit ein dem Zielkonstrukt ähnliches, aber doch anderes Konstrukt misst. Auch Tests, die hohe Reliabilität dadurch erreichen, dass sie mit nahezu inhaltsgleichen Items arbeiten, büßen Validität ein, weil das Zielkonstrukt dann nicht in seiner Breite abgedeckt wird.

Leider ist die Validierung eines Tests theoretisch und methodisch sehr viel anspruchsvoller als die Reliabilitätsüberprüfung. Grundsätzlich stehen in traditionellem Verständnis – wie es in der ersten Auflage der internationalen Test-Standards formuliert worden war (APA, AERA, & NCME, 1954) – **drei verschiedene Methoden der Validierung** zur Verfügung: die Inhalts-, Konstrukt- und Kriteriumsvalidierung (◘ Tab. 10.17). Eine gründliche Validierung erfordert in der Regel mehrere empirische Untersuchungen mit dem betreffenden Test, welche dann auch in die Berechnung von statistischen **Validitätskoeffizienten** (typischerweise Korrelationskoeffizienten) münden.

Von der Inhaltsvalidität im Sinne einer Expertenbeurteilung (inklusive ggf. Verbesserungsvorschlägen im Zuge der Testentwicklung; ◘ Abb. 10.17) ist die **Augenschein-Validität** („face validity“) abzugrenzen, die sich darauf bezieht, ob bzw. wie stark Laien bzw. Testpersonen anhand des Testmaterials auf das zu testende Merkmal rückschließen können. Bei hoher Augenscheinvalidität wirkt ein Test auf die Testpersonen plausibel und glaubwürdig, was z. B. die Bereitschaft zur Untersuchungsteilnahme und zu ehrlichem Antworten steigern mag. Augenscheinvalidität kann jedoch problematisch sein, weil damit die Verfälschbarkeit des Tests steigt (mögliche Gegenmaßnahme: „objektive Persönlichkeitstests“ mit geringer Augenscheinvalidität; ▶ Abschn. 10.4.3).

Tabelle 10.17 Das Testgütekriterium der Validität (Gültigkeit) und seine Unterformen

Testgütekriterium	Definition	Überprüfung
Validität (Gültigkeit, „validity“)	Der Testwert misst tatsächlich das Merkmal, das er laut Testbezeichnung bzw. Testbeschreibung zu messen beansprucht und primär kein anderes.	Die Validität eines Tests wird bei der Testentwicklung anhand mindestens eines der drei Validierungsmethoden geprüft. Dabei werden typischerweise Faktorenanalysen und Korrelationskoeffizienten herangezogen.
– **Inhaltsvalidität** (Logische Validität, „content validity“)	Die einzelnen Testitems spiegeln das Zielkonstrukt in seinen inhaltlichen Bedeutungsaspekten vollständig und sinngemäß wider.	Voraussetzung einer Inhaltsvalidierung ist eine genaue Konzeptspezifikation auf der Basis einer Bedeutungs- und/oder Dimensionsanalyse (► Abschn. 8.2). Den inhaltlichen Aspekten des Konstruktes werden dann die einzelnen Testitems zugeordnet, um zu beurteilen, a) ob alle Aspekte des Konstruktes berücksichtigt wurden und b) ob die Items inhaltlich einzeln und in ihrer Gesamtheit die theoretisch vorgegebenen Bedeutungen treffen. Die Inhaltsvalidierung erfolgt theoretisch-argumentativ und gestützt durch Urteile von Fachexperten. Sie wird nicht in einem numerischen Validitätskoeffizienten ausgedrückt.
– **Konstruktvalidität** („construct validity“)	Der Testwert korreliert inhaltlich und theoretisch begründet hypothesenkonform mit anderen theoretischen Konstrukten. – **Konvergente Validität** („convergent validity“): Der zu validierende Test soll mit eng verwandten Konstrukten bzw. konstruktnahen Verfahren relativ hoch korrelieren (auf jeden Fall höher als mit konstruktfremden Verfahren). – **Diskriminante Validität** (divergente Validität; „discriminant validity“, „divergent validity“): Der zu validierende Test soll mit entfernter verwandten Konstrukten bzw. konstruktfremden Verfahren gar nicht oder nur gering korrelieren (auf jeden Fall geringer als mit konstruktnahen Verfahren). Zur Konstruktvalidität gehört auch die Überprüfung der Dimensionalität. – **Faktorielle Validität** („factorial validity“) liegt vor, wenn bei einem mehrdimensionalen Test sich die inhaltlich zu einer Subdimension gehörenden Items jeweils auch empirisch zu einem Faktor bündeln lassen. Im Falle eines eindimensionalen Tests müssen sich sämtliche Items auf einen einzigen Faktor vereinigen lassen.	Eine Konstruktvalidierung ist nur möglich, wenn für das durch den Test erfasste Zielkonstrukt ein theoretisch und/oder empirisch gut fundiertes Netz an Hypothesen (idealerweise ein „ **nomologisches Netz**“) vorliegt, welches Richtung und Enge der Zusammenhänge mit anderen Konstrukten im Sinne konvergenter und diskriminanter Validität vorgibt (Cronbach & Meehl, 1955; Hartig et al. 2012, S. 153f.) Entsprechende Zusammenhangsstrukturen können durch bivariate und v. a. multivariate Verfahren (z. B. multiple Regressionsanalysen) und Strukturgleichungsmodelle geprüft werden. Ein spezieller Ansatz zur Überprüfung der Konstruktvalidität ist die Multtrait-Multimethod-Methode (MTMM-Methode; Abschnitt ► Abschn. 10.4.4). Die faktorielle Validität wird mittels explorativer oder konfirmatorischer Faktorenanalysen untersucht.

Es ist unzulässig, einen Test pauschal für „valide“ zu erklären. Da Testwerte in ganz unterschiedlichen Kontexten verwendet werden, müssen die zur Validierung herangezogenen empirischen Evidenzen auf den jeweiligen Einsatzzweck des Tests bezogen werden, damit man aus ihnen schlüssig ableiten kann, dass eben diese konkreten Testwert-bezogenen Aussagen valide bzw. gültig sind (Hartig et al., 2012). Beispiele: Ein Berufseignungstest, der für diagnostische Zwecke verwendet wird und vor allem dazu dienen soll, den Berufserfolg der Testpersonen vorherzusagen, sollte vor allem eine hohe prognostische Kriteriumsvalidität aufweisen. Ob und wie er mit anderen Konstrukten korreliert ist für valide Testwert-basierte Aussagen zum zukünftigen Berufserfolg irrelevant. Eine Testskala, die Empathie mit Medienpersonen erfasst und zur Theorieprüfung im Rahmen der Medienforschung eingesetzt wird, sollte dagegen vor allem eine hohe Konstruktvalidität aufweisen, denn zur theoriebezogenen Interpretation der Medienempathie-Werte ist entscheidend, ob und wie stark sie mit verwandten Konstrukten zusammenhängen oder sich von ihnen unterscheiden (z. B. Empathie mit realen Personen; Identifikation mit Medienpersonen). Wieder anders gelagert sind die Validitätsanforderungen, wenn ein Schulleistungstest dazu dienen soll, den Wissensstand in einem Schulfach abzuprüfen. Hier ist dann vor allem die Inhaltsvalidität angesprochen,

Tabelle 10.17 (Fortsetzung)

Testgütekriterium	Definition	Überprüfung
– **Kriteriumsvalidität** („criterion validity“)	Der Testwert korreliert positiv mit inhaltlich korrespondierenden manifesten Merkmalen außerhalb der Testsituation **(Außenkriterien),** die für diagnostische Entscheidungen bedeutsam sind. Dabei sind je nach dem Zeitpunkt der Erfassung des Außenkriteriums drei unterschiedliche Varianten der Kriteriumsvalidität zu unterscheiden: – **Retrospektive Validität** („retrospective validity“): Der Kriteriumswert wurde zeitlich vor dem zu validierenden Testwert erhoben (z. B. ein Schulleistungstest wird über zurückliegende Schulnoten validiert). – **Konkurrente Validität** (Übereinstimmungsvalidität, „concurrent validity“): Der Kriteriumswert wird zum gleichen Messzeitpunkt wie der zu validierende Testwert erhoben (z. B. ein Schulleistungstest wird mit den aktuellen Schulnoten validiert). – **Prognostische Validität** (Vorhersagevalidität, prädiktive Validität; „prognostic validity“, „predictive validity“): Der Kriteriumswert wird zeitlich nach dem zu validierenden Testwert erhoben (z. B. ein Berufseignungstest wird über später erhobenen Berufserfolg validiert). Zudem soll der zu validierende Test hinsichtlich seiner (retrospektiven, konkurrenten und prospektiven) Kriteriumsvalidität einem herkömmlichen Test überlegen sein: – **Inkrementelle Validität** („incremental validity“): Der zu validierende Test korreliert enger mit dem betrachteten Außenkriterium als ein herkömmlicher Test bzw. durch Hinzuziehung des neuen Tests kann das Außenkriterium signifikant besser vorhergesagt werden als mit dem herkömmlichen Test.	Eine Kriteriumsvalidierung ist nur möglich, wenn mindestens ein reliables und valides Außenkriterium vorliegt. Als Validitätskoeffizienten dienen Korrelationskoeffizienten. Die inkrementelle Validität wird selten geprüft. Hier kommen die hierarchische multiple Regressionsanalyse oder Strukturgleichungsmodelle zum Einsatz.

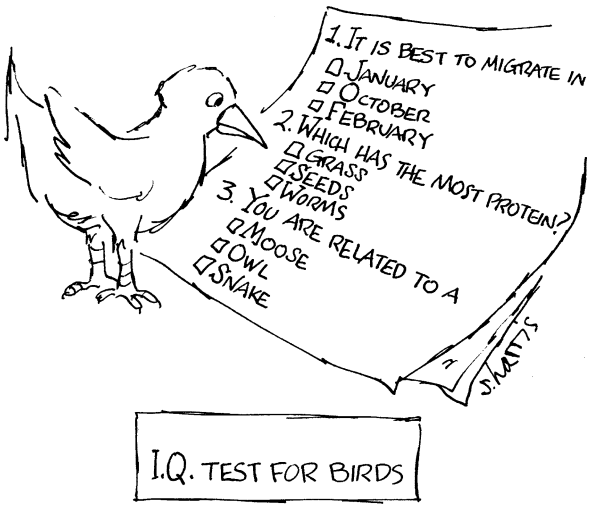


Abb. 10.17 Inhaltsvalidität: Erfassen die Testitems aus Expertensicht die inhaltlich relevanten Aspekte des zu messenden Konstruktes? (© Sidney Harris/Search ID: shr1300, Rights Available from CartoonStock.com)

d. h. die Testaufgaben müssen nachweislich das Spektrum der Inhalte des betreffenden Schulfachs in seiner Breite angemessen abbilden.

Aus validitätstheoretischer Sicht ist somit systematisch zu differenzieren, auf **welche Interpretation eines Testergebnisses sich die Validität beziehen** soll (Kane, 2001; Hartig et al., 2012, S. 144):

1. das **Bewerten** des Testergebnisses,
2. das **Verallgemeinern** des Testergebnisses,
3. das **Extrapolieren** des Testergebnisses auf andere Bereiche,
4. das (kausale) **Erklären** eines Testergebnisses und/oder
5. das Fällen von weiterführenden **Entscheidungen** als Konsequenz aus dem Testergebnis.

Auf der anderen Seite ist zu spezifizieren, auf welche **Quellen von Evidenz** man sich stützt, um die Gültigkeit der Testwerte bzw. der auf sie bezogenen Interpretatio-

nen zu begründen (AERA, APA, & NCME, 1999). Die klassische Dreiteilung der Validität in Inhalts-, Konstrukt- und Kriteriumsvalidität mit ihren jeweils spezifischen Evidenzquellen (Tab. 10.17), lässt sich noch weiter ausdifferenzieren. So unterscheidet Samuel Messick (1995, S. 745) **sechs Evidenzquellen** für die Gültigkeit der Testwerte, wobei er als Oberbegriff für alle die **Konstruktvalidität** wählt:

1. Evidenz auf der Basis der **Testinhalte** (traditionell bezeichnet als Inhaltsvalidität)
2. Evidenz auf der Basis von **Antwortprozessen bei der Testbearbeitung** (z. B. Erfassung der kognitiven Prozesse bei der Bearbeitung von Intelligenztestaufgaben mit der **Methode des lauten Denkens,** ▶ Abschn. 10.2.2)
3. Evidenz auf der Basis der **internen Struktur des Tests** (traditionell bezeichnet als faktorielle Validität)
4. Evidenz auf der Basis der **Verallgemeinerbarkeit der Testergebnisse** auf andere Populationen, Situationen und Aufgaben (z. B. Überprüfung der Anwendbarkeit eines Tests in unterschiedlichen Altersgruppen oder kulturellen Kontexten)
5. Evidenz auf der Basis von **Relationen des Tests zu anderen Merkmalen** (traditionell bezeichnet als konvergente und diskriminante Konstruktvalidität sowie als Kriteriumsvalidität)
6. Evidenz auf der Basis von **Konsequenzen des Testergebnisses** (in traditioneller „Gütekriterien-Sprechweise" bezeichnet als Consequential Validity).

Die letztgenannte Evidenzquelle wird in der Praxis der Testvalidierung bislang kaum systematisch genutzt (Cizek, Rosenberg, & Koons, 2008), spielt aber doch eine wichtige Rolle. Ein Beispiel für sog. **Consequential Validity**, also die individuellen und gesellschaftlichen Folgen der Testanwendung und daraus resultierende Validitätseinbußen, ist das Phänomen des sog. **Teaching-to-the-Test** in Bildungseinrichtungen: Ein Schulleistungstest kann nicht mehr die im Unterricht vermittelten Inhalte messen, wenn – als Konsequenz der regelmäßigen Durchführung von Tests – der Unterricht von den Lehrkräften weitgehend darauf umgestellt wurde, zielgerichtet nur noch auf den bevorstehenden Test vorzubereiten (Hartig et al., 2012, S. 167).

Aufgrund der inhaltlichen **Komplexität des Validitätskonzepts** ist es einleuchtend, dass alle zur Validierung berechneten statistischen Kennwerte – z. B. im Zuge einer Kriteriumsvalidierung wird die Korrelation des Tests mit einem Außenkriterium als **Validitätskoeffizient** herangezogen, oder im Zuge einer diskriminanten Konstruktvalidierung wird die Korrelation des Tests mit dem Testwert eines konstruktfremden Merkmals bestimmt – jeweils kontextspezifisch zu interpretieren sind (▶ Interpretation von Validitätskoeffizienten). Faustregeln, denen gemäß Validitätskoeffizienten > .60 **als hoch** und > .40 **als mittel oder ausreichend** einzustufen sind (Fisseni, 1990; Bühner, 2011, S. 81), sind nur mit großer Vorsicht zu betrachten. Für spezielle Validierungsansätze wie die Multitrait-Multimethod-Methode stehen weitere Hinweise zur Validitätsbewertung zur Verfügung.

Interpretation von Validitätskoeffizienten

Die **Validität** („validity") gibt an, ob ein Test tatsächlich das Merkmal misst, das er zu messen vorgibt. Traditionell werden zur Validitätsprüfung drei empirische Vorgehensweisen und damit auch drei Typen der Validität unterschieden: **Inhalts-, Konstrukt- und Kriteriumsvalidität**. Die Validität ist das wichtigste psychometrische Hauptgütekriterium. Die anderen beiden Hauptgütekriterien Objektivität und Reliabilität sind notwendige (aber nicht hinreichende) Voraussetzungen der Validität.

Aus validitätstheoretischer Sicht wird inzwischen noch genauer differenziert, für **welche Art von Aussagen die Testwerte** genutzt werden und mit welchen **empirischen Evidenzen** genau diese Aussagen bzw. Interpretationen hinsichtlich ihrer Gültigkeit am besten gestützt werden können. Dabei lassen sich neben den Verfahren der klassischen Inhalts-, Konstrukt- und Kriteriumsvalidierung weitere Evidenzen heranziehen, insbesondere auch Evidenzen, die sich auf Prozesse bei der Aufgabenbearbeitung oder auf Konsequenzen der Testanwendung beziehen.

Die statistischen Kennwerte einer Validierung sind kontextspezifisch im Hinblick auf das Konstrukt und die angestrebte Testverwendung zu interpretieren. Allgemeingültige **quantitative Normen** für geringe, ausreichende oder hohe Validität sind nicht angebbar.

Schließlich ist zu beachten, dass die Validität als Gütekriterium eines Messinstruments – bzw. genauer: als Gütekriterium für wissenschaftliche Aussagen auf der Basis von Testwerten – nicht vom Datenerhebungsinstrument (d. h. dem psychologischen Test) allein, sondern von weiteren **Merkmalen der empirischen Studie** abhängt, welche die Testwerte hervorgebracht hat. So können die Testwerte eines an sich qualitativ hochwertigen Tests von zahlreichen Verzerrungen und Verfälschungen betroffen sein bzw. zu falschen Schlussfolgerungen führen, wenn es Probleme bei der Umsetzung des Untersuchungsdesigns gibt und z. B. Versuchsleiter-Erwartungseffekte wirken oder ungewollt auch die Kontrollgruppe durch das zu prüfende Treatment beeinflusst wird. Im Kontext der Behandlung wissenschaftlicher Gütekriterien quantitativer Studien (▶ Abschn. 3.2.2) hatten wir festgehalten, dass sich die Konstruktvalidität in der Campbell-Tradition auf die **Gültigkeit der Interpretation empirischer Daten als Indikatoren für theoretische Konstrukte bezieht** und nicht nur vom Messinstrument, sondern von diversen Merkmalen des Untersuchungsdesigns und seiner Umsetzung abhängt.

Tabelle 10.18 Sieben Nebengütekriterien psychologischer Tests

Nebengütekriterium	Definition	Überprüfung
Skalierung bzw. **Skalierbarkeit** („scalability“)	Ein Test erfüllt das Kriterium der Skalierung, wenn der Testwert durch eine gültige Verrechnungsvorschrift aus den Test-Items gebildet wird, d. h. der Testwert muss als echter Messwert die empirischen Merkmalsausprägungen adäquat abbilden.	In der Klassischen Testtheorie werden die Items meist zu einem ungewichteten Summen- oder Durchschnittsscore zusammengefasst, ohne Prüfung der messtheoretischen Eigenschaften (zur Messtheorie ▶ Abschn. 8.4.1). Demgegenüber wird die Skalierbarkeit aller Items in der Probabilistischen Testtheorie bzw. Item-Response-Theorie empirisch genau geprüft und nur modellkonforme Items werden verwendet (▶ Abschn. 10.4.4).
Normierung/Eichung („standardization“)	Ein Test ist normiert, wenn aktuelle Testnormen (durchschnittliche Testergebnisse repräsentativer Vergleichsstichproben) vorliegen, die eine Einordnung individueller Testwerte erlauben (normorientiertes Testen). Der Testrohwert wird anhand der Normwerte in einen standardisierten Testwert überführt (z. B. z-Wert, Prozentrang, Stanine-Wert etc.).	Die Normierung oder Eichung eines Tests anhand von repräsentativen Eichstichproben (Daumenregel $n > 300$) unterschiedlicher Bevölkerungsgruppen (typischerweise Geschlechts-, Alters- und Bildungsgruppen) sollte laut DIN 33430 spätestens alle acht Jahre aktualisiert werden.
Testökonomie („economical testing“)	Ein Test ist ökonomisch wenn er in Relation zum Erkenntnisgewinn – eine kurze Durchführungszeit beansprucht, – wenig Material verbraucht, – einfach zu handhaben ist, – als Gruppentest durchführbar und – schnell und bequem auszuwerten ist.	Die Einschätzung der Ökonomie kann nur im Vergleich zwischen unterschiedlichen Parallel-Tests erfolgen. Generell sollte die Aussagekraft eines Tests höher gewichtet werden als die Aufwandsminimierung. Ein längerer, aber validerer Test ist deswegen z. B. einem kürzeren, aber weniger validen Test vorziehen. Besonders ökonomisch sind computergestützte Tests.
Nützlichkeit („utility“)	Ein Test ist nützlich, wenn er ein für Praxis und/oder Forschung relevantes Merkmal misst, für das bislang überhaupt kein Test oder nur ein Test mit beschränkter Testgüte vorlag.	Die Relevanz des gemessenen Konstruktes ist inhaltlich zu begründen. Die Nützlichkeit eines Tests kann jederzeit schwinden, wenn alternative Tests entwickelt werden.
Zumutbarkeit („reasonableness“)	Ein Test ist zumutbar, wenn er die Testpersonen in Relation zum Erkenntnisgewinn nicht übermäßig in – zeitlicher, – körperlicher und/oder – psychischer Hinsicht belastet.	Die Einschätzung der Zumutbarkeit kann nur im Vergleich zwischen unterschiedlichen Parallel-Tests erfolgen. Es handelt sich im Grunde um ein ethisches Kriterium, da es die Forschungsethik grundsätzlich verbietet, Untersuchungsteilnehmer über Gebühr zu belasten oder gar zu beeinträchtigen (▶ Abschn. 4.1).
Nicht-Verfälschbarkeit („non-fakeability“)	Ein Test ist nicht-verfälschbar, wenn es Testpersonen kaum gelingen kann, absichtlich ein besonders gutes oder besonders schlechtes Testergebnis zu erzeugen, das nicht als unplausibel oder gefälscht auffällt.	Der Grad der Verfälschbarkeit von Tests wird durch sog. Faking-Good- und Faking-Bad-Instruktionen empirisch untersucht und in der Testbeschreibung bzw. Testrezension angegeben.
Testfairness („test fairness“)	Ein Test ist fair, wenn er allen Personengruppen, für die er anwendbar sein soll, gleiche Chancen bietet bzw. wenn er keine Testpersonen systematisch aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit benachteiligt. Das ethische Gütekriterium der Testfairness wird in den aktuellen „Standards for Educational and Psychological Testing“ in seiner Bedeutung gleichrangig wie Reliabilität und Validität als notwendiges Fundament jeden Tests betrachtet.	Kulturfaire bzw. kulturfreie Intelligenztests sollen z. B. die Intelligenzleistung unabhängig von der Schulbildung erfassen. Das Ausmaß der Fairness bzw. Unfairness wird durch Gruppen-Vergleiche empirisch untersucht. Die Testfairness einzelner Items kann im Rahmen der Probabilistischen Testtheorie z. B. über das Rasch-Modell mittels einer sog. DIF-Analyse („differential item functioning“) geprüft werden (▶ Abschn. 10.4.4).

Nebengütekriterien psychologischer Tests

Abgesehen von den drei Hauptgütekriterien psychometrischer Tests existieren noch **sieben Nebengütekriterien** (Bühner, 2011; Moosbrugger & Kelava, 2012a), die im Folgenden kursorisch vorgestellt werden (Tab. 10.18).

Dabei stellt die **Normierung bzw. Eichung** – also die Bereitstellung von bevölkerungsrepräsentativen Vergleichswerten zur Einschätzung individueller Testwerte – das charakteristische Abgrenzungskriterium zwischen einem psychologischen Test, der auch für die Diagnostik nutz-

bar ist, und einem standardisiertem Fragebogen, der nur forschungsbezogene Auswertungen auf Gruppenebene erlaubt, dar. Es ist zu beachten, dass Fragen der Testgüte – also Definition, Gewichtung und Prüfung von Haupt- und Nebengütekriterien – fortlaufender kritischer Diskussion und Überarbeitung unterliegen. Eine zentrale internationale Referenz sind dabei die **Standards for Educational and Psychological Testing**, die seit 1954/1955 von drei führenden Fachgesellschaften herausgegeben werden, der American Educational Research Association (AERA), der American Psychological Association (APA) und dem National Council on Measurement in Education (NCME). Die vierte Fassung (AERA, APA, & NCME, 1999) wurde erneut überarbeitet und bezieht mit der 2014 neu erschienenen Auflage jetzt u. a. technische Aspekte des computerbasierten Testens stärker mit ein und hat zudem das ethische Kriterium der Testfairness aufgewertet, das jetzt neben Reliabilität und Validität als drittes Fundament jeden psychologischen Tests gilt (http://www.apa.org/science/programs/testing/; zur Forschungsethik ► Abschn. 4.1).

10

10.4.2 Projektive Tests

Projektive Testverfahren sind nicht- oder teilstrukturierte Verfahren. Sie können somit auch zu den **qualitativen Datenerhebungsmethoden** gezählt werden, da die Testpersonen keine vollstandardisierten Testinstrumente ausfüllen, bei denen die einzelnen Aufgaben sowie die Antwortmöglichkeiten vorgegeben sind, sondern sich **frei in eigenen Worten** – oder auch in Form von Zeichnungen – äußern. Dazu wird ihnen zum Beispiel in der Gruppe der **Formdeuteverfahren** vieldeutiges Reizmaterial vorgelegt (z. B. Tintenkleckse), das die Testpersonen dann in eigenen Worten deuten sollen. Projektive Testverfahren gehen auf der Basis **psychoanalytischer Theorie** (Freud, 1911) davon aus, dass die Testpersonen bei der Interpretation der uneindeutigen Reizvorlagen unbewusst ihre eigenen Motive, Wünsche, Konflikte, Einstellungen, Eigenschaften etc. in das Material „hineinprojizieren". Was eine Person in einem amorphen Tintenklecks zu sehen meint (z. B. einen Schmetterling, ein kopulierendes Paar, ein Tierfell mit Blutflecken), soll dementsprechend Rückschlüsse auf die Persönlichkeit zulassen. Projektive Testverfahren lassen sich gemäß der **Merkmale der Testaufgaben** in drei Gruppen einteilen, neben den **Formdeuteverfahren** existieren **verbal-thematische Verfahren** sowie **zeichnerische und Gestaltungsverfahren**.

Da das **psychoanalytische Projektions-Konzept**, das den Testverfahren ihren Namen verleiht, teilweise unterschiedlich verwendet und kontrovers diskutiert wird, schlagen manche Autoren vor, statt von projektiven Tests lieber von **Persönlichkeits-Entfaltungstests** zu sprechen (Brähler et al., 2002a, S. XII), um zum Ausdruck zu bringen, dass diese Tests nicht auf die Messung eines einzelnen Merkmals hinauslaufen, sondern durch wenig strukturierte Reize oder globale Aufgabenstellungen den Testpersonen viel Freiraum lassen, um sich individuell auszudrücken bzw. ihre Persönlichkeit in mehreren Fassetten zur Entfaltung kommen zu lassen.

Projektive Testverfahren – Projektive Testverfahren (Persönlichkeits-Entfaltungstests, „projective tests") basieren auf der psychoanalytischen Annahme, dass Testpersonen unbewusste psychische Prozesse in eine freie Interpretation von uneindeutigen Reizvorlagen (in der Regel Bildmaterial) oder in die freie Gestaltung von Geschichten und Zeichnungen hinein projizieren. Projektive Testverfahren sind unstrukturierte bzw. teilstrukturierte Verfahren, die mit offenen Aufgabenstellungen den Testpersonen viel Spielraum lassen, um ihre Persönlichkeit darzustellen. Das von den Testpersonen produzierte nicht-numerische Datenmaterial wird von geschulten Testanwendern ausgewertet, um Rückschlüsse auf die Persönlichkeit der Testpersonen zu ziehen.

Die Auswertung eines projektiven Testverfahrens besteht im Grunde in einer spezifischen **qualitativen Datenanalyse** (► Abschn. 12.1): Die verbalen Beschreibungen des Reizmaterials (z. B. Beschreibung von Tintenklecksen) oder die selbst erstellten Zeichnungen (z. B. eines Baumes) werden hinsichtlich bestimmter Merkmale nach gewissen Regeln analysiert und interpretiert, um daraus Rückschlüsse auf die Persönlichkeit der Testperson zu ziehen. Die qualitativ-interpretierende Testauswertung ist somit sehr aufwändig und nur von ausgesprochen erfahrenen Psychologinnen und Psychologen zu leisten. Dennoch schneiden die Ergebnisse projektiver Tests typischerweise schlecht ab, wenn man z. B. ihre Reliabilität oder Validität empirisch prüft. Deswegen wurden teilweise Ansätze entwickelt, um die qualitativen Daten, die bei der Anwendung projektiver Tests entstehen, einer **standardisierten Auswertung** zu unterziehen (z. B. das Exner-Coding-System für den Rorschach-Test; Exner, 1995). Doch auch die psychometrische Qualität dieser Verfahren wird teilweise kritisch gesehen. Zudem kann man argumentieren, dass eine nachträgliche Standardisierung und quantifizierende Auswertung der ursprünglichen Intention eines projektiven Testverfahrens und seiner Verwurzelung in der psychoanalytischen Theorie widerspricht.

In der sozial- und humanwissenschaftlichen Forschung nehmen projektive Testverfahren heute eine Sonderrolle ein. Sie werden in der Praxis teilweise noch eingesetzt, allerdings nie als einzige Informationsquelle, sondern oft als Explorationshilfe, z. B. bei der Diagnostik oder Forschung mit Kindern.

Formdeuteverfahren

Bei **Formdeuteverfahren** wird den Testpersonen ein relativ unstrukturiertes, **vieldeutiges Reizmaterial** (z. B. Tintenkleckse) vorgelegt, das von ihnen beschrieben und gedeutet werden soll. Die Beschreibungen werden als Indikatoren latenter Persönlichkeitsmerkmale aufgefasst und qualitativ (zuweilen auch quantitativ) ausgewertet.

Der wohl in der breiten Bevölkerung bekannteste projektive psychologische Test aus der Gruppe der Formdeuteverfahren ist der **Rorschach-Test** (RO-T; „rorschach inkblot test"), der die Testpersonen zur Deutung von **zehn verschiedenen Tintenklecksmustern** auffordert (Abb. 10.18) und von dem Schweizer Psychoanalytiker Herrmann Rorschach in den 1920er-Jahren entwickelt wurde (Rorschach, 1927). Die **Testdurchführung** besteht in zwei Phasen: Zunächst betrachten die Testpersonen der Reihe nach jeden der 10 Tintenklecks-Tafeln und geben jeweils mündlich in eigenen Worten an, was sie sehen (freie Assoziations-Phase). Anschließend werden alle Tafeln erneut durchgegangen und die Testpersonen werden zu ihren ursprünglichen Deutungen befragt, insbesondere geht es darum, wo sie welches Detail im Klecksmuster gesehen haben (Befragungs-Phase). Bei der Testdurchführung notiert der Psychologe bzw. die Psychologin alle Äußerungen der Testperson.

Die **Testauswertung** erfolgt **qualitativ** in Form einer Interpretation der Äußerungen der Testpersonen oder **quantitativ** anhand eines in den 1960er-Jahren entwickelten standardisierten Systems, des sog. Exner Scoring Systems (Exner, 1995). Bei der Auswertung spielen neben dem **Inhalt** des Gesehenen (z. B. Mensch oder Tier) vier weitere Aspekte eine wichtige Rolle: die **Lokalisierung** (welche Teile der Tafel wurden zur Deutung herangezogen), die **Determinanten** (welche Merkmale der Tafel wurden beachtet, z. B. Form, Farbe, Schattierung), die **Häufigkeit** (z. B. originelle versus häufige Deutungen) sowie sonstige **besondere Phänomene** (z. B. sehr langes Überlegen vor dem Antworten, mehrfaches Drehen der Tafel etc.). Sowohl die nicht-standardisierte Testdurchführung als auch die Testauswertung erfordern eine gründliche Schulung und viel Erfahrung. Der RO-T ist also kein Verfahren, das im Rahmen einer eigenen Studie einfach eingesetzt werden könnte.

Haupteinsatzfelder des Rorschach-Tests sind klinische Forschung und Praxis sowie Persönlichkeitsdiagnostik und auch die Forensik (z. B. Begutachtung von Straftätern). Die führende Fachzeitschrift heißt *Rorschachiana: Journal of the International Society for the Rorschach*. Aktuelle Studien, die den Rorschach-Test teils interpretativ, teils standardisiert auswerten, befassen sich z. B. mit der Tauglichkeit des Verfahrens zur Diagnose posttraumatischer Belastungsstörungen (Arnon, Maoz, Gazit, & Klein, 2011), zur Differenzierung unterschiedlicher Formen von Essstörungen (Guinzbourg, 2011), zur Vorhersage von vorzeitigem Therapieabbruch (Charnas, Hilsenroth, Zodan, & Blais, 2010), zur Messung von Aggressivität (Katko, Meyer, Mihura, & Bombel, 2010) oder zur Erfassung der interpersonalen Verbundenheit von Kindern (Rivera & Viglione, 2010).



Abb. 10.18 Tafel 1 des Rorschach-Tests. (Aus Hermann Rorschach, Rorschach©-Test. Psychodiagnostik Tafeln. © Verlag Hans Huber, Hogrefe AG, Bern, Schweiz, 1921, 1948, 1994, 2009)

Die Testgüte (v. a. die **Validität**) des Rorschach-Tests wird auf der Basis vorliegender Einzelstudien und Metaanalysen nach wie vor **kontrovers diskutiert** (vgl. Garb, Wood, Nezworski, Grove, & Stejskal, 2001), teilweise wird vom Einsatz dieses Tests abgeraten. Validitätsstudien stammen u. a. aus dem klinischen und forensischen Bereich (für eine forensische Fallstudie s. z. B. Weizmann-Henelius, Kivilinna & Eronen, 2010; und für eine Metaanalyse z. B. Wood et al., 2010) sowie aus dem Bereich der Personalauswahl (z. B. Identifikation von Führungskräften; Del Giudice, 2010; Kottke, Olson & Shultz, 2010).

Verbal-thematische Verfahren

Bei den **verbal-thematischen** Verfahren liefern die Testpersonen mehr oder minder umfangreiche Erzählungen, zu denen sie durch **themenbezogenes Reizmaterial** (z. B. Erzählaufforderung, Fotos, Satzanfänge) angeregt werden.

Das bekannteste Verfahren aus dieser Testgruppe ist der **Thematische Apperzeptionstest TAT** („thematic apperception test"). Er wurde in den 1940er-Jahren von dem amerikanischen Psychologen Henry Murray (1943) zusammen mit Christiana Morgan an der Harvard Universität entwickelt und dient der Erfassung von Persönlichkeitsmerkmalen, Bedürfnissen und Motiven (für eine ausführliche Darstellung des Verfahrens s. Cramer, 2004). Den Testpersonen werden als Stimulus-Material meist 10



Abb. 10.19 Zeichnung von Christiana Morgan aus dem TAT. (Aus Murray, 1938, S. 406. Published in Explorations in personality: A clinical and experimental study of fifty men of college age. By permission of Oxford University Press, www.oup.com)

verschiedene **Bildtafeln** vorgelegt, auf denen **Menschen in alltäglichen oder ungewöhnlichen Situationen** zu sehen sind (Abb. 10.19). Die Testpersonen werden aufgefordert, zu jedem Bild eine kurze (ca. fünf Minuten) **Geschichte** zu erzählen, aus der hervorgeht, was zu der auf dem Bild gezeigten Situation führte, was gerade geschieht, was die beteiligten Personen fühlen und denken und wie die Geschichte ausgeht. Während im Rorschach-Test zu unstrukturierten Tintenklecksen Beschreibungen (was sehen Sie?) erfragt werden, geht es im TAT anhand themengebundener szenischer Bilder also ausdrücklich um Narrationen.

Dem TAT liegt ebenso wie dem RO-T die Theorie zugrunde, dass Menschen unbewusst ihre eigenen Motive, Konflikte und Persönlichkeitseigenschaften in die Bilder hineinprojizieren. Die **qualitative Auswertung** der erzählten und vom Testleiter notierten – oder von den Testpersonen selbst aufgeschriebenen (die Methode ist auch schriftlich anwendbar) – Geschichten läuft darauf hinaus, formale und inhaltliche Aspekte der Erzählungen (z. B. wer ist der Held, die Heldin der jeweiligen Geschichte, wie werden die sozialen Beziehungen zwischen den dargestellten Personen beschrieben, wie geht die Geschichte aus) mit Bezug zur Biografie der Testperson psychologisch zu interpretieren.

Angesichts methodischer Kritik an der qualitativen Auswertung wurden auch für den TAT verschiedene quantitative **standardisierte Kodierungsverfahren** entwickelt, z. B. das Defense Mechanisms Manual (DMM; Cramer, 1991; zur Validierung s. Porcerelli, Cogan, Kamoo & Miller, 2010) und die Social Cognition and Object Relations SCOR Scale (Westen, 1991).

Der TAT wird primär in zwei Forschungsfeldern eingesetzt: Zum einen in psychoanalytischer Tradition in der **klinischen Forschung und Praxis** (z. B. Whipple & Fowler, 2011) und zum anderen sozialpsychologisch ausgerichtet im Bereich der **Leistungs- und Motivationsforschung** (z. B. Spangler, 1992). Ebenso wie beim Rorschach-Test gibt es auch beim TAT Kontroversen um die Validität des Verfahrens (z. B. Keiser & Prather, 1990) sowie um angemessene Validierungsverfahren im Sinne einer Anwendung psychometrischer Kriterien (Cramer, 2004). Die anhaltende Popularität der Methode hängt möglicherweise auch damit zusammen, dass narrative Verfahren und narrative Analysen insgesamt in der qualitativen Sozialforschung an Bedeutung gewinnen. Zudem wird das Grundprinzip der TAT-Methode auch außerhalb von diagnostischen Testsettings intensiv eingesetzt, etwa im Rahmen **projektiver Fragebögen** (Hofer, Busch, Bender, Ming, & Hagemeyer, 2010). Eine verwandte Methode ist die **Vignetten-Technik**, sofern sie zur Generierung von Geschichten eingesetzt wird (Barter & Renold, 1999).

Neben dem TAT existieren noch andere **narrative projektive Test-Verfahren** (Teglasi, 2010), z. B. die auf Kinder zugeschnittenen Tests **Roberts-2** und **TEMAS** (Tell-Me-a-Story).

Zeichnerische und Gestaltungsverfahren

Bei **zeichnerischen Verfahren** werden die Testpersonen aufgefordert, Zeichnungen zu erstellen. Dabei können entweder formale Vorgaben gemacht oder Themen vorgegeben werden. Ein Beispiel für einen projektiven Test mit a-thematischen formalen Vorgaben ist der **Wartegg-Zeichen-Test (WZT)**, der von dem Leipziger Psychologen Ehrig Wartegg entwickelt wurde (Wartegg, 1939). Die Testpersonen erhalten acht Vorlagen, auf denen bestimmte Formen (z. B. Punkte, Striche, Kreise) bereits eingetragen sind. Jede Vorlage muss zu einer Zeichnung vervollständigt werden.

Ein Beispiel für einen thematischen Zeichentest ist der von dem Heilpädagogen Hermann Ziler (1997) entwickelte **Mann-Zeichen-Test (MZT)**, bei dem die Testpersonen (Kinder zwischen 4 und 14 Jahren) aufgefordert werden, einen Mann zu zeichnen. Die Auswertung erfolgt hierbei standardisiert, indem für einzelne gezeichnete Details (z. B. Haare, Mund, Zähne) Punkte vergeben und zu einem Gesamtwert verrechnet werden, der Rückschlüsse auf Entwicklungsstand, Schulreife sowie besonderen Förderbedarf des Kindes erlauben soll. Ein weiterer thematischer projektiver Zeichentest für Kinder ist der **Baum-Test** (Koch, 2003), bei dem die Testpersonen sich einen Baum vorstellen und diesen zeichnen sollen. Der Baum-

test wird als diagnostisches Hilfsmittel z. B. im Bereich der Heilpädagogik, Schulpsychologie oder Kinderpsychiatrie eingesetzt. Deutungen können sich dabei z. B. auf die Ausprägungen der Wurzeln, des Stammes und der Krone des Baumes beziehen. Die Validität des Baum-Tests ist strittig.

Andere Gestaltungsverfahren, die z. B. im Rahmen der Diagnostik und in der Psychotherapie mit Kindern, aber auch mit Erwachsenen eingesetzt werden, arbeiten mit Puppen und anderen Gegenständen, die von den Testpersonen zur Darstellung von Szenen verwendet werden. Sehr bekannt ist der 1938 von der Ärztin Gerdhild von Staabs entwickelte **Scenotest** (Staabs, 1997), der Hinweise auf absichtlich verschwiegene oder den Testpersonen nicht bewusste Konflikte aus ihrem Beziehungs- und Familienleben liefern soll. Das Testmaterial besteht aus biegbaren Puppenfiguren, die z. B. Mutter, Vater, Großeltern, Kinder, Arzt etc. repräsentieren, ergänzt um Tiere, Fahrzeuge, Lebensmittel, Alltagsgeräte (z. B. Fernseher, Schrubber), Symbolfiguren (z. B. Engel, Prinzessin) etc. Das an eine Puppenstube erinnernde Testmaterial hat einen hohen Aufforderungscharakter und lädt förmlich zum Spielen ein. In der ersten Phase des Tests werden die Testpersonen gebeten, einfach etwas mit dem Material zu bauen und Bescheid zu geben, wenn sie fertig sind. In der zweiten Phasen sollen sie dann erläutern, was die gestaltete Szene zu bedeuten hat. Gemäß der Logik des projektiven Testens sollen in den gestalteten Szenen unbewusste Konflikte, Wünsche, Ängste, Beziehungsmuster etc. zu Tage treten. Für eine speziellere Exploration könnte man das Kind auch bitten, mit dem Material zunächst die eigene Familie – so wie sie ist – darzustellen, und danach die Szene so zu bauen, wie das Kind sich die Familie wünschen würde. In der Kinder- und Jugendpsychiatrie in Deutschland kommt neben psychometrischen Tests (allen voran Intelligenztests) auch der projektive Scenotest recht oft zum Einsatz (vgl. Bölte, Adam-Schwebe, Englert, Schmeck, & Poustka, 2000).

Analog der Empfehlung für die diagnostische Praxis, projektive Verfahren – ergänzend zu anderen Testverfahren – eher **explorativ und hypothesenbildend einzusetzen** (vgl. Schaipp & Plaum, 2000), ist auch ihre Verwendung in der Forschung unter Vorbehalt zu sehen. Selbst wenn die notwendige Expertise im Forschungsteam vorhanden ist, um die projektiven Tests durch geschulte Psychologinnen und Psychologen sachgerecht anzuwenden und auszuwerten (was naturgemäß anspruchsvoller ist als bei vollstandardisierten Tests), bleibt eben oft unklar, wie aussagekräftig die Befunde sind. Es werden in der Forschung aber auch immer wieder Versuche unternommen, die Logik projektiver Tests fruchtbar zu machen, etwa wenn Szenen, die Spielerinnen und Spieler im Computergame „Die Sims 2“ gestalten, analog einem Scenotest interpretiert werden (Griebel, 2006).

10.4.3 Psychometrische Tests

Bei der Klassifizierung der psychometrischen Testverfahren steht die Art des zu messenden Merkmals im Mittelpunkt. So existiert einerseits die Gruppe der **Leistungstests**, in denen die Testpersonen ihre maximalen Fähigkeiten unter Beweis stellen müssen. Andererseits existiert die Gruppe der **Persönlichkeitstests**, in denen die Testpersonen über ihre persönlichen Eigenschaften, Einstellungen, Vorlieben etc. Selbstauskünfte erteilen und dabei ihr typisches Erleben und Verhalten darstellen sollen. Eine Sondergruppe innerhalb bzw. neben den Persönlichkeitstests bilden die **klinischen Tests**, in denen die Testpersonen um Selbstauskünfte über Symptome psychologischer Störungen und Probleme gebeten werden. Jede dieser drei Testgruppen lässt sich noch in einige Untergruppen ausdifferenzieren, die jeweils spezielle Merkmale bzw. Merkmalsbündel erfassen. Im Folgenden werden die einzelnen Testgruppen kurz vorgestellt.

Leistungstests

In **Leistungs-** bzw. **Fähigkeitstests** soll die maximale Leistungsfähigkeit herausgefunden werden. Um Leistungsfähigkeit und Leistungsgrenzen der Testpersonen zu ermitteln, wird entweder die Bearbeitungszeit bewusst knapp bemessen (**Speed-Test**) und/oder das Niveau der Aufgaben sukzessive gesteigert (**Power-Test**). Die Gruppe der Leistungstests lässt sich weiter unterteilen in (Amelang & Schmidt-Atzert, 2006; Brähler et al., 2002a): 1.) allgemeine Leistungstests, 2.) Intelligenztests, 3.) Entwicklungstests, 4.) Schultests und 5.) spezielle Funktionsprüfungs- und Eignungstests.

Leistungs- bzw. Fähigkeitstests – Leistungs- bzw. Fähigkeitstests („achievement tests“) sind vollstrukturierte psychologische Testverfahren, mit denen Leistungsmerkmale (Ausprägungen von Fähigkeiten und Fertigkeiten) gemessen werden. Diese Tests zeichnen sich dadurch aus, dass Testaufgaben zu bearbeiten sind, für die es objektiv richtige und falsche Lösungen gibt und mittels derer die Testpersonen ihre maximalen Fähigkeiten unter Beweis stellen sollen. Als Leistungsindikatoren werden z. B. Lösungsmenge (Anzahl richtiger Lösungen) und Lösungsgüte (Qualität der Lösungen) herangezogen. Leistungstests werden meist unter Zeitdruck durchgeführt (Speed-Test) und/oder arbeiten mit sukzessiver Steigerung der Aufgabenschwierigkeit (Power-Test).

Leistungstests basieren nicht auf Selbstauskunfts-Items, sondern auf Testaufgaben, die objektiv richtig oder falsch gelöst werden können. Bevor wir auf die fünf Unterformen von Leistungstests eingehen, sollen zunächst die

speziellen Item-Varianten in Leistungstests vorgestellt werden.

Item-Varianten in Leistungstests

In Anlehnung an Rütter (1973) wird zwischen Test-Items mit offener Beantwortung, mit halboffener Beantwortung und mit Antwortvorgaben unterschieden: offene, halboffene und geschlossene Items. Diese drei Arten von Test-Items lassen sich in Leistungstests mit unterschiedlichen Antwortmöglichkeiten kombinieren (▶ Antwortmodalitäten für Test-Items in Leistungstests).

Offene Items Items mit offener Beantwortung überlassen es der Testperson, wie sie die gestellte Aufgabe löst. Die Aufgabenlösung kann verbal (oder auch spielerisch oder bildnerisch) frei gestaltet werden, sie kann die Auslegung, Interpretation oder Deutung bestimmter Reizvorlagen bzw. freie Assoziationen zu sprachlichen, optischen oder akustischen Reizen fordern. Die offene Aufgabenstellung und auch die wenig geregelte Auswertung widersprechen der Logik psychometrischer Test. Sie kommen deswegen in Leistungstests nicht vor, sind jedoch im Rahmen der Vorbereitung einer Testkonstruktion nützlich, etwa um halboffene oder geschlossene Items zu entwickeln und dafür typische Antworten von Testpersonen zu kennen.

Halboffene Items Auch halboffene Items überlassen die Antwortformulierung der Testperson; die gestellte Aufgabe sollte jedoch im Unterschied zu einem offenen Item so präzise sein, dass **nur eine Antwort richtig** ist. Erst dann lässt sich ein Test mit halboffenen Items vollständig objektiv auswerten. Üblicherweise bereitet die Auswertung halboffener Items jedoch Probleme. Oftmals sind es nur Formulierungsnuancen, die bei der Auswertung zweifeln lassen, ob die Testperson tatsächlich die richtige Antwort meinte. Mit unterschiedlichen Punktbewertungen versucht man dann auch „weniger richtigen" Antworten gerecht zu werden (zur Gewichtungsproblematik vgl. Stanley & Wang, 1970). Dennoch muss man bei Tests mit halboffenen Items meistens Objektivitätseinbußen (Auswertungsobjektivität) in Kauf nehmen. Man kann bei Items mit halboffener Beantwortung verschiedene Konstruktionsformen unterscheiden (▶ Antwortmodalitäten für Test-Items in Leistungstests): Einfachantworten (eine Frage und eine Antwort), Mehrfachantworten (mehrere Fragen und mehrere Antworten), Reihenantworten (eine Frage und mehrere Antworten) sowie Sammelantworten (mehrere Fragen und eine Antwort).

Antwortmodalitäten für Test-Items in Leistungstests

1. **Items mit offener Beantwortung**
   - a) Freie Gestaltung
     Beispiel: Was halten Sie von Horoskopen? Begründen Sie Ihre Ansicht!
   - b) Freie Deutung
     Beispiel: Was sagt Ihnen dieses Röntgenbild?
   - c) Freie Assoziation
     Beispiel: Bilde möglichst viele Sätze zu folgenden Wortanfängen: H-H-G-V
2. **Items mit halboffener Beantwortung**
   - a) Einfachantworten
     Beispiel: Was versteht man unter dem Begriff „Metamorphose"?
   - b) Mehrfachantworten
     Beispiel: An welchen Flüssen liegen die folgenden Städte?

Ingolstadt	.....	Nürnberg	.....
Hameln	.....	Heilbronn	.....
Emden	.....	Hannover	.....

   - c) Reihenantworten
     Beispiel: Welche Holzblasinstrumente sind Dir bekannt?
   - d) Sammelantworten
     Beispiel: Welches deutsche Verb trifft mehr oder weniger präzise auf die folgenden Vokabeln zu: „to test", „examine", „try", „inspect", „investigate", „audit", „check".
3. **Items mit Antwortvorgaben**
   - a) Alternativantworten
     Beispiel: Unter Anamnese versteht man die Vorgeschichte einer Erkrankung.
     Richtig ☐ Falsch ☐
   - b) Auswahlantworten
     Beispiel: Ein Grundstück ist 48 m breit und 149 m lang und kostet € 79 400. Was kostet ein Quadratmeter?
     - A: addiere und multipliziere
     - B: multipliziere und dividiere
     - C: subtrahiere und dividiere
     - D: addiere und subtrahiere
     - E: dividiere und addiere
   - c) Umordnungsantworten
     Beispiel: Ordne – mit dem kleinsten beginnend – die folgenden Brüche nach ihrer Größe!
     A: $\frac{4}{9}$ B: $\frac{3}{4}$ C: $\frac{2}{3}$ D: $\frac{7}{12}$ E: $\frac{5}{6}$
   - d) Zuordnungsantworten
     Beispiel: Welches Verb gehört zu welchem Substantiv?

A: einen Vortrag	a) erzählen
B: eine Geschichte	b) machen
C: eine Erklärung	c) halten
D: ein Gespräch	d) abgeben
E: einen Vorschlag	e) führen

   - e) Ergänzungsantworten
     Beispiel: Blitz verhält sich zu Hören wie Donner zu .....
     a) Gewitter, b) Sehen, c) Regen, d) Fühlen, e) Wolken.

Geschlossene Items Die dritte Kategorie (Items mit Antwortvorgaben) ist in der modernen Testkonstruktion vorherrschend. Items mit **Alternativantworten** (z. B. richtig, falsch) sowie **Auswahlantworten** (Multiple Choice Items) mit mehr als zwei Antwortalternativen sind vorherrschend. Bei Multiple-Choice-Fragen ist noch zu differenzieren, ob **nur eine Antwortoption richtig** ist („single answer question"; manchmal auch: „single choice") oder ob mehrere der vorgegebenen Antwortmöglichkeiten zutreffend sind („multiple answer question"). Da für die Testauswertung genau bekannt ist, welche Antwortmöglichkeit/en richtig ist/sind, bereitet die Auswertung keine Schwierigkeiten: Tests, die aus Items mit Antwortvorgaben bestehen, sind **auswertungsobjektiv**, d. h., sie ermöglichen eine intersubjektiv eindeutige Auswertung. Für Multiple-Choice-Aufgaben sind rund drei Antwortvorgaben optimal (Bruno & Dirkzwager, 1995; Rogers & Harley, 1999).

Das Auffinden geeigneter Alternativantworten ist oftmals ein mühsames, zeitaufwändiges Unterfangen. Die verschiedenen Antwortmöglichkeiten müssen so geartet sein, dass eine uninformierte Testperson sämtliche Antwortoptionen mit möglichst gleicher Wahrscheinlichkeit für richtig hält, d. h., sie müssen die Aufmerksamkeit der Testperson von der richtigen Antwortalternative ablenken bzw. „zerstreuen". Erfüllen die Antwortvorgaben diese Forderung, bezeichnet man sie als **gute Distraktoren**. Die Konstruktion geeigneter Distraktoren macht erhebliche empirische Vorarbeiten (wie z. B. Pretests mit offenen Items) erforderlich (einen formalen Ansatz zur Auswahl von Distraktoren beschreibt Wilcox, 1981; weitere Hinweise zu diesem Thema findet man bei Green, 1984, bzw. Haladyna & Downing, 1990a, 1990b). Bei häufiger Verwendung eines Tests kommt neben der Auswertungsobjektivität ein weiterer Vorteil geschlossener Items zum Tragen: die Ökonomie im Sinne geringen Zeit- und Kostenaufwandes für die Testauswertung (maschinelle Auswertung über Belegleser, Auswertung mit Schablonen oder computergestützte Auswertung).

Diesen Vorteilen geschlossener Items stehen allerdings einige Nachteile gegenüber: Multiple-Choice-Fragen fordern von Testpersonen jeweils eine Wiedererkennungsleistung, die gegenüber der Reproduktionsleistung bei halboffenem Antwortformat als weniger anspruchsvolle Fähigkeit anzusehen ist. Ein weiteres Problem liegt eher in den Merkmalen der Testpersonen und betrifft damit die Testfairness: Manche Testpersonen haben mehr Mut zum Raten und können deshalb höhere Punktwerte erzielen als Testpersonen, die nur dann eine Antwortvorgabe ankreuzen, wenn sie von deren Richtigkeit überzeugt sind (► Maßnahmen zur Kontrolle und Korrektur des Ratens in Leistungstests). Eine ausführlichere Kritik von Multiple-Choice-Items findet man bei Kubinger und Wolfsbauer (2010).

Bei **Umordnungsaufgaben** hat die Testperson vorgegebene Elemente so umzuordnen, dass sich eine richtige oder sinnvolle Abfolge ergibt. Auch dieser Item-Typ zählt zu den geschlossenen Aufgaben, denn die Testpersonen formulieren ihre Lösungen ausschließlich aus vorgegebenen Elementen. Auswertungsschwierigkeiten ergeben sich bei diesem Item, wenn prinzipiell mehrere Reihenfolgen richtig sind bzw. nur einige Elemente richtig geordnet wurden.

Für das Abfragen homogener Wissensbereiche sind auch **Zuordnungsaufgaben** geeignet. Die Aufgaben enthalten zwei oder mehr Serien von Elementen, und die Testperson hat nach vorgegebenen Regeln die Elemente der einen Serie den Elementen der anderen Serie(n) zuzuordnen. Ein Nachteil dieser Item-Form ist darin zu sehen, dass Testpersonen, die alle Zuordnungen richtig vornehmen, von Testpersonen, die alle Zuordnungen bis auf eine beherrschen, nicht unterschieden werden können, weil sich die letzte Zuweisung zwangsläufig ergibt. Dieses Problem kann jedoch weitgehend behoben werden, wenn die Anzahl der Elemente in den Vergleichsserien ungleich ist.

Die letzte Item-Variante, die **Ergänzungsaufgabe**, umfasst alle Auswahlaufgaben, die anstelle von Fragen oder Behauptungen Informationslücken enthalten und dann ein Angebot von Ergänzungen zur Auswahl vorgeben. Diese Item-Art eignet sich besonders zur Überprüfung der Fähigkeit, die interne Logik einer Abfolge von Begriffen, Zahlen, Zeichnungen oder Symbolen zu erkennen.

■ Allgemeine Leistungstests

Allgemeine kognitive Leistungstests erfassen in erster Linie **Aufmerksamkeit und Konzentration**. Ein Test der allgemeinen Leistungsfähigkeit fordert mit seinen Aufgaben von den Testpersonen „anhaltende Konzentration bei geistiger Tempoarbeit" (Brähler et al., 2002a, S. XIV). Eine detaillierte Klassifikation von Aufmerksamkeits- und Konzentrationstests ist bei Amelang und Schmidt-Atzert (2006) zu finden. Die Konzentration ist eine notwendige Voraussetzung für komplexere kognitive Leistungen. Wer also z. B. schlechte Ergebnisse in einem Intelligenztest erzielt, dem mangelt es möglicherweise nicht an Intelligenzleistung, sondern an der notwendigen Konzentration, die Intelligenztest-Aufgaben genau zu lesen und sorgfältig zu beantworten. Konzentrationstest werden – im Kontext anderer Verfahren – u. a. eingesetzt, um die in den letzten Jahren stärker beachtete Aufmerksamkeits-Defizit-/Hyperaktivitäts-Störung (ADHS) zu diagnostizieren.

Einer der bekanntesten Vertreter der allgemeinen Leistungstests ist der **Aufmerksamkeits-Belastungstest d2** (Brähler et al., 2002a, S. 300ff.) Bei ihm sollen alle Buchstaben **„d"** auf dem Testbogen, die oben und/oder

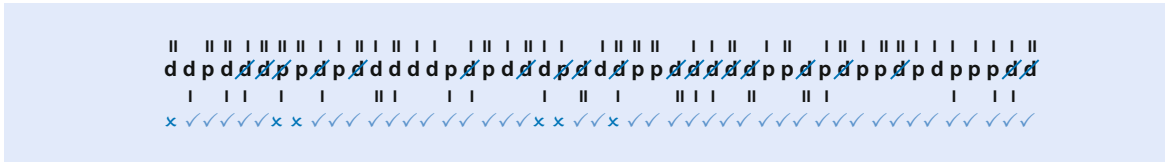


Abb. 10.20 d2 Aufmerksamkeits-Belastungstest. (Mod. nach Brickenkamp, 1994. d2-R © Hogrefe Verlag GmbH & Co. KG, Göttingen. Mit freundlicher Genehmigung)

unten **mit insgesamt zwei Strichen** markiert sind, durchgestrichen werden. Diese Aufgabe erfordert deswegen hohe Konzentration, weil der Testbogen neben „d" auch den Buchstaben „p" enthält und jeder Buchstabe mit 1–4 Strichen markiert sein kann. Ein Testbogen besteht aus 14 Zeilen mit jeweils 47 Zeichen. Eine Zeile kann als ein Test-Item aufgefasst werden (Abb. 10.20). Pro Zeile stehen 20 Sekunden Bearbeitungszeit zur Verfügung. Bei der Auswertung des d2-Tests kann nicht nur die Gesamtfehlerzahl betrachtet werden, sondern auch der Konzentrationsverlauf über die 14 Zeilen hinweg. Reliabilität und Validität des Verfahrens sind hoch.

10

Intelligenztests

Bei den Intelligenztests („intelligence tests") werden Tests unterschieden, welche die **allgemeine Intelligenz** oder **mehrere Intelligenzkomponenten** (z. B. sprachliches oder räumliches Denken) erfassen (Amelang & Schmidt-Atzert, 2006; S. 201f.). Die letzte Gruppe wird auch als Intelligenz-Strukturtests bezeichnet, wenn sie viele Komponenten berücksichtigt und Unterschiede zwischen den Untertests interpretiert werden. Außerdem können Intelligenztests in **kulturgebundene Tests** und **kulturfreie (oder kulturfaire) Tests** eingeteilt werden. Bei kulturgebundenen Intelligenztests benötigt man zur Lösung der Testaufgaben sprachliche Kompetenz und kulturspezifisches Hintergrundwissen (sog. Allgemeinbildung), z. B. zur Bedeutung von Fremdwörtern oder zu geschichtlichen Ereignissen. Damit wird Wissen als Teil der Intelligenz definiert. Kulturfreie Intelligenztests versuchen demgegenüber, Denkleistungen unabhängig von der schulischen Vorbildung zu erfassen, indem sie z. B. vollkommen nonverbal aufgebaut sind und nur mit geometrischen Mustern arbeiten (Vorgabe eines unvollständigen Musters, für welches das passende „Puzzleteil" gefunden werden muss). Allerdings scheinen auch solche geometrisch-analytischen Aufgaben nicht in allen Kulturen verbreitet zu sein.

Aufgrund von relativ hohen Korrelationen (ca. .50) zwischen Intelligenzleistung einerseits und Ausbildungs- und Berufserfolg andererseits werden Intelligenztests oft zur **Eignungsdiagnostik** eingesetzt (Schmidt & Hunter, 1977). Die entwicklungspsychologische Forschung widmet sich u. a. der Veränderung der Intelligenz über die **Lebensspanne** hinweg inklusive der Entstehung von „Weisheit" als einer auf Lebenserfahrung basierenden speziellen Form von Intelligenz (Oswald & Gunzelmann, 1991). Besonders kontrovers ist nach wie vor die Frage nach dem spezifischen Einfluss von **Anlage** (Genen) einerseits und **Umwelt** (Lernerfahrungen) andererseits auf die Intelligenz.

Ein sehr häufig eingesetzter Intelligenztest ist die 1949 von dem amerikanischen Psychologen David Wechsler entwickelte „Wechsler Adult Intelligence Scale" (WAIS). Wechsler definierte Intelligenz als „die zusammengesetzte oder globale Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen." (Wechsler, 1964, S. 13). Sein Intelligenztest besteht aus 11 Untertests, die in Verbaltests (z. B. Allgemeines Wissen, Wortschatztest) und Handlungstests (z. B. Bilder ergänzen, Bilder ordnen) eingeteilt werden. Die aktuellen deutschen Testfassungen sind der **WIE-III** (Wechsler Intelligenztest für Erwachsene; Aster, Neubauer & Horn, 2006) sowie der **HAWIK-IV** (Hamburg Wechsler Intelligenztest für Kinder; Petermann & Petermann, 2010).

Entwicklungstests

Entwicklungstests („developmental tests") messen den momentanen Leistungsstand eines Kindes in Relation zu seinem Lebensalter (Amelang & Schmidt-Atzert, 2006, S. 232). Durch Vergleiche mit Normwerten von Gleichaltrigen kann der Entwicklungsstand des Kindes eingeschätzt werden. Hierbei sind allgemeine und spezielle Entwicklungstests zu unterscheiden. **Allgemeine Entwicklungstests** erfassen das gesamte Verhaltensspektrum über Subtests. **Spezielle Entwicklungstests** fokussieren hingegen auf einen bestimmten Entwicklungsbereich wie z. B. Kreativität oder motorische Entwicklung bei Vorschulkindern. Häufig werden auch Intelligenztests, wie der oben genannte HAWIK zur Entwicklungsdiagnostik eingesetzt. Die Intelligenztests erfassen dabei, genau wie die speziellen Entwicklungstests, nur einen Teilbereich des Entwicklungsstandes.

Allgemeine und spezielle Entwicklungstests werden in der Praxis und Forschung verwendet, um frühzeitig Hinweise auf eine verlangsamte Entwicklung zu erhalten. Sie werden nicht nur bei Kindern, sondern auch bei

Jugendlichen und Erwachsenen mit Intelligenzminderung eingesetzt (Amelang & Schmidt-Atzert, 2006, S. 232f.). Voraussetzung hierfür ist, dass Normwerte für diese Gruppen vorliegen. Ein typisches Beispiel für einen allgemeinen Entwicklungstest ist der **Wiener Entwicklungstests** (**WET**; Kastner-Koller & Deimann, 2002). Dieser Test erfasst den Entwicklungsstand in den sechs Funktionsbereichen Motorik, visuelle Wahrnehmung, Lernen, Gedächtnis, kognitive Entwicklung, Sprache und sozialemotionale Entwicklung.

Schultests

Schultests beziehen sich auf schulische Leistungen und lassen sich grundsätzlich in drei Bereiche einteilen (Brähler et al., 2002a):

- **Schuleingangstests** beschäftigen sich mit der Schulreife von Kindern und werden eingesetzt, um zu überprüfen, ob ein Kind voraussichtlich in der Lage sein wird, dem Unterricht zu folgen (Wottawa & Hossiep, 1997). Dabei werden Test-Aufgaben verwendet, welche die Anforderungen des Unterrichts abbilden. Ein typisches Beispiel ist das Mannheimer Schuleingangsdiagnostikum (Jäger, Beetz, Erler, & Walter, 1994).
- **Schulleistungstests** sollen die Leistung in bestimmten Schulfächern möglichst objektiv, zuverlässig und gültig messen (Ingenkamp, 1995, S. 505). Sie werden als Ergänzung zu den in der Schule üblichen mündlichen und schriftlichen Prüfungen eingesetzt. Die Interpretation der Ergebnisse kann entweder bezugsgruppenorientiert (normorientiert) oder kriteriumsorientiert (lernzielorientiert) erfolgen. Bei der Bezugsgruppenorientierung werden erzielte Ergebnisse in Relation zu anderen Lernenden interpretiert und auf diese Weise überregional vergleichbar. Für die kriteriumsorientierte Interpretation wird vorher festgelegt, wie viele Aufgaben gelöst werden müssen, damit das Lernziel als erreicht gilt. Ein Beispiel für diese Art von Test ist die „Würzburger Leise Leseprobe – Revision“ (Schneider, Blanke, Faust, & Küspert, 2011). Ebenso sind hier die Schulleistungstests der bereits erwähnten **PISA-Studie** einzuordnen (http://www.oecd.org/de/PISA).
- **Schulübertrittstests**: Sie sollen die (Miss-)Erfolgswahrscheinlichkeit beim Übergang in die weiterführende Schule abschätzen helfen (Amelang & Schmidt-Atzert, 2006, S. 239). Ihre prognostische Validität ist jedoch meist gering, so dass sie nur in Kombination mit Schulleistungstests eingesetzt werden sollten.

Spezielle Funktionsprüfungs- und Eignungstests

Diese Gruppe von Testverfahren ist sehr heterogen und umfasst Berufseignungstests, Instrumente zur Analyse von Arbeitsbedingungen, wie auch spezielle Verfahren zur Testung von Funktionen wie z. B. Psychomotorik (Brähler et al., 2002a). Das wichtigste Einsatzgebiet dieser Testgruppe ist die Feststellung der Berufseignung bzw. die Analyse von Arbeitsbedingungen. Beispiele aus dieser Gruppe von Testverfahren sind die Kontrastive Aufgabenanalyse im Büro (KABA; Dunkel, Volpert, Zölch, Pleiss, & Hennes, 1993), die Drahtbiegeprobe, bei der die Testpersonen zur Messung ihrer Feinmotorik ein Stück Draht unter Zeitdruck möglichst akkurat in eine vorgegebene Form biegen müssen (Lienert, 1967) oder die Qualitative Führungsstilanalyse (QFA; Fennekels, 1995).

Persönlichkeitstests

Bei **Persönlichkeitstests** spielen objektive Beurteilungsmaßstäbe keine Rolle. Ein starkes Interesse für Musik zu bekunden, kann nicht in derselben Weise objektiv als „richtig“ oder „falsch“ bewertet werden wie die korrekte Lösung einer Rechenaufgabe im Intelligenztest. Im Zusammenhang mit Persönlichkeitstests wird das Konstrukt „Persönlichkeit“ eng ausgelegt. Biologisch-physiologische Personenmerkmale sowie der Intelligenz- und Leistungsbereich werden ausgeklammert; stattdessen konzentriert man sich auf Persönlichkeitsdimensionen im Sinne situations- und zeitübergreifender Merkmale des emotionalen und motivationalen Erlebens sowie des (sozialen) Handelns. Es geht also um Einstellungen, Motive, Präferenzen, Interessen, Werthaltungen sowie psychische Gesundheit. Persönlichkeitstests basieren auf **Selbstauskunfts-Items** und werden oft (selbst wenn sie normiert sind) auch als **Persönlichkeitsfragebögen** oder **Persönlichkeitsinventare** bezeichnet.

Bei der kleinen Sondergruppe der **objektiven Persönlichkeitstests** bleibt den Testpersonen die Messintention verborgen (geringe Augenscheinvalidität des Tests), der Rückschluss vom Verhalten zum latenten Merkmal wird vom Testanwender vorgenommen, während die üblichen „subjektiven“ Persönlichkeitstests mit Selbsteinschätzungen arbeiten und deswegen eher „verfälschbar“ sind (Amelang & Schmidt-Atzert, 2006, S. 306).

Persönlichkeitstests – Persönlichkeitstests („personality tests“) sind vollstrukturierte psychologische Testverfahren zur Messung von Persönlichkeitsmerkmalen. Hierbei werden individuelle Merkmale anhand von Selbstauskunfts-Items gemessen, die sich auf situations- und zeitüberdauernde Aspekte des emotionalen, motivationalen und sozialen Verhaltens und Erlebens beziehen. Im Unterschied zu Leistungstests gibt es bei den Test-Items in Persönlichkeitstests keine richtigen oder falschen Lösungen und es geht auch nicht um maximale Fähigkeiten, sondern um Selbstauskünfte über typisches Erleben und Verhalten der betreffenden Person.

Man unterscheidet **Persönlichkeits-Struktur-Tests**, die mehrere zentralen Dimensionen der Persönlichkeit erfassen, von Test, die sich auf einzelne Merkmalsbereiche

oder Anwendungsfelder konzentrieren, etwa **Interessen- und Motivationstests**. Schließlich existieren auch Testverfahren, die sich ergänzend zu den üblicherweise erfassten Persönlichkeitseigenschaften auf **aktuelle Zustände** beziehen (Amelang & Schmidt-Atzert, 2006; Brähler et al., 2002a, 2002b).

Persönlichkeits-Struktur-Tests

Bei der Erfassung von Persönlichkeit wird zwischen der Messung von Eigenschaften („traits“) und aktuellen Zuständen („states“) unterschieden. Das **Eigenschaftsmodell menschlichen Verhaltens** basiert auf der Annahme, dass Verhalten in erster Linie durch Persönlichkeitseigenschaften determiniert wird. Eigenschaften sind dabei zeitstabile und situationsübergreifende Merkmale der Persönlichkeit. So geht z. B. das **Big Five Modell der Persönlichkeit** von Allport und Odbert (1936) davon aus, dass die Persönlichkeitsstruktur eines Menschen durch die fünf unabhängigen Dimensionen Neurotizismus/emotionale Labilität (Gegenpol: emotionale Stabilität), Extraversion, Offenheit für neue Erfahrungen, soziale Verträglichkeit und Gewissenhaftigkeit beschrieben werden kann. Ein Test, der mehrere Dimensionen der Persönlichkeit erfasst, wird als Persönlichkeits-Struktur-Test bezeichnet.

Ein bekannter Persönlichkeits-Struktur-Test ist das NEO Five-Factor Inventory (**FFI**, Costa & McCrae, 1992) – in aktueller Fassung als **NEO-FFI** (Borkenau & Ostendorf, 2008; ▶ Die „Neurotizismus“-Subskala aus dem NEO-FFI) verbreitet. Es beruht auf dem Big Five Modell. Während das FFI **fünf Subdimensionen** enthält, beinhaltet das Freiburger Persönlichkeitsinventar (**FPI**) von Fahrenberg, Hamel und Selg (2010) sogar **12 Subskalen**, um theoriebasiert ein Gesamtbild der Persönlichkeit zu messen. Das FPI nutzt 10 Standardskalen (Lebenszufriedenheit, Soziale Orientierung, Leistungsorientierung, Gehemmtheit, Erregbarkeit, Aggressivität, Beanspruchung, Körperliche Beschwerden, Gesundheitssorgen, Offenheit) sowie 2 Zusatzskalen (Extraversion, Neurotizismus) zur Beschreibung der Persönlichkeit.

Zu diagnostischen Zwecken werden Persönlichkeits-Struktur-Tests sowohl im klinischen Bereich als auch im Rahmen der Personalauswahl eingesetzt. Die Messung von Persönlichkeitsmerkmalen ist in den Sozialwissenschaften weit über die Persönlichkeits- und Differentielle Psychologie hinaus von Bedeutung. Denn Persönlichkeitseigenschaften spielen als unabhängige Variablen oder Moderatorvariablen in vielen Bereichen menschlichen Erlebens und Verhaltens eine wichtige Rolle, etwa beim Leistungs-, Gesundheits- und Sozial- und Mediennutzungsverhalten. Vollständige Persönlichkeitstests oder einzelne Subskalen werden für Forschungszwecke oft in umfangreichere Fragebögen integriert.

Bei der Nutzung etablierter Testskalen oder Subskalen für eigene Studien mag man versucht sein, diese zu modifizieren. Denn mit konkreten Testskalen konfrontiert, stören sich viele Menschen spontan an der Formulierung einzelner Items. Es ist aus psychometrischer Sicht **unwissenschaftlich**, Test-Items ad hoc auf der Basis des persönlichen Eindrucks zu bewerten. Dies kann zu gravierenden Fehleinschätzungen führen. Der Verzicht auf **merkwürdig wirkende Items** oder das willkürliche Hinzufügen sinnvoll erscheinender Items kann z. B. die Validität einer Skala reduzieren.

Will man eine Testskala – wie z. B. ▶ Die „Neurotizismus“-Subskala aus dem NEO-FFI – für eigene Studien übernehmen, sollte wenn möglich die **vollständige Skala** verwendet werden, da die Operationalisierung des theoretischen Konstruktes sowie die publizierten psychometrischen Kennwerte sich auf die Gesamtskala beziehen. Zudem sind eigene Studien-Ergebnisse direkt mit früheren Studien vergleichbar, wenn das Originalinstrument verwendet wird.

Soll die Skala angepasst werden (Eliminierung von Items, Umformulierung von Items, Hinzufügen von Items), so sollte die Skalenrevision auf der Basis einer empirischen **Itemanalyse** (▶ Abschn. 10.4.4) anhand einer aussagekräftigen Stichprobe aus der Zielpopulation erfolgen. Auch sollten die Reliabilität und Validität der revidierten Skala empirisch geprüft werden. Ist eine entsprechend umfassende Vorstudie zur **statistischen Überprüfung der Skalenrevision** nicht möglich, so ist die Skalenrevision zumindest einem ausführlichen **qualitativen Pretest** zu unterziehen. Dabei werden Verständlichkeit und Bedeutungsnuancen der Items im Sinne des Zielkonstruktes sowie die Augenscheinvalidität der gesamten Skala mit einzelnen Testpersonen oder in der Gruppe diskutiert (zu Fragebogen-Pretests ▶ Abschn. 10.3.3). Nur Items, die laut Pretest gravierende Schwächen aufweisen (z. B. unklare Bedeutung durch komplizierte oder mehrdeutige Formulierung, Irritation durch ungewöhnliche oder antiquierte Wortwahl) sind zu eliminieren oder umzuformulieren. Die höhere Akzeptanz der revidierten Skala bei potenziellen Testpersonen sollte ihre Teilnahmebereitschaft erhöhen und somit der Qualität der Studie zugutekommen. Allerdings sind durch Skalenmodifikationen etablierter Instrumente, die nur auf qualitativen Pretests basieren und nicht auf eine empirische Überprüfung der psychometrischen Item- und Skaleneigenschaften des Instruments zurückgreifen, Qualitätseinbußen bei den gewonnenen Messwerten zu befürchten. Entsprechende Entscheidungen sind im Forschungsbericht transparent zu machen und zu begründen.

Die „Neurotizismus"-Subskala aus dem NEO-FFI

Beschreibung der Testskala

Neurotizismus bzw. emotionale Labilität ist eine Persönlichkeitsdisposition, die sich auf die Tendenz eines Menschen bezieht, sich leicht und häufig zu ärgern, ängstlich, unsicher oder unglücklich zu fühlen. Es handelt sich um eine Dimension der fünf Dimensionen innerhalb des **Big-Five-Modells der Persönlichkeit**.

Das NEO-FFI (Neurotizismus, Extraversion, Offenheit) – Fünf-Faktoren-Inventar (Borkenau & Ostendorf, 2008) als Persönlichkeits-Struktur-Test enthält fünf Subdimensionen, um die fünf zentralen Persönlichkeitsdispositionen laut Big-Five-Modell zu messen. Neben Neurotizismus sind das Extraversion, Offenheit für neue Erfahrungen, Gewissenhaftigkeit und soziale Verträglichkeit. Die Neurotizismus-Subskala umfasst **12 Items**, die jeweils auf einer **5-stufigen Ratingskala** (starke Ablehnung bis starke Zustimmung) zu beantworten sind. Einige Items sind dabei positiv – d. h. in Richtung Neurotizismus – gepolt (z. B. „Ich bin leicht beunruhigt.") andere Items sind entgegensetzt bzw. negativ gepolt (z. B. „Ich fühle mich selten einsam und traurig").

Datenerhebung und Messung mit der Testskala

Für die Skalenpunkte der Antwortskala werden Punktwerte vergeben (starke Ablehnung: 0, Ablehnung: 1, neutral 2; Zustimmung: 3; starke Zustimmung: 4), wobei negative Items umzupolen sind, d. h. bei einem negativen Item werden bei starker Ablehnung 4 Punkte und bei starker Zustimmung 0 Punkte vergeben. Der **Rohwert der Skala** setzt sich als **additiver ungewichteter Index** aus den korrekt gepolten Punktwerten der 12 Items zusammen (zur Indexbildung ► Abschn. 8.7). Dieser **Summenscore** hat somit einen Wertebereich von **0 bis 48 Punkte** (minimaler bis maximaler Neurotizismus).

Im Folgenden sind die 12 Items der Neurotizismus-Subskala aufgelistet und negativ gepolte Items markiert (N).

1. Ich bin leicht beunruhigt.
2. Ich fühle mich anderen oft unterlegen.
3. Wenn ich unter starkem Stress stehe, fühle ich mich manchmal, als ob ich zusammenbräche.
4. Ich fühle mich selten einsam und traurig. (N)
5. Ich fühle mich oft angespannt und nervös.
6. Manchmal fühle ich mich völlig wertlos.
7. Ich empfinde selten Furcht oder Angst. (N)
8. Ich ärgere mich oft darüber, wie Leute mich behandeln.
9. Zu häufig bin ich entmutigt und will aufgeben, wenn etwas schief geht.
10. Ich bin selten traurig oder deprimiert. (N)
11. Ich fühle mich oft hilflos und wünsche mir eine Person, die meine Probleme löst.
12. Manchmal war mir etwas so peinlich, dass ich mich am liebsten versteckt hätte.

Während man zu Forschungszwecken direkt mit den Rohwerten arbeiten kann, ist es im Rahmen der Diagnostik notwendig, den Rohwert mit Normwerten zu vergleichen, um eine genaue Aussage über die Merkmalsausprägung einer konkreten Person zu treffen. So könnten Sie z. B. Ihren eigenen Neurotizismus-Wert mit dem Durchschnittswert für die Bevölkerung in Deutschland vergleichen, der für die Altersgruppe der 18- bis 33-jährigen Männer bei $M = 18.20$ ($SD = 7.35$) und für Frauen bei $M = 21.07$ ($SD = 7.30$) liegt (Körner et al., 2008).

Da der Rohwert eines Tests immer von der Anzahl der Items und der vergebenen Punkte abhängig und somit nicht direkt interpretierbar ist, arbeitet man in der Diagnostik und oft auch in der Forschung bevorzugt mit **standardisierten Werten** bzw. **Normskalen**. Der Vorteil der Normskalen besteht darin, dass – unabhängig von der Punktzahl im ursprünglichen Test – in einheitlicher Weise die Merkmalsausprägung angegeben werden kann, wobei sofort deutlich ist, ob und um wie viele Standardabweichungen eine Person über oder unter dem Durchschnittswert ihrer Bezugsgruppe liegt. Anhand der empirischen Ergebnisse von Normstichproben werden die Rohwerte in unterschiedliche Varianten von standardisierten Werten transfomiert. Verbreitet sind als Normskalen u. a. ***z*-Werte** ($M = 0$, $SD = 1$), ***T*-Werte** ($M = 50$, $SD = 10$), **Z-Werte** ($M = 100$, $SD = 10$) oder **Stanine-Werte** ($M = 5$, $SD = 2$).

Die Vorgehensweise bei der Umrechnung von Rohwerten in standardisierte Werte sei für Stanine-Werte kurz beschrieben (Borkenau & Ostendorf, 2008; Körner et al., 2008): Die **Stanine-Skala** ist eine auf der Normalverteilung basierende standardisierte Normierungsskala mit einem Mittelwert von $M = 5$ und einer Standardabweichung $SD = 2$. Zur Berechnung der Stanine-Werte werden die Test-Rohwerte der Normstichprobe von den niedrigsten bis zu den höchsten aufsteigend in eine Rangreihe gebracht **(Prozentrangwerte)**. Die niedrigsten 4 Prozent bekommen den Stanine-Wert von 1 zugeteilt, die folgenden 7 % den Stanine-Wert von 2 und so weiter (Tab. 10.19). Erzielt man in einem Test einen Stanine-Wert von 5, liegt man im mittleren Bereich der Normwerte. 20 % der Normierungsstichprobe haben den gleichen Stanine-Wert erzielt. Die Umrechnung der NEO-FFI-Testwerte in Stanine-Werte erfolgt gemäß Tab. 10.20.

Lesebeispiel: Erzielt ein Mann im Alter von 25 Jahren auf der Neurotizismus-Skala einen **Testrohwert von 24**, erreicht er damit einen **Stanine-Wert von 7**. Im Vergleich zur Normierungsstichprobe liegt seine emotionale Labilität also im überdurchschnittlichen Bereich. Nur 11 % der männlichen Bevölkerung dieser Altersgruppe in Deutschland haben höhere Neurotizismus-Werte auf der Stanine-Skala, nämlich Stanine-Werte von 8 oder 9 (Tab. 10.19).

Tabelle 10.19 Stanine-Skala: Verteilung des prozentualen Anteils der individuellen Testscores auf die Stanine-Werte (gerankt von niedrigsten bis zu den höchsten Werten)

Prozentrang	4 %	11 %	23 %	40 %	60 %	77 %	89 %	96 %	100 %
Testwert-Ranking	Niedrigste 4 %	7 %	12 %	17 %	20 %	17 %	12 %	7 %	Höchste 4 %
Stanine-Wert	1	2	3	4	5	6	7	8	9

Tabelle 10.20 Umrechnung der Rohwerte der Neurotizismus-Subskala aus dem NEO-FFI (Borkenau & Ostendorf, 2008) in Stanine-Werte für Männer und Frauen in der Altersgruppe der 18- bis 33-Jährigen. (Aus Körner et al., 2008, S. 142)

NEO-FFI-Rohwerte	NEO-FFI-Rohwerte	Stanine-Werte
Männer	Frauen	
≤ 5	≤ 8	1
6–9	9–11	2
10–12	12–15	3
13–16	16–19	4
17–20	20–23	5
21–23	24–26	6
24–27	27–30	7
28–31	31–33	8
32–48	34–48	9

Interessen- und Motivations-Tests

Interessen- und (Leistungs-)Motivations-Tests („interest and motivation tests“) werden häufig in der Berufsberatung und bei der Personalauswahl eingesetzt (Amelang & Schmidt-Atzert, 2006, S. 292). Sie lassen sich in normative und ipsative Verfahren einteilen (Scheurer, 1995, S. 259). **Normative Verfahren** erlauben einen interindividuellen Vergleich, denn jede Person bewertet jede vorgegebene Tätigkeit auf einer Ratingskala z. B. hinsichtlich ihres Interesses. Bei **ipsativen Verfahren** werden Forced-Choice-Items verwendet, d. h., die Testpersonen müssen alle vorgegebenen Tätigkeiten nach ihren Präferenzen in eine persönliche Rangreihe bringen. Die ipsativen Interessentests ermöglichen intraindividuelle Vergleiche im Hinblick auf verschiedene Interessensausrichtungen. Die **generelle Interessenskala** erfasst 16 Interessenbereiche, wie z. B. Musik, Politik, Naturwissenschaften und Sport, sowie drei Verhaltensmodalitäten (Rezeptivität, Reproduktivität, Kreativität). Der **Berufs-Interessentest II** (Irle & Allehoff, 1984) besteht aus Forced-Choice-Aufgaben, bei denen vier Tätigkeiten aus vier verschiedenen Interessensbereichen miteinander verglichen werden sollen (Brähler et al., 2002a). Zur Messung der Leistungsmotivation kann z. B. das **Leistungsmotivationsinventar** (LMI) eingesetzt werden (Schuler & Prochaska, 2001), das insgesamt 170 Items umfasst (17 Skalen mit jeweils 10 Items).

Verfahren zur Erfassung aktueller Zustände

Der Erfassung von aktuellen Zuständen **(„states“)** – im Unterschied zu überdauernden Persönlichkeitseigenschaften („traits“) – liegt das **situationistische Modell** menschlichen Verhaltens zugrunde, demgemäß aktuelles Verhalten oft stärker durch situative Einflüsse als durch Persönlichkeitseigenschaften geprägt ist (Jäger, 1995, S. 346f.). Im Vergleich zu den Traits verändern sich States oder Zustände definitionsgemäß sehr viel schneller. Dies ist z. B. bei der Diagnostik von angstauslösenden Situationen der Fall (zur situativen Angstmessung existiert z. B. das **State-Trait-Angst-Inventar** (STAI-S) s. Laux, Glanzmann, Schaffner, & Spielberger, 1981, mit Items wie „Ich bin nervös“, „Ich bin verkrampft“, „Ich bin zappelig“, die jeweils auf einer 4-stufigen Ratingskala zu beantworten sind). Die Messung von States kann in vielen sozialwissenschaftlichen Forschungsbereichen von Interesse sein, in der Gesundheitsforschung ebenso wie z. B. in der Medienforschung, weil aktuelle emotionale Zustände u. a. das Essverhalten oder die Medienwahl beeinflussen und umgekehrt gesundheitsbezogene Verhaltensweisen oder Medieninhalte wiederum auf den emotionalen Zustand zurückwirken. Ebenso bestehen enge Verbindungen zwischen emotionalen Zuständen und kognitiver Informationsverarbeitung.

Klinische Tests

Klinische Tests („clinical psychometric tests“) stellen eine Sondergruppe der Persönlichkeitstests dar und basieren meist auf Selbstauskünften zu Symptomen **psychologischer Störungen**. Sie sollen bei der Abgrenzung von **Normalität** und **Psychopathologie** helfen. Mithilfe klinischer Tests lassen sich Notwendigkeit, Verlauf und Erfolg von entsprechenden psychotherapeutischen und/oder medizinischen Interventionen kontrollieren sowie wissenschaftlich evaluieren (Schumacher & Brähler, 2000). Klinische Tests stellen im gesamten Spektrum der psychometrischen Tests die größte Gruppe dar. Zum einen wegen des großen Anwendungsbedarfs, zum anderen wegen der starken Ausdifferenzierung unterschiedlichster psychologischer Störungsbilder.

Der Fachverlag Hogrefe gliedert die von ihm in der Testzentrale (http://www.testzentrale.de) angebotenen klinischen Testverfahren in vier Gruppen:

1. **Klinische Verfahren für Erwachsene**: Diese Gruppe versammelt Testverfahren für Erwachsene zur Messung psychologischer Erkrankungen wie z. B. Depressions-, Angst-, Zwangserkrankungen.
2. **Klinische Verfahren für Kinder und Jugendliche**: In dieser Gruppe geht es um Testverfahren, die kinder-

und jugendtypische Ausprägungsformen psychologischer Erkrankungen erfassen, z. B. Aufmerksamkeits-Defizit-/Hyperaktivitäts-Störungen (ADHS), Essstörungen, Angststörungen oder Lernstörungen und zudem hinsichtlich der Item-Formulierungen auf die junge Zielgruppe zugeschnitten sind.

3. **Neuropsychologische Verfahren**: In dieser Gruppe befinden sich u. a. Tests zur Messung von Demenz, Alzheimer Erkrankung, Sprachstörungen (Aphasie) und Störungen der Zahlenverarbeitung (Akalkulie), bei denen jeweils neurologische Grundlagen vermutet werden. Diese Verfahren basieren teilweise nicht auf Selbstauskünften, sondern auf Testaufgaben.
4. **Medizinpsychologische Verfahren**: Zu dieser sehr heterogenen Gruppe gehören Testverfahren, die mit Wissen, Erleben und Verhalten rund um Gesundheit zusammenhängen. Etwa Fragebögen über körperliche Beschwerden, Essverhalten oder Alkoholkonsum bis hin zum Wissen über Diabetes.

10.4.4 Testtheorie

Eine Pumpe füllt einen Behälter, der 40 Liter fasst, in 5 Minuten. Wie lange benötigt die Pumpe, um einen Behälter mit 64 Liter zu füllen? Auf diese Frage gibt eine Testperson die richtige Antwort: 8 Minuten. Kann man aufgrund dieser einen Antwort behaupten, die Testperson verfüge über eine gute mathematische Denkfähigkeit? Sicherlich nicht. Es leuchtet intuitiv ein, dass diese Informationsbasis nicht ausreicht, um entscheiden zu können, ob diese Frage „mathematische Denkfähigkeit" oder etwas anderes misst. Es bleibt offen, wie viel Zeit die Testperson zur Lösung dieser Aufgabe beanspruchte, ob sie nur zufällig eine richtige Schätzung abgab, ob sie ähnliche oder auch schwerere Aufgaben lösen könnte und vieles mehr.

Solchen Problemstellungen widmen sich **Testtheorien** („test theory", „statistical test theory"). Allgemein formuliert befasst sich eine Testtheorie damit, in welchem Zusammenhang das zu testende latente Merkmal mit dem in einem Test tatsächlich beobachtbaren Testverhalten (d. h. den Testantworten) steht. Zudem werden Annahmen getroffen, wie aus den empirisch ermittelten Testergebnissen Aussagen über die latenten Merkmale der getesteten Personen getroffen werden können. Auch die Frage der Anforderungen, denen ein Test genügen muss, um aufgrund eines Testergebnisses auf die tatsächliche Ausprägung des getesteten Merkmales schließen zu können, ist Gegenstand der Testtheorie.

Ein Test besteht gewöhnlich aus mehreren unterschiedlich schweren Aufgaben oder Fragen (**Test-Items**), die die Testperson lösen oder beantworten muss. Als Testergebnis resultiert eine Anzahl richtig bzw. im Sinne des Konstruktes zustimmend beantworteter Items, woraus sich verschiedene Schlüsse ableiten lassen.

Die an einem naturwissenschaftlichen Messmodell orientierte **Klassische Testtheorie** (**KTT;** „classical test theory", auch: „true score theory", TST) nimmt an, dass das Testergebnis direkt dem wahren Ausprägungsgrad des untersuchten Merkmals entspricht, dass aber jede Messung oder jedes Testergebnis zusätzlich von einem Messfehler überlagert ist. Der Testwert repräsentiert damit die **wahre Merkmalsausprägung** zuzüglich einer den Testwert vergrößernden oder verkleinernden **Fehlerkomponente** (z. B. aufgrund mangelnder Konzentration, ungeeigneter Items, Übermüdung, schlechter Untersuchungsbedingungen etc.). Die wahre Merkmalsausprägung kann jedoch nur erschlossen werden, wenn der Testfehler bekannt ist. Hierin liegt das Problem der Klassischen Testtheorie. Die Aussagekraft eines Tests ist nur bestimmbar, wenn wahre Merkmalsausprägung und Fehleranteil getrennt zu ermitteln sind.

Im Unterschied dazu basiert der Grundgedanke der **Probabilistischen Testtheorie** (**PTT;** „probabilistic test theory", auch: „modern test theory", „latent trait theory", „strong true score theory") – die oft auch als **Item-Response-Theorie** (**IRT**; „item response theory") bezeichnet wird – auf der Annahme, dass die Wahrscheinlichkeit einer bestimmten Antwort auf ein einzelnes Item von Merkmalen des Items (Item-Parameter) und latenten Merkmalen der Person (Personen-Parameter) abhängt. Eine Person mit besserer mathematischer Denkfähigkeit löst die eingangs gestellte Aufgabe mit höherer Wahrscheinlichkeit als eine Person mit schlechterer mathematischer Denkfähigkeit (Moosbrugger, 2012a). Die Klassische Testtheorie ist deterministisch: Das Testergebnis entspricht – abgesehen von Messfehlern – direkt der Merkmalsausprägung. Ein Probabilistisches Testmodell hingegen ermittelt diejenigen Merkmalsausprägungen, die für verschiedene Arten der Item-Beantwortung jeweils am wahrscheinlichsten sind.

Die Klassische Testtheorie ist in erster Linie eine Messfehlertheorie und konzentriert sich auf die Bestimmung der Reliabilität. Durch Erweiterungen wie z. B. den Multitrait-Multimethod-Ansatz sind im Rahmen der KTT auch Validitätsprüfungen theoriebasiert möglich. Die KTT ist in Forschung und Praxis bis heute dominierend. Dass der neue probabilistische Ansatz aufwändiger, komplizierter und nicht in gängige Statistikprogramme integriert war, trug nicht zu seiner Popularisierung bei, zumal Ergebnisse von Tests zum selben Merkmal, die gemäß KTT und IRT konstruiert wurden, oft sehr gut übereinstimmten (Rost, 1999). Allerdings lassen sich die Testwerte eines Tests auf der Basis der IRT sehr viel besser interpretieren, weil zusätzliche Testgütemerkmale

prüfbar sind (v. a. Skalierbarkeit und damit auch das Skalenniveau). Die internationalen Bildungsstudien IGLU, TIMSS und PISA basieren auf der Probabilistischen Testtheorie. Generell haben IRT-Modelle den großen Vorteil, dass sie eine viel stringentere Überprüfung der Modellannahmen erlauben als die KTT und somit methodisch überzeugender sind. Eigenschaften beider Testmodelle sollen im Folgenden dargestellt werden.

Klassische Testtheorie – Die Klassische Testtheorie (KTT; „classical test theory", auch: „true score theory", TST) nimmt an, dass das Testergebnis direkt dem wahren Ausprägungsgrad des untersuchten Merkmals entspricht, dass aber jede Messung oder jedes Testergebnis zusätzlich von einem Messfehler überlagert ist.

Probabilistische Testtheorie – Die Probabilistische Testtheorie (PTT; „probabilistic test theory", auch: „modern test theory", „latent trait theory", „strong true score theory") – international meist als **Item-Response-Theorie (IRT;** „item response theory") bezeichnet – basiert auf der Annahme, dass die Wahrscheinlichkeit einer bestimmten Antwort auf ein einzelnes Item von Merkmalen des Items (Item-Parameter) und latenten Merkmalen der Person (Personen-Parameter) abhängt.

Grundkenntnisse in Testtheorie sind notwendig, um Testergebnisse aus publizierten empirischen Studien korrekt interpretieren sowie um sachgerecht einen qualitativ hochwertigen Test für eine eigene Studie auswählen zu können. **Vertiefte Kenntnisse in Testtheorie** sind erforderlich, wenn man selbst einen Test modifizieren und die resultierende Testrevision überprüfen will (z. B. Übertragung eines englischsprachigen Tests ins Deutsche; Ausgliederung einer Kurzskala aus einem umfassenden Test; Aktualisierung eines veralteten Tests) oder einen ganz neuen psychologischen Test entwickeln möchte.

Die Kriterien der Klassischen wie der Probabilistischen Testtheorie lassen sich sowohl auf **Tests** im engeren Sinne – die auch eine Individualdiagnose anhand von Normwerten anstreben – als auch auf standardisierte Fragebögen anwenden, die eher die Funktion von Forschungsinstrumenten haben und bei denen statt individueller Werte primär Aggregatwerte (v. a. Gruppenmittelwerte) interessieren (vgl. Mummendey, 1999, S. 17ff.). Fragebögen, die den Kriterien der Testtheorie genügen, nennt man auch **psychometrische Fragebögen** oder Skalen.

Man beachte, dass es für die Anwendung der Klassischen wie der Probabilistischen Testtheorie notwendig ist, dass jeweils mehrere Items (Aufgaben, Fragen) verwendet werden, um gemeinsam dasselbe **latentes Merkmal** zu erfassen. Die einzelnen Items, die von den Testpersonen bearbeitet werden, gelten dabei als **manifeste Merkmale** bzw. **Indikatoren**, die zusammengefasst zu einem Test- oder Skalenwert Hinweise auf die Ausprägung des latenten Merkmals geben sollen (zum Indikatorkonzept ► Abschn. 8.3.1 und ► Abschn. 17.2.1). Wenn wir z. B. das Persönlichkeitsmerkmal „Schüchternheit" erfassen möchten und dazu eine Testskala bzw. einen Kurzfragebogen mit 10 Items verwenden, die alle Schüchternheit in verschiedenen Facetten betreffen („Mir fällt es schwer, mit Fremden ins Gespräch zu kommen", „Ich bin eher ein schüchterner Mensch", „In Gruppen fühle ich mich oft unwohl" etc.), dann kann mittels Testtheorie analysiert werden, wie gut die einzelnen Schüchternheitsitems zusammenpassen und wie gut sie gemeinsam das Merkmal Schüchternheit messen. Auf Fragebögen und Tests, in denen jede einzelne Frage ein eigenständiges Merkmal erfasst (**Einzelindikatoren**; z. B. „Wie zufrieden sind Sie insgesamt mit Ihrem Leben?", „Wie oft pro Woche treiben Sie Sport?", „Haben Sie ein Ehrenamt?", „Mögen Sie Krimis?" etc.) lässt sich die Testtheorie nicht sinnvoll anwenden (zu Einzelindikatoren ► Abschn. 8.5).

Die Klassische sowie die Probabilistische Testtheorie lassen sich nur anwenden, wenn *mehrere Items* verwendet werden, die alle *dasselbe Konstrukt* messen und zu einem Test- bzw. Skalenwert zusammengefasst werden. Nicht anwendbar ist die Testtheorie, wenn als Messinstrumente Einzelitems bzw. Einzelindikatoren verwendet werden.

Bei Indikatoren unterscheidet man zwischen reflektiven Indikatoren und formativen Indikatoren (Bühner, 2011, S. 37). In der Praxis sehr verbreitet sind **reflektive Indikatoren**, bei denen man davon ausgeht, dass sie durch das latente Merkmal kausal bedingt werden.

Beispiel: Dadurch, dass Menschen sich im Grad ihrer Schüchternheit unterscheiden **(latentes Merkmal als Ursache der Indikatoren)** beantworten sie Items wie „Mir fällt es schwer, mit Fremden ins Gespräch zu kommen" oder „In Gruppen fühle ich mich oft unwohl" unterschiedlich. Die Operationalisierung läuft hier auf ein sog. Domain-Sampling hinaus (Churchill, 1979), d. h. aus der Menge aller möglichen Indikatoren bzw. inhaltlich passenden Items wird eine sinnvolle Auswahl getroffen. Reflektive Indikatoren bilden inhaltlich ähnliche Testitems, die hoch miteinander korrelieren sollten. Bei einer Ausprägungsänderung des latenten Merkmals (z. B. Verringerung der Schüchternheit) sind auch entsprechende Merkmalsänderungen aller Indikatoren zu erwarten (d. h. allen Schüchternheits-Items wird weniger zugestimmt). Die Klassische Testtheorie setzt ein reflektives Messmodell voraus.

Bei sog. **formativen Indikatoren** ist es genau umgekehrt, hier werden die Indikatoren im Zuge der Konzeptspezifikation (► Abschn. 8.2) als Determinanten bzw. Ursachen des latenten Merkmals aufgefasst. Beispiel: Wenn eine Person einen geringen Bildungsstand hat und/oder einen Beruf mit geringem Prestige ausübt und/oder über

ein unterdurchschnittliches Einkommen verfügt (drei Indikatoren als Determinanten), dann hat sie definitionsgemäß einen niedrigen sozio-ökonomischen Status (**latentes Merkmal als Effekt der Indikatoren**). Formative Indikatoren sind inhaltlich oft sehr heterogen, so dass dann nicht unbedingt hohe Item-Korrelationen zu erwarten sind. Auch sind formative Indikatoren meist nicht so leicht wie reflektive Indikatoren durch Alternativen ersetzbar, da sie in spezifischer Weise das Konstrukt konstituieren. Eine Veränderung eines formativen Indikators (z. B. Einkommenssteigerung) kann die anderen Indikatoren (z. B. Bildungsstand) unbeeinflusst lassen und dennoch zu einer Veränderung des latenten Merkmals führen (z. B. gestiegener sozioökonomischer Status). Bei einem formativen Messmodell ist die KTT nicht anwendbar, hier ist aber z. B. eine Indexbildung gemäß dem **C-OAR-SE-Modell** (Rossiter, 2010) möglich (zur Indexbildung ► Abschn. 8.7). Bereits im Zuge der Konzeptspezifikation sollte man sich darüber klar werden, ob man ein reflektives oder formatives Messmodell anzulegen hat und die Konsequenzen z. B. für die Spezifikation von Strukturgleichungsmodellen beachten, die sowohl reflektive als auch formative Messmodelle abbilden können (Binder & Eberl, 2005, zu Strukturgleichungsmodellen ► Kap. 17). In der Forschungspraxis (z. B. Marketingforschung) dominieren reflektive Messmodelle bei psychometrischen Fragebögen und Tests, obwohl eine kritische inhaltlich-theoretische Betrachtung der Zielkonstruktive häufiger auch formative Messmodelle nahelegen würde (Eberl, 2004).

Reflektives Messmodell – Bei einem reflektiven Messmodell greift man zur Erfassung des latenten Merkmals auf Indikatoren zurück, die durch das latente Merkmal verursacht werden. Diese reflektiven Indikatoren („reflective indicators", „effect indicators") sind einander in der Regel sehr ähnlich.

Formatives Messmodell – Bei einem formativen Messmodell greift man zur Erfassung des latenten Merkmals auf Indikatoren zurück, die das latente Merkmal erzeugen. Diese formativen Indikatoren („formative indicators", „cause indicators") können einander auch sehr unähnlich sein.

Klassische Testtheorie

Kennzeichnend für eine Testtheorie sind ihre Annahmen über die Testwerte. Die Klassische Testtheorie konzentriert sich dabei auf die Messfehler, die in die Testwerte eingehen. Sie ist deswegen auch als Messfehlertheorie zu kennzeichnen. Die grundlegenden **Annahmen der KTT** werden im Folgenden kursorisch vorgestellt. Bei der praktischen Anwendung der KTT steht die Feststellung der Messgenauigkeit mittels **Reliabilitätskoeffizienten** für einzelne Tests oder Untertests im Zentrum, sie wird anschließend behandelt. Zur Ermittlung des Testgütekriteriums der **Validität** sind Ergänzungen der KTT notwendig, insbesondere ein Rückgriff auf den Multitrait-Multimethod-Ansatz, den wir kurz vorstellen. Neben der Abschätzung der Reliabilität und Validität eines gesamten Tests oder Untertests ist es in der Praxis im Zuge der Testentwicklung und Testrevision auch wichtig, die einzelnen Items eines Tests auf den Prüfstand zu stellen. Dazu dienen Techniken der statistischen **Itemanalyse**. Denn durch zielgerichtete Modifikation einzelner Items können Validität und Reliabilität eines Tests maßgeblich beeinflusst werden, wie wir an Beispielen zeigen werden.

Axiome und Annahmen der Klassischen Testtheorie

Die Klassische Testtheorie besteht im Kern aus einem Satz von Annahmen über die Messfehler von Testwerten, die theoretisch gesetzt und nicht empirisch prüfbar sind. Sie wurde seit Beginn des 20. Jahrhunderts entwickelt. Frühe Darstellungen stammen aus den 1950er- und 1960er-Jahren (Gulliksen, 1950; Lord & Novick, 1968). In der Literatur besteht bis heute keine Einigkeit darüber, wie die Annahmen stringent zu gliedern sind. Manche Autoren gehen z. B. von „zwei Definitionen", „sechs Folgerungen" sowie „einer Zusatzannahme" aus (Bühner, 2011), andere von „drei Axiomen" und „zwei Zusatznahmen" (Moosbrugger, 2012b). Auch sind die Notationen sehr uneinheitlich. Auf die umfangreiche Methodenliteratur, welche die KTT darstellt und kritisch würdigt, kann hier nur exemplarisch verwiesen werden (Crocker & Algina, 2006; Lienert & Raatz, 1998; Steyer & Eid, 2001; Rost, 2004). Die folgende, bewusst anschauliche und weniger formalisierte Darstellung lehnt sich an Moosbrugger (2012b, S. 104f.) an und gliedert die Grundannahmen der KTT in drei Axiome, die durch zwei Zusatzannahmen über die Unabhängigkeit der Messfehler ergänzt werden.

1. **Verknüpfungsaxiom:** Der Testwert einer Person m, der mit einem bestimmten Test oder Item i (z. B. Intelligenztest) zu einem bestimmten Zeitpunkt erhoben wird (**Test Score** x_{mi}), setzt sich jeweils additiv zusammen aus der wahren Ausprägung des gemessenen Merkmals (z. B. der als konstant angenommenen Intelligenz der Person; **True Score** τ_{mi} [tau]) sowie einem situationsbedingt wechselnden, unsystematischen Messfehler (z. B. Müdigkeit, äußere Störeinflüsse wie Lärm; **Error Score** ε_{mi} [epsilon]):

$$x_{mi} = \tau_{mi} + \varepsilon_{mi}$$

2. **Existenzaxiom:** Der wahre Wert des Merkmals einer Person, der mit einem Test gemessen werden soll (z. B. wahrer Intelligenzwert τ), existiert als Erwartungs- bzw. Mittelwertwert des empirischen Testwertes: Würde man dieselbe Person den Intelligenztest sehr oft durchführen lassen, dann würde der Mittelwert ihrer Testwerte ihrem wahren Intel-

ligenzwert entsprechen: $E(x_{mi}) = \tau_{mi}$. Denn die unsystematischen Messfehler würden sich im Mittel zu Null ausgleichen: $E(\varepsilon_{mi}) = 0$. Beispiel: Bei der ersten Testdurchführung ist der Testwert der Person vielleicht durch besonders gute Konzentration höher, beim nächsten Durchgang ist der Testwert durch Müdigkeit geringer, beim übernächsten Testdurchgang herrscht ablenkender Lärm etc. Bei wiederholten Testdurchläufen mitteln sich diese zufälligen Schwankungen laut Grundannahme der KTT heraus. Man kann also das Verknüpfungsaxiom $x_{mi} = \tau_{mi} + \varepsilon_{mi}$ für wiederholte Testdurchgänge umformen: $E(x_{mi}) = E(\tau_{mi}) + E(\varepsilon_{mi})$. Dadurch, dass $E(\varepsilon_{mi}) = 0$ (Messfehler gleichen sich aus) und $E(\tau_{mi}) = \tau_{mi}$ (der wahre Wert einer Person ist eine Konstante), ergibt sich eben:

$$\boldsymbol{E(x_{mi}) = \tau_{mi}}$$

3. **Unabhängigkeitsaxiom:** Die bei wiederholten Messungen auftretenden unsystematischen Fehlereinflüsse durch Tagesform und äußere Störfaktoren (Motivation, Wachheit, Lärm etc.) sind unabhängig von der wahren Merkmalsausprägung der Person, d. h., Messfehler durch Müdigkeit treten unabhängig davon auf, ob eine Person nun z. B. geringe oder hohe Intelligenz aufweist. Eine Korrelation (Corr) zwischen dem wahren Wert einer beliebigen Person und den situativen Messfehlern eines beliebigen Items liegt definitionsgemäß nicht vor:

$$\boldsymbol{Corr(\tau_{mi}, \varepsilon_{mi}) = 0}.$$

4. **Zusatzannahme der Unabhängigkeit der Messfehler zwischen Items:** Die Messfehler zweier beliebiger Items i und j sind jeweils unkorreliert (lokale stochastische Unabhängigkeit), wenn eine Person m die beiden Items wiederholt bearbeitet. Beispiel: Ob die Person bei der Bearbeitung des einen Items gerade durch Lärm abgelenkt ist, bei einem anderen Item aber die Aufgabenstellung missversteht, ist unabhängig voneinander:

$$\boldsymbol{Corr(\varepsilon_{mi}, \varepsilon_{mj}) = 0.}$$

5. **Zusatzannahme der Unabhängigkeit der Messfehler zwischen Personen:** Die Messfehler zweier beliebiger Personen m und v sind jeweils unkorreliert, wenn sie dasselbe Item i wiederholt bearbeiten. Beispiel: Die Müdigkeits- oder Konzentrationsschwankungen einer Testperson sind unabhängig von der Befindlichkeit einer anderen Testperson:

$$\boldsymbol{Corr(\varepsilon_{mi}, \varepsilon_{vi}) = 0.}$$

Die Klassische Testtheorie bildet eine bewährte und gut etablierte Grundlage der Entwicklung und Überprüfung der Testgüte (insbesondere der Reliabilität) psychologischer Tests. Aus den Grundannahmen der KTT lässt sich zum einen ableiten, dass sich ein aus mehreren Items (im Sinne von reflektiven Indikatoren) additiv gebildeter **Testwert** einer Person als Schätzung der wahren Merkmalsausprägung der Person auffassen lässt. Ebenso können aus den Annahmen der KTT Varianten der Reliabilitätsbestimmung abgeleitet werden.

Eine Erweiterung der KTT bildet die **Generalisierbarkeitstheorie** (Brennan, 2001), dergemäß sich Testwerte nicht nur aus der wahren Merkmalsausprägung der Testpersonen sowie zufälligen Messfehlern zusammensetzen, sondern weitere Faktoren wie Beurteiler- oder Methodeneffekte hinzukommen.

Die Klassische Testtheorie weist jedoch – wie bereits angesprochen – auch **Schwächen** auf, die v. a. die Testgütekriterien der Skalierung und der Konstruktvalidität betreffen (Grubitzsch, 1991; Moosbrugger, 2012b, S. 115f.):

- Probleme mit der **Skalierung** ergeben sich dadurch, dass es sich bei Axiomen grundsätzlich um Festsetzungen bzw. Definitionen handelt und nicht um empirische Tatsachen. Ob sich wahrer Wert und Fehlerwert tatsächlich „in Wirklichkeit" **additiv verknüpfen**, ist nicht beweisbar. Damit kann im Rahmen der KTT das Testgütekriterium der Skalierung (siehe „Nebengütekriterien" in ▶ Abschn. 10.4.1) nicht geprüft werden. Ebenso ist eine Überprüfung des zugrunde gelegten **Intervallskalen-Niveaus der Items** nicht prüfbar.
- Probleme mit der **Validität** ergeben sich dadurch, dass die **Homogenität der Items** (ihre sog. lokale stochastische Unabhängigkeit) nicht empirisch geprüft wird, sondern ersatzweise nur die Item-Interkorrelationen oder Itemtrennschärfen betrachtet werden. Zudem sind die auf der Basis der KTT zur Testbeurteilung bestimmten **Kennwerte** (v. a. Itemschwierigkeiten, Itemtrennschärfen und Reliabilitätskoeffizienten), die Voraussetzung für valide Testwerte sind, ihrerseits **stichprobenabhängig**, so dass ein verallgemeinerbarer Nachweis der Testgüte problematisch ist.

Diese Schwächen der KTT sollen durch die moderne Probabilistische Testtheorie bzw. Item-Response-Theorie überwunden werden. Im Unterschied zur deterministischen KTT postuliert die IRT einen probabilistischen Zusammenhang zwischen den Merkmalsausprägungen der Testpersonen und den Messwerten.

Schätzung der Reliabilität

Die Qualität eines psychometrischen Tests bzw. Fragebogens lässt sich v. a. an den drei Hauptgütekriterien Objek-

tivität, Reliabilität und Validität festmachen (► Abschn. 10.4.1). Während Objektivität bei einem standardisierten Test vorauszusetzen ist, werden Reliabilität und Validität in der Regel anhand von empirisch bestimmten Kennwerten bzw. Koeffizienten geschätzt. Zur Bestimmung von **Reliabilitätskoeffizienten** liegen vier verschiedene Methoden **(Retest-, Paralleltest-, Testhalbierungs- und Interne Konsistenz-Methode)** vor (Schermelleh-Engel & Werner, 2012). Die Koeffizienten sind kontextspezifisch im Hinblick auf notwendige **Reliabilitätsanforderungen** zu bewerten.

Reliabilitätskoeffizienten und Reliabilitätsanforderungen

Die Reliabilität (Zuverlässigkeit) gibt im traditionellen Verständnis der Testgütekriterien den Grad der Messgenauigkeit (Präzision) eines Instrumentes an. Die Reliabilität ist umso höher, je kleiner der zu einem Messwert X gehörende Fehleranteil E ist. Perfekte Reliabilität würde gemäß den Axiomen der KTT bedeuten, dass ein Test in der Lage ist, den wahren Wert der getesteten Personen ohne jeden Messfehler zu erfassen – und zwar idealerweise bei allen Testaufgaben sowie beim Gesamttestergebnis. Dieser Idealfall tritt in der Praxis leider nicht auf, da sich Fehlereinflüsse durch situative Störungen, Müdigkeit der Testpersonen, Missverständnisse oder Raten nie ganz ausschließen lassen.

Reliabilität – Die Reliabilität (Messgenauigkeit, Präzision, „reliability") eines Tests kennzeichnet den Grad der Genauigkeit bzw. Messfehlerfreiheit, mit dem das geprüfte Merkmal gemessen wird. Ein reliabler psychologischer Test liefert Messwerte, die wenig von Messfehlern belastet sind.

Wie kann man nun die Messgenauigkeit bzw. Reliabilität eines Tests quantifizieren, wenn doch stets nur messfehlerbehaftete Werte für die Test-Items und den Gesamttest verfügbar und die „wahren" Werte unbekannt sind? Wie will man erkennen, ob in einer Messwertreihe mit Intelligenztestergebnissen ein großer Fehleranteil (= unreliable Messung) oder ein kleiner Fehleranteil (= reliable Messung) steckt?

Zur Lösung dieses Problems greifen wir auf die bereits dargestellten Axiome der Klassischen Testtheorie zurück. Ein vollständig reliabler Test müsste nach wiederholter Anwendung bei denselben Personen zu exakt den gleichen Ergebnissen führen (perfekte Korrelation beider Messwertreihen), sofern der „wahre" Wert unverändert ist (was bei zeitstabilen Eigenschaften vorausgesetzt werden kann). Weichen die Ergebnisse wiederholter Tests voneinander ab bzw. sind sie unkorreliert, so werden hierfür Messfehler verantwortlich gemacht. Da Messfehler sowohl von den wahren Werten T, von anderen Merkmalen T' als auch voneinander unabhängig sind (Unabhängigkeitsaxiom plus Zusatzannahmen der KTT), können die Messungen nur unsystematische Abweichungen zwischen den Messwerten zweier Messzeitpunkte erzeugen. Diese unsystematischen Abweichungen konstituieren die sog. **Fehlervarianz** $\mathrm{Var}(\varepsilon)$. Je größer die Fehlervarianz, umso mehr Messfehler fließen in die Testwerte ein.

Umgekehrt spricht eine niedrige Fehlervarianz für hohe Messgenauigkeit: Je größer die Ähnlichkeit bzw. der korrelative Zusammenhang zwischen beiden Messwertreihen, umso höher ist der Anteil der systematischen, gemeinsamen Variation der Werte und umso geringer ist gleichzeitig der Fehleranteil. Messwertunterschiede sind dann nicht „zufällig", sondern systematisch; sie gehen auf unterschiedliche „wahre" Merkmalsausprägungen zurück und konstituieren die sog. **wahre Varianz** $\mathrm{Var}(\tau)$. Gemäß Grundannahmen der KTT gilt, dass sich in der Population die Varianz des Testwertes aus der Varianz des wahren Wertes und der Varianz des Fehlerwertes additiv zusammensetzt: $\mathrm{Var}(x) = \mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon)$.

Allgemein lässt sich die Reliabilität (Rel) als Anteil der wahren Varianz an der beobachteten Varianz des Testwertes $\mathrm{Var}(x)$ definieren. Je größer der Anteil der wahren Varianz (maximal 1 bzw. 100 %), umso geringer ist der Fehleranteil in den Testwerten.

$$\mathrm{Rel} = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(x)} = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon)} \qquad (10.1)$$

Will man für einen Test die Reliabilität berechnen, so benötigt man neben der empirisch ermittelbaren Varianz der Testwerte noch eine Schätzung für die (unbekannte) wahre Varianz. Je nach Art der Schätzung der Reliabilität sind vier Methoden zu unterscheiden, mit denen die Reliabilitätskoeffizienten von eindimensionalen Testskalen (oder Subskalen) berechnet werden können: Retest-Reliabilität, Paralleltest-Reliabilität, Testhalbierungs-Reliabilität und interne Konsistenz.

Hinsichtlich der **Reliabilitätsanforderungen** werden konventionell Reliabilitätskoeffizienten über .90 als hoch und über .80 als ausreichend betrachtet, wobei diese groben Faustregeln immer kontextspezifisch (Welche Art von Merkmal? Welcher Typ von Reliabilitätskoeffizient? Verfügbarkeit besserer Testverfahren?) und nicht mechanisch anzuwenden sind (Bühner, 2011, S. 81; Schermelleh-Engel & Werner, S. 2012, S. 135ff.; ► Abschn. 10.4.1).

Retest-Reliabilität Zur Bestimmung der Retest-Reliabilität **(Stabilität)** wird derselbe Test derselben Stichprobe zweimal vorgelegt, wobei das zwischen den Messungen (x_{t_1}: Testwert erste Messung, x_{t_2}: Testwert zweite Messung) liegende Zeitintervall variiert werden kann (in der Regel sind es mehrere Wochen). Die Retest-Reliabilität ist definiert als **Korrelation** beider Messwertreihen. Diese Korrelation (mit 100 % multipliziert) gibt an, wie

viel Prozent der Gesamtunterschiedlichkeit der Testergebnisse auf „wahre“ Merkmalsunterschiede zurückzuführen sind. Eine Retest-Reliabilität von Rel = .76 lässt darauf schließen, dass 76 % der Merkmalsvarianz auf „wahre“ Merkmalsunterschiede zurückgehen und nur 24 % auf Fehlereinflüsse.

$$\begin{aligned} \mathrm{Rel}_{\mathrm{Retest}} &= \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(x)} = \frac{\mathrm{Cov}(x_{t_1}, x_{t_2})}{\mathrm{SD}(x_{t_1}) \cdot \mathrm{SD}(x_{t_2})} \\ &= r_{x_{t_1} x_{t_2}} \end{aligned} \qquad (10.2)$$

Die Korrelation der Testwerte zu Messzeitpunkt 1 und Messzeitpunkt 2 kann der Reliabilitätsschätzung dienen, weil gemäß Axiomen und Annahmen der KTT aus $x_{t_1} = \tau + \varepsilon_1$ und $x_{t_2} = \tau + \varepsilon_2$ abzuleiten ist, so dass $\mathrm{Cov}(x_{t_1}, x_{t_2})$ eben genau der $\mathrm{Var}(\tau)$ entspricht bzw. dass die Reliabilität mit der Korrelation beider Testwerte gleichzusetzen ist (für eine schrittweise Herleitung s. Moosbrugger, 2012b, S. 107ff.).

Bei der Reliabilitätsbestimmung nach der Testwiederholungsmethode besteht die Gefahr, dass die Reliabilität eines Tests überschätzt wird, wenn die Lösungen der Testaufgaben erinnert werden, womit v. a. bei kurzen Tests mit inhaltlich interessanten Items zu rechnen ist. Die Wahrscheinlichkeit von **Erinnerungseffekten** nimmt jedoch mit wachsendem zeitlichem Abstand zwischen den Testvorgaben ab. Mit wachsendem **Retest-Intervall** verringern sich Erinnerungseffekte, allerdings wächst der logistische Aufwand, die ursprünglichen Testpersonen noch einmal zu erreichen und zur Teilnahme zu motivieren. In der Praxis sind Retest-Intervalle von wenigen Wochen bis zu mehreren Monaten zu finden, die entsprechende Entscheidung ist inhaltlich zu treffen und zu begründen.

Wenig brauchbar ist die Retest-Methode bei Tests, die **instabile bzw. zeitabhängige Merkmale** erfassen. Hierbei wäre dann unklar, ob geringe Test-Retest-Korrelationen für geringe Reliabilität des Tests oder für geringe Stabilität des Merkmals sprechen. Beispiel: Ein Test soll Stimmungen (States) erfassen (z. B. Angespanntheit, Müdigkeit), die typischerweise sehr starken intra-individuellen Schwankungen unterliegen. Die Reliabilitätsschätzung mittels Retest-Methode ergibt z. B. Rel = .34. Dies würde einem Anteil von 34 % „wahrer“ Varianz in den Messwerten entsprechen (bzw. 66 % Fehlervarianz). Es wäre jedoch verfehlt, den Test nun wegen vermeintlich fehlender Messgenauigkeit abzulehnen, da in diesem Fall unsystematische Messwertedifferenzen zwischen t_1 und t_2 nicht nur Fehlereffekte, sondern auch „echte“ Veränderungen darstellen.

Ein weiterer Nachteil der Retest-Methode besteht in ihrem relativ großen zeitlichen und untersuchungstechnischen Aufwand. Da dieselben Testpersonen nach einem festgelegten Zeitintervall erneut kontaktiert und zur Teilnahme motiviert werden müssen, ist mit größeren Ausfallzahlen zu rechnen. Diese Testpersonenverluste („drop out“) sind bereits bei der Untersuchungsplanung einzukalkulieren, indem eine besonders **große Startstichprobe** gezogen wird. Das Problem, dass bei **systematischen** Drop-outs (es fallen z. B. überwiegend Testpersonen mit schlechten Testergebnissen aus) die Stichprobe verzerrt wird, ist damit allerdings nicht gelöst.

Bei der ersten Testung fordert man üblicherweise die Testpersonen auf, sich ein persönliches Kennwort auszudenken bzw. nach einer bestimmten Regel einen **Identifizierungs-Code** zu bilden (z. B. aus Initialen, Geburtsdaten, Adressdaten o. ä.) und sich diesen zu merken. Dieses Kennwort dient zur Wahrung der Anonymität als Namensersatz und wird von den Testpersonen bei der ersten und zweiten Testung auf dem Lösungsbogen notiert, so dass personenweise eine eindeutige Zuordnung der Messwiederholungen möglich ist.

Paralleltest-Reliabilität Die Ermittlung der Paralleltest-Reliabilität **(Äquivalenz)** ist ebenso wie die Bestimmung der Retest-Reliabilität mit einigem untersuchungstechnischen Aufwand verbunden. Zunächst werden zwei Testversionen entwickelt, die beide Operationalisierungen desselben Konstruktes darstellen. Die Testpersonen bearbeiten diese sog. Paralleltests in derselben Sitzung kurz hintereinander. Je ähnlicher die Ergebnisse beider Tests ausfallen, umso weniger Fehlereffekte sind offensichtlich im Spiel, d. h., die wahre Varianz wird hier als Kovarianz zwischen den Testwerten einer Personenstichprobe auf beiden Paralleltests geschätzt.

Das Ergebnis einer Reliabilitätsprüfung nach der Paralleltestmethode sind stets zwei Testformen, die sich entweder beide als reliabel oder beide als unreliabel erweisen. Der mit der Erstellung von zwei Parallelformen verbundene Aufwand ist v. a. dann gerechtfertigt, wenn für praktische Zwecke tatsächlich zwei (oder auch mehr) äquivalente Testformen benötigt werden. Dies ist z. B. bei Gruppentestungen im Leistungsbereich der Fall, wo durch den Einsatz von Testversion A und B unerwünschtes Abschreiben verhindert werden kann.

Die Konstruktion von zwei Paralleltests erfolgt in vier Schritten:

1. **Itempool:** Auf der Grundlage von Theorie und Empirie wird eine Liste von Items zusammengestellt (Itempool), die allesamt reflektive Indikatoren des Zielkonstruktes darstellen. Der Itempool enthält mindestens doppelt so viele Items wie für eine Testform angestrebt wird.
2. **Itemanalyse:** Der Itempool wird einer Personenstichprobe vorgelegt und anschließend einer Itemanalyse

unterzogen. Ziel dieser Analyse ist die Kennzeichnung aller Items durch ihre jeweiligen Schwierigkeitsindizes und Trennschärfekoeffizienten.

3. **Itemzwillinge:** Je zwei Items mit vergleichbarer Schwierigkeit und Trennschärfe werden zu „Itemzwillingen" zusammengestellt.
4. **Paralleltests:** Die beiden Paralleltests A und B entstehen, indem je ein „Zwilling" zufällig der einen, und der andere „Zwilling" der anderen Testform zugeordnet wird.

Bearbeitet nun eine neue Stichprobe beide Paralleltests A und B, so lässt sich die Reliabilität folgendermaßen als Korrelation der Testwerte von Test A (x_A) und Test B (x_B) bestimmen:

$$\text{Rel}_{\text{Paralleltest}} = \frac{\text{Var}(\tau)}{\text{Var}(x)} = \frac{\text{Cov}(x_A, x_B)}{\text{SD}(x_A) \cdot \text{SD}(x_B)} = r_{x_A x_B} \tag{10.3}$$

Mit der hier beschriebenen Vorgehensweise erhält man zwei Tests, die man als **nominell parallel** bezeichnet, denn die Parallelität ist statistisch nicht nachgewiesen. **Echte Paralleltests** liegen vor, wenn bei Testform A und Testform B nachweislich sowohl die wahre Varianz und als die Fehlervarianz jeweils gleich sind. Bei sog. τ **(tau)-äquivalenten Tests** sind die wahren Varianzen gleich, die Fehlervarianzen können variieren. Auch dann ist das Verfahren noch einsatzbar. Zur Prüfung der Parallelität bzw. τ -Äquivalenz von zwei Testformen A und B sowie gleichzeitig auch der Paralleltest-Reliabilität ist die **konfirmatorische Faktorenanalyse** einsetzbar (Schermelleh-Engel & Werner, 2012, S. 127; ▶ Kap. 17).

Bei der Datenerhebung ist zu beachten, dass die Korrelation der Paralleltests eben nur dann die Messgenauigkeit schätzt, wenn es sich tatsächlich um parallele Tests und parallele Testsituationen handelt. Werden die Paralleltest-Versionen denselben Testpersonen in derselben Testsitzung oder in kurzem zeitlichem Abstand präsentiert, so können **Übertragungseffekte** (z. B. Lerneffekte) zu einer Überschätzung der Reliabilität führen.

Testhalbierungs-Reliabilität Die Testhalbierungs-Reliabilität **(Split-Half-Reliabilität)** erfordert im Unterschied zur Retest- und Paralleltest-Methode keinerlei untersuchungstechnischen Mehraufwand, da der zu untersuchende Test einer Stichprobe einmalig zur Bearbeitung vorgelegt wird. Anschließend werden pro Testperson zwei Testwerte berechnet, die jeweils auf der Hälfte aller Items beruhen, wobei diese Testhalbierung bei der Auswertung unterschiedlich realisiert werden kann (Zufallsauswahl aus allen Test-Items; erste und zweite Testhälfte; Items mit gerader und ungerader Nummer: Odd-Even-Methode etc.). Da die Testhälften quasi „Paralleltests" mit halber Länge darstellen, kann man die Testhalbierungsmethode als **Sonderform der Paralleltestmethode** auffassen. Entscheidend ist deswegen, dass die Testhalbierung so vorgenommen wird, dass die Testhälften tatsächlich inhaltlich dasselbe messen.

Die gemeinsame Varianz der inhaltlich vergleichbaren Testhälften repräsentiert die messfehlerfreie „wahre" Varianz, d. h., die Testhalbierungs-Reliabilität entspricht der Korrelation der Testwerte der Testhälften ($x_{1/2}$, $x'_{1/2}$).

$$\text{Rel}_{\text{Testhalbierung}} = \frac{\text{Var}(\tau)}{\text{Var}(x)} = \frac{\text{Cov}(x_{1/2}, x'_{1/2})}{\text{SD}(x_{1/2}) \cdot \text{SD}(x'_{1/2})} = r_{x_{1/2}x'_{1/2}} \tag{10.4}$$

Die Reliabilität eines Tests nimmt – sieht man von Ermüdungseffekten etc. ab – mit der Anzahl seiner Items zu. Sie nähert sich mit wachsender Itemzahl asymptotisch einem Präzisionsmaximum. Demzufolge unterschätzt eine Methode, die nur die halbe Testlänge berücksichtigt, die Reliabilität des Gesamttests. Mittels der sog. **Spearman-Brown-Prophecy-Formula** kann ein mit der Testhalbierungs-Methode gewonnener Reliabilitätskoeffizient jedoch nachträglich um den Betrag, der durch die Testhalbierung verloren ging, aufgewertet werden (vgl. Spearman, 1910, zit. nach Lienert & Raatz, 1994, S. 185):

$$\text{Rel}_{\text{Testhalbierung (korrigiert)}} = \frac{2 \cdot \text{Rel}_{\text{Testhalbierung}}}{1 + \text{Rel}_{\text{Testhalbierung}}} \tag{10.5}$$

Wenn Testhalbierungs-Reliabilitäten angegeben werden, so handelt es sich in der Regel um die in dieser Weise korrigierten Reliabilitätskoeffizienten. Auf Probleme der Spearman-Brown-Prophecy-Formula, die nur unter sehr strengen Voraussetzungen gültig ist, geht Yousfi (2005) ein.

Interne Konsistenz Die Bestimmung der Reliabilität gemäß Testhalbierungs-Methode hängt stark von der Art der zufälligen Testhalbierung ab. Zu stabileren Schätzungen der Reliabilität führt die Berechnung der internen Konsistenz (Konsistenzanalyse). Interne Konsistenzschätzungen stellen eine Erweiterung der Testhalbierungs-Methode dar, und zwar nach der Überlegung, dass sich ein Test nicht nur in zwei Testhälften, sondern in so viele „kleinste" Teile zerlegen lässt, wie er vergleichbare Items enthält. Es kann also praktisch jedes einzelne Item wie ein „Paralleltest" behandelt werden. Voraussetzung ist jedoch, dass die Items dasselbe Merkmal in vergleichbarer Weise messen (was mittels Strukturgleichungsmodellen bzw. konfirmatorischen Faktorenanalysen zu prüfen ist; ▶ Kap. 17):

- Strikt **parallele Items** haben allesamt die gleichen wahren Werte und die gleichen Fehlerwerte und somit die gleichen Itemschwierigkeiten.
- Bei unterschiedlichen Itemschwierigkeiten muss zumindest von τ **(tau-)äquivalenten Items** auszugehen sein (dieselben wahren Werte, aber verschiedene Fehlerwerte).
- Auch diese Voraussetzung kann noch gelockert werden: Letztlich reichen sog. **essenziell** τ **(tau-)äquivalente Items** (wahre Werte unterscheiden sich um eine additive Komponente; Fehlerwerte differenzieren) als Voraussetzung einer Reliabilitätsschätzung mittels interner Konsistenzanalyse.

Die Korrelationen zwischen parallelen oder (essenziell) τ-äquivalenten Items spiegeln die „wahre" Varianz des Tests wider und sind damit zur Schätzung der Messgenauigkeit geeignet. Die Berechnung der internen Konsistenz kann über die sog. **Kuder-Richardson-Formel** erfolgen (vgl. Richardson & Kuder, 1939, zit. nach Lienert & Raatz, 1994, S. 192).

Sehr viel gebräuchlicher ist jedoch **Cronbachs Alpha-Koeffizient** (einen Vergleich von Cronbachs Alpha mit anderen Maßen der internen Konsistenz findet man bei Osburn, 2000). Formal entspricht der Alpha-Koeffizient der mittleren Testhalbierungs-Reliabilität eines Tests für alle möglichen Testhalbierungen. Insbesondere bei heterogenen bzw. mehrdimensionalen Tests unterschätzt der Alpha-Koeffizient allerdings die Reliabilität, wenn er für den gesamten Test berechnet wird. Deswegen sollten bei einem mehrdimensionalen Test jeweils separat die **internen Konsistenzen der Subtests** bestimmt werden.

Der Alpha-Koeffizient ist umso höher, je mehr Items ein Test oder Untertest enthält (p = Anzahl der Items; *xi:* Itemwerte von 1 bis p; x: Testwert als Summe aller Itemwerte) und je höher die Item-Interkorrelationen ausfallen. Alpha wird folgendermaßen berechnet (vgl. Cronbach, 1951; Bortz & Schuster, 2010, S. 429f., Gl. 23.77; Schermelleh-Engel & Werner, 2012, S. 131f.):

$$\text{Rel}_{\text{interne Konsistenz}} = \alpha = \frac{p}{p-1} \cdot \left(1 - \frac{\sum_{i=1}^{p} \text{Var}_{x_i}}{\text{Var}_x}\right) \tag{10.6}$$

Welchen Einfluss einzelne Items auf die Höhe des Alpha-Koeffizienten haben, ist gängigen Statistikprogrammen (z. B. SPSS) im Zuge einer Reliabilitäts- bzw. **Itemanalyse** zu entnehmen. Signifikanztests für den Alpha-Koeffizienten findet man bei Feldt, Woodruff und Salih (1987). Ein Verfahren, mit dem die Äquivalenz zweier unabhängiger Alpha-Koeffizienten geprüft werden kann, haben Feldt und Ankenmann (1999) sowie Alsawalmeh und Feldt (2000) vorgeschlagen. Über „optimale" Stichprobenumfänge (▶ Abschn. 14.3) für den statistischen Vergleich zweier Alpha-Koeffizienten berichten Feldt und Ankenmann (1998). Zum Thema „Missing Data" im Rahmen der Bestimmung von Alpha-Koeffizienten findet man Informationen bei Enders (2003).

Die Berechnung von Cronbachs Alpha setzt voraus, dass die Fehleranteile der Items wechselseitig unkorreliert und die Items mindestens essenziell τ-äquivalent sind, d. h. denselben Anteil an wahrer Varianz aufweisen. In der Praxis sind diese beiden Voraussetzungen jedoch meist verletzt:

- Voraussetzung unkorrelierter Messfehler: Wenn die Berechnung von Cronbachs Alpha nur anhand einer einmaligen Messung von Testpersonen erfolgt, ist mit korrelierten Fehleranteilen zu rechnen (z. B. eine unkonzentrierte Person wird jedes Item unkonzentriert beantworten). Entsprechende korrelierte Fehler **(Transient Error)** führen zu überhöhten Alpha-Koeffizienten. Mit diesem Problem befasst sich eine Arbeit von Green (2003). Nach einer Analyse der einschlägigen Literatur zur Thematik entwickelte der Autor einen **Alpha-Koeffizienten für Test-Retest-Daten**, mit dem die „wahre" Reliabilität eines Tests genauer geschätzt werden kann als mit dem Alpha-Koeffizienten auf der Basis eines einmal erhobenen Datensatzes (zum Problem des „Transient Error" beim Alpha-Koeffizienten s. auch Becker, 2000).
- Voraussetzung (essenziell) τ-äquivalenter Items: Auch die Voraussetzung der τ-Äquivalenz ist meist verletzt, denn in der Regel messen nicht alle Items dasselbe Konstrukt mit demselben (oder allenfalls durch eine additive Konstante differierenden) Anteil an wahrer Varianz. Deswegen wurde inzwischen ein Maß für **kongenerische Items** entwickelt, also reflektive Indikatoren, die dasselbe Konstrukt in unterschiedlichem Ausmaß messen: **McDonalds Omega**. Dieser Kennwert kann auf unterschiedliche Arten berechnet werden (für Details zu Berechnungsvorschriften, Anwendungsvoraussetzungen und Unterschieden zwischen McDonalds Omega und Cronbachs Alpha s. Zinbarg, Revelle, Yovel, & Li, 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006).

Zusammengefasst ist zu konstatieren, dass die Reliabilitätsschätzung mittels interner Konsistenzmethode in der Praxis am beliebtesten ist, da sie den **geringsten Aufwand** erzeugt. Allerdings muss beachtet werden, dass der Alpha-Koeffizient auf die häufig vorkommenden Voraussetzungsverletzungen **progressiv reagiert** und die **Reliabilität überschätzt**. Da man sich mit progressiven Ergebnissen selbst begünstigt und empirische Ergebnisse in ungerechtfertigt positivem Licht erscheinen, hat die Problematik der Voraussetzungsverletzung zu einer starken Kritik am Alpha-Koeffizienten und u. a. zu den oben skiz-

zierten Alternativvorschlägen geführt. Manche Autoren raten inzwischen davon ab, Cronbachs Alpha überhaupt noch zu verwenden (Rauch & Moosbrugger, 2011). Wer nicht auf Alternativen zurückgreift, sondern weiterhin Cronbachs Alpha nutzt, muss sich also darüber im Klaren sein, dass die wahre Reliabilität des untersuchten Tests überschätzt werden kann, wenn die Voraussetzungen verletzt sind.

Reliabilität von Untertests Die vier beschriebenen Methoden der Reliabilitätsschätzung gehen von eindimensionalen Tests aus, deren Items als reflektive Indikatoren allesamt dasselbe globale Konstrukt erfassen und somit hoch interkorrelieren. Demgegenüber haben **mehrdimensionale Tests** die Aufgabe, Teilaspekte eines komplexen Merkmals mittels sog. Untertests (bzw. Teilskalen, Faktoren oder Dimensionen bzw. Subdimensionen) separat zu messen (zum Problem der Heterogenität von Testmodellen s. Rauch & Moosbrugger, 2011, S. 26ff.). Bei mehrdimensionalen Tests korrelieren jeweils die zu einem Untertest gehörenden Items hoch, während die Teiltests weniger oder gar nicht miteinander korrelieren. Es ist folglich sinnvoll, die interne Konsistenz über einen **Cronbachs Alpha-Koeffizienten für jede Subskala** einzeln zu bestimmen, statt für alle Items gemeinsam einen Alpha-Koeffizienten zu berechnen.

Zur Reliabilitätsbestimmung von unkorrelierten Untertests schlägt Cliff (1988) vor, statt des Alpha-Koeffizienten einen Kennwert zu berechnen, der auf den Ergebnissen einer Faktorenanalyse beruht und in den der Eigenwert des Faktors – lambda (λ) – zusammen mit der durchschnittlichen Item-Interkorrelation ($\bar{r}_{ij}$) der zum Faktor gehörenden Items eingeht:

$$\mathrm{Rel}_{\mathrm{Subskala}} = \frac{\lambda_{\mathrm{Subskala}} - \left(1 - \bar{r}_{ij}\right)}{\lambda_{\mathrm{Subskala}}} \qquad (10.7)$$

Wenn die Items perfekt interkorrelieren, erreicht der Teiltest unabhängig von der Höhe des Eigenwertes λ eine perfekte Reliabilität von 1 (Bortz & Schuster, 2010, S. 429f.; Gl. 23.78).

Im Falle von korrelierten Subtests kann zur Reliabilitätsschätzung **Cronbachs Alpha** für das hierarchische Faktormodell verwendet werden (sofern die Voraussetzungen erfüllt sind), zudem steht **McDonalds Omega_h** zur Verfügung (Zinbarg et al., 2005; Zinbarg et al., 2006).

▪ Schätzung der Validität

Zur Abschätzung der Validität eines Tests werden im traditionellen Testgüte-Verständnis die bereits eingeführten Methoden der Inhalts-, Kriteriums- und Konstruktvalidierung genutzt. Nur bei der **Kriteriums- und Konstruktvalidität** werden statistische **Validitätskoeffizienten** berechnet. Diese sind kontextspezifisch im Hinblick auf notwendige **Validitätsanforderungen** zu bewerten.

Validitätskoeffizienten und Validitätsanforderungen Die Validität (Gültigkeit) gilt im traditionellen Verständnis als das wichtigste Testgütekriterium, da sie angibt, ob ein Test das misst, was er messen soll bzw. was er zu messen beansprucht (d. h., ein Intelligenztest sollte tatsächlich Intelligenz messen und nicht z. B. Allgemeinbildung). Eine hohe Objektivität und eine hohe Reliabilität sind notwendige – aber nicht hinreichende – Voraussetzungen für hohe Validität. Beachte: Das testtheoretische Kriterium der Validität, das die Qualität von Messinstrumenten – bzw. die Gültigkeit von Aussagen auf der Basis von Testwerten – betrifft, ist **nicht zu verwechseln** mit den Kriterien der **internen und externen Validität**, die als Gütekriterien von Untersuchungsdesigns bzw. von wissenschaftlichen Aussagen über Studienergebnisse gelten (► Abschn. 3.2.2).

Validität – Die Validität („validity“) als psychometrisches Gütekriterium eines Tests gibt an, wie gut der Test in der Lage ist, genau das zu messen, was er laut Testbezeichnung zu messen beansprucht. Ein valider psychologischer Test liefert Messwerte, die sich zielgenau auf das interessierende Merkmal beziehen.

Von diesem traditionellen Validitätsverständnis abzugrenzen ist eine übergeordnete Definition von Validität bzw. Konstruktvalidität, die sich auf die **Gültigkeit von Interpretationen basierend auf Testergebnissen** bezieht.

Dabei sind sechs verschiedene Evidenzquellen zur Aussagekraft der Testwerte einzubeziehen (**Konstruktvalidität in der Messick-Tradition**: Messick, 1995) sowie – vor allem bei einer Testverwendung im Kontext (quasi-)experimenteller Studien – auch diverse Merkmale des Untersuchungsdesigns (**Konstruktvalidität in der Campbell-Tradition** ► Abschn. 3.2.2).

Im Vergleich zu Objektivität und Reliabilität ist die Erfassung und Überprüfung der Validität eines Tests sehr viel aufwändiger. Wir unterscheiden traditionell drei Hauptarten von Validität: Inhalts-, Konstrukt- und Kriteriumsvalidität, wobei nur für Konstrukt- und Kriteriumsvalidität Koeffizienten berechnet werden. Neuere Validitätskonzeptionen in der Messick-Tradition (Messick, 1995) differenzieren bis zu **sechs verschiedene Evidenzquellen** für die Gültigkeit von wissenschaftlichen Aussagen, die auf Testwerten basieren (zu Hauptgütekriterien psychologischer Tests ► Abschn. 10.4.1). Eine ausdifferenzierte Darstellung von Validierungsmöglichkeiten findet sich in den *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Bei **Validitätskoeffizienten**, die für Kriteriumsvalidität sowie konvergente und diskriminante Konstruktvalidität berechnet werden, handelt es sich überwiegend um Korrelationskoeffizienten. Dabei strebt man für die Kriteriums- und konvergente Konstruktvalidität Korrela-

tionskoeffizienten an, die statistisch bedeutsam größer als Null und **möglichst nahe bei Eins** liegen. Nach Weise (1975, S. 219) gelten Validitätskoeffizienten für Kriteriums- oder konkordante Konstruktvalidität zwischen .40 und .60 als mittelmäßig und Validitätskoeffizienten über .60 als hoch (ebenso Fisseni, 1990, Bühner, 2011, S. 81). Derartige Richtwerte sind jedoch – wie mehrfach betont – nur mit großer Vorsicht zu verwenden. Letztlich müssen Validitätskoeffizienten ebenso wie Reliabilitätskoeffizienten oder Effektgrößenmaße immer inhaltlich mit Blick auf die jeweiligen Merkmale und Untersuchungsbedingungen bewertet werden. Anstelle absolut gesetzter Richtwerte sind hier Vergleichswerte aus dem jeweiligen Forschungsfeld (z. B. typische Höhe der Korrelation zwischen Intelligenz und Berufserfolg) aussagekräftiger.

Auch mit sorgfältigen testtheoretischen Validierungen lassen sich keine unzweifelbar „gültigen" Tests konstruieren. Von theoretischen und methodischen Einschränkungen ist jeder Validierungsversuch betroffen. Dennoch lässt sich der Einsatz eines psychometrischen Tests generell **pragmatisch** rechtfertigen, wenn die Entscheidungen und Vorhersagen, die auf der Basis des Tests getroffen werden, tauglicher sind als Entscheidungen und Vorhersagen, die ohne den Test möglich wären – es sei denn, der mit dem Test verbundene Aufwand steht in keinem Verhältnis zum Informationsgewinn. Dieser **Minimalanspruch an die Validität eines Tests** ist einleuchtend, wenn man bedenkt, wie viele Personalentscheidungen z. B. allein aufgrund des persönlichen Eindrucks, zweifelhafter Gutachten oder gar der Handschrift vorgenommen werden – also aufgrund von Informationen, deren Validität in vielen Fällen nicht erwiesen ist bzw. niedriger sein dürfte als die Validität eines psychometrischen Tests (Fahrenberg, 2002, Kap. 10). Es wäre illusionär, Tests zu fordern, die perfekte oder nahezu perfekte Validitäten aufweisen und perfekte Entscheidungen gewährleisten. Der Wert eines Tests lässt sich letztlich nur an seinem Beitrag messen, den Nutzen testgestützter Entscheidungsstrategien in der Praxis sowie Forschungsaktivitäten zu optimieren (Cronbach & Gleser, 1965; Kubinger, 1996; Wottawa & Hossiep, 1987).

Zu beachten ist, dass Validitätskoeffizienten grundsätzlich in ihrer Höhe durch die **Reliabilität des Testwerts** sowie die **Reliabilität des zur Validierung genutzten Merkmals** (z. B. Außenkriterium) limitiert werden. Denn wenn die korrelierten Werte durch unsystematische Messfehler belastet sind, kann sich kein perfekter Zusammenhang zeigen. So kann die Kriteriumsvalidität maximal nur den Wert des geometrischen Mittels (Bortz & Schuster, 2010, S. 29) aus der Reliabilität des Tests und der Reliabilität des Kriteriums erreichen. Hieraus folgt, dass die Kriteriumsvalidität bei einem perfekt reliablen Kriterium nicht größer sein kann als die Wurzel aus der Reliabilität bzw. dass allgemein gilt (vgl. Fisseni, 1990, S. 102):

$$\text{Val} < \sqrt{\text{Rel}}. \tag{10.8}$$

Ist die Reliabilität des Kriteriums nicht größer als die des Tests (was in der Forschungspraxis häufig vorkommt), kann die Validität nicht größer sein als die Reliabilität (Val $\leq$ Rel; zur Begründung vgl. Rost, 2004, S. 390).

Die in der Praxis nie perfekten Reliabilitäten des Tests und des Kriteriums „mindern" also die Validität des Tests. Will man erfahren, wie hoch die Validität (r_{xy}) bei perfekter Reliabilität von Test ($r_{xx} = 1$) und Kriterium ($r_{yy} = 1$) wäre, kann man die sog. **Minderungskorrekturformel** („correction for attenuation") einsetzen zur Bestimmung der minderungskorrigierten Validität r_{xyc}:

$$r_{xyc} = \frac{r_{xy}}{\sqrt{r_{xx}} \cdot \sqrt{r_{yy}}}. \tag{10.9}$$

Diese Korrelation schätzt die Korrelation zwischen fehlerfreien Test- und Kriteriumswerten. Durch Verwendung von Strukturgleichungsmodellen (▶ Kap. 17) sind Minderungskorrekturen heute jedoch nicht mehr so aktuell. Wie man Konfidenzintervalle für die „wahre" Validität bestimmt, erläutert Charles (2005).

Inhaltsvalidität Die Inhaltsvalidierung erfolgt nicht empirisch-statistisch, sondern vor allem inhaltlich-theoretisch. Sie mündet nicht in die Berechnung und Bewertung von Validitätskoeffizienten. Entscheidend sind vielmehr der Rückgriff auf Theorien, sorgfältiges Arbeiten mit Begriffsdefinitionen und Begriffs-Relationen sowie Einschätzungen von Fachexperten. Es sei darauf hingewiesen, dass die Inhaltsvalidierung in ihrer Bedeutung traditionell unterschätzt wird und stärker beachtet werden sollte (Rossiter, 2008).

Kriteriumsvalidität Kriteriumsvalidität (kriterienbezogene Validität) liegt vor, wenn das Ergebnis eines Tests zur Messung eines latenten Merkmals bzw. Konstruktes (z. B. Berufseignung) mit Messungen eines korrespondierenden manifesten Merkmals außerhalb der Testsituation (Außenkriterium) übereinstimmt, das für diagnostische Zwecke relevant ist (z. B. beruflicher Erfolg). Die Kriteriumsvalidität ist definiert als Korrelation zwischen den Testwerten und den Kriteriumswerten einer Stichprobe. Die formale Definition lautet folgendermaßen: Die Validität eines Tests X entspricht seiner Korrelation mit einem reliablen und validen Außenkriterium Y.

Nicht selten handelt es sich bei dem Außenkriterium um einen Beobachtungssachverhalt, der erst zu einem späteren Zeitpunkt gemessen werden kann. Ob ein Schulreife-Test wirklich „Schulreife" erfasst, wird sich in der Praxis erst zeigen, nachdem der Test durchgeführt wurde

und das Verhalten des Kinds in der Schule beobachtbar ist. Die Validität eines Tests bemisst sich dann daran, ob der Testwert das spätere Verhalten korrekt vorhersagt. Diese Form der Kriteriumsvalidität nennt man **prognostische Validität** im Unterschied zur **Übereinstimmungsvalidität** (konkurrenten Validität), bei der Testwert und Kriteriumswert zum selben Messzeitpunkt erhoben werden, und zur **retrospektiven Validität**, bei der auf Messwerte zurückgegriffen wird, die zu einem früheren Zeitpunkt erfasst und dokumentiert wurden. So kann zur Validierung einer Deliquenz-Skala auf die Anzahl der aktenkundigen Vorstrafen zurückgegriffen werden. Bei Entwicklungs- und Leistungstests (z. B. Berufseignung, Schuleignung), sind verhaltensbezogene Außenkriterien inhaltlich oft durch das Testziel vorgegeben (z. B. Bewährung im Berufs- oder Bildungssystem).

Anders ist das bei Persönlichkeitstests. Hier wird der Testwert teilweise einfach mit der **Selbsteinschätzung der betreffenden Eigenschaft** als Außenkriterium korreliert. Eine überzeugendere Variante zur Bestimmung der Kriteriumsvalidität ist die **Technik der bekannten Gruppen** (Known Groups). Das Kriterium ist hierbei die Zugehörigkeit zu Gruppen, für die Unterschiede in der Ausprägung des zu messenden Konstruktes erwartet werden (Schnell et al., 2008, S. 156). So könnte man einen Einsamkeitsfragebogen z. B. dadurch validieren, dass man ihn einer „normalen“ und einer isolierten Gruppe (z. B. Strafgefangene) vorlegt. Höhere Einsamkeitswerte der isolierten Gruppe wären ein Indiz für die Validität des Fragebogens.

Leider ist die Kriteriumsvalidierung in ihrem Anwendungsbereich dadurch stark eingeschränkt, dass vielfach kein inhaltlich adäquates Außenkriterium für einen Test benannt werden kann. Neben der Schwierigkeit, überhaupt ein angemessenes Außenkriterium zu finden, stellt sich auch die Frage nach der Operationalisierung des Kriteriums. Sind Kriteriumswerte invalide oder unreliabel erfasst, so ist natürlich jede Validierung mit diesem Kriterium unbrauchbar. Weiterhin ist zu beachten, dass Korrelationen zwischen Testwert und Kriterium in unterschiedlichen Populationen verschieden ausfallen können **(differenzielle Validität).** Auf der Itemebene ist gelegentlich festzustellen, dass einzelne Items in verschiedenen Gruppen unterschiedliche Validitäten aufweisen. Weitere Einzelheiten zu diesem als **„Differential Item Functioning“** (DIF) bezeichneten Sachverhalt findet man z. B. bei Holland und Wainer (1993).

Konstruktvalidität Ursprünglich geprägt wurde das Konzept der Konstruktvalidität in den 1950er-Jahren (Cronbach & Meehl, 1955) als Ergänzung zur Inhalts- und Kriteriumsvalidität. Zunehmend wird heute in der Tradition von Messick (1980, 1995) die Bezeichnung Konstruktvalidität als Oberbegriff für alle mit Tests in Verbindung stehenden Validierungsansätze verwendet.

Ein Test ist konstruktvalide, wenn aus dem zu messenden Zielkonstrukt theoretisch und/oder empirisch gut fundierte Hypothesen ableitbar sind, die anhand der Testwerte bestätigt werden können. Anstatt ein einzelnes manifestes Außenkriterium zu benennen, formuliert man ein **Netz von Hypothesen** über das Konstrukt und seine Relationen zu anderen manifesten und latenten Variablen.

Beispiel: Ein Fragebogen zur Erfassung von subjektiver Einsamkeit soll validiert werden. Aus der Einsamkeitstheorie ist bekannt, dass Einsamkeit mit geringem Selbstwertgefühl und sozialer Ängstlichkeit einhergeht und bei Geschiedenen stärker ausgeprägt ist als bei Verheirateten. Diese inhaltlichen Hypothesen anhand der Testwerte zu prüfen, wäre Aufgabe einer Konstruktvalidierung. Die zur Konstruktvalidierung genutzten Hypothesen können und sollen sich sowohl darauf beziehen, mit welchen anderen Variablen der zu validierende Test eng zusammenhängt **(konvergente Validität)** und mit welchen er nicht oder nur sehr gering zusammenhängt **(divergente Validität)**.

Der Umstand, dass Testwerte so ausfallen, wie es die aus Theorie und Empirie abgeleiteten Hypothesen vorgeben, kann als Indiz für die Konstruktvalidität des Tests gewertet werden. Eine Konstruktvalidierung ist nur dann erfolgversprechend, wenn neben dem zu prüfenden Test für die anderen betrachteten Merkmale ausschließlich **gut gesicherte Instrumente** verwendet werden und inhaltlich auf **gut gesicherte Hypothesen** zurückgegriffen wird. Man spricht auch vom **nomologischen Netz**, um auszudrücken, dass es sich bei den der Konstruktvalidierung zugrunde liegenden Annahmen um ein Netz von empirischen Gesetzmäßigkeiten (und nicht nur um spekulative Hypothesen) handelt (Cronbach & Meehl, 1955). Fallen unter diesen Umständen die anhand des zu prüfenden Tests gewonnenen Ergebnisse hypothesenkonträr aus, so ist die Konstruktvalidität des Instrumentes anzuzweifeln. Eine Konstruktvalidierung ist umso überzeugender, je mehr gut gesicherte Hypothesen einbezogen werden und je besser die Testdaten die Hypothesen dann auch stützten.

Methodisch gibt es bei einer Konstruktvalidierung unterschiedliche Herangehensweisen. **Logisch-inhaltliche Analysen der Test-Items** können Hinweise geben, ob tatsächlich das fragliche Konstrukt (z. B. subjektive Einsamkeit) oder ein alternatives Konstrukt (z. B. Unzufriedenheit) erfasst wird. Mit **experimentellen Methoden** kann man herausfinden, ob die Variation von Merkmalen, die für das Konstrukt essenziell sind, zu unterschiedlichen Testwerten führt (die systematische Variation der Anzahl sozialer Kontakte sollte unterschiedliche Einsamkeitstestwerte nach sich ziehen). In der Praxis werden

Die Multitrait-Multimethod-Analyse

Der auf Campbell und Fiske (1959) zurückgehende Multitrait-Multimethod-Ansatz (**MTMM**) stellt eine besondere Variante der Konstruktvalidierung dar. Diese Validierungsstrategie erfordert es, dass mehrere Konstrukte („multi trait“) jeweils durch mehrere Erhebungsmethoden („multi method“) erfasst werden (Schermelleh-Engel & Schweizer, 2012). Eine systematische, regelgeleitete Analyse der wechselseitigen Beziehungen zwischen Konstrukten und Methoden erlaubt es, die Höhe der Konstruktvalidität abzuschätzen. Der Innovationsgehalt der MTMM-Analyse bestand in den 1950er-Jahren darin, dass neben der konvergenten, nun auch die **diskriminante Validität** in den Blick genommen wurde. Zudem war die Überlegung wichtig, dass es neben den von der Klassischen Testtheorie berücksichtigten unsystematischen Messfehlern, eben auch systematische Messfehler – sog. **Methodeneffekte** – gibt, die auf die Art der Erhebungsmethode (z. B. Test, Interview oder Beobachtung zur Erfassung von der Eignung für einen bestimmten Beruf) zurückgehen.

- **Konvergente Validität** („convergent validity“): Diese liegt vor, wenn mehrere Methoden dasselbe Konstrukt übereinstimmend (konvergent) messen, d. h., wenn verschiedene Operationalisierungen desselben Konstruktes auch zu ähnlichen Ergebnissen führen (Monotrait-Heteromethod-Korrelationen müssen hoch sein).
  Beispiele: Die Kooperationsfähigkeit einer Person (Trait) wird durch Kollegen (Method 1) und Vorgesetzte (Method 2) erfasst (Abb. 10.22). Die Einsamkeit einer Person (Trait) wird durch einen Einzelindikator (Method 1) und durch einen psychometrischen Test (Method 2) erfasst.
- **Diskriminante Validität** („discriminant validity“): Dieses Kriterium fordert, dass sich die Messungen des Zielkonstruktes von Messungen anderer Konstrukte unterscheidet (Heterotrait-Monomethod- sowie Heterotrait-Heteromethod-Korrelationen müssen gering sein).
  Beispiel: Die Kooperationsfähigkeit einer Person (Trait 1) und ihre Kreativität (Trait 2) wird a) jeweils durch Kollegen (Method 1) eingeschätzt (Abb. 10.23) oder b) Trait 1 wird durch Kollegen (Method 1) und Trait 2 durch Vorgesetzte (Method 2) eingeschätzt (Abb. 10.24).

derartige experimentelle Ansätze aber oft an ethische und forschungspraktische Grenzen stoßen. **Korrelationsstatistisch** wären Zusammenhänge zwischen den für ein Konstrukt relevanten Merkmalen bzw. Unabhängigkeit von irrelevanten Merkmalen nachzuweisen. Hypothesengemäß sollte z. B. Einsamkeit mit sozialer Ängstlichkeit hoch, aber mit Intelligenz nur wenig korrelieren.

Die Multitrait-Multimethod-Analyse Für eine besonders umfassende Konstruktvalidierung kann die **Multitrait-Multimethod-Analyse** eingesetzt werden, die mit eigenen Validierungskriterien und -anforderungen arbeitet und deswegen gesondert behandelt wird (▶ Die Multitrait-Multimethod-Analyse).

Multitrait-Multimethod-Ansatz – Der Multitrait-Multimethod-Ansatz (MTMM-Ansatz; „multitrait multimethod approach“) überprüft systematisch, mit welcher Übereinstimmung verschiedene Methoden dasselbe Konstrukt erfassen (konvergente Validität) und wie gut verschiedene Konstrukte durch eine Methode differenziert werden (diskriminante Validität). Der MTMM-Ansatz spielt für die Konstruktvalidierung von Tests eine zentrale Rolle. Zu unterscheiden sind die korrelationsbasierte MTMM-Analyse (mittels bivariater Korrelationskoeffizienten) und die konfirmatorische MTMM-Analyse (mittels Strukturgleichungsmodellen).

Mithilfe der Multitrait-Multimethod-Technik lassen sich zur Konstruktvalidierung die diskriminante und konvergente Validität anhand von Zusammenhangsmaßen systematisch abschätzen. Dabei werden die wechselseitigen Zusammenhänge zwischen Merkmalen und Methoden in einer speziellen Korrelationsmatrix, der sog. **Multitrait-Multimethod-Matrix** (kurz: **MTMM-Matrix**) dargestellt. Die statistische Auswertung kann dabei einerseits auf der Basis bivariater Korrelationen erfolgen – dieser Ansatz wird der Anschaulichkeit halber im Folgenden vorgestellt – oder als konfirmatorische MTMM-Analyse

		Kollege/in			Vorgesetzte/r		
		Koop	Kreat	Leist	Koop	Kreat	Leist
Kollege/in	Koop	1.0					
	Kreat		1.0				
	Leist			1.0			
Vorgesetzte/r	Koop				1.0		
	Kreat					1.0	
	Leist						1.0

Abb. 10.21 MTMM-Matrix. Eingetragene Werte: Monotrait-Monomethod-Korrelationen. *Koop* Kooperationsfähigkeit, *Kreat* Kreativität, *Leist* Leistungsfähigkeit

		Kollege/in			Vorgesetzte/r		
		Koop	Kreat	Leist	Koop	Kreat	Leist
Kollege/in	Koop						
	Kreat						
	Leist						
Vorgesetzte/r	Koop	.63					
	Kreat		.83				
	Leist			.58			

Abb. 10.22 MTMM-Matrix. Eingetragene Werte: Monotrait-Heteromethod-Korrelationen. *Koop* Kooperationsfähigkeit, *Kreat* Kreativität, *Leist* Leistungsfähigkeit

		Kollege/in			Vorgesetzte/r		
		Koop	Kreat	Leist	Koop	Kreat	Leist
Kollege/in	Koop						
	Kreat	.44					
	Leist	.55	.52				
Vorgesetzte/r	Koop						
	Kreat				.41		
	Leist				.64	.51	

Abb. 10.23 MTMM-Matrix. Eingetragene Werte: Heterotrait-Monomethod-Korrelationen. *Koop* Kooperationsfähigkeit, *Kreat* Kreativität, *Leist* Leistungsfähigkeit

		Kollege/in			Vorgesetzte/r		
		Koop	Kreat	Leist	Koop	Kreat	Leist
Kollege/in	Koop						
	Kreat						
	Leist						
Vorgesetzte/r	Koop		.19	.42			
	Kreat	.14		.37			
	Leist	.29	.29				

Abb. 10.24 MTMM-Matrix. Eingetragene Werte: Heterotrait-Heteromethod-Korrelationen. *Koop* Kooperationsfähigkeit, *Kreat* Kreativität, *Leist* Leistungsfähigkeit

auf der Basis von Strukturgleichungsmodellen (Schermelleh-Engel & Schweizer, 2012; ► Kap. 17).

Die MTMM-Matrix und ihre Elemente Die Entwicklung einer MTMM-Matrix wird im Folgenden an einem fiktiven Beispiel demonstriert: Im Kontext der Personalauswahl interessiert man sich dafür, wie kooperativ, kreativ und leistungsfähig potenzielle Mitarbeiterinnen und Mitarbeiter sind. Die genannten drei Konstrukte Kooperationsfähigkeit („Koop“), Kreativität („Kreat“) und Leistungsfähigkeit („Leist“) sollen einfachheitshalber anstelle von Tests durch Fremdbeurteilungen erfasst werden. Dabei

Tabelle 10.21 Ergebnisse der Messung von 3 Merkmalen (Traits; *Koop* Kooperationsfähigkeit, *Kreat* Kreativität, *Leist* Leistungsfähigkeit) mit 2 Methoden (Methods; Beurteilung durch Kollege/in; Vorgesetzte/r)

Zu beurteilende Person	Koop (Kollege/in)	Kreat (Kollege/in)	Leist (Kollege/in)	Koop (Vorgesetzte/r)	Kreat (Vorgesetzte/r)	Leist (Vorgesetzte/r)
1	4	2	3	5	3	4
2	6	6	7	4	5	5
...	...	...	...	...	...	...
n	7	3	9	7	6	7

werden sowohl die Urteile eines ehemaligen Arbeitskollegen als auch der/des letzten Vorgesetzten herangezogen (Urteile auf einer Ratingskala von 1: überhaupt nicht kooperativ/kreativ/leistungsfähig bis 10: sehr kooperativ/kreativ/leistungsfähig). Mit der MTMM-Technik kann getestet werden, ob sich die drei Zielkonstrukte tatsächlich unterscheiden (diskriminante Validität) und wie gut sich die beiden „Test"- bzw. Urteilsformen zur Operationalisierung der Konstrukte eignen (konvergente Validität).

Die Daten, auf denen die MTMM-Matrix beruht, bestehen zunächst aus einer Liste von Messwerten (hier Ratings) für die zu beurteilenden Personen (Tab. 10.21).

Diese Messwerte werden nun spaltenweise miteinander korreliert, so dass sich die in Abb. 10.21 aufgeführte MTMM-Matrix ergibt (zunächst ohne Einträge).

Diese MTMM-Matrix zerfällt in **vier Teilmatrizen**: Zwei Monomethod-Matrizen (links oben: Kollege-Kollege; rechts unten: Vorgesetzter-Vorgesetzter) und zwei Heteromethod-Matrizen (links unten und rechts oben: Vorgesetzter-Kollege, Kollege-Vorgesetzter; diese beiden Heteromethod-Matrizen sind identisch). Die MTMM-Matrix insgesamt, aber auch die beiden Monomethod-Teilmatrizen sind symmetrisch, d. h., oberhalb und unterhalb der Diagonale befinden sich dieselben Zelleneinträge. Es genügt also, jeweils nur die untere Dreiecksmatrix zu betrachten. Innerhalb der Teilmatrizen sind insgesamt vier unterschiedliche „Blöcke" von Zellen zu unterscheiden:

Monotrait-Monomethod-Block (Diagonale der Gesamtmatrix): Ein Konstrukt (Monotrait) wird mit einer Methode (Monomethod) gemessen. Korreliert man diese Werte mit sich selbst, ergeben sich perfekte Korrelationen ($r = 1.0$; Abb. 10.21). Manchmal werden die Diagonalelemente nicht dargestellt, oder es werden die Reliabilitätskoeffizienten eingetragen.

Monotrait-Heteromethod-Block (Diagonale der Heteromethod-Teilmatrix): Ein Konstrukt (Monotrait) wird mit mehreren Methoden (Heteromethod) gemessen (Abb. 10.22). Beispiel: Die Kooperationsfähigkeit der Personen wird durch einen ehemaligen Kollegen und den ehemaligen Vorgesetzten eingeschätzt. Die Übereinstimmung beider Einschätzungen ($r = .63$) ist indikativ für die konvergente Validität. Der Durchschnitt der Monotrait-Heteromethod-Korrelationen für die drei Konstrukte gilt als Maß für die konvergente Validität und sollte statistisch signifikant und bedeutsam größer als Null sein (mittlere bis große Effektgröße).

Heterotrait-Monomethod-Block (Dreiecksmatrix der Monomethod-Teilmatrix): Mehrere Konstrukte (Heterotrait) werden mit derselben Methode (Monomethod) gemessen und die Messwerte anschließend korreliert (Abb. 10.23). Beispiel: Die Kreativitätseinschätzungen durch den Kollegen werden mit den Kooperationsfähigkeitseinschätzungen durch den Kollegen korreliert ($r = .44$). Da hier unterschiedliche Konstrukte erfasst werden, sollten die Korrelationen nicht allzu groß sein, denn hohe Korrelationen würden auf Redundanzen in den Konstrukten oder auf unsensible Messungen hindeuten.

Heterotrait-Heteromethod-Block (Heteromethod-Teilmatrix ohne Diagonale): Mehrere Konstrukte (Heterotrait) werden mit unterschiedlichen Methoden (Heteromethod) gemessen und miteinander korreliert (Abb. 10.24). Beispiel: Die Kreativitätseinschätzungen durch den Kollegen werden mit den Kooperationsfähigkeitseinschätzungen durch den Vorgesetzten korreliert ($r = .19$). Hier werden die geringsten Korrelationen erwartet, da weder methodische noch inhaltliche Übereinstimmungen vorliegen.

Kriterien für konvergente und diskriminante Validität

Campbell und Fiske (1959) schlagen vier Kriterien vor, anhand derer bei korrelationsbasierten MTMM-Analysen über das Vorliegen von konvergenter und diskriminanter Validität entschieden wird, wobei diese Kriterien teilweise recht viel Spielraum für subjektive Interpretationen lassen:

- Kriterium 1 für konvergente Validität: Konvergente Validität liegt vor, wenn die konvergenten Va-

		Kollege/in			Vorgesetzte/r		
		Koop	Kreat	Leist	Koop	Kreat	Leist
Kollege/in	Koop	1.0					
	Kreat	.44	1.0				
	Leist	.55	.52	1.0			
Vorgesetzte/r	Koop	.63	.19	.42	1.0		
	Kreat	.14	.83	.37	.41	1.0	
	Leist	.29	.29	.58	.64	.51	1.0

Abb. 10.25 Vollständige MTMM-Matrix mit allen vier Blöcken. *Koop* Kooperationsfähigkeit, *Kreat* Kreativität, *Leist* Leistungsfähigkeit

liditätskoeffizienten (Monotrait-Heteromethod-Korrelationen, Abb. 10.22) bzw. ihr Mittelwert signifikant größer als Null sind.

- Kriterium 2 für diskriminante Validität: Die Heterotrait-Monomethod-Korrelationen (Abb. 10.23) sollten signifikant kleiner sein als die Monotrait-Heteromethod-Korrelationen. Dies bedeutet, dass Differenzierungen zwischen verschiedenen Konstrukten (Heterotrait) nicht durch die Verwendung derselben Methode (Monomethod) verwischt werden dürfen. Trotz Verwendung derselben Operationalisierungsform (z. B. Einschätzung durch einen Kollegen) müssen die Konstrukte Kreativität und Kooperation „diskriminierbar" sein.
- Kriterium 3 für diskriminante Validität: Die Heterotrait-Heteromethod-Korrelationen (Abb. 10.24) sollten signifikant kleiner sein als die Monotrait-Heteromethod-Korrelationen. Insgesamt ist zu erwarten, dass die Heterotrait-Heteromethod-Korrelationen am kleinsten sind.
- Kriterium 4 für Konstruktvalidität: Konvergente und diskriminante Validität sind Voraussetzungen für eine gute Konstruktvalidität. Indikativ für das gemeinsame Vorliegen von konvergenter und diskriminanter Validität sind identische Muster von Traitinterkorrelationen in allen Monomethod- und Heteromethodteilmatrizen, d. h., die Rangreihe der Traitinterkorrelationen sollte in allen Teilmatrizen identisch sein (man muss also die vollständige Matrix betrachten, Abb. 10.25). In der oben dargestellten vollständigen MTMM-Matrix ist die Korrelation zwischen Leistung und Kooperation jeweils am größten, gefolgt von Leistung und Kreativität und schließlich Kreativität und Kooperation. Diese interne „Replizierbarkeit" der Rangreihe spricht dafür, dass hier „wahre" Varianz gemessen wird bzw. eine „wahre" Korrelationsstruktur zwischen den Traits besteht, die mit den betrachteten Methoden valide gemessen werden können.

  Ein weiterer Hinweis auf Konstruktvalidität wäre z. B. der Umstand, dass man die gefundene Rangreihe zumindest im Nachhinein auf der Basis von theoretischem und empirischem Hintergrundwissen plausibel machen kann. Es ist zu beachten, dass auch beim Nachweis konvergenter und diskriminanter Validität nie zweifelsfrei sichergestellt ist, dass tatsächlich das angezielte Konstrukt erfasst wird. Obwohl die Urteile von Kollegen und Vorgesetzten den Regeln der MTMM-Analyse entsprechen, könnten sie dennoch beide grundlegend verzerrt sein, etwa wenn übereinstimmend Kooperationsfähigkeit als Unterwürfigkeit missdeutet wird.

Multitrait-Multimethod-Analysen sind sehr aufwändig; einfacher ist eine reduzierte Variante, bei der statt gänzlich verschiedener Methoden lediglich mehrere Indikatoren (Items) für dasselbe Konstrukt erhoben werden (vgl. Schnell et al., 2008, S. 160). Neuere Auswertungsmethoden für MTMM-Matrizen – etwa auf der Basis von **Strukturgleichungsmodellen** (▶ Kap. 17) – sowie weiterführende Literatur findet man bei Schmitt und Stults (1986); Eid (2000), Eid, Lischetzke, Nussbeck und Trierweiler (2003); Grayson und Marsh (1994), Kiers, Takane und Ten Berge (1996) und Schermelleh-Engel & Schweizer (2012). Eine kurze Zusammenstellung der wichtigsten quantitativen Auswertungstechniken ist Lance, Noble und Scullen (2002) zu entnehmen.

Itemanalyse

Die Qualität eines psychometrischen Tests oder Fragebogens bzw. der aus den Testwerten abgeleiteten Interpretationen ist nicht zuletzt abhängig von der Art und der Zusammensetzung der Items, aus denen der Test besteht.

Die Itemanalyse (Aufgabenanalyse, „item analysis“) ist deswegen ein zentrales Instrument der Testkonstruktion und Testbewertung, in deren Verlauf die Item-Eigenschaften als Kennwerte bestimmt und kontextspezifisch beurteilt werden. Grundlage der Itemanalyse sollte nach Möglichkeit eine sog. **Eichstichprobe** sein, d. h. ein Miniaturabbild genau jener Population, für die der Test konzipiert ist. So führt man die Itemanalyse für einen Test zur Gedächtnisleistung im Alter am besten an einer repräsentativen Stichprobe älterer Testpersonen durch und nicht etwa an Studierenden.

Itemanalyse – Im Rahmen der Itemanalyse („item analysis“) wird anhand einer möglichst repräsentativen Stichprobe der Zielpopulation des Tests die Verteilung der Rohwerte der Items und des Gesamttests betrachtet. Zudem werden die Itemschwierigkeiten und Itemtrennschärfen berechnet und bewertet. Auch die Dimensionalitätsprüfung – also die Frage, ob sich alle Items eines Messinstruments auf eine Skala zusammenfassen lassen (eindimensionaler Test) oder ob Subskalen zu bilden sind (mehrdimensionaler Test) – wird der Itemanalyse zugerechnet. Ziel der Itemanalyse ist es, problematische Items zu entdecken und durch ihre Modifikation oder Eliminierung die Testgüte des Tests (v. a. Reliabilität und Validität) zu erhöhen.

Der Begriff „Itemanalyse“ ist in der Literatur nicht eindeutig festgelegt. Meistens werden – bei Testkonstruktionen gemäß Klassischer Testtheorie – die Berechnung von **Itemschwierigkeiten** und **Itemtrennschärfen** sowie die **Dimensionalitätsprüfung** zur Itemanalyse gezählt (zu Details der Durchführung einer Itemanalyse s. Bühner, 2011, Kap. 5.1 bis Kap. 5.3 sowie Kelava & Moosbrugger, 2012). Auch die **Rohwerteverteilung** von Items und Gesamttest wird betrachtet. Für Tests, die nach einem Probabilistischen Testmodell konstruiert werden, erübrigt sich eine entsprechende Itemanalyse, da hier stattdessen Modelltests vorgenommen werden, die die Verträglichkeit der Items mit den Annahmen des jeweiligen Testmodells detailliert überprüfen.

Rohwerteverteilung Die Häufigkeitsverteilung der Items sowie des Gesamttestwertes (grafisch darstellbar als Histogramm) vermittelt einen ersten Überblick über das Antwortverhalten der untersuchten Testpersonen. Am Histogramm ist z. B. abzulesen, wie stark die Testergebnisse streuen, d. h., ob sie den gesamten Wertebereich ausfüllen oder sich um bestimmte Werte konzentrieren. Häufig interessiert man sich dafür, ob die Rohwerteverteilung des Testscores einer Normalverteilung entspricht. Normalverteilte Testwerte sind erstrebenswert, weil viele inferenzstatistische Verfahren normalverteilte Werte voraussetzen. Ob die empirisch gefundene Verteilung überzufällig von einer Normalverteilung abweicht oder nicht, kann mit dem sog. **Goodness-of-Fit-Chi-Quadrat-Test** (vgl. Bortz & Schuster, 2010) oder mit dem **Kolmogoroff-Smirnov-Test** (Bortz & Lienert, 2008, S. 240ff.; Bortz, Lienert, & Boehnke, 2008, S. 299ff.) überprüft werden.

Intelligenztests z. B. sind extra so angelegt, dass sie normalverteilte Testergebnisse produzieren, was in Einklang steht mit der inhaltlichen Vorstellung, dass die meisten Menschen mittlere Intelligenz aufweisen, während extrem hohe oder extrem niedrige Intelligenz nur selten auftritt. Nicht bei allen Konstrukten ist eine Normalverteilung zu erwarten. Bei der Erfassung von Lebenszufriedenheit zeigt sich z. B. in vielen Kontexten, dass die Befragten nicht etwa mittelmäßig, sondern eher zufrieden sind.

Stellt sich heraus, dass die Rohwerteverteilung von einer Normalverteilung abweicht, sind folgende Konsequenzen in Erwägung zu ziehen:

- Sofern aus theoretischer Sicht normalverteilte Merkmalsausprägungen zu erwarten sind, modifiziert man die Item-Zusammensetzung des Tests in der Weise, dass die revidierte Version normal verteilte Ergebnisse produziert.
- Ist die Nicht-Normalverteilung der Testwerte theoriekonform, kann der Test unverändert bleiben. Allerdings muss die statistische Auswertung (z. B. Gruppenvergleiche) auf die Verletzung der Normalverteilungsvoraussetzung abgestimmt werden. Zwei Strategien sind möglich: Entweder man operiert mit größeren Stichproben (ab ca. 30 Untersuchungsobjekten), wodurch sich gemäß dem zentralen Grenzwerttheorem die Forderung nach normalverteilten Messwerten als Voraussetzung statistischer Analysen in der Regel erübrigt (Bortz & Schuster, 2010, S. 85ff.), oder man verwendet (v. a. bei kleinen Stichproben) statt der „normalen“ (verteilungsgebundenen) statistischen Verfahren die sog. verteilungsfreien Analysetechniken (Bortz & Lienert, 2008).

Über mögliche Ursachen nicht normal verteilter Testwerte und nachträgliche Normalisierungsverfahren berichten z. B. Lienert und Raatz (1994, Kap. 8 und Kap. 12).

Itemschwierigkeiten Items besitzen unterschiedliche Lösungs- bzw. Zustimmungsraten, die als Itemschwierigkeiten (Itemschwierigkeitsindizes) quantifizierbar sind. Schwierige Items werden nur von wenigen Testpersonen bejaht bzw. richtig gelöst. Bei leichten Items kommen dagegen fast alle Testpersonen zum richtigen Ergebnis. Die Itemschwierigkeiten beeinflussen also ganz wesentlich die Verteilung der Testwerte. Der Schwierigkeitsindex wird für jedes Item eines Tests bzw. eines Itempools einzeln berechnet, wobei die Berechnung vom Skalenniveau abhängt.

Itemschwierigkeit – Die Schwierigkeit eines Items („item difficulty") wird durch einen Schwierigkeitsindex ausgedrückt. Ein Item ist umso schwieriger, je geringer seine Zustimmungs- bzw. Lösungsrate in der betrachteten Stichprobe ausfällt. Umgekehrt wird ein leichtes Item von fast allen Testpersonen in Schlüsselrichtung beantwortet bzw. korrekt gelöst. Bei **nominalskalierten Items** wird der Schwierigkeitsindex über den **Prozentsatz der Zustimmer/Richtiglöser** berechnet. Bei **intervallskalierten Items** entspricht der Schwierigkeitsindex dem Item-Mittelwert, der wiederum in einen Prozentwert umgerechnet werden kann.

Es ist zu beachten, dass nach dieser Definition eine hohe inhaltliche Schwierigkeit des Items (d. h. es wird von nur wenigen Testpersonen richtig gelöst bzw. bejaht) mit einem geringen numerischen Schwierigkeitsindex einhergeht (z. B. $p = 6\,\%$; $M = 1.12$). Ein hoher numerischer Schwierigkeitsindex (z. B. $p = 98\,\%$; $M = 4.77$) deutet dagegen auf ein leichtes Item hin (das von fast allen Testpersonen richtig gelöst bzw. zustimmend beantwortet wurde). Es handelt sich also eigentlich eher um einen **„Leichtigkeitsindex"**.

Bei dichotomen Antwortalternativen erhält man die Schwierigkeit (p_i) von Item i, indem die Anzahl der richtigen Lösungen bzw. Zustimmungen (R) durch die Gesamtzahl der Antworten (N) dividiert wird; der **Anteil der „Richtiglöser" oder „Zustimmer"** p_i ist damit der **Schwierigkeitsindex für dichotome Items**:

$$p_i = \frac{R_i}{N_i}. \tag{10.10}$$

Ein Schwierigkeitsindex von $p_i = .50$ bei dichotomem Antwortformat besagt, dass das Item von 50 % der Untersuchungsteilnehmenden richtig gelöst (bzw. bejaht) und von 50 % falsch beantwortet (bzw. verneint) wurde (Fisseni, 1990, S. 30ff.; Lienert & Raatz, 1994).

Handelt es sich um kardinalskalierte Items, bei denen das Antwortformat z. B. eine Ratingskala ist, so wird als **Schwierigkeitsindex** der **arithmetische Mittelwert** des **intervallskalierten Items** berechnet (vgl. Bühner, 2011; Rost, 2004). Je höher der Item-Mittelwert und damit der numerische Schwierigkeitsindex, umso stärker wird dem Item zugestimmt, d. h. umso leichter ist dieses Item inhaltlich (Gl. 10.11).

$$\overline{x_i} = \frac{\sum_{v=1}^{N} x_{vi}}{N} \tag{10.11}$$

$\bar{x}$ = Mittelwert des Items
x_{vi} = Itemrohwert der Person v in Item i
N = Stichprobengröße

Anstelle des Mittelwertes wird auch bei kardinalskalierten Variablen der Schwierigkeitsindex gern als Prozentwert angegeben. Dieser **prozentuale Schwierigkeitsindex** P_i eines **intervallskalierten Items** i ist (sofern die Antwortskala mit dem Messwert 0 beginnt) einfach der Quotient aus der bei diesem Item tatsächlich erreichten Punktsumme aller n Testpersonen und der maximal erreichbaren Punktsumme, multipliziert mit 100 (Kelava & Moosbrugger, 2012, S. 76f.) und berechnet sich nach (Gl. 10.12)

$$P_i = \frac{\sum_{v=1}^{n} x_{vi}}{n \cdot \max(x_i)} \cdot 100. \tag{10.12}$$

$\sum x_{vi}$ = Punktsumme aller Testpersonen beim Item i
$\max(x_i)$ = maximal erreichbare Punktsumme

Beginnt die Antwortskala des Items nicht mit Wert 0, sondern mit einem anderen Wert (z. B. 1), so muss eine verallgemeinerte Umrechnungsformel genutzt werden, um kardinalskalierte Itemantworten in einen prozentualen Schwierigkeitsindex zu überführen (Kelava & Moosbrugger, 2012, S. 77):

$$P_i = \frac{\sum_{v=1}^{n} [x_{vi} - \min(x_i)]}{n \cdot [\max(x_i) - \min(x_i)]} \cdot 100. \tag{10.13}$$

x_{vi} = Punktsumme aller Testpersonen beim Item i
$\max(x_i)$ = maximal erreichbare Punktsumme
$\min(x_i)$ = minimal erreichbare Punktsumme

Extrem schwierige Items, denen kaum jemand zustimmt (z. B. nur 0 %–20 % Zustimmer/Löser), oder extrem leichte Items, die von fast allen Testpersonen gelöst werden (z. B. 80 %–100 % Zustimmer/Löser), sind wenig informativ, da sie keine Personenunterschiede sichtbar machen. Damit ein Test die verschiedenen Testpersonen mit ihren unterschiedlichen Fähigkeiten annähernd gleich gut differenziert, ist darauf zu achten, dass die Items eine möglichst breite Schwierigkeitsstreuung aufweisen, das im Test also leichte bis schwierige Items vertreten sind. Im Allgemeinen werden (angesichts des Verzichts auf extrem schwierige und extrem leichte Items) schwere bis leichte Items mit **Schwierigkeitsindizes zwischen .20 und .80** bevorzugt. Zur Kennzeichnung eines gesamten Tests kann die **durchschnittliche Itemschwierigkeit** angegeben werden (für ein Beispiel s. Tab. 10.22).

Itemtrennschärfen Die Trennschärfe bzw. der Trennschärfekoeffizient gibt an, wie gut ein einzelnes Item das Zielkonstrukt des Tests misst bzw. wie hoch das Item mit dem Gesamttestwert korreliert. Bei solchen Trennschärfen handelt es sich um sog. **Eigentrennschärfen**. Diese werden im Folgenden behandelt. Von Eigentrennschärfen sind Fremdtrennschärfen, d. h. Itemkorrelationen mit Außenkriterien im Sinne der Item-Validität, abzugrenzen (Bühner, 2011).

Tabelle 10.22 Itemschwierigkeiten der Neurotizismus-Subskala des NEO-FFI ($n = 611$). *M* Mittelwert, *SD* Standardabweichung, *P* prozentualer Schwierigkeitsindex

Items	*M*	*SD*	*P*
1. Ich bin leicht beunruhigt.	2.07	.94	51.76
2. Ich fühle mich anderen oft unterlegen.	1.64	1.01	40.92
3. Wenn ich unter starkem Stress stehe, fühle ich mich manchmal, als ob ich zusammenbräche.	1.98	1.11	49.51
4. Ich fühle mich selten einsam und traurig.	1.82	1.12	45.46
5. Ich fühle mich oft angespannt und nervös.	1.84	1.07	45.91
6. Manchmal fühle ich mich völlig wertlos.	1.64	1.24	40.96
7. Ich empfinde selten Furcht oder Angst.	1.96	.97	48.90
8. Ich ärgere mich oft darüber, wie Leute mich behandeln.	2.10	.97	52.41
9. Zu häufig bin ich entmutigt und will aufgeben, wenn etwas schief geht.	1.52	1.05	37.97
10. Ich bin selten traurig oder deprimiert.	1.91	1.09	47.79
11. Ich fühle mich oft hilflos und wünsche mir eine Person, die meine Probleme löst.	1.64	1.10	40.96
12. Manchmal war mir etwas so peinlich, dass ich mich am liebsten versteckt hätte.	2.29	1.07	57.24
Skala	**22.39**	**8.74**	**46.65**

Die Trennschärfe wird für jedes Item einer Skala berechnet und ist definiert als die Korrelation der Beantwortung dieses Items mit dem Summenwert der übrigen Skala (Fisseni, 1990, S. 40f.; Lienert & Raatz, 1994). Der zu berechnende Korrelationskoeffizient richtet sich nach dem Skalenniveau der Testwerte. Bei intervallskalierten Testwerten wählt man als Trennschärfe (r_{it}) die Produkt-Moment-Korrelation zwischen den Punktwerten pro Item i und dem korrigierten Gesamtskalenwert t (Summen- oder Durchschnittsscore der Testskala ohne Berücksichtigung des betrachteten Items):

$$r_{it} = \frac{\mathrm{cov}(i,t)}{s_i \cdot s_t}. \qquad (10.14)$$

Der Begriff „Trennschärfe" ist so zu verstehen, dass Personen, die im Gesamtergebnis der Skala einen hohen Wert erreichen, auf einem trennscharfen Einzelitem ebenfalls eine hohe Punktzahl aufweisen. Umgekehrtes gilt für Personen mit niedrigem Testergebnis. Nach diesem Verständnis lässt sich an einem trennscharfen Einzelitem bereits ablesen, welche Personen bezüglich des betrachteten Konstruktes hohe oder niedrige Ausprägungen besitzen. Beide Gruppen werden durch das Item also gut voneinander „getrennt".

Itemtrennschärfe – Die Trennschärfe eines Items („item discriminability") gibt an, wie gut die Beantwortung des betreffenden Items das Gesamttestergebnis vorhersagt. Bei trennscharfen Items gilt, dass Personen mit hohem (niedrigem) Testwert auch bei dem betreffenden Item eine hohe (niedrige) Punktzahl erreichen. Der Trennschärfe-Index berechnet sich als Korrelation des Items mit dem Gesamttestwert (als Zusammenfassung aller Items ohne das betrachtete Item).

Bei eindimensionalen Tests sind hohe Trennschärfen erstrebenswert: Trennschärfekoeffizienten haben einen korrelationstypischen Wertebereich von $-1 \leq r_{it} \leq +1$, dabei werden positive Werte zwischen .30 und .50 zuweilen als mittelmäßig und Werte größer als .50 hoch eingeordnet (Weise, 1975, S. 219; für ein Beispiel s. Tab. 10.23). Entsprechende Bewertungskriterien sind jedoch niemals mechanisch anwendbar, sondern müssen immer inhaltlich interpretiert werden. Items mit geringer Trennschärfe, die Informationen generieren, die nicht mit dem Gesamtergebnis übereinstimmen, sind nur bei einem **engen Konstrukt** bzw. einem homogenen Test als schlechte Indikatoren des angezielten Konstruktes zu betrachten und aus einem eindimensional angelegten Test bzw. einer Subskala zu entfernen. Soll ein **breites Konstrukt** über viele Verhaltensausschnitte operationalisiert werden, so kann eine sehr hohe Trennschärfe sogar inhaltlich problematisch sein, wenn sie darauf hindeutet, dass es sich um ein redundantes Item handelt. Hier würde man geringere Trennschärfen erwarten und zulassen zugunsten einer größeren inhaltlichen Breite der Items.

Es ist zu beachten, dass die Trennschärfe eines Items von seiner Schwierigkeit abhängt: Je extremer die Item-

Tabelle 10.23 Itemtrennschärfen (korrigierte Item-Skala-Korrelationen) der Neurotizismus-Subskala des NEO-FFI ($n = 661$)

	Korrigierte Item-Skala-Korrelation (r_{it})	Cronbachs Alpha, wenn Item weggelassen
1. Ich bin leicht beunruhigt.	.63	.89
2. Ich fühle mich anderen oft unterlegen.	.64	.89
3. Wenn ich unter starkem Stress stehe, fühle ich mich manchmal, als ob ich zusammenbräche.	.56	.89
4. Ich fühle mich selten einsam und traurig.	.55	.89
5. Ich fühle mich oft angespannt und nervös.	.71	.88
6. Manchmal fühle ich mich völlig wertlos.	.75	.88
7. Ich empfinde selten Furcht oder Angst.	.54	.89
8. Ich ärgere mich oft darüber, wie Leute mich behandeln.	.53	.89
9. Zu häufig bin ich entmutigt und will aufgeben, wenn etwas schief geht.	.69	.89
10. Ich bin selten traurig oder deprimiert.	.62	.89
11. Ich fühle mich oft hilflos und wünsche mir eine Person, die meine Probleme löst.	.70	.88
12. Manchmal war mir etwas so peinlich, dass ich mich am liebsten versteckt hätte.	.43	.90

Schwierigkeit, desto geringer die Item-Trennschärfe. Bei sehr leichten und sehr schweren Items wird man deshalb Trennschärfeeinbußen in Kauf nehmen müssen. Items mit mittleren Schwierigkeiten besitzen die höchsten Trennschärfen (► Itemanalyse der Neurotizismus-Subskala aus dem NEO-FFI).

Dimensionalität von Test und Items Generell unterscheiden wir eindimensionale Tests (alle Items messen dasselbe globale Konstrukt und gehen in einen gemeinsamen Gesamttestwert ein) und mehrdimensionale Tests (die Items messen verschiedene Subdimensionen des Konstruktes und gehen in verschiedene Subskalen ein, für die jeweils separate Subskalenwerte berechnet werden). Im Zuge der Dimensionalitätsprüfung wird empirisch untersucht, ob die Anzahl der Dimensionen sowie – bei mehrdimensionalen Tests – die Zuordnung der Items zu den Dimensionen den theoretischen Erwartungen gemäß Konzeptspezifikation (► Abschn. 8.2) des zu messenden Merkmals entspricht. Die Frage der theoretisch vorgegebenen und empirisch nachweisbaren Dimensionalität eines Tests wird oft auch als sog. **faktorielle Validität** bezeichnet und als Unteraspekt der Konstruktvalidität betrachtet (► Abschn. 10.4.1). Wir behandeln die Dimensionalität hier im Kontext der Itemanalyse, da sie auf die Betrachtung von Kennwerten für alle Items eines Tests hinausläuft, um zu beurteilen, ob und wie gut die einzelnen Items jeweils zu einem Gesamttest-Wert oder zu mehreren Untertest-Werten zusammenzufassen sind. Die Dimensionalitätsprüfung der Items sowie des Tests basiert statistisch auf Korrelationsanalysen (Item-Interkorrelationen) sowie auf exploratorischen oder konfirmatorischen Faktorenanalysen.

Dimensionalität eines Tests – Die Dimensionalität eines Tests („test dimensionality“) gibt an, ob er mit seinen verschiedenen Test-Items nur ein globales Konstrukt erfasst (eindimensionaler Test), oder ob er mehrere Teilkonstrukte operationalisiert (mehrdimensionaler Test mit zwei oder mehr Untertest). Im Zuge der Dimensionalitäts-Überprüfung wird mit Korrelationsanalysen bzw. vor allem mit exploratorischen oder konfirmatorischen Faktorenanalysen geprüft, ob a) die Anzahl der Dimensionen sowie b) bei mehrdimensionalen Tests auch die Zuordnung der Items zu den verschiedenen Subtests („item dimensionality“) den inhaltlichen Vorgaben der Konzeptspezifikation entspricht.

Dimensionalitätsprüfung bei eindimensionalen Tests. Bei einem eindimensionalen Test messen alle Items dasselbe Konstrukt. Der Testwert wird als additiver Summen- oder Durchschnittswert aus allen Items berechnet. Dabei kann der Testwert eines eindimensionalen Tests ein ungewichteter additiver Index oder ein gewichteter additiver Index sein (z. B. indem man alle Items mit ihren Faktorladungen gewichtet und dann aufsummiert; zur Indexbildung ► Abschn. 8.7). Bei einem eindimensionalen Test erwarten wir eine hohe **Homogenität** („test homo-

Itemanalyse der Neurotizismus-Subskala aus dem NEO-FFI

Die **Neurotizismus-Subskala** aus dem Persönlichkeitstest NEO-FFI (Borkenau & Ostendorf, 2008; ▶ Abschn. 10.4.3) erfasst mit emotionaler Labilität eine der fünf zentralen Persönlichkeitseigenschaften. Alle Items sind jeweils auf einer **5-stufigen Ratingskala** zu beantworten (starke Ablehnung: 0, Ablehnung: 1, neutral 2; Zustimmung: 3; starke Zustimmung: 4). Die mittels Ratingskala erhobenen Messwerte werden als **intervallskaliert** aufgefasst (▶ Abschn. 8.4.4; mithilfe der Probabilistischen Testtheorie könnte diese Skalierungs-Annahme auch empirisch geprüft werden).

Im Folgenden soll an einem empirischen Datensatz demonstriert werden (Datensatz aus Pöschl, 2010), wie die Schwierigkeiten und Trennschärfen der einzelnen Items der Testskala ausfallen und inwiefern die Itemanalyse zur Verbesserung der Testgüte genutzt werden kann.

Die Neurotizismus-Subskala wurde einer Stichprobe von $n = 611$ **Testpersonen** (Alter $M = 36.21$, SD $= 11.92$; 59 % Frauen, 41 % Männer; 77 % Berufstätige, 13 % Studierende/Schüler/in, Ausbildung, 10 % anderes) vorgelegt. Will man nun die Itemschwierigkeiten der 12 Skalen-Items bestimmen, so muss angesichts des Intervallskalenniveaus der Messungen als **Schwierigkeitsindex jeweils der Mittelwert** (M) inklusive Standabweichung (SD) pro Item berechnet werden. Die Itemschwierigkeiten können angesichts der verwendeten Ratingskala (0–4) zwischen dem **Mittelwert 0** (höchste Schwierigkeit: alle Testpersonen lehnen die Aussage des Items stark ab) und dem **Mittelwert 4** (geringste Schwierigkeit: alle Testpersonen stimmen der Aussage des Items stark zu) variieren. Zusätzlich zu den Mittelwerten werden die umgerechneten **prozentualen Schwierigkeitsindizes** P (nach Gl. 10.12) angegeben (Tab. 10.22). Es zeigt sich, dass in der verwendeten Stichprobe von $n = 611$ Testpersonen weder sehr leichte Items (Mittelwerte von 3 oder 4 bzw. Prozentwerte von 70 % oder höher), noch sehr schwierige Items (Mittelwerte von 0 oder 1 bzw. Prozentwerte von 20 % oder geringer) vorkommen, sondern dass die meisten Items mittlere Schwierigkeitsindizes aufweisen (Mittelwerte um 2; Prozentwerte um 50 %). Die meisten Testpersonen in der Stichprobe schätzen ihre Neigung zum Neurotizismus bzw. zu emotionaler Labilität also eher neutral oder tendenziell gering ein. Das für die Stichprobe schwierigste Item mit der vergleichsweise geringsten Zustimmung war Item 9 mit $M = 1.52$ bzw. $P = 37.97$ und das leichteste Item mit der vergleichsweise stärksten Zustimmung war Item 12 mit $M = 2.29$ bzw. $P = 57.24$. Es ist also festzuhalten, dass die 12 Items der betrachteten Neurotizismus-Skala anhand der untersuchten Stichprobe $n =$ **durchgängig mittlere Schwierigkeiten** zeigen.

Zu beachten ist: Die im Kontext der Klassischen Testtheorie berechneten Gütekriterien und Kennwerte sind immer **stichprobenabhängig**. Das Ergebnis kann also auf den Test (er differenziert mit seinen verschiedenen Items die untersuchten Testpersonen offenbar wenig hinsichtlich ihrer Neurotizismus-Ausprägungen) oder auf die Stichprobe (in ihr ist Neurotizismus offenbar durchgängig mit mittlerer Intensität ausgeprägt) bezogen werden. Für eine umfassende Interpretation würde man für andere Stichproben – am besten repräsentative Normstichproben – publizierte Itemschwierigkeiten heranziehen.

Betrachten wir nun die Messgenauigkeit der Skala, ermittelt über **Cronbachs Alpha-Koeffizienten** (SPSS → Analysieren → Skalierung → Reliabilitätsanalyse → Modell: Alpha).

Mit einem Reliabilitätskoeffizienten von $\alpha = .897$ – berechnet mit der internen Konsistenzmethode gemäß Gl. 10.6 – kann die Messgenauigkeit dieser homogenen, aus reflektiven Indikatoren bestehenden Neurotizismus-Testskala als sehr hoch eingestuft werden.

Über die Berechnung der Itemschwierigkeiten kann nun geprüft werden, wie hoch die Items mit dem Skalenwert korrelieren bzw. ob sich die Reliabilität der Skala durch das Herausnehmen eines gering korrelierenden Items ggf. noch steigern ließe. Um sich die Itemschwierigkeiten in SPSS anzeigen zu lassen, muss bei der Reliabilitätsbestimmung in SPSS Analysieren → Skalierung → Reliabilitätsanalyse → Modell: Alpha und unter dem Menü „Statistiken“ die Option „Deskriptive Statistiken für Item“ angewählt werden. Die Itemtrennschärfen werden als „korrigierte Item-Skala-Korrelationen“ ausgegeben (wenn im Menü „Statistiken“ → „Deskriptive Statistiken“ für „Skala“ sowie „Skala wenn Item gelöscht“ ausgewählt werden) und es wird direkt angegeben, wie sich der Alpha-Koeffizient verändert, wenn man das Item aus der Skala herausnehmen würde (Tab. 10.23).

Die Items der Neurotizismus-Skala weisen hohe Trennschärfen ($r_{it} > .50$) auf. Somit kann auch kein „unpassendes“ Item identifiziert werden, dessen Herausnahme aus der Skala die Reliabilität nennenswert steigern würde. Das Item mit der höchsten Trennschärfe ist Item 6 mit $r_{it} = .75$, d. h. Testpersonen, die der Aussage „Manchmal fühle ich mich völlig wertlos“ zustimmen, haben tendenziell auch einen hohen Neurotizismus-Wert. Das Item mit der geringsten Trennschärfe ist Item 12, das sich auf das Erleben von Scham bezieht. Die vergleichsweise geringe Itemschwierigkeit von $r_{it} = .43$ zeigt an, dass das Item nicht sehr gut zwischen neurotischen und weniger neurotischen Personen differenziert. Dieses Ergebnis ist auch daraus verständlich, dass Item 12 das vergleichsweise leichteste Item ist, ihm wird also in der Stichprobe tendenziell zugestimmt, und zwar von Personen mit unterschiedlicher Neurotizismus-Ausprägung.

Fazit: Bei der vorliegenden Neurotizismus-Subskala aus dem NEO-FFI hat die Analyse der Itemschwierigkeiten und der Itemtrennschärfen keinen Änderungsbedarf ergeben; die interne Konsistenz der Skala ist sehr hoch.

geneity“), d. h. alle Items des Tests müssen inhaltlich ähnlich (homogen) sein und empirisch untereinander relativ hoch korrelieren, denn sie messen ja (meist als reflektive Indikatoren) ein und dasselbe Konstrukt. Es ist zu beachten, dass die Homogenitätsabschätzung auf der Basis der Interkorrelationen der Items relativ ungenau ist und Itemhomogenität im engeren Sinne (also lokale stochastische Unabhängigkeit) nur im Rahmen der Item-Response-Theorie empirisch zu ermitteln ist (Moosbrugger, 2012a, S. 230ff.).

Für eine grobe Begutachtung der Dimensionalität kann im einfachsten Fall die **Korrelationsmatrix** der Items herangezogen werden, die alle $k(k-1)/2$ bivariaten Item-Interkorrelationen ($r_{ii'}$) enthält. Die Item-Interkorrelationen sollten bei einem eindimensionalen Test alle positiv ausfallen und signifikant von Null abweichen. Die Ho-

mogenität des Tests kann über den Mittelwert der Item-Interkorrelationen ($\bar{r}_{ii'}$) quantifiziert werden (zur Berechnung einer durchschnittlichen Korrelation vgl. Bortz & Schuster, 2010, S. 161). Briggs und Cheek (1986, S. 115) schlagen zur Bewertung von Gesamttest-Homogenitäten einen Akzeptanzbereich von .20 bis .40 vor. Innerhalb dieses Bereiches soll eine hinreichende Homogenität eines eindimensionalen Tests gewährleistet sein, ohne dass gleichzeitig die inhaltliche Bandbreite des gemessenen Konstruktes durch übermäßige Redundanz zu sehr eingeschränkt wird. Die Item-Interkorrelationen gehen auch in den zur Reliabilitätsschätzung verwendeten Alpha-Koeffizienten von Cronbach ein. Zuweilen wird deshalb der Alpha-Koeffizient auch als Homogenitätsindex bezeichnet. Es ist zu beachten, dass sich Cronbachs Alpha nicht nur mit wachsenden Item-Interkorrelationen, sondern auch mit steigender Anzahl der Items erhöht. Eine Homogenität von .50 produziert z. B. bei 10 Items ein Alpha von .90 (Schnell et al., 2008, S. 153). Items, die wegen auffallend geringer itemspezifischer Homogenität (durchschnittliche Korrelation eines Items mit allen anderen Items) offensichtlich etwas anderes messen als die übrigen Items, sollten aus einem eindimensionalen Test entfernt werden. Die nahezu identische Information bietet jedoch auch der oben dargestellte Trennschärfe-Index, weshalb separate Homogenitätsanalyen kaum durchgeführt werden.

Anstelle einer Homogenitätsbetrachtung auf der Basis von Korrelationsanalysen haben sich zur Dimensionalitätsprüfung bei ein- und mehrdimensionalen Tests heute Faktorenanalysen etabliert. Während die **explorative/exploratorische Faktorenanalyse** („exploratory factor analysis“ EFA) die Items gemäß ihrer Interkorrelationen induktiv zu Faktoren bündelt und dafür (je nach gewählter Extraktions- und Rotationsmethode sowie gewählten Entscheidungskriterien) mehrere deskriptive Lösungen anbietet, operiert die **konfirmative/konfirmatorische Faktorenanalyse** („confirmatory factor analysis“, CFA) modellprüfend, indem sie deduktiv ein theoretisch vorgegebenes Faktormodell (z. B. eindimensionaler Test oder dreidimensionaler Test mit genauer Item-Zuordnung zu den drei Dimensionen) inferenzstatistisch mit den Daten vergleicht. Die modellprüfende konfirmatorische Faktorenanalyse basiert statistisch auf Strukturgleichungsmodellen (▶ Kap. 17; zum Einsatz der konfirmatorischen Faktorenanalyse im Zuge der Testkonstruktion s. auch Bühner, 2011). Explorative Faktoranalysen produzieren u. a. für jedes Item und jeden Faktor eine sog. **Faktorladung**. Eindimensionalität liegt vor, wenn die Item-Interkorrelationen auf einen einzigen Faktor (sog. **Generalfaktor**) reduziert werden können, auf dem sie alle hoch „laden“ (d. h., mit dem sie hoch korrelieren). Der Faktor repräsentiert inhaltlich das „Gemeinsame“, das in allen Items ausgedrückt wird und steht für das zu messende Konstrukt. Sind die Faktorladungen homogen, d. h. sehr einheitlich, ist die Berechnung eines ungewichteten additiven Gesamtwerts gerechtfertigt. Variieren die Faktorladungen (Symbol a) innerhalb ihres theoretischen Wertebereiches von $-1 \leq a \leq +1$ deutlich, so sind sie bei der Berechnung eines Gesamttest-Wertes als Gewichte zu verwenden. Items mit geringen Faktorladungen (Faustregel: Faktorladungen mit Beträgen unter 0.6) sind aus dem Test bzw. Fragebogen zu entfernen (zum Problem „bedeutsamer“ Faktorladungen vgl. Bortz & Schuster, 2010, S. 422; Briggs & Cheek, 1986; Fürntratt, 1969).

Dimensionalitätsprüfung bei mehrdimensionalen Tests. Eine methodisch saubere Konstruktion mehrdimensionaler Tests geht von einer theoretisch begründeten, genau festgelegten Zahl inhaltlich klar umrissener Teilkomponenten (Faktoren) des Zielkonstruktes aus, die als Subtests operationalisiert werden. Bei einem mehrdimensionalen Test wird für jeden Faktor ein separater (gewichteter oder ungewichteter) Testwert („subtest score“) berechnet. Ob die theoretisch festgelegte Zuordnung der Items zu den Subdimensionen/Faktoren des Tests sich empirisch bestätigen lässt, ist mit der **explorativen Faktorenanalyse** zu „prüfen“ und gilt deskriptiv als bestätigt, wenn ein Faktormodell mit entsprechender Faktorzahl hohe Varianzaufklärung besitzt und die den Faktoren zugeordneten Items hohe Faktorladungen auf einem Faktor aufweisen, dagegen keine oder geringe Faktorladungen auf anderen Faktoren (▶ Dimensionalitätsprüfung der Neurotizismus-Subskala aus dem NEO-FFI). Noch überzeugender ist die Dimensionalitätsprüfung jedoch mithilfe einer **konfirmatorischen Faktorenanalyse** (▶ Kap. 17), welche die Anpassung des Messmodells an die Daten inferenzstatistisch mit Modellanpassungstests prüft.

Ursprünglich eindimensional intendierte psychometrische Tests und Fragebögen erweisen sich nicht selten bei späteren empirischen Dimensionalitäts-Überprüfungen als mehrdimensional. Wie viele Faktoren zu extrahieren und wie diese angemessen zu interpretieren sind, ist dabei jedoch keineswegs immer eindeutig, da die Technik der explorativen Faktorenanalyse erhebliche Interpretationsspielräume offenlässt. Die spätere Ausdifferenzierung eindimensionaler Tests hat in erster Linie explorativen Wert; sie dient der Verfeinerung theoretischer Annahmen über das Konstrukt und regt neue Testentwicklungen an.

Die Klassische Testtheorie ist in der Konzeption ihrer Test- und Itemkennwerte auf eindimensionale Tests bzw. Skalen aus reflektiven Indikatoren zugeschnitten. Bei der Übertragung dieser Kennwerte auf mehrdimensionale Tests oder Fragebögen bieten sich – sofern die **Subtests** eine ausreichende Zahl an Items enthalten –

Dimensionalitätsprüfung der Neurotizismus-Subskala aus dem NEO-FFI

Aus insgesamt 12 Items besteht die Neurotizismus-Subskala aus dem Persönlichkeitstest NEO-FFI (Borkenau & Ostendorf, 2008; ▶ Abschn. 10.4.3), die emotionale Labilität als eine der fünf zentralen Persönlichkeitseigenschaften des Menschen erfassen soll. Hinweise zu Aufbau und Auswertung der Skala (▶ Die „Neurotizismus"-Subskala aus dem NEO-FFI) sowie Ausschnitte aus einer Itemanalyse dieser Skala (▶ Itemanalyse der Neurotizismus-Subskala aus dem NEO-FFI) haben wir schon vorgestellt. Jetzt geht es um die Dimensionalitätsprüfung, also um die Frage, ob die einzelnen Items tatsächlich alle gemeinsam dasselbe latente Merkmal erfassen.

Würde man anhand eines aussagekräftigen Datensatzes mittels Faktorenanalyse empirische Hinweise auf die Eindimensionalität (in traditioneller Sprechweise „faktorielle Validität") finden, so könnte man schlussfolgern, dass die Items mit hoher Wahrscheinlichkeit dasselbe Merkmal erfassen. Allerdings ist damit noch nicht bewiesen, dass es sich um das Merkmal Neurotizismus handelt (dazu müssten weitere empirische Evidenzen zur Validierung herangezogen werden).

Für eine inferenzstatistische Dimensionalitätsprüfung ist die konfirmatorische Faktorenanalyse indiziert (▶ Abschn. 17.2.1). Für eine deskriptive Betrachtung kann die exploratorische Fakorenanalyse genutzt werden. Bei der explorativen Faktorenanalyse werden die Items auf der Basis ihrer Interkorrelationen zu Faktoren gebündelt, ein Signifikanztest wird dabei nicht durchgeführt. Bei der Anwendung der exploratorischen Faktorenanalyse muss man sich für eine **Extraktionsmethode** sowie für eine **Rotationsmethode** und nicht zuletzt für eine **Regel zur Auswahl des Faktor-Modells** entscheiden (Bortz & Schuster, 2010, Kap. 23).

Im vorliegenden Beispiel wird die Dimensionalität der Neurotizismus-Subskala anhand einer Stichprobe von $n = 611$ (Alter $M = 36.21$, $SD = 11.92$; 59 % Frauen, 41 % Männer; Datensatz aus Pöschl, 2010) durchgeführt. Als Extraktionsmethode wird hier die **Hauptkomponentenanalyse** gewählt. Zur Umsetzung mit SPSS sind Analysieren → Dimensionsreduzierung → Faktorenanalyse und unter dem Menü „Extraktion" als Methode „Hauptkomponenten" zu wählen. Bei der Faktorextraktion werden aus den 12 Items im ersten Schritt 12 Faktoren gebildet und jeweils durch ihren Eigenwert lambda (λ; Varianz des Faktors) gekennzeichnet. Für die extrahierten Faktoren werden die Eigenwerte betrachtet und in ihrem Verlauf grafisch als Scree-Plot dargestellt, um ein Faktor-Modell auszuwählen (◘ Abb. 10.26).

Als Faustregel wählt man die stärksten Faktoren aus, also diejenigen **Faktoren mit Eigenwerten größer als 1** (sog. Kaiser-Guttman-Kriterium) und/oder mit **Eigenwerten, die sich deutlich von denen der anderen Faktoren abheben**, was an einem „Knick" im Screeplot erkennbar ist (◘ Abb. 10.26). (Andere Methoden zur Wahl des Faktor-Modells wie z. B. die **Parallelanalyse nach Ward**, die den empirischen Eigenwerteverlauf mit Zufallseigenwerten vergleicht, werden hier nicht betrachtet.)

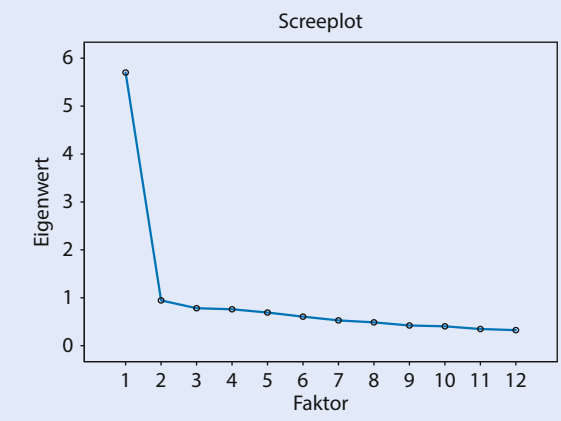
Im vorliegenden Fall hat nur der **erste Faktor mit $\lambda = 5.70$** einen Eigenwert größer als 1. Und nur dieser Faktor hebt sich im Eigenwert deutlich von allen anderen ab, wie der starke Knick im Scree-Plot zeigt. Wir entscheiden uns somit für ein **1-Faktor-Modell**, werten die empirische Evidenz also als Hinweis auf die Eindimensionalität der Skala. Mit einem Eigenwert von $\lambda = 5.70$ kann der erste Faktor $5.70/12 = 4.75$ % der Varianz der 12 Items erklären

Will man nun wissen, wie gut jedes einzelne Item zu dem Faktor passt, so sind die **Faktorladungen a** der Items zu betrachten, diese sind in der Komponentenmatrix dargestellt. Das Vorzeichen der Faktorladung gibt jeweils die Richtung und ihr Betrag die Enge des Zusammenhanges jedes Skalen-Items mit der Gesamtskala an. Es zeigt sich, dass alle Items positive Faktorladungen beträchtlicher Größe aufweisen und somit gut zu dem Generalfaktor passen (Bei einer Mehr-Faktor-Lösung würde man zur besseren Interpretierbarkeit der Faktorladungen zusätzlich ein Rotationsverfahren einsetzen (◘ Tab. 10.24).)

separate Itemanalysen sowie Objektivitäts-, Reliabilitäts- und Validitätsbeurteilungen für die einzelnen Teiltests bzw. Subdimensionen an. Gelegentlich interessieren bei der Itemanalyse auch die Reliabilitäten und Validitäten einzelner Items (vgl. Lienert & Raatz, 1994, Abschn. 2.2).

Probabilistische Testtheorie

Die Probabilistische Testtheorie PTT bzw. Item-Response-Theorie IRT ist statistisch anspruchsvoller und bietet die Möglichkeit, wichtige Gütekriterien eines Tests tatsächlich empirisch zu prüfen, die im Rahmen der KTT einfach vorausgesetzt werden. Beide Testansätze sollten nicht einem Konkurrenz- sondern besser in einem **Ergänzungsverhältnis** zueinander betrachtet werden (vgl. Moosbrugger & Kelava, 2012b, S. 229). Wenn mit der etablierten KTT gearbeitet wird, sind damit Vorteile, aber auch die ausgeführten Schwächen verbunden. Wenn man



◘ **Abb. 10.26** Screeplot der explorativen Faktorenanalyse der Neurotizismus-Subskala des NEO-FFI ($n = 611$)

Tabelle 10.24 Faktorladungen (a) der Items der Neurotizismus-Subskala des NEO-FFI (explorative Faktorenanalyse, 1-Faktor-Modell, $n = 611$)

	Faktor 1
1. Ich bin leicht beunruhigt.	.705
2. Ich fühle mich anderen oft unterlegen.	.713
3. Wenn ich unter starkem Stress stehe, fühle ich mich manchmal, als ob ich zusammenbräche.	.639
4. Ich fühle mich selten einsam und traurig.	.625
5. Ich fühle mich oft angespannt und nervös.	.773
6. Manchmal fühle ich mich völlig wertlos.	.811
7. Ich empfinde selten Furcht oder Angst.	.610
8. Ich ärgere mich oft darüber, wie Leute mich behandeln.	.602
9. Zu häufig bin ich entmutigt und will aufgeben, wenn etwas schief geht.	.756
10. Ich bin selten traurig oder deprimiert.	.694
11. Ich fühle mich oft hilflos und wünsche mir eine Person, die meine Probleme löst.	.774
12. Manchmal war mir etwas so peinlich, dass ich mich am liebsten versteckt hätte.	.505

diese Schwächen vermeiden möchte und es gelingt, einen inhaltlich passenden vorliegenden Test auf der Basis der IRT für die eigene Studie zu finden oder sogar selbst einen solchen zu konstruieren, können zusätzliche Vorteile in Anspruch genommen werden: Im Rahmen der IRT können das Gütekriterium der **Skalierung** und die **Dimensionalität der Skala** (als Element der Konstruktvalidität) geprüft werden, d. h. man kann feststellen, ob Itemhomogenität bezüglich des zu messenden Konstruktes vorliegt und das Antwortverhalten bei reflektiven Items tatsächlich nur von dem latenten Merkmal abhängt, so dass die Zusammenfassung der Items zu einem Testwert messtheoretisch sinnvoll ist. Zudem kann das **Skalenniveau des zu messenden Konstruktes** (z. B. der oft strittige Intervallskalencharakter) empirisch geprüft werden. Nicht zuletzt lassen sich mit der IRT Itemparameter bestimmen, die **nicht stichprobenabhängig** sind.

Im Rahmen der IRT werden zwei Arten von Variablen ausdrücklich unterschieden:

1. **Manifeste Variablen** („manifest variables“): Diese stellen das **beobachtbare Antwortverhalten** auf verschiedene Test-Items dar (z. B. das Lösen oder Nicht-Lösen einer Rechenaufgabe in einem Intelligenztest). Hierbei ist das Itemformat wichtig (z. B. dichotome versus geordnet polytome Items), aus dem sich unterschiedliche Möglichkeiten für Antwortmuster ergeben.
2. **Latente Variablen** („latent variables“): Die mit dem Test zu messenden **nicht direkt beobachtbaren Konstrukte** (Fähigkeiten oder Persönlichkeitseigenschaften) werden als latente Merkmale aufgefasst (z. B. Rechenfähigkeit), welche das manifeste Verhalten bestimmen. Wir unterscheiden kardinalskalierte latente Merkmale (**Latent Trait**; z. B. Intelligenz) und norminalskalierte latente Merkmale (**Latent Class**; z. B. Zugehörigkeit zur Gruppe der Normal-, Mittel- oder Extremkreuzer, d. h. Personen, die im Persönlichkeitstest normale Antwortmuster zeigen, bevorzugt die Mittelkategorien oder bevorzugt die Extremkategorien ankreuzen).

Probabilistische Testmodelle bestimmen Zusammenhänge zwischen der individuellen latenten Merkmalsausprägung einerseits und der Wahrscheinlichkeit für das Auftreten bestimmter manifester Itemantworten andererseits. Dies spiegelt sich auch in den synonym verwendeten Bezeichnungen „Probabilistische Testtheorie“ (Bezug zu Zustimmungs- bzw. Lösungswahrscheinlichkeiten) und „Item-Response-Theorie“ (Bezug zum manifesten Antwortverhalten der Testpersonen) wider (vgl. Rost & Spada, 1983).

Probabilistische Testtheorie bzw. Item Response Theorie – Die Probabilistische Testtheorie (PTT; „probabilistic test theory“) bzw. Item

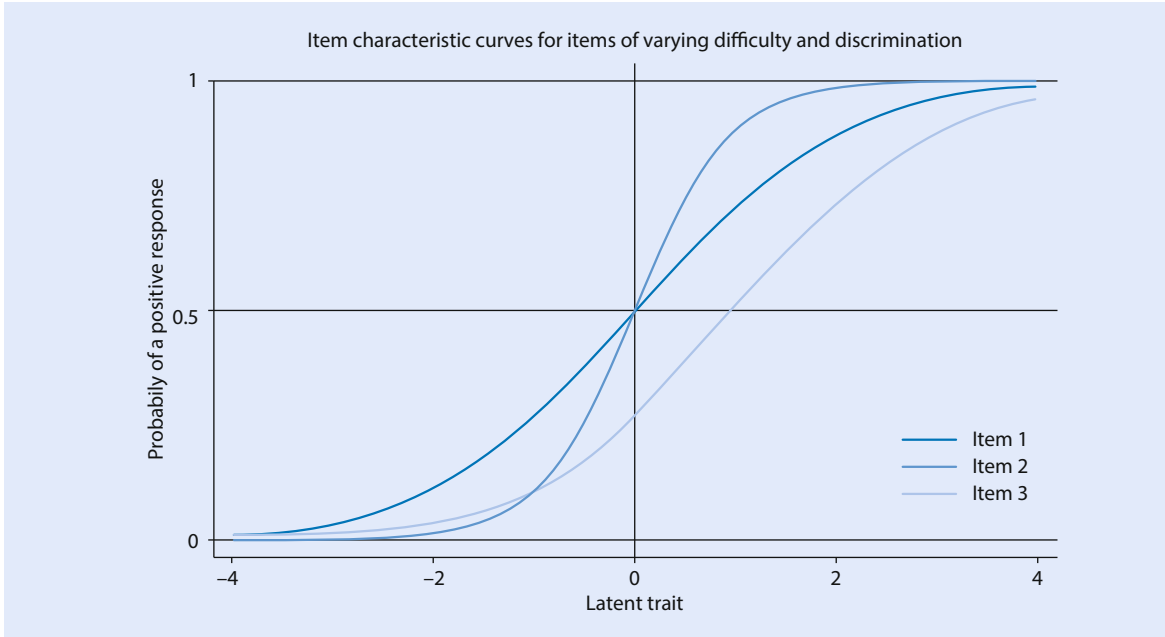


Abb. 10.27 Itemcharakteristische Kurven für drei dichotome Items gemäß probabilistischem Latent-Trait-Modell, wobei Item 3 eine höhere Schwierigkeit (geringere Lösungswahrscheinlichkeit bei mittlerer Fähigkeit) aufweist als Item 1 und 2 und Item 2 eine höhere Trennschärfe *(steilere Kurve)* besitzt als Item 1. (Aus Mattson, Campbell, Karabatsos, Agot, Ndinya-Achola, Moses, & Bailey, 2010)

Response Theorie (IRT; „item response theory") beschreibt die Zusammenhänge zwischen einem zu testenden Merkmal (latente Variable) und dem in einem Test tatsächlich beobachtbaren Antwortverhalten (manifeste Variable). Während die Klassische Testtheorie (KTT) das Testergebnis unmittelbar als messfehlerbelastete Ausprägung des zu messenden Konstruktes auffasst, ordnet die PTT/IRT dem als Indikator interpretierten Testergebnis mit bestimmter Wahrscheinlichkeit latente Merkmalsausprägungen zu.
Die Lösungs- bzw. Zustimmungswahrscheinlichkeiten eines Items bzw. einer Item-Kategorie werden in probabilistischen Testmodellen stets in Abhängigkeit von der Ausprägung des latenten Merkmals sowie – je nach Testmodell – von weiteren Einflussgrößen (z. B. Itemschwierigkeit, Itemtrennschärfe, Ratewahrscheinlichkeit, Antworttendenz) berechnet.

Probabilistische Testmodelle treffen Annahmen darüber, wie eine Antwort auf ein Item zustande kommt. Unterscheiden sich zwei Personen hinsichtlich ihrer Ausprägung eines latenten Merkmals (**Personenparameter** theta/θ bzw. xi/ξ), wird ein – das Konstrukt messendes – Item von der Person mit der höheren Merkmalsausprägung (im Folgenden soll vereinfachend von höherer Fähigkeit dieser Person gesprochen werden) mit größerer Wahrscheinlichkeit gelöst als von einer Person mit geringerer Fähigkeit. Außerdem wird eine Person mit bestimmter Fähigkeit von zwei Items dasjenige mit größerer Wahrscheinlichkeit lösen, dessen Lösung weniger Fähigkeit voraussetzt, das also leichter ist (**Itemschwierigkeitsparameter** sigma/σ). Die Bestimmung der Itemlösungswahrscheinlichkeiten in Abhängigkeit von der Ausprägung des zu messenden Konstruktes sowie weiterer Einflussgrößen wie Itemschwierigkeit, Itemtrennschärfe, Ratewahrscheinlichkeit oder Gruppenzugehörigkeit einer Person erfolgt bei dichotomen Items über **itemcharakteristische Funktionen** (IC-Funktionen, Itemcharakteristik, „item characteristic function"), die grafisch als **itemcharakteristische Kurven** („item characteristic curve", ICC) dargestellt werden.

Generell sollten itemcharakteristische Funktionen und Kurven **monoton** sein, d. h. mit zunehmender Ausprägung des latenten Merkmals sollte die Itemlösungswahrschlichkeit steigen. Unter den möglichen Funktionstypen, die die Bedingung der Monotonie erfüllen, haben sich logistische Funktionen im Rahmen der IRT besonders etabliert. Abb. 10.27 zeigt die auf einer **logistischen IC-Funktion** basierenden S-förmigen IC-Kurven von drei dichotomen Items. Auf der X-Achse ist die Ausprägung des latenten Merkmals bzw. des **Personenparameters** abgetragen: Negative Werte zeigen geringe Merkmalsausprägung und positive Werte hohe Merkmalsausprägung. Auf der Y-Achse stehen die Lösungswahrscheinlichkeiten der Items von 0 bis 1 bzw. 0 % bis 100 %. Eine Person mit geringer Fähigkeit kann die Items mit sehr geringer Wahrscheinlichkeit lösen, eine Person mit hoher Fähigkeit dagegen hat eine hohe Lösungswahrscheinlichkeit. Eine Person mit dem Personenparameter $\theta = 0$ weist eine mittlere Merkmalsausprägung auf und hat somit eine mittlere Lösungswahrscheinlichkeit, wenn die Itemschwierigkeit ebenfalls im mittleren Bereich liegt. Bei Item 1 und Item 2

beträgt die Lösungswahrscheinlichkeit einer Person mit $\theta = 0$ genau 50 %, während sie bei Item 3 niedriger liegt, etwa bei 25 %. Dementsprechend weist **Item 3 eine höhere Itemschwierigkeit** auf als Item 1 und 2. Betrachten wir nun die Kurven von Item 1 und Item 2, die beide dieselbe Schwierigkeit aufweisen, so fällt auf, dass sie sich in der Steigung unterscheiden. Die Kurve von Item 2 steigt steiler als die von Item 1. Dementsprechend weist **Item 2 eine höhere Trennschärfe** auf, es differenziert deutlicher zwischen Personen mit geringer und hoher Fähigkeit. Da die IRT darauf hinausläuft, die Lösungswahrscheinlichkeiten der Items zu berechnen und mit der Ausprägung des latenten Merkmals (sowie weiterer Einflussgrößen) in Verbindung zu setzen, stellen die IC-Kurven eine wichtige Hilfe zur Veranschaulichung dar.

Bei einem Probabilistischen Test lässt sich das Testergebnis einer Person (**Testwert**; z. B. $x = 18$ Punkte) einem **Personenparameter** (Ausprägung des latenten Merkmals; z. B. $\theta = -1.4$) zuordnen. Werden die Personenparameter auf der Basis eines logistischen Modells berechnet, sind sie auf (logarithmiertem) **Verhältnisskalenniveau** angesiedelt, dem informationsreichsten Skalenniveau mit gleichabständigen Ausprägungen und absolutem Nullpunkt: Eine Person mit dem Personenparameter $\theta = -1.4$ weist somit z. B. eine dreimal geringere latente Merkmalsausprägung auf als eine Person mit $\theta = +2.8$.

Auf der Basis des Personenparameters lassen sich – gemäß dem verwendeten und auf Anwendbarkeit geprüften probabilistischen Testmodell – dann die **Lösungswahrscheinlichkeiten** für die einzelnen Items des Tests bestimmen. Somit kann z. B. für die Items des PISA-Tests, der Kompetenzen 15-jähriger Schülerinnen und Schüler misst, genau angegeben werden, wie hoch die Lösungswahrscheinlichkeit eines konkreten Items bei Schülern mit einem bestimmten Testwert ist, damit wird die Bedeutung des Testwertes konkretisiert. Ein weiterer Vorteil probabilistischer Tests besteht darin, dass sie ökonomischer anwendbar sind, indem den Testpersonen nicht alle, sondern nur Items mit passender Schwierigkeit vorgelegt werden (adaptives Testen).

Für die Modelltestung gemäß IRT wurde bislang eine Vielzahl an **Software-Tools** entwickelt. Diese unterscheiden sich im Funktionsumfang erheblich, noch gibt es kein allumfassendes Gesamtpaket und die gängigen Statistik-Programmpakete wie SPSS enthalten keine IRT-Funktionen. Empfehlenswert sind die Programme **ConQuest**, **MULTIRA** und **WINMIRA**, die über eine Vielzahl an Auswertungsoptionen verfügen. Eine aktuelle Übersicht über Software-Tools zur Probabilistischen Testtheorie und deren Vor- und Nachteilen findet sich bei Zhao und Hambleton (2009). Verschiedene Lehrbuch- und Handbuchdarstellungen zur Testtheorie, die auch IRT mit Anwendungsbeispielen behandeln, liegen vor (Bühner, 2011; Moosbrugger & Kelava, 2012b). Einen umfassenden Überblick über unterschiedliche Testmodelle bietet Rost (2004), weitere Modelle, neuere Entwicklungen und Anwendungen findet man zudem u. a. bei Baker und Kim (2004), Fischer und Molenaar (1995), van der Linden und Hambleton (1997) sowie auf http://www.rasch.org. Als Versuch einer Integration von Statistik, Klassischer und Probabilistischer Testtheorie sei McDonald (1999) empfohlen.

Im Folgenden wird zunächst eine **Übersicht Probabilistischer Testmodelle** und ihrer Bezeichnungen gegeben, um eine Orientierung im Feld zu ermöglichen. Anschließend wird exemplarisch das **dichotome Rasch-Modell** beschrieben, das für Tests aus dichotomen Items indiziert ist und die Itemlösungswahrscheinlichkeit aus der Fähigkeit der Testperson sowie zusätzlich mit einem weiteren Parameter – der Itemschwierigkeit – berechnet. Abschließend wird das **adaptive Testen** vorgestellt, bei dem anstelle eines vollständigen Tests den Testpersonen nur die jeweils auf ihr individuelles Fähigkeitsniveau abgestimmten, maximal informativen Items präsentiert werden. Die für eine bestimmte Person zu leichten oder zu schwierigen Items bekommt sie also gar nicht erst vorgelegt, was den Testaufwand reduziert.

Übersicht Probabilistischer Testmodelle

Probabilistische Testmodelle lassen sich in drei Gruppen einteilen hinsichtlich des **gemessenen latenten Merkmals**:

- **Latent-Trait-Modelle** („latent trait models“) messen ein kontinuierliches, kardinalskaliertes latentes Merkmal (z. B. Ausprägung von Intelligenz)
- **Latent-Class-Modelle** („latent class models“) messen ein kategoriales, nominalskaliertes latentes Merkmal (z. B. Einteilung von Spielsüchtigen in drei Patienten-Gruppen).
- **Mischverteilungs-Modelle** („mixed Rasch models“) messen ein kardinalskaliertes latentes Merkmal differenziert für die durch ein latentes nominalskaliertes Merkmal gebildeten Gruppen. Mischerverteilungs-Modelle stellen also eine Kombination aus Latent-Trait- und Latent-Class-Modellen dar.

Ein weiteres wichtiges Kriterium zu Klassifikation von Probabilistischen Testmodellen ist das **Itemformat**. Da im Rahmen der IRT die Antwortwahrscheinlichkeiten für einzelne Items bzw. einzelne Itemausprägungen berechnet werden, existieren unterschiedliche probabilistische Testmodelle in Abhängigkeit vom Item- bzw. Antwortformat (z. B. dichotome Items: **dichotome Modelle** versus geordnet polytome bzw. mindestens ordinalskalierte Items: **ordinale Modelle**).

Probabilistische Testmodelle

Probabilistische Testmodelle unterscheiden sich dahingehend, ob das durch den Test zu messende latente Merkmal als kardinalskaliert **(Latent-Trait-Modelle)** oder als nominalskaliert **(Latent-Class-Modelle)** aufgefasst wird. **Mischverteilungs-Modelle** (Mixed-Rasch-Modelle) kombinieren beide Ansätze. Welches **Itemformat** für die Messung des latenten Merkmals verwendet wird, ist ebenfalls ein wichtiges Merkmal zur Kennzeichnung probabilistischer Testmodelle:

- Bei **dichotomen Modellen**, die auf dichotomen Items basieren, wird die Zustimmungs-/Lösungswahrscheinlichkeit des Items in Abhängigkeit von der Ausprägung des latenten Merkmals (sowie weitere Einflussgrößen) über eine itemcharakteristische Funktion (**IC-Funktion**) berechnet, die grafisch als itemcharakteristische Kurve **(IC-Kurve)** darstellbar ist.
- Bei **ordinalen Modellen**, die auf mindestens ordinalskalierten Items basieren, muss für jede Ausprägung des Items eine eigene kategoriencharakteristische Funktion („category characteristic function", **CC-Funktion**) bestimmt werden, die sich als kategoriencharakteristische Kurve („category characteristic curve", **CC-Kurve**) grafisch veranschaulichen lässt.

Zudem unterscheiden sich Testmodelle dahingehend, wie viele **Einflussgrößen** auf das Antwortverhalten sie berücksichtigen. Die Personeneigenschaft (Persönlichkeitsmerkmal oder Fähigkeit) wird immer als zentrale Einflussgröße berücksichtigt, ebenso die Itemschwierigkeit. Darüber hinaus können auch Trennschärfen und bei Leistungstests Verfälschungen durch Raten berücksichtigt werden. Personengruppen mit unterschiedlichem Antwortmuster – z. B. Mittel- oder Extremkreuzer (Personen, die grundsätzlich eher Mittelkategorien oder Randkategorien ankreuzen) – lassen sich im Rahmen von Mixed-Rasch-Modellen einbeziehen (▶ Probabilistische Testmodelle).

Im Unterschied zur Klassischen Testtheorie, deren Annahmen als ungeprüfte Axiome zugrunde gelegt werden, postuliert die Probabilistische Testtheorie **Test- bzw. Messmodelle**, die anhand der mit einem Test erhobenen Daten inferenzstatistisch geprüft werden. Nur wenn die Daten aller Items mit dem Messmodell übereinstimmen, was durch entsprechende **Modelltests** überprüfbar ist, werden die Testwerte als echte Messwerte des latenten Konstruktes akzeptiert und zu diagnostischen oder Forschungszwecken verwendet. Die Anforderungen, die im Rahmen der IRT an **modellkonforme Items** gestellt werden, sind dabei deutlich höher als die Anforderungen, die im Rahmen der KTT an Items gestellt werden.

Latent-Trait-Modelle Probabilistische Latent-Trait-Modelle sind indiziert, wenn das durch den Test gemessene latente Merkmal (Fähigkeit, Persönlichkeitsdimension) als **kontinuierliches** bzw. **kardinalskaliertes latentes Merkmal** mit sehr vielen graduellen Abstufungen zwischen minimaler und maximaler Ausprägung aufgefasst wird (z. B. Intelligenz, Extraversion, Depressivität). Die meisten Tests erfassen kontinuierliche Merkmale, so dass Latent-Trait-Modelle verbreiteter sind als ihr Gegenstück, die Latent-Class-Modelle.

Eine Übersicht über verbreitete Latent-Trait-Modelle ist ◘ Tab. 10.25 zu entnehmen, wobei v. a. die Unterscheidung gemäß Itemformat wichtig ist (dichotome versus ordinale Latent-Trait-Modelle mit ihren jeweiligen Unterformen).

Um den Unterschied zwischen dichotomen und ordinalen Latent-Trait-Modellen zu veranschaulichen, wollen wir noch einmal auf die IC-Kurven zurückgreifen: Bei dichotomen Items wird **pro Item** eine IC-Funktion bzw. eine IC-Kurve berechnet (◘ Abb. 10.27). Bei ordinalen Items muss **pro Ausprägung/Kategorie** des Items eine eigene **Kategorienfunktion** (CC-Funktion) mit zugehöriger **CC-Kurve** ermittelt werden (◘ Abb. 10.28). Betrachten wir ein ordinalskaliertes Item mit vier Ausprägungen (z. B. stimmt gar nicht – stimmt wenig – stimmt ziemlich – stimmt völlig), so sind vier Kategorienfunktionen bzw. vier CC-Kurven zu bestimmen, die gemeinsam die IC-Funktion bilden. Bei einer Person mit sehr geringer latenter Merkmalsausprägumg (z. B. $\theta = -4$) ist die Wahrscheinlichkeit, dass Antwortkategorie 1 „stimmt gar nicht" angekreuzt wird, über 90 %, die Wahrscheinlichkeit, dass Antwortkategorie 4 „stimmt völlig" angekreuzt wird, dagegen nahe Null. Umgekehrt ist bei einer Person mit sehr starker latenter Merkmalsausprägung (z. T. $\theta = +4$) die Wahrscheinlichkeit, dass Antwortkategorie 4 „stimmt völlig" angekreuzt wird, nahe 100 %.

Bei der Betrachtung aller Kategorienfunktionen eines ordinalen Items ist nun entscheidend, an welcher Stelle sich die Kategorienkurven kreuzen. Diese Schnittpunkte werden als **Schwellenparameter** bezeichnet. Bei vier Kategorien und vier CC-Kurven ergeben sich drei Schnittpunkte. Diese Schwellenparameter liegen in ◘ Abb. 10.28 bei $\theta = -1$ (stimmt gar nicht/stimmt wenig), $\theta = +0.5$ (stimmt wenig/stimmt ziemlich) und bei $\theta = +1$ (stimmt ziemlich/stimmt völlig). Bei einer Person mit der latenten Merkmalsausprägung $\theta = +1$ ist also – da es sich hier genau um den Wert des Schwellenparameters handelt – die Wahrscheinlichkeit, dass sie „stimmt ziemlich" oder „stimmt völlig" ankreuzt, identisch. Erst wenn die Schwelle $\theta = +1$ überschritten ist, kreuzt eine Person mit höherer Wahrscheinlichkeit die Kategorie „stimmt völlig" an. Ordinale Latent-Trait-Modelle unterscheiden sich nun dahingehend, welche Restriktionen sie zu den Schwellenparametern vorgeben. Das ordinale Rasch-Modell verlangt **geordnete Schwellenparameter**, das Äqui-

Tabelle 10.25 Übersicht gängiger Latent-Trait-Modelle im Kontext der probabilistischen Testtheorie

Probabilistische Testmodelle für kardinalskalierte latente Merkmale	Erläuterung
1. Dichotome Latent-Trait-Modelle	Dichotome Latent-Trait-Modelle sind indiziert, wenn ein kardinalskaliertes latentes Merkmal (Latent Trait) mit einem Test erfasst werden soll, der nur aus **dichotomen Items** besteht (ja/nein im Persönlichkeitstest; richtig/falsch im Leistungstest). Die Modelle lassen sich noch genauer als „dichotome logistische Latent-Trait-Modelle" kennzeichnen, denn die Itemlösungswahrscheinlichkeit wird jeweils auf der Basis einer **logistischen IC-Funktion** (itemcharakteristische Funktion) bzw. einer **S-förmigen IC-Kurve** (itemcharakteristische Kurve) mit dem latenten Merkmal verknüpft, so dass Personen mit geringer Merkmalsausprägung bzw. Fähigkeit eine geringe Lösungswahrscheinlichkeit (nahe 0), Personen mit mittlerer Fähigkeit eine mittlere Lösungswahrscheinlichkeit (um .50) und Personen mit hoher Fähigkeit eine hohe Lösungswahrscheinlichkeit (nahe 1 bzw. 100 %) zugeordnet wird (Abb. 10.27).
1.1 Das 1-PL-Modell (Ein-Parameter-Logistisches Modell, dichotomes Rasch-Modell)	Das 1PL-Modell ist ein dichotomes logistisches Latent-Trait-Modell, bei dem die Itemlösungswahrscheinlichkeit berechnet wird auf der Basis der Personenfähigkeit (Personenparameter theta θ) – dieser Parameter wird immer berücksichtigt – sowie zusätzlich einem weiteren Parameter: der Itemschwierigkeit **(Itemschwierigkeitsparameter** sigma σ). Dieses Modell ist das sparsamste Latent-Trait-Modell, es wird im Text näher beschrieben.
1.2 Das 2-PL-Modell (Zwei-Parameter-Logistisches Modell, Birnbaum-Modell)	Das 2PL-Modell ist ein dichotomes logistisches Latent-Trait-Modell, bei dem die Itemlösungswahrscheinlichkeit berechnet wird auf der Basis der Personenfähigkeit (Personenparameter θ) sowie zusätzlich zwei weiteren Parametern: der Itemschwierigkeit (Itemschwierigkeitsparameter σ) und der Trennschärfe **(Trennschärfeparameter** beta β). Ein trennschärferes Items grenzt fähige und weniger fähige Personen deutlicher voneinander ab und besitzt eine steilere itemcharakteristische Kurve (Abb. 10.27). Die Lösung eines trennscharfen Items wird im Modell stärker gewichtet als die Lösung eines weniger trennscharfen Items.
1.3 Das 3-PL-Modell (Drei-Parameter-Logistisches Modell)	Das 3PL-Modell ist ein dichotomes logistisches Latent-Trait-Modell, bei dem die Itemlösungswahrscheinlichkeit berechnet wird auf der Basis der Personenfähigkeit (Personenparameter θ) sowie zusätzlich drei weiteren Parametern: der Itemschwierigkeit (Itemschwierigkeitsparameter σ), der Trennschärfe (Trennschärfeparameter β) und der Ratewahrscheinlichkeit (**Rateparameter** gamma γ). Ein auf der Basis von Raten lösbares Item besitzt eine itemcharakteristische Kurve, bei der die minimale Lösungswahrscheinlichkeit auch bei sehr geringer Fähigkeit größer als Null ist.
2. Ordinale Latent-Trait-Modelle (Partial-Credit-Modell)	Ordinale Latent-Trait-Modelle sind indiziert, wenn ein kardinalskaliertes latentes Merkmal mit einem Test erfasst werden soll, der aus **geordnet polytomen** bzw. **mindestens ordinalskalierten Items** besteht (z. B. Antwortformate wie „Ablehnung – neutral – Zustimmung"; „nie – selten – oft", „stimmt gar nicht – wenig – teils-teils – ziemlich – völlig"). Die Streitfrage, ob eine Ratingskala als Antwortformat Ordinal- oder Intervallskalenniveau aufweist (▶ Abschn. 8.4.4) lässt sich mithilfe eines ordinalen Latent-Trait-Modells für einen konkreten psychologischen Test empirisch überprüfen. Ordinale Latent-Trait-Modelle werden auch als **Partial-Credit-Modelle** bezeichnet, weil sie bei Leistungstests teilrichtige Antworten berücksichtigen können, indem die Itemauswertung ordinal erfolgt („richtige – teilweise richtige – falsche Antwort"). Ebenso wie die dichotomen Latent-Trait-Modelle basieren auch die gängigen ordinalen Latent-Trait-Modelle auf logistischen Funktionen, wobei zu beachten ist, dass für jede einzelne Antwortwortkategorie eines Items eine **Kategorienfunktion (CC-Funktion)** bestimmt wird, die sich als kategoriencharakteristische Kurve **(CC-Kurve)** grafisch darstellen lässt (s. (Abb. 10.28). Bei einem ordinalen Item ergibt sich die IC-Funktion aus den CC-Funktionen aller Kategorien. Für die Kennzeichnung eines ordinalen Latent-Trait-Modells sind die Schnittpunkte der CC-Kurven – die **Schwellenparameter** – entscheidend.
2.1 Ordinales Rasch-Modell	Unter den ordinalen Latent-Trait-Modellen ist das ordinale Rasch-Modell am populärsten (ebenso wie bei den dichotomen Latent-Trait-Modellen das dichotome Rasch-Modell das meistverwendete ist). Das ordinale Rasch-Modell verlangt **geordnete (ordinale) Schwellenparameter**, nicht jedoch gleichabständige.
2.2 Andere ordinale Latent-Trait-Modelle	Neben dem ordinalen Rasch-Modell existieren weitere ordinale logistische Latent-Trait-Modelle, die sich hinsichtlich ihrer Modellannahmen zu den Schwellenparametern unterscheiden (z. B. Ratingskalen-Modell, Äquidistanz-Modell; vgl. Rost, 2004).
3. Weitere Latent-Trait-Modelle	Es existieren zahlreiche weitere Latent-Trait-Modelle (z. B. sog. **facettentheoretische, mehrmodale** und **mehrdimensionale Verallgemeinerungen**), die jeweils Zusatzannahmen in das Modell aufnehmen und zur Beantwortung sehr spezifischer Fragen im Rahmen der Testentwicklung und Eigenschaftsmessung zum Tragen kommen (zur Vertiefung s. Rost, 1999).

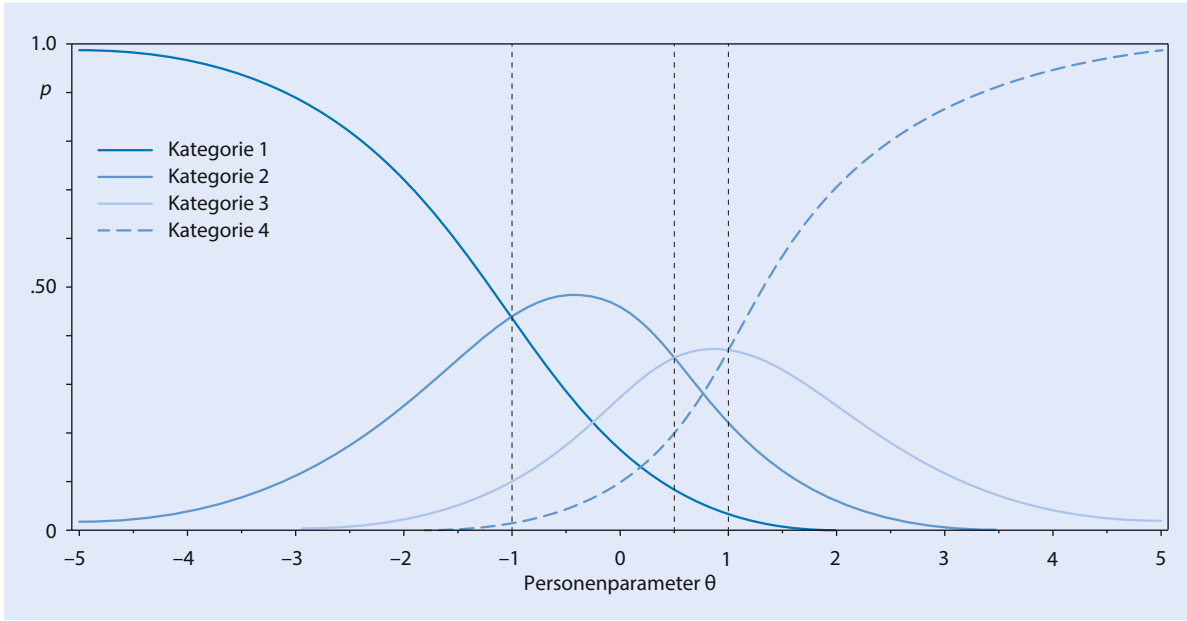


Abb. 10.28 CC-Kurven für die vier Ausprägungen/Kategorien eines ordinalen Items (z. B. *Kategorie 1:* stimmt gar nicht – *Kategorie 2:* stimmt wenig – *Kategorie 3:* stimmt ziemlich – *Kategorie 4:* stimmt völlig. (Mod. nach Rost, 2004. Mit freundlicher Genehmigung von Huber, Bern)

distanz-Modell darüber hinaus **gleichgroße Abstände bzw. gleichgroße Differenzen zwischen den Schwellenparametern**. Die Bedingung der Gleichabständigkeit (Äquidistanz), die für Intervallskalenniveau vorausgesetzt wird, ist im Beispiel in Abb. 10.28 nicht erfüllt. Denn der mittlere Schwellenparameter liegt bei $\theta = +0.5$ und nicht bei $\theta = 0$. Die Analyse der Items mit dem probabilistischen Testmodell liefert also konkrete empirische Hinweise darauf, ob der mit dem Test gebildete Skalenwert als ordinal- oder intervallskaliert einzustufen ist.

Latent Class-Modelle Das durch den Test gemessene latente Konstrukt wird als **kategoriales** bzw. **nominalskaliertes Merkmal** mit sehr wenigen Ausprägungen bzw. Kategorien aufgefasst (z. B. drei Subgruppen von Patienten mit einem bestimmten Störungsbild). Psychologische Tests zielen seltener auf die Messung kategorialer Merkmale ab, weshalb Latent-Class-Modelle und die dafür notwendige latente Klassenanalyse (Latent-Class-Analyse, **LCA**) seltener zum Einsatz kommen. Ziel der latenten Klassenanalyse ist die Zuordnung von Personen zu Gruppen mit jeweils spezifischem Antwortmuster. Bei einem Latent-Class-Modell werden Klassenzugehörigkeit und Antwortwahrscheinlichkeit durch **diskrete Itemfunktionen** bzw. **Itemprofile** einander zugeordnet (Abb. 10.29).

Analog der Unterscheidung zwischen dichotomen und ordinalen Latent-Trait-Modellen anhand des Itemformats werden auch Latent-Class-Modelle für dichotome und ordinale Items/Daten differenziert.

Die Latent-Class-Analyse (LCA) wurde bereits 1968 von Lazarsfeld und Henry unter der Bezeichnung „Latent Structure Analysis" entwickelt. Als Einführung in die LCA seien McCutcheon (1987), Formann (1984) und Rost (2004, Abschn. 3.1.2.2) empfohlen. Die für die LCA erforderliche Software mit Übungsbeispielen hat Rost (2004) auf einer CD zusammengestellt. Einen Überblick und weitere Literatur findet man bei Langeheine und Rost (1996).

Mischverteilungs-Modelle Mischverteilungsmodelle bzw. Mixed Rasch-Modelle kombinieren ein Latent-Trait-Modell mit der Latent-Class-Analyse: Die Ausprägung eines kontinuierlichen latenten Merkmals wird gemessen, wobei die Testpersonen in homogene Untergruppen (Klassen) eingeteilt werden. Im Mixed-Rasch-Modell werden zwei Personenmesswerte bestimmt: Die **Klassenzugehörigkeit der Person** sowie die **Merkmalsausprägung innerhalb der Klasse**. Das Mixed Rasch-Modell ist weniger restriktiv und breiter anwendbar, weil es unterschiedliche Itemschwierigkeiten pro Klasse zulässt bzw. die Heterogenität von Testpersonen abbilden kann. Mixed-Rasch-Modelle weisen wiederum diverse Unterformen auf, etwa in Abhängigkeit vom Itemformat (dichtom, ordinal).

Die für Latent-Trait-Modelle geltende Annahme der Stichprobenunabhängigkeit, also die Annahme, dass die Itemparameter für verschiedene Stichproben konstant seien, wird im Mixed-Rasch-Modell aufgehoben. Man denke z. B. an Tests zum räumlichen Vorstellungsvermögen, die

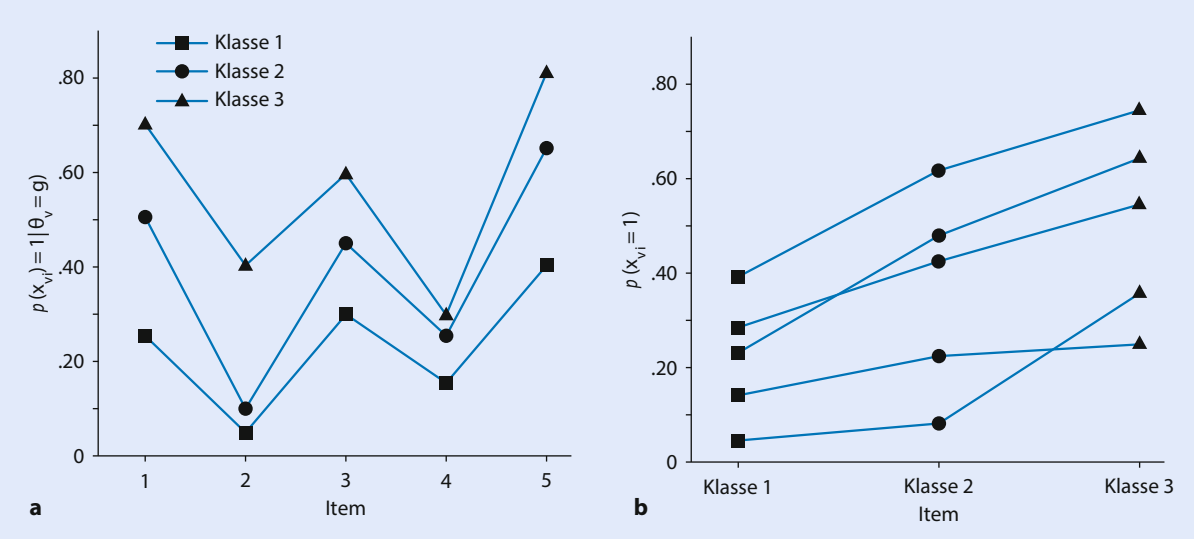


Abb. 10.29 Itemprofile der drei Klassen (**a**) und entsprechende diskrete Itemfunktionen der fünf Items gemäß probabilistischem Latent-Class-Modell (**b**). Die Itemfunktionen aller fünf Items sind monoton steigend, d. h. Klasse 1 besitzt die niedrigste und Klasse 3 die höchste Merkmalsausprägung. (Aus Ma, 2010, http://www.docstoc.com/docs/25282676/Analyse-latenter-Klassen-und-das-Mixed-Raschmodel)

von unterschiedlichen Personengruppen mit unterschiedlichen Lösungsstrategien bearbeitet werden, so dass gruppenspezifische Itemparameter sehr wahrscheinlich sind. Genau dieses Problem wird mit den Mischverteilungs-Modellen aufgegriffen. Die Modelle suchen nach homogenen Teilstichproben, zwischen denen die Itemparameter maximal unterschiedlich sind. Sie stellen deshalb eine wichtige Bereicherung für die persönlichkeitspsychologische Forschung dar, in der es u. a. um die Bildung von Typologien bezüglich spezifischer Persönlichkeitsmerkmale (Motivation, Attributionsstile, Intelligenz etc.) geht.

Für jede Person mit einem bestimmten Antwortmuster wird errechnet, mit welcher Wahrscheinlichkeit sie den einzelnen latenten Klassen angehört, wobei die Anzahl der latenten Klassen hypothetisch vorgegeben werden muss. Personen innerhalb einer latenten Klasse sind homogen in Bezug auf die Lösungswahrscheinlichkeiten für die Items, und verschiedene latente Klassen sollen in Bezug auf die Lösungswahrscheinlichkeiten möglichst unterschiedlich sein.

Das dichotome Rasch-Modell Das dichotome logistische Rasch-Modell wurde 1960 von dem dänischen Statistiker Georg Rasch für die Analyse von Tests mit dichotomen Antwortvorgaben zur Messung von Latent Traits entwickelt. Es wird auch als Ein-Parameter-Logistisches Modell (**1-PL-Modell;** Tab. 10.25) bezeichnet (Moosbrugger, 2012a, S. 236ff.). Es kann auf **dichotome Items**, d. h. Items mit zwei Antwortmöglichkeiten angewendet werden (Selbstauskunftsitems mit „ja"/"nein"-Antwortformat; Leistungsaufgaben mit den Bewertungsstufen „gelöst"/"nicht gelöst") und dient der Messung eines latenten Traits. Die Itemcharakteristiken werden über eine **logistische IC-Funktion** bestimmt.

Nach diesem Modell wird die Zahl möglicher monotoner IC-Funktionstypen erheblich eingegrenzt, wenn ein Test die folgenden fünf Annahmen erfüllt:

1. Der Test besteht aus einer endlichen Menge von Items.
2. Der Test ist homogen in dem Sinne, dass alle Items dasselbe Merkmal messen.
3. Die Itemcharakteristiken sind monoton steigend.
4. Es wird lokale stochastische Unabhängigkeit vorausgesetzt: Ob eine Person ein Item löst oder nicht, hängt ausschließlich von ihrer Fähigkeit und der Schwierigkeit des Items ab.
5. Die Anzahl der gelösten Aufgaben stellt eine **„erschöpfende Statistik"** für die Fähigkeit einer Person dar, d. h., es interessiert nicht, welche Aufgaben gelöst wurden, sondern lediglich wie viele.

Nimmt man nun für ein beliebiges Item eine **logistische Funktion** als Itemcharakteristik an (Abb. 10.27), folgt bei Zutreffen der oben genannten Annahmen, dass alle übrigen Items ebenfalls Itemcharakteristiken in Form logistischer Funktionen aufweisen (Fischer, 1974, S. 193ff.).

Eine logistische Funktion wird allgemein durch folgende Gleichung beschrieben (Gl. 10.15):

$$y = \frac{e^x}{1 + e^x} \quad (e = 2.718). \qquad (10.15)$$

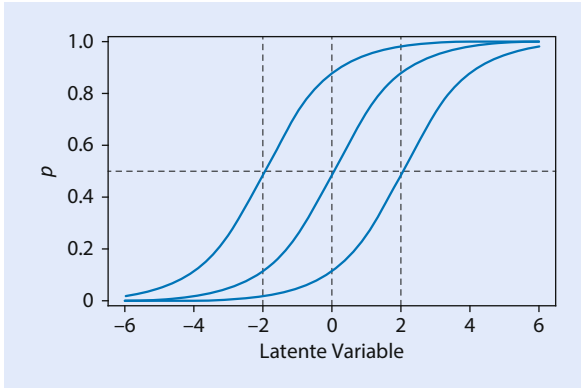


Abb. 10.30 Itemcharakteristiken des dichotomen Rasch-Modells. (Aus Schnell, Hill, & Esser, 2008, S. 198. Mit freundlicher Genehmigung von Oldenbourg, München)

Auf die gesuchte Itemcharakteristik angewendet bedeutet dies: Die Wahrscheinlichkeit (p), ein Item zu lösen, hängt ausschließlich von der Fähigkeit der Person (Personenparameter θ) und der Schwierigkeit des Items (Itemparameter σ) bzw. von der Differenz $\theta - \sigma$ ab. Die Art der Beantwortung eines Items ist also davon unabhängig, welche anderen Items die Person bereits bearbeitet hat (Prinzip der lokalen stochastischen Unabhängigkeit; eine ausführliche Erläuterung findet sich in Moosbrugger, 2012a; Rost, 2004). Die Itemcharakteristik wird somit über folgende Gleichung (Gl. 10.16) bestimmt:

$$p = \frac{e^{(\theta-\sigma)}}{1 + e^{(\theta-\sigma)}}. \qquad (10.16)$$

Die Abb. 10.30 enthält drei Itemcharakteristiken für Items mit unterschiedlichen Schwierigkeiten bzw. Itemparametern.

Auf der x-Achse (latente Variable) werden sowohl die Personen als auch die Items skaliert. Für eine durchschnittliche Ausprägung der latenten Variablen wurde hier der Wert Null angenommen. Sind der Personen- und der Itemparameter identisch ($\theta - \sigma = 0$) erhält man eine Lösungswahrscheinlichkeit von .50 (Gl. 10.16).

$$p = \frac{e^0}{1 + e^0} = .50$$

Oder umgekehrt: Wenn Personen mit der Fähigkeit θ ein Item mit der Schwierigkeit σ mit einer Wahrscheinlichkeit von 50 % lösen, sind Personen- und Itemparameter identisch. Die drei Items in Abb. 10.30 haben demnach Itemparameter von -2, 0 und $+2$. Eine Person mit $\theta = 4$ wird das Item mit $\sigma = 2$ mit einer Wahrscheinlichkeit von 88 % lösen (Gl. 10.16).

$$p = \frac{e^{(4-2)}}{1 + e^{(4-2)}} = .88$$

Eine weniger befähigte Person ($\theta = -2$) löst dieses Item nur mit einer Wahrscheinlichkeit von 2 %.

$$p = \frac{e^{(-2-2)}}{1 + e^{(-2-2)}} = .02$$

Auf der Basis des dichotomen logistischen Modells können **Personenparameter** (Fähigkeiten) und **Aufgaben-/Itemparameter** (Schwierigkeiten) ermittelt werden. In diesen Beispielen haben wir vorausgesetzt, dass die Personen- und Itemparameter bekannt sind. Tatsächlich müssen diese jedoch mit aufwändigen iterativen Algorithmen geschätzt werden, deren Darstellung über den Rahmen dieses Buches hinausgeht. Im Folgenden sei lediglich die prinzipielle Vorgehensweise skizziert.

Die Schätzungen basieren auf Summenstatistiken (Anzahl der gelösten Aufgaben pro Person bzw. Anzahl lösender Personen pro Item) als erschöpfende Statistik (d. h. entscheidend ist die Menge der richtig gelösten Items, wobei es egal ist, welche Items gelöst wurden). Bei den Personenparametern interessieren diejenigen, die die Wahrscheinlichkeit für das Zustandekommen der individuell erreichten Anzahl gelöster Aufgaben maximieren. Aufgrund des **Prinzips der lokalen stochastischen Unabhängigkeit** geht man davon aus, dass die Art der Beantwortung eines Items davon unabhängig ist, welche anderen Items die Person bereits bearbeitet hat. Psychologisch gesehen bedeutet diese Forderung, dass die Itembeantwortungen von Übungs-, Ermüdungs- oder Positionseffekten unabhängig sind. Formal hat dieses Prinzip zur Konsequenz, dass sich die Wahrscheinlichkeit für die Gesamtanzahl gelöster Items für eine Person mit bestimmter Fähigkeit aus dem Produkt der Wahrscheinlichkeiten für die Lösung der einzelnen Items ergibt (genauer hierzu z. B. Amelang & Zielinski, 2002, Abschn. 2.1.2.1; Rost, 2004, Abschn. 2.3.4).

Die Schätzung der Itemparameter (Schwierigkeiten) erfolgt in ähnlicher Weise. Die Wahrscheinlichkeit, dass ein Item von einer bestimmten Anzahl von Personen richtig beantwortet wird, ergibt sich aus dem Produkt der Wahrscheinlichkeiten, mit denen die einzelnen Personen dieses Item richtig beantworten. Gesucht werden diejenigen Itemparameter, die die Wahrscheinlichkeit für das Zustandekommen der jeweils erzielten Lösungshäufigkeiten maximieren.

Die rechnerische Ermittlung der Personen- und Itemparameter macht von der Theorie **erschöpfender Statistiken** Gebrauch, die in diesem Falle besagt, dass es für die Schätzung der Personenparameter nicht darauf ankommt, welche Items gelöst wurden. Die Anzahl aller gelösten Items enthält sämtliche für die Schätzung eines Personenparameters relevanten Informationen, d. h., Personen mit unterschiedlichen Antwortmustern (z. B. $++--+$ und $+-+-+$) werden nicht unterschieden, wenn die An-

zahl aller gelösten Items übereinstimmt. Entsprechendes gilt für die Schätzung der Itemparameter: Auch hier interessiert nur die Anzahl der Personen, die ein Item lösten, und nicht, welche Personen das Item lösten.

Die Bestimmung der Personen- und Itemparameter ist rechnerisch anspruchsvoll und kann nur computergestützt erfolgen. Die resultierenden Testwerte der Personen (Personenparameter) und die Itemparameter sind als Maßzahlen einer Differenz- bzw. Verhältnisskala zu interpretieren (zur Metrik einer Rasch-Skala vgl. Conrad et al., 1976a, 1976b; Österreich, 1978).

Wie bei der Guttman-Skala (► Abschn. 8.6.2) sind bei einem modellkonformen Itemsatz die Personenparameter davon unabhängig, welche Items aus der Population aller möglichen Items, die das Merkmal repräsentieren, ausgewählt wurden. Sie sind auch davon unabhängig, wie die Stichprobe, die aus der Population derjenigen Personen gezogen wurde, für die die Skala gilt, zusammengesetzt ist. Entsprechendes trifft auf die Itemparameter zu: Sie sind ebenfalls stichprobenunabhängig. Die Bedeutung dieses als **spezifische Objektivität** bezeichneten Faktums wird bei Fischer (1974, Kap. 19) ausführlich diskutiert. Die spezifische Objektivität bzw. die **Stichprobenunabhängigkeit** ermöglichen die Entwicklung von Modelltests, mit denen die Modellannahmen eines nach dem Rasch-Modell konstruierten Tests überprüft werden können. Sind sämtliche Items homogen im Sinne des Rasch-Modells und treffen auch die übrigen Annahmen zu, müsste die Bestimmung der Personenparameter auf der Basis verschiedener zufälliger Itemstichproben zu identischen Resultaten führen. Entsprechendes gilt für die Bestimmung der Itemparameter.

Wenn also die Itemparameter aufgrund verschiedener Stichproben geschätzt werden, erwartet man identische oder nur zufällig voneinander abweichende Schätzungen, unabhängig von der Stichprobe. Diese Forderung lässt sich auch grafisch überprüfen: Trägt man die Itemparameter, die in einer Stichprobe 1 geschätzt werden, auf der x-Achse eines Koordinatensystems ab und die Itemparameter auf der Basis einer Stichprobe 2 auf der y-Achse, müssten alle Items idealerweise auf der Winkelhalbierenden des Koordinatensystems liegen. Statistische Tests zur Überprüfung der Modellkonformität wurden von Andersen (1973) sowie Fischer und Scheiblechner (1970) entwickelt. Man beachte jedoch die Problematik, die sich bei diesem Test dadurch ergibt, dass die Nullhypothese (Modellkonformität) die „Wunschhypothese" ist (► Abschn. 15.3).

Weichen die Parameterschätzungen bedeutsam voneinander ab, sind einige oder mehrere Items nicht modellkonform, d. h., sie müssen aus dem Test ausgeschieden werden. Die inhaltliche Analyse der selegierten und der modellkonformen Items liefert häufig interessante Aufschlüsse über das eigentlich getestete Merkmal und erleichtert die Formulierung neuer Items, deren Modellkonformität allerdings in weiteren Modelltests nachzuweisen ist. Da sich bei der Konstruktion einer Rasch-Skala in der Regel viele Items als nicht modellkonform erweisen, sollte der ursprüngliche Itemsatz erheblich mehr Items enthalten als die angestrebte Endform (ca. 20 Items reichen im allgemeinen für die Testendform aus).

Ausführliche Informationen zur Mathematik des dichotomen logistischen Modells findet man u. a. bei Fischer (1974), Fischer und Molenaar (1995), Guthke, Böttcher und Sprung (1990), Kubinger (1996), Krauth (1995), Rost (2004) sowie Steyer und Eid (1993, Kap. 16–18). Über Anwendungen probabilistischer Modelle in der Sozialpsychologie berichtet Kempf (1974). Für praktische Anwendungen seien Rost (2004) sowie seine Programmpakete WINMIRA und MULTIRA empfohlen, die die Entwicklung einfacher Rasch-Skalen, aber auch komplexere Skalierungen für unterschiedliche Modelle der Item-Response-Theorie ermöglichen.

▪ Adaptives Testen

Eine spezielle Anwendungsvariante der IRT ist das – v. a. bei Leistungstests relevante – adaptive Testen („adaptive testing"). Bei herkömmlicher Testvorgabe bearbeiten alle Testpersonen nacheinander alle Testaufgaben, was unökonomisch ist, weil in der Regel viel redundante Information gewonnen wird: Eine Testperson mit mittlerer Fähigkeit wird sehr leichte Items mit hoher Wahrscheinlichkeit und sehr schwere Items mit geringer Wahrscheinlichkeit lösen. Ihr derartige – hinsichtlich Itemschwierigkeit von ihrem latenten Fähigkeitsniveau stark abweichende – Items vorzulegen, bietet keinen Informationswert. Denn man weiß nahezu sicher im Vorfeld schon, dass die Testperson die für sie zu leichten Items korrekt beantworten und an den für sie zu schwierigen Items scheitern wird. Eine Vorlage dieser überflüssigen Items soll beim adaptiven Testen vermieden werden, so dass die Testdurchführung ökonomischer wird.

Ist über die Fähigkeit der zu testenden Person nichts bekannt, beginnt das adaptive Testen mit einem mittelschweren Item, um dann – je nachdem, ob das Item gelöst oder nicht gelöst wurde – mit dem schwierigsten oder leichtesten Item fortzufahren. Nach Beantwortung der ersten beiden Items ist eine vorläufige Schätzung des Personenparameters möglich, die dann durch die Vorgabe weiterer Items mit „maximaler Information" sukzessive präzisiert wird. **Items mit maximaler Information** haben eine Lösungswahrscheinlichkeit von 50 %, d. h., die Schwierigkeit der sukzessiv zu bearbeitenden Items sollte jeweils der zuletzt ermittelten Fähigkeit entsprechen. Ob ein derartiges Item vorhanden ist, hängt natürlich von der Größe des (vorgetesteten bzw. kalibrierten) Itempools

ab. Nach Wild (1986, zit. nach Kubinger, 1996) reichen hierfür in der Regel 60–70 Items aus, wobei der Personenparameter nach ca. 15 Items hinreichend genau geschätzt werden kann.

Nach Kubinger (1996) unterscheidet man das sog. „Tailored Testing" und das „Branched Testing", wobei das **Tailored Testing** im Wesentlichen der oben beschriebenen Vorgehensweise entspricht. Es basiert üblicherweise auf dem dichotomen logistischen Modell, was unter praktischen Gesichtspunkten den Einsatz eines Computers erfordert (Guthke & Caruso, 1989). Aufwändigere Modelle berücksichtigen einen weiteren Itemparameter, den sog. Diskriminationsparameter, und einen weiteren Personenparameter, den Rateparameter.

Das **Branched Testing** kann auch als Paper-Pencil-Variante eingesetzt werden. Hierbei werden die Items zu **homogenen Itemgruppen** (z. B. mit 5 Items pro Gruppe) zusammengefasst, die sukzessiv leistungsabhängig zu bearbeiten sind. Man beginnt mit einer Itemgruppe mittlerer Schwierigkeit und fährt mit einer leichteren bzw. schwierigeren Itemgruppe fort, wenn weniger oder mehr als ca. 50 % der Items einer Gruppe (höchstens 1 Item bzw. mindestens 4 Items) gelöst wurden. Liegt die Anzahl der gelösten Items bei ca. 50 %, bleibt die Schwierigkeit der Items des nächsten Blocks auf demselben Niveau etc. Bei dieser Vorgehensweise kann auf eine wiederholte Schätzung des Personenparameters verzichtet werden. Anders als beim Tailored Testing, bei dem sich aufgrund der itemspezifischen Verzweigungen sehr viele individuelle Pfade ergeben, ist die Anzahl der möglichen Pfade beim Branched Testing deutlich geringer, so dass sämtliche pfadspezifischen Personenparameter vorab errechnet und tabellarisch aufbereitet werden können. Das Branched Testing reduziert sich also auf die leistungsabhängige Vorgabe von 35 Itemblöcken mit der anschließenden Entnahme des Fähigkeitsparameters aus einer vorgefertigten Tabelle.

Ausführliche Informationen zum adaptiven Testen findet man bei Frey (2012), Hornke (1993), Kubinger (1995, 1996), van der Linden und Glas (2010; 2000), Meijer und Neving (1999) sowie Wainer (1990). Ferner sei auf das Special Issue „Computerized Adaptive Testing" in der Zeitschrift *Applied Psychological Measurement* (23/3, 1999) hingewiesen. Als Beispiel für adaptive Tests ist u. a. FAKT II – die 2. Auflage des Frankfurter Adaptiven Konzentrationsleistungstests (Moosbrugger & Goldhammer, 2006) – zu nennen, der u. a. in der psychiatrischen und psychologischen Grundlagenforschung sowie in Anwendungsfächern wie der Arbeits- und Betriebspsychologie, der Klinischen und Pädagogischen Psychologie, der Sport- und Verkehrspsychologie zum Einsatz kommt. Weitere Beispiele sind der ANF (Adaptiver Test zur Messung numerischer Flexiblität; Arendasy, Sommer, & Hergovich, 2005), der ATEM (Adaptiver Test zur Erfassung von Emotionsregulation; Knapp-Rudolph, Steinmayr, & Amelang, 2008) oder das AID 2 (Adaptives Intelligenz Diagnostikum, 2. Auflage, Kubinger & Wurst, 2000).

Lernquiz

1. Nennen Sie die Haupt- und Nebengütekriterien psychometrischer Tests!
2. Welche Techniken zur Abschätzung der Reliabilität eines psychometrischen Tests werden verwendet?
3. Was versteht man unter einer Item-Charakteristik und wie wird sie dargestellt?
4. Bei der Korrelation eines Intelligenztests mit der aktuellen Schulleistung spricht man von: a) Vorhersagevalidität, b) Übereinstimmungsvalidität, c) Konstruktvalidität, d) Augenscheinvalidität?
5. Die Korrelation zwischen zwei Testhälften ist eine a) Unterschätzung, b) exakte Schätzung, c) Überschätzung der Reliabilität des Tests.
6. Welche Maßnahmen werden zur Kontrolle bzw. Verhinderung sozial erwünschten Antwortens in Persönlichkeitstests diskutiert und welche werden im Normalfall eingesetzt?
7. Sie planen eine Untersuchung zur sozialen Kompetenz von Polizisten. In der Literatur finden Sie zwei Kompetenz-Tests. Der eine hat eine Reliabilität von .76 und eine Kriteriumsvalidität von .48. Der andere weist eine Reliabilität von .41 und eine Kriteriumsvalidität von .75 auf. Welchen Test wählen Sie und warum?
8. Wie ist die Item-Trennschärfe definiert und in welcher Relation steht sie zur Reliabilität?
9. Welche Klassifikationskriterien sind zur Einteilung psychometrischer Tests einschlägig und welche Gruppen von Tests werden durch sie gebildet?
10. Wozu dient das 3PL-Modell?

Übungsaufgaben

1. Recherchieren Sie verschiedene Intelligenztests für Erwachsene – z. B. aus dem Brickenkamp Handbuch psychologischer und pädagogischer Tests (Brähler et al., 2002a, 2002b) oder bei der Hogrefe Testzentrale (http://www.testzentrale.de). Vergleichen Sie deren Validität, Reliabilität, Testdauer, Verfügbarkeit von Normwerten sowie Kosten und treffen Sie eine Entscheidung, welchen Test Sie anhand dieser Kriterien in einer eigenen Studie einsetzen würden. Begründen Sie Ihre Auswahl!
2. Schauen Sie sich auf den Verlags-Websites die aktuellen Hefte der Zeitschriften *Diagnostica* (Hogrefe) und *Psychological Assessment* (APA) an.

Stellen Sie anhand der Abstracts der Zeitschriftenartikel fest, a) welche inhaltlichen Typen von Tests (Leistungstests, Persönlichkeitstests, klinische Tests) behandelt werden, b) welche Testgütekriterien diskutiert werden (Reliabilität, verschiedene Evidenzen für Validität und/oder verschiedene Formen der Validität) und c) ob auf KTT und/oder auf IRT zurückgegriffen wird.

3. Eine neu entwickelte psychometrische Skala misst die Intensität der Erfahrungen von Jugendlichen mit dem Phänomen Mobbing auf zwei Subdimensionen: 1.) Grad der Viktimisierung durch Mobbing oder 2.) Grad der Beteiligung am Mobbing anderer. Der Test soll genutzt werden, um die Verbreitung von Mobbing sowie die Merkmale der von Mobbing Betroffenen bzw. an Mobbing Beteiligten unter Jugendlichen zu ermitteln. Erläutern Sie verschiedene Evidenzquellen, die im Sinne einer Konstruktvalidierung nach Messick (1995) herangezogen werden können, um die Gültigkeit entsprechender Aussagen auf der Basis der Werte der Mobbing-Skala zu belegen. Vergleichen Sie Ihre Überlegungen mit der Vorgehensweise von Shaw, Dooley, Cross, Zubrick und Waters (2013) bei der Entwicklung ihrer zweidimensionalen Mobbing-Skala! Diskutieren Sie die Begriffsverwendung von Validität und Unterformen der Validität in dem Paper vor dem Hintergrund aktueller Diskussionen zum Verständnis von Testqualität!

Weiterführende Literatur

- Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg: Springer.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.
- Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. Thousand Oaks: Sage.
- Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl.). Heidelberg: Springer.
- Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion* (2. Aufl.). Bern: Huber.

Literatur

Adams, J., Parkinson, L., Sanson-Fisher, R. W., & Walsh, R. A. (2008). Enhancing Self-Report of Adolescent Smoking: The Effects of Bogus Pipeline and Anonymity. *Addictive Behaviors, 33*(10), 1291–1296.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1995). *Standards for educational and psychological testing* (4th ed.). Washington: AERA, APA, NCME.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing* (5th ed.). Washington: AERA, APA, NCME.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing* (6th ed.) Washington: AERA, APA, NCME.

American Psychological Association (APA), American Educational Research Association (AERA), & National Council on Measurement in Education (NCME). (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington DC: APA.

Aguinis, H. & Handelsman, M. M. (1997). Ethical issues in the use of the bogus pipeline. *Journal of Applied Social Psychology 27*(7), 557–573.

Aguinis, H. & Henle, C. A. (2001). Empirical assessment of the ethics of the bogus pipeline. *Journal of Applied Social Psychology, 31*(2), 352–375.

Aguinis, H., Pierce, C. A., & Quigley, B. M. (1993). Conditions under which a bogus pipeline procedure enhances the validity of self-reported cigarette smoking: A meta-analytic review. *Journal of Applied Social Psychology, 23*(5), 352–373.

Aiken, L. R. & Williams, E. N. (1978). Effects of instructions, option keying, and knowledge of test material on seven methods of scoring two-options items. *Educational and Psychological Measurement, 38*(1), 53–59.

Alexander, M. G. & Fisher, T. D. (2003). Truth and consequences: using the bogus pipeline to examine sex differences in self-reported sexuality. *Journal of Sex Research, 40*(1), 27–35.

Allport, G. & Odbert, H. (1936). Trait.names: A psycho-lexical study. *Psychological Monographs*, 47(211).

Alsawalmeh, Y. M. & Feldt, L. S. (2000). A test of the equality of two related α coefficients adjusted by the Spearman-Brown formula. *Applied Psychological Measurement, 24*(2), 163–172.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg: Springer.

Amelang, M. & Zielinski, W. (2002). *Psychologische Diagnostik und Intervention*. Heidelberg: Springer.

Anastasi, A. (1963). *Psychological testing*. New York: MacMillan.

Andersen, E. B. (1973). *Conditional interference and models for measurement*. Kopenhagen: Mentalhygiejnisk Forlog.

Arendasy, M., Sommer, M., & Hergovich, A. (2005). *ANF: Adaptiver Test zur Messung numerischer Flexibilität (Test & Manual)*. Mödling: Schuhfried.

Arnon, Z., Maoz, G., Gazit, T., & Klein, E. (2011). Rorschach indicators of PTSD: A retrospective study. *Rorschachiana, 32*(1), 5–26.

Aster, M. G., Neubauer, A., & Horn, R. (2006). *Wechsler-Intelligenztest für. Erwachsene (WIE-III)*. Frankfurt: Harcourt Test Services.

Baker, F. B. & Kim, S. H. (2004). *Item Response Theory. Parameter Estimation Techniques* (2. Aufl.). New York: Marcel Dekker.

Barter, C. & Renold, E. (1999). The use of vignettes in qualitative research. *Social Research Update, 25*. Retrieved February 7, 2014, from http://sru.soc.surrey.ac.uk/SRU25.html

Barth, N. (1973). Modelle zur Ratewahrscheinlichkeit bei Mehrfach-Antwort-Aufgaben. *Zeitschrift für erziehungswissenschaftliche Forschung, 7*, 63–70.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5*(3), 370–37.

Bierhoff, H. W. & Rudinger, G. (1996). Quasi-experimentelle Untersuchungsmethoden. In E. Erdfelder, R. Mausfeld & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 47–58). Weinheim: Psychologie Verlags Union.

Binder, H. & Eberl, M. (2005). Statistisch unterstützte Spezifikationsprüfung: Die Performance von Tetrad-Test und SEM. *Schriftenreihe zur Empirischen Forschung und Quantitativen Unternehmensplanung der Ludwig-Maximilians-Universität München, 23*. Abgerufen 21. Februar, 2014, unter http://www.imm.bwl.unimuenchen.de/forschung/schriftenefo/ap_efoplan23.pdf

Bölte, S., Adam-Schwebe, S., Englert, E., Schmeck, K., & Poustka, F. (2000). Zur Praxis der psychologischen Testdiagnostik in der deutschen Kinder- und Jugendpsychiatrie: Ergebnisse einer Umfrage. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 28*(3), 151–161.

Borkenau, P. & Ostendorf, F. (2008). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen: Hogrefe.

Bortz, J. & Lienert, G. A., (2008). *Kurzgefasste Statistik für die klinische Forschung. Ein praktischer Leitfaden für die Analyse kleiner Stichproben* (3. Aufl.). Heidelberg: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer

Brähler, E., Holling, H., Leutner, D., & Petermann, F. (Hrsg.). (2002a). *Brickenkamp Handbuch psychologischer und pädagogischer Tests. Bd. 1* (3. Aufl.). Göttingen: Hogrefe.

Brähler, E., Holling, H., Leutner, D., & Petermann, F. (Hrsg.). (2002b). *Brickenkamp Handbuch psychologischer und pädagogischer Tests. Bd. 2* (3. Aufl.). Göttingen: Hogrefe.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brickenkamp, R. (1994). *Test d2 Aufmerksamkeits-Belastungs-Test* (8. Aufl.). Göttingen: Hogrefe.

Briggs, S. R. & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Personality and Social Psychology, 54*(1), 106–148.

Bruno, J. E. & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item. An information theoretic perspective. *Educational and Psychological Measurement, 55*(6), 959–966.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktual. Aufl.). München: Pearson.

Bühner, M. & Ziegler, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69*(4), 548–565.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 103*(56), 2.

Cattell, R. B. & Warburton, F. W. (1967). *Objective personality and motivation tests*. Urbana: University of Illinois Press.

Charles, E. P. (2005). The correction for attenuation due to measurement error. Clarifying concepts and creating confidence sets. *Psychological Methods, 10*(2), 206–226.

Charnas, J. W., Hilsenroth, M. J., Zodan, J., & Blais, M. A. (2010). Should I stay or should I go? Personality assessment inventory and Rorschach indices of early withdrawal from psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 47*(4), 484–499.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*(1), 64–73.

Cizek, G. J., Rosenberg, S., & Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397–412.

Cliff, N. (1988). The eigenvalue-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103*(2), 276–279.

Conrad, W., Bollinger, G., Eberle, G., Kurdorf, B., Mohr, V., & Nagel, B. (1976a). Beiträge zum Problem der Metrik von subjektiven Persönlichkeitsfragebögen, dargestellt am Beispiel der Skalen E und N des HANES, KJI. *Diagnostica, 22*, 13–26.

Conrad, W., Bollinger, G., Eberle, G., Kurdorf, B., Mohr, V., & Nagel, B. (1976b). Erstellung von Rasch-Skalen für den Angst-Fragebogen FS 5–10 und KAT. *Diagnostica, 22*, 110–125.

Costa, P. T. & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI–R) and NEO Five Factor Inventory (NEO-FFI) Professional Manual*. Odessa: Psychological Assessment Resources.

Cramer, P. (1991). *The development of defense mechanisms: Theory, research, and assessment*. New York: Springer.

Cramer, P. (2004). *Storytelling, narrative, and the thematic apperception test*. New York: Guilford.

Crocker, L. & Algina, J. (2006). *Introduction to classical and modern test theory*. Belmont: Wadsworth.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J. (1960). *Essentials of psychological testing*. New York: Harper.

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(281–302).

Crowne, D. P. & Marlowe, D. (1964). *The approval motive*. Hoboken: Wiley.

Del Giudice, M. J. (2010). What might this be? Rediscovering the Rorschach as a tool for personnel selection in organizations. *Journal of Personality Assessment, 92*(1), 78–89.

Dunkel, H., Volpert, W., Zölch, M., Pleiss, C., & Hennes, K. (1993). *Kontrastive Aufgabenanalyse im Büro. Der KABA-Leitfaden Grundlagen und Manual*. Zürich: Verlag der Fachvereine.

Eberl, M. (2004). Formative und reflektive Indikatoren im Forschungsprozess: Entscheidungsregeln und die Dominanz des reflektiven Modells. *Schriften zur Empirischen Forschung und Quantitativen Unternehmensplanung der Ludwig–Maximilians–Universität München, 19*. Abgerufen 17. Dezember, 2013, unter http://www.imm.bwl.uni-muenchen.de/forschung/schriftenefo/ap_efoplan_19.pdf.

Edwards, A. L. (1953). *Edwards personal preference schedule*. New York: Psychological Corperations.

Edwards, A. L. (1957). *The social desirability variable in personality research*. New York: Dryden.

Edwards, A. L. (1970). *The measurement of personality traits by scales and inventories*. New York: Holt, Rinehart & Winston.

Eheim, W. P. (1977). Zur Beeinflußbarkeit der Schwierigkeit von Mehrfachwahl-Aufgaben. *Diagnostica, 23*, 193–198.

Eid, M. (2000). A Multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*(2), 241–261.

Eid, M., Gollwitzer, M., & Schmitt, M. (2011). *Statistik und Forschungsmethoden.* (2. Aufl.). Weinheim: Beltz.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models. A multiple indicator CT-C (M-1) Model. *Psychological Methods, 8*(1), 38–60.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8*(3), 322–337.

Exner, J. E. (1995). The Rorschach: A Comprehensive System. Vol 1: Basic Foundations. Hoboken: Wiley.

Fahrenberg, J. (2002). *Psychologische Interpretation*. Bern: Huber.

Fahrenberg, J., Hampel, R. & Selg, H. (2010). *FPI–R Freiburger Persönlichkeitsinventar* (8. Aufl.). Göttingen: Hogrefe.

Fay, E. (2005). *Tests unter der Lupe. Aktuelle psychologische Testverfahren – kritisch betrachtet* (Bd. 5). Göttingen: Vandenhoeck & Ruprecht.

Feger, H. (1983). Planung und Bewertung von wissenschaftlichen Beobachtungen. In H. Feger & J. Bredenkamp (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich B: Methodologie und Methoden, Serie I, Bd. 2: Datenerhebung* (S. 1–75). Göttingen: Hogrefe.

Feldt, L. S. & Ankenmann, R. D. (1998). Appropriate sample sizes for comparing alpha reliabilities. *22*(2), 170–178.

Feldt, L. S. & Ankenmann, R. D. (1999). Determining sample size for a test of the equality of alpha coefficients when the number of part–tests is small. *Psychological Methods, 4*(4), 366–377.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*(1), 93–103.

Fennekels, G. P. (1995). *Qualitative Führungsstilanalyse (QFA)*. Göttingen: Hogrefe.

Fidler, D. S. & Kleinknecht, R. E. (1977). Randomized response versus direct questioning: Two data collection methods for sensitive information. *Psychological Bulletin, 84*(5), 1045–1046.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fischer, G. H. & Molenaar, J. W. (Eds.). (1995). *Rasch models. Foundations, recent developments and applications.* New York: Springer.

Fischer, G. H. & Scheiblechner, H. H. (1970). Algorithmen und Programme für das Probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12*, 23–51.

Fisseni, H. J. (1974). Zur Zuverlässigkeit von Interviews. *Archiv für Psychologie, 126*, 71–84.

Fisseni, H. J. (1990). *Lehrbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.

Formann, A. K. (1984). *Die Latent-Class-Analyse*. Weinheim: Beltz.

Fox, J. A. & Tracey, P. E. (1986). *Randomized response. A method for sensitive surveys*. Thousand Oaks: Sage.

Freud, S. (1911). *Psychoanalytische Bemerkungen über einen autobiographisch beschribenen Fall von Paranoia*. Frankfurt/Main: Fischer.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 143–172). Heidelberg: Springer.

Fürntratt, E. (1969). Zur Bestimmung der Anzahl interpretierbarer gemeinsamer Faktoren in Faktorenanalysen psychologischer Daten. *Diagnostica, 15*, 62–75.

Gannon, T. A., Keown, K., & Polaschek, D. L. L. (2007). Increasing Honest responding on cognitive distortions in child molesters: The bogus pipeline revisited. *Sexual Abuse: Journal of Research and Treatment, 19*(1), 5–22.

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment, 13*(4), 433–448.

Gordon, M. E. & Gross, R. H. (1978). A critique of methods for operationalizing the concept of fakeability. *Educational and Psychological Measurement, 38*(3), 771–782.

Grayson, D. & Marsh, H. W. (1994). Identification with deficient rank loading matrices in confirmatory factor analysis. Multitrait-multimethods models. *Psychometrika, 59*(1), 121–134.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement, 44*(3), 551–561.

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8*(1), 88–101.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.

Griebel, T. (2006). Self-portrayal in a simulated life: Projecting personality and values in the sims 2. *Game Studies The International Journal of Computer Game Research, 6*(1). Retrieved February 7, 2014, from http://gamestudies.org/0601/articles/griebel

Grubitzsch, S. (1991). *Testtheorie – Testpraxis. Psychologische Tests und Prüfverfahren im kritischen Überblick.* (2. Aufl.). Reinbek Rowohlt.

Guinzbourg, M. (2011). Eating disorders. A current concern: Similarities and differences among the anorexia, bulimia, and EDNOS categories. *Rorschachiana, 32*(1), 27–45.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Guthke, J., Bötcher, H. R., & Sprung, L. (1990). *Psyochodiagnostik* (Bd. 1). Berlin: Deutscher Verlag der Wissenschaften.

Guthke, J. & Caruso, M. (1989). Computer in der Psychodiagnostik. *Psychologische Praxis, 3*, 203–222.

Haladyna, T. M. & Downing, S. M. (1990a). A Taxonomy of Multiple–Choice Item–Writing Rules. *Applied Measurement in Education, 2*(1), 37–50.

Haladyna, T. M. & Downing, S. M. (1990b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51–78.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 143–172). Heidelberg: Springer.

Hergovich, A., Arendasy, M. E., M., S., Bognar, B., & Olbrich, A. (2004). Zur Dimensionalität und Konstruktvalidität eines videobasierten, objektiven Persönlichkeitstests zur Messung der Riskobereitschaft im Straßenverkehr. *Zeitschrift für Differentielle und Diagnostische Psychologie, 25*, 79–86.

Hofer, J., Busch, H., Bender, M., Ming, L., & Hagemeyer, B. (2010). Arousal of achievement motivation among student samples in three different cultural contexts: Self and social standards of evaluation. *Journal of Cross-Cultural Psychology, 41*(5–6), 758–775.

Holland, P. W. & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale: Erlbaum.

Hornke, L. F. (1993). Mögliche Einspareffekte beim computergestützten Testen. *Diagnostica, 39*, 109–199.

Hsu, L. M. (1979). A comparison of three methods of scoring true-false tests. *Educational and Psychological* Measurement, 39(4), 785–790.

Ingenkamp, K. (1995). Pädagogische Diagnostik. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3. Aufl., S. 495–510). Weinheim: Beltz.

Irle, M. & Allehoff, W. (1984). *Berufs-Interessen-Test II (BIT II).* Göttingen: Hogrefe.

Jäger, R. S., Beetz, E., Erler, R., & Walter, R. (1994). MSD – Mannheimer Schuleingangsdiagnostikum (4. Aufl.). Göttingen: Hogrefe.

Jäger, R. S. (1995). Die Datenbasis. In R. S. Jäger & F. Petermann (Eds.), *Psychologische Diagnostik* (3. Aufl., S. 345–350). Weinheim: Beltz.

Jäger, R. S. & Petermann, F. (Hrsg.). (1995). *Psychologische Diagnostik. Ein Lehrbuch* (3. Aufl.). Weinheim: Beltz.

Jaradad, D. & Tollefson, N. (1988). The impact of alternative scoring procedures for multiple-choice on test reliability, validity, and grading. *Educational and Psychological Measurement, 48*(3), 627–635.

Johnson, D. (2009). *A compendium of psychosocial measures: Assessment of people with serious mental illness in the community*. New York: Springer.

Jones, E. & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76*(5), 349–364.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kastner-Koller, U. & Deimann, P. (2002). *Wiener Entwicklungstest (WET)*. Göttingen: Hogrefe.

Katko, N. J., Meyer, G. J., Mihura, J. L., & Bombel, G. (2010). A principal components analysis of Rorschach aggression and hostility variables. *Journal of Personality Assessment, 92*(6), 594–598.

Keillor, B., Owens, D., & Pettijohn, C. (2001). A cross-cultural/cross-national study of influencing factors and socially desirable response biases. *International Journal of Market Research, 43*(1), 63–84.

Keiser, R. E. & Prather, E. N. (1990). What is the TAT? A review of ten years of research. *Journal of Personality Assessment, 55*(3–4), 800–803.

Kelava, A. & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 77–102). Heidelberg: Springer.

Kempf, W. (1974). *Probabilistische Modelle in der Sozialpsychologie*. Bern: Huber.

Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen. In M. Kersting (Hrsg.), *Qualitätssicherung in der Diagnostik und Personalauswahl – der DIN Ansatz* (S. 141–210). Göttingen: Hogrefe.

Kiers, H. A. L., Takane, Y., & Ten Berge, J. M. F. (1996). The analysis of multitrait–multimethod matrices via constraint components analysis. *Psychometrika, 61*(4), 601–628.

Knapp-Rudolph, M., Steinmayr, R., & Amelang, M. (2008). Konstruktion und Erprobung des Adaptiven Tests zur Erfassung von Emotionsregulation (ATEM) als einer Facette von Emotionaler Intelligenz. In W. Sarges & D. Scheffer (Hrsg.), *Innovationen in der Eignungsdiagnostik* (S. 145–154). Göttingen: Hogrefe.

Koch, K. (2003). *Der Baumtest. Der Baumzeichenversuch als psychodiagnostisches Hilfsmittel* (10. Aufl.). Bern: Huber.

Körner, A., Geyer, M., Roth, M., Drapeau, M., Schmutzer, G., Albani, C., et al. (2008). Persönlichkeitsdiagnostik mit dem NEO-Fünf-Faktoren-Inventar: Die 30-Item-Kurzversion (NEO-FFI-30). *Psychotherapie Psychosomatik Medizinische Psychologie, 58*(6), 238–245.

Kottke, J. L., Olson, D. A., & Shultz, K. S. (2010). The devil is in the details: A comment on „What might this be? Rediscovering the Rorschach as a tool for personnel selection in organizations“ (Del Giudice, 2010). *Journal of Personality Assessment, 92*(6), 610–612.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Beltz.

Kubinger, K. & Wolfsbauer, C. (2010). On the risk of certain psycho-technological response otions in multiple-choice tests: does a particular personality handicap examinees? *European Journal of Psychological Assessment, 26*, 302–308.

Kubinger, K. D. (1995). Objektive Diagnostik. In K. Pawlik (Hrsg.), *Enzyklopädie der Psychologie, Differentielle Psychologie 1, Grundlagen und Methoden* (S. 507–541). Göttingen: Hogrefe.

Kubinger, K. D. (1996). Methoden der Psychologischen Diagnostik. In E. Erdfelder, R. Mausfeld & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 567–576). Weinheim: Beltz.

Kubinger, K. D. (1997). Zur Renaissance der objektiven Persönlichkeitstests sensu R. B. Cattell. In H. Mandl (Hrsg.), *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996* (S. 755–761). Göttingen: Hogrefe.

Kubinger, K. D. & Wurst, E. (2000). *Adaptives Intelligenz Diagnostikum – Version 2.1 (AID 2)*. Göttingen: Beltz.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*(2), 228–244.

Langeheine, R. & Rost, J. (1996). Latent-Class-Analyse. In E. Erdfelder, R. Mausfeld & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 315–348). Weinheim: Beltz.

Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. D. (1981). *Das State-Trait-Angstinventar (Testmappe mit Handanweisung, Fragebogen STAI-G Form X 1 und Fragebogen STAI-G Form X 2)*. Weinheim: Beltz.

Lienert, G. A. (1967). *Drahtbiegeprobe (DBP)* (2. Aufl.). Göttingen: Hogrefe.

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Beltz.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Lück, H. E. & Timaeus, E. (1969). Skalen zur Messung manifester Angst (MAS) und sozialer Wünschbarkeit (SD-E und SD-CM). *Diagnostica, 15*, 134–141.

Ma, L. (2010). Analyse latenter Klassen und das Mixed Raschmodell. Abgerufen 29. Februar, 2012, unter http://www.docstoc.com/docs/25282676/AnalyselatenterKlassenunddasMixedRaschmodell

Macha, T. & Petermann, F. (2013). Objektivität von Entwicklungstests. Zur Standardisierung der entwicklungsdiagnostischen Befunderhebung. *Diagnostica, 59*(4), 183–191.

Markus, H. & Nurusius, P. (1986). Possible selves. *American Psychologist, 41*(4), 858–866.

Mattson, C. L., Campbell, R. T., Karabatsos, G., Agot, K., Ndinya-Achola, J. O., Moses, S., & Bailey, R. C. (2010). Scaling sexual behavior or „sexual risk propensity“ among men at risk for HIV in Kisumu, Kenya, *AIDS and Behavior, 14*(1), 162-172.

McCutcheon, A. L. (1987). *Latent class analysis*. Thousand Oaks: Sage.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Erlbaum.

Meijer, R. R. & Neving, M. L. (1999). Computerized adaptive testing. Overview and introduction. *Applied Psychological Measurement, 23*, 187–194.

Messick, S. J. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027.

Messick, S. J. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 33–48). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Messick, S. J. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. [DOI:10.1037/0003–066X.50.9.741]. *American Psychologist, 50*(9), 741–749.

Middleton, K. L. & Jones, J. L. (2000). Socially desirable response sets. The impact of country culture. *Psychology & Marketing, 17*(2), 149–163.

Moosbrugger, H. (2012a). Item–Response–Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 227–274). Heidelberg: Springer.

Moosbrugger, H. (2012b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 103–118). Heidelberg: Springer

Moosbrugger, H. & Goldhammer, F. (2006). *Frankfurter Adaptiver Konzentrationsleistungs-Test. Grundlegend neu bearbeitete und neu normierte 2. Aufl. des FAKT von Moosbrugger und Heyden (1997) (FAKT-II)*. Bern: Huber.

Moosbrugger, H. & Kelava, A. (2012a). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 7–26). Heidelberg: Springer.

Moosbrugger, H. & Kelava, A. (2012b). *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl.). Heidelberg: Springer.

Mummendey, H. D. (1990). *Psychologie der Selbstdarstellung*. Göttingen: Hogrefe.

Mummendey, H. D. (1999). *Die Fragebogenmethode*. Göttingen: Hogrefe.

Murray, H. A. (1938). Conference. In H. A. Murray (Ed.), *Explorations in personality: A clinical and experimental study of fifty men of college age* (pp. 399–412). New York: Oxford University Press.

Murray, H. A. (1943). *Thematic Apperception Test*. Cambridge: Harvard University Press.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2008). Psychological testing on the Internet: New problems, old issues. In D. N. Bersoff (Eds.), *Ethical conflicts in psychology* (4th ed., pp. 306–312). Washington, DC: American Psychological Association.

Newton, P. & Shaw, S. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301–319.

Ortmann, R. (1973). Zur Gewichtung von Testaufgaben nach ihrer Schwierigkeit. Diskussion eines von E. Rützel vorgeschlagenen Bewertungsverfahrens. *Psychologie und Praxis, 17*, 87–89.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*(3), 343–355.

Österreich, R. (1978). Welche der sich aus der Rasch-Skalierung ergebenden Personenkennwerte sind für statistische Auswertungen geeignet? *Diagnostica, 24*, 341–349.

Oswald, W. D. & Gunzelmann, T. (1991). Altern, Gedächtnis und Leistung-Veränderungen und Interventionsmöglichkeiten. In E. Lang & K. Arnold (Hrsg.), *Altern und Leistung: Medizinische, psychologische und soziale Aspekte* (S. 272–281). Stuttgart: Enke.

Petermann, F. & Petermann, U. (Hrsg.). (2010). *Hamburg-Wechsler-Intelligenztest für Kinder IV: HAWIK-IV; Manual* (3. Aufl.). Bern: Huber.

PISA-Konsortium Deutschland (Hrsg.). (2004). *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland.* Münster: Waxmann.

Porcerelli, J. H., Cogan, R., Kamoo, R., & Miller, K. (2010). Convergent validity of the defense mechanisms manual and the defensive functioning scale. *Journal of Personality Assessment, 92*(5), 432–438.

Pöschl, S. (2010). *Die Handhabung mobiler Erreichbarkeit. Experimentelle Überprüfung eines interdependenz- und persönlichkeitstheoretischen Modells*. Stuttgart: Kohlhammer.

Rauch, W. A. & Moosbrugger, H. (2011). Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich B, Methodologie und Methoden.* Serie II, *Psychologische Diagnostik. Bd. 2, Methoden der psychologischen Diagnostik* (S. 1–87). Göttingen: Hogrefe.

Rey, E.-R. (Hrsg.). (1977). *Allgemeine Probleme Psychologischer Tests*. Zürich: Kindler.

Richardson, M. W. & Kuder, G. F. (1939). The calculations of test reliability coefficients based on the method of rational equivalence. *Journal of Educational Psychology, 30*(9), 681–687.

Rivera, B. A. & Viglione, D. J. (2010). Conceptualization of children's interpersonal relatedness with the Rorschach. A qualitative multiple case study. *Journal of Personality Assessment, 92*(5), 377–389.

Roese, N. J. & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114*(2), 363–375.

Rogers, W. T. & Harley, D. (1999). An empirical comparison of three and four choice items and tests. Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234–247.

Rorschach, H. (1927). *Rorschach test – psychodiagnostic plates*. Cambridge: Hogrefe.

Rorschach, H. (1941). *Psychodiagnostik. Methoden und Ergebnisse eines wahrnehmungspsychologischen Experiments* (4. Aufl.). Bern: Huber.

Rossiter, J. R. (2008). Content validity of measures of abstract constructs in management and organizational research. *British Journal of Management, 19*, 380–388.

Rossiter, J. R. (2010). *Measurement for the social sciences: The C-OAR-SE method and why it must replace psychometrics*. New York: Springer.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau, 50*, 140–156.

Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion* (2., vollständig überarbeitete Aufl.). Bern: Huber.

Rost, J. & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten. *Zeitschrift für Differentielle und Diagnostische Psychologie, 1*(4), 29–49.

Rütter, T. (1973). *Formen der Testaufgabe*. München: Beck.

Rützel, E. (1972). Zur Gewichtung von Testaufgaben nach Schwierigkeit. *Psychologie und Praxis, 16*, 128–133.

Sarges, W., Wottawa, H., & Roos, C. (2010). *Handbuch wirtschaftspsychologischer Testverfahren. Bd. II: Organisationspsychologische Instrumente*. Legerich: Pabst.

Schaipp, C. & Plaum, E. (2000). Sogenannte projektive Techniken: Verfahren zwischen Psychometrie, Hermeneutik und qualitativer Heuristik. *Journal für Psychologie, 8*(1), 29–44.

Schermelleh-Engel, K. & Schweizer, K. (2012). Multimethod-Analysen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 345–362). Heidelberg: Springer.

Schermelleh-Engel, K. & Werner, C. S. (2012). Reliabilität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. u. überarb. Aufl., S. 119–142). Heidelberg: Springer.

Scheurer, H. (1995). Diagnostik als Testung. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3. Aufl., S. 257–263). Weinheim: Beltz.

Schmidt-Atzert, L., Bühner, M., Rischen, S., & Warkentin, V. (2004). Erkennen von Simulation und Dissimulation im Test d2. *Diagnostica, 50*(3), 124–133.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*(5), 529–540.

Schmidt, L. R. (1975). *Objektive Persönlichkeitsmessung in diagnostischer und klinischer Psychologie*. Weinheim: Beltz.

Schmitt, N. & Stults, D. M. (1986). Methodology review. Analysis of multi-trait-multimethod matrices. *Applied Psychological Measurement, 10*(1), 1–22.
Schneider, W., Blanke, I., Faust, V., & Küspert, P. (2011). *WLLP–R – Würzburger Leise Leseprobe*. Göttingen: Hogrefe.
Schnell, R., Hill, P. B., & Esser, E. (1999). *Methoden der empirischen Sozialforschung* (6. Aufl.). München: Oldenbourg.
Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der empirischen Sozialforschung*. München: Oldenbourg.
Schuler, H. & Prochaska, M. (2001). *Leistungsmotivationsinventar (LMI)*. Göttingen: Hogrefe.
Schumacher, J. & Brähler, E. (2000). Testdiagnostik in der Psychotherapie. In W. Senf & M. Broda (Hrsg.), *Praxis in der Psychotherapie. Ein integratives Lehrbuch: Psychoanalyse, Verhaltenstherapie, Systemische Therapie* (2. neu bearbeitete und erweiterte Aufl., S. 116–128). Stuttgart: Thieme.
Shaw, T., Dooley, J. J., Cross, D., Zubrick, S. R., & Waters, S. (2013). The forms of bullying scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in adolescence. *Psychological Assessment, 25*(4), 1045–1057.
Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin, 112*(1), 140–154.
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295.
Staabs, G. von (1997). *Scenotest* (8. Aufl.). Göttingen: Hogrefe.
Stanley, J. C. & Wang, M. D. (1970). Weighting test items and test-item options: An overview of the analytical and empirical literature. *Educational and Psychological Measurement, 30*(1), 21–35.
Sternberg, R. J. (1992). Psychological bulletin's top 10 "hit parade". *Psychological Bulletin, 112*(3), 387–388.
Steyer, R. & Eid, M. (1993). *Messen und Testen*. Heidelberg: Springer.
Steyer, R. & Eid, M. (2001). *Messen und Testen.* (2. Aufl.). Heidelberg: Springer.
Sullivan, J. L. & Feldman, S. (1979). *Multiple indicators. An introduction*. Thousand Oaks: Sage.
Teglasi, H. (2010). *Essentials of TAT and other storytelling assessments. Essentials of psychological assessment series* (2nd ed.). Hoboken: Wiley.
Testkuratorium. (2010). TBS-TK – Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. *Psychologische Rundschau, 61*(1), 52–56.
Van der Linden, W. & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York: Springer.
Van der Linden, W. J. & Glas, A. W. (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer.
Van der Linden, W. J. & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
Wainer, H. (1990). *Computerized adaptive testing. A primar.* Hillsdale: Erlbaum.
Warner, S. L. (1965). Randomized responses: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*(309), 63–69.
Wartegg, E. (1939). *Gestaltung und Charakter. Ausdrucksdeutung zeichnerischer Gestaltung und Entwurf einer charakterologischen Typologie.* Universität Leipzig, Leipzig.
Wechsler, D. (1964). *Die Messung der Intelligenz Erwachsener* (3. Aufl.). Bern: Huber.
Weise, G. (1975). *Psychologische Leistungstests.* Göttingen: Hogrefe.
Weizmann-Henelius, G., Kivilinna, E., & Eronen, M. (2010). The utility of Rorschach in forensic psychiatric valuations. A case study. *Nordic Psychology, 62*(3), 36–49.
Westen, D. (1991). Clinical assessment of object relations using the TAT. *Journal of Personality Assessment, 56*(1), 56–74.
Westhoff, K., Hagemeister, C., Kersting, M., Lang, F., Moosbrugger, H., Reimann, G., et al. (Hrsg.). (2010). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (3. Aufl.). Lengerich: Pabst.
Whipple, R. & Fowler, J. C. (2011). Affect, relationship schemas, and social cognition: Self-injuring borderline personality disorder in patients. *Psychoanalytic Psychology, 28*(2), 183–195.
Wilcox, R. R. (1981). Analyzing the distractors of multiple-choice test items or partitioning multinomial cell probabilities with respect to a standard. *Educational and Psychological Measurement, 41*(4), 1051–1068.
Wild, B. (1986). *Der Einsatz adaptiver Teststrategien in der Fähigkeitsmessung. Unveröffentlichte Dissertationsschrift.* Institut für Psychologie der Universität Wien, Wien.
Wood, J. M., Lilienfeld, S. O., Nezworski, M. T., Garb, H. N., Allen, K. H., & Wildermuth, J. L. (2010). Validity of Rorschach inkblot scores for discriminating psychopaths from nonpsychopaths in forensic populations: A meta-analysis. *Psychological Assessment, 22*(2), 336–349.
Wottawa, H. & Hossiep, R. (1987). *Grundlagen psychologischer Diagnostik*. Göttingen: Hogrefe.
Wottawa, H. & Hossiep, R. (1997). *Anwendungsfelder psychologischer Diagnostik*. Göttingen: Hogrefe.
Yousfi, S. (2005). Mythen und Paradoxien der Klassischen Testtheorie (I). *Diagnostica, 5*, 1–11.
Zhao, Y. & Hambleton, R. (2009). Software for IRT analyses: Descriptions and features. Retrieved February 29, 2012, from http://www.umass.edu/remp/software/CEA-652.ZH-IRTSoftware.pdf
Ziler, H. (1997). *Der Mann-Zeichen-Test*. Münster: Aschendorff.
Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133.
Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement, 30*(2), 121–144.

10.5 Physiologische Messungen[1]

Lernziele

- Wissen, was die physiologische Messung als wissenschaftliche Datenerhebungsmethode auszeichnet;
- Verschiedene Möglichkeiten physiologischer Messungen im Hinblick auf Herangehensweise und Anwendungsgebiet voneinander abgrenzen können.
- Potenzielle Messprobleme und Gegenmaßnahmen kennen.
- Vor- und Nachteile von physiologischen Messungen kennen und gegeneinander abwägen können.
- Einfache physiologische Messungen im Rahmen eigener Studien planen und durchführen können, sofern praktische Übungsmöglichkeiten gegeben sind.

Das Spektrum der Datenerhebungsmethoden wird durch physiologische Messungen um eine neue Qualität erweitert, weil hier körperliche Prozesse auf der Ebene unterschiedlicher Organsysteme sehr detailliert und z. T. auch in Echtzeit erfasst werden können. Isoliert betrachtet stellen physiologische Variablen (z. B. Blutdruck, Muskelaktivität) meist keine sozialwissenschaftlich besonders relevanten Informationen dar. Erst wenn physiologische Maße als Indikatoren psychischer Konstrukte verwendet werden – man spricht dann auch von **psychophysiologischen Messungen** – gewinnen sie für sozialwissenschaftliche Fragestellungen an Aussagekraft. Dabei wird entweder per Induktionsschluss von physiologischen Maßen auf psychische Prozesse geschlossen und/oder es werden parallel zu den physiologischen Messungen auch psychologische Variablen erhoben und die Daten miteinander in Beziehung gesetzt.

In diesem Kapitel wird zunächst erklärt, wodurch sich wissenschaftliche physiologische Messungen als empirische Datenerhebungsmethoden auszeichnen (► Abschn. 10.5.1). Anschließend werden in den Sozialwissenschaften gebräuchliche Verfahren erläutert und anhand von Studienbeispielen illustriert: Nicht-invasive Messverfahren für die **Aktivitäten von Gehirn, Herz-Kreislauf-System, Haut, Muskeln** und **Augen** stehen im Mittelpunkt.

Das Gehirn hat eine herausgehobene Position im menschlichen Organismus. Zusammen mit dem Rückenmark bildet es das **zentrale Nervensystem** (ZNS), das für die Steuerung vieler lebenserhaltender Prozesse (z. B. Atmung), aber auch für motorische, emotionale, motivationale und kognitive Prozesse notwendig ist. Physiologische Messungen von Hirnaktivitäten sind deswegen für die Forschung besonders aussagekräftig. Die in den 1920er-Jahren entwickelte Elektroenzephalografie **(EEG)** ist bis heute die wichtigste Methode zur Messung von Hirnaktivität im wachen oder schlafenden Zustand des Menschen. Von besonderem Interesse ist die Messung **ereigniskorrelierter Potenziale** (EKP), also von Veränderungen der Hirnaktivität im Zuge von Sinneswahrnehmungen oder kognitiven Prozessen. Die Elektroenzephalografie verfügt durch ihre direkte elektrische Messung neuronaler Aktivität über eine sehr **hohe zeitliche Auflösung** im Bereich von wenigen Millisekunden, dafür ist ihre räumliche Auflösung beschränkt: Es können nur neuronale Aktivitäten an der Oberfläche des Gehirn gemessen werden und auch diese nur relativ großflächig.

Eine wichtige Ergänzung zum EEG stellt deswegen die in den 1980er- und 1990er-Jahren entwickelte funktionelle Magnetresonanztomografie **(fMRT)** dar. Bei dieser Methode wird die neuronale Aktivität unterschiedlicher Hirnareale indirekt anhand des Sauerstoffgehaltes des Blutes erfasst (sog. **BOLD-Effekt**, „blood oxygenation level dependent"). **Hohe räumliche Auflösung** geht bei dieser technisch sehr aufwändigen physiologischen Messmethode mit geringer zeitlicher Auflösung einher, denn Durchblutungsveränderungen dauern zumindest mehrere Hundert Millisekunden.

Will man Daten zur Hirnaktivität sowohl mit hoher zeitlicher als auch hoher räumlicher Auflösung gewinnen, so ist eine **Kombination von EEG und fMRT** indiziert. Zudem kann anstelle der Elektroenzephalografie auch die Magnetoenzephalografie **(MEG)** zum Einsatz kommen. Diese Methode misst die neuronale Aktivität nicht elektrisch wie das EEG, sondern magnetisch und ist dadurch einerseits technisch aufwändiger, andererseits aber auch räumlich präziser. Als Alternative zur fMRT wird in jüngster Zeit die Nahinfrarotspektrografie **(NIR)** betrachtet. Sie misst optisch den Sauerstoffgehalt im Blut und kann dadurch mit hoher räumlicher Auflösung aktive Hirnareale identifizieren. Diese Methode ist für Probanden angenehmer als das fMRT. Derartige Methoden einordnen zu können, ist relevant, da sie inzwischen immer häufiger in der Fachliteratur auftauchen. Nur wenige Studierende und Forschende dürften indessen Gelegenheit haben, diese teuren Technologien für eigene Studien zu nutzen.

Das **periphere Nervensystem** umfasst nervöses Gewebe außerhalb des Zentralnervensystems. Es enthält Anteile des vegetativen Nervensystems (Sympathikus, Parasympathikus, Darmnervensystem), das die Aktivität der inneren Organe und Drüsen reguliert, und des somatischen Nervensystems (sensorische und motorische Systeme), über das der menschliche Organismus mit seiner Umwelt interagiert. Sowohl das vegetative als auch das somatische Nervensystem haben zentral-

[1] Der vorliegende Abschnitt wurde für die 5. Auflage grundlegend überarbeitet und erweitert. Wir danken Herrn Dr. Marcus Ising (Max Planck-Institut für Psychiatrie München) für die frühere Fassung des Abschnitts, die in den Auflagen 2, 3 und 4 erschienen ist und aus der einzelne Teile übernommen wurden.

nervöse Anteile und sind auf der Ebene des ZNS eng miteinander verknüpft. Im vorliegenden Kapitel werden Merkmale von Hirnaktivitäten als Indikatoren des zentralen Nervensystems (► Abschn. 10.5.2) sowie Merkmale von kardiovaskulären (► Abschn. 10.5.3), elektrodermalen (► Abschn. 10.5.4), muskulären (► Abschn. 10.5.5) und Augenaktivitäten (► Abschn. 10.5.6) als Indikatoren des peripheren Nervensystems behandelt. Hier handelt es sich teilweise um weniger technologieintensive Verfahren, die deswegen auch eher von Studierenden in eigenen Studien eingesetzt werden können (z. B. Messung von Blutdruck und Hautleitfähigkeit).

Das **endokrine System** (Hormonsystem) sowie das **Immunsystem** können im vorliegenden Kapitel aus Platzgründen nicht näher erläutert werden. Es ist jedoch zu beachten, dass hormonelle und immunologische Indikatoren in der Forschung immer stärkere Bedeutung gewinnen. Ihre Messung erfolgt v. a. über die Auswertung von Speichel-, Urin- und Blutproben.

Der letzte Abschnitt dieses Kapitels wendet sich der **Messung sexueller Erregung** zu (► Abschn. 10.5.7), denn an diesem Beispiel kann das spannungsreiche Verhältnis zwischen psychologischen und physiologischen Indikatoren samt der zugehörigen Interpretationsprobleme besonders gut veranschaulicht werden. Die Messung sexueller Erregung stellt innerhalb der Biopsychologie einen Spezialfall dar, spielt aber in diversen grundlagen- und anwendungswissenschaftlichen Forschungsfeldern der Sozial- und Humanwissenschaften durchaus eine wichtige Rolle.

10.5.1 Wissenschaftliche physiologische Messung

Um nachzuvollziehen, wodurch sich physiologische Messungen als Datenerhebungsverfahren auszeichnen und von anderen Erhebungsmethoden unterscheiden, beschäftigen wir uns zunächst mit ihrer Definition und Indikation und grenzen sie zudem von nicht-wissenschaftlichen Alltagsbeobachtungen physiologischer Signale ab. Anschließend gehen wir kurz auf methodische Grundlagen physiologischer Messungen ein (d. h. auf Messprinzipien, Messprobleme und Analysestrategien).

Definition und Indikation der physiologischen Messung

Bei dem Versuch, physiologische Messungen zu definieren, stößt man auf zahlreiche Definitionen und Begrifflichkeiten. Um die hier vorgestellten Datenerhebungsmethoden besser einordnen zu können, sollen zunächst Forschungsdisziplinen innerhalb der Psychologie voneinander abgegrenzt werden, die maßgeblich mit physiologischen Indikatoren arbeiten (Schandry, 2003).

Die Messung physiologischer Indikatoren ist als Datenerhebungsmethode in den Sozial- und Humanwissenschaften neben der Medizin und Biologie insbesondere in der Psychologie von Bedeutung. Grundlegend ist für letztere, dass ein biologischer Zugang zur Beschreibung und Erklärung menschlichen Erlebens und Verhaltens gewählt wird, der v. a. die **Biopsychologie** bzw. biologische Psychologie („biological psychology“) prägt. Die Biopsychologie untergliedert sich nach ihren Schwerpunktsetzungen wiederum in drei Teilbereiche:

1. Die **physiologische Psychologie** („physiological psychology“) als Teildisziplin der Biopsychologie erklärt in möglichst direkter Weise Verhaltensphänomene durch Prozesse des Zentralnervensystems, insbesondere durch Hirnprozesse. Sie ist stark grundlagenwissenschaftlich ausgerichtet und zielt auf Theoriebildung ab. Anwendungsaspekte stehen im Hintergrund. Abhängige Variablen stellen in der Regel Verhaltensparameter dar, während auf der biologischen Ebene experimentell manipuliert wird (z. B. durch elektrische Stimulation, neurochemische Manipulation oder chirurgische Eingriffe als experimentelle unabhängige Variablen). Das Vorgehen der physiologischen Psychologie ist dementsprechend in weiten Teilen **tierexperimentell** (Animalforschung). Sie hat jedoch stets das Ziel, auf Basis der so errungenen Befunde menschliches Verhalten zu erklären.
2. Die **Neuropsychologie** („neuropsychology“) stellt die zweite Subdisziplin der Biopsychologie dar. Diese studiert den Zusammenhang zwischen menschlichem Verhalten und Gehirnprozessen. Hierbei werden u. a. Verhaltensauffälligkeiten bei **Patienten und Patientinnen mit Störungen der Gehirnfunktion** untersucht. Solche Schädigungen können durch Krankheiten, Verletzungen und neurochirurgische Eingriffe verursacht worden sein. Die Zuordnung von Verhaltensdefiziten zu den lokalisierten Läsionen (Störungen oder Schädigungen) lässt Rückschlüsse über die Bedeutung der beeinträchtigten Hirnregion für Verhalten und Erleben beim gesunden Menschen zu. Während die klinische Neuropsychologie sehr stark anwendungsbezogen ausgerichtet ist und das Anliegen hat, die Situation der betroffenen Patienten zu verbessern, wendet sich die Neuropsychologie insgesamt – spätestens seit der Entwicklung funktioneller Bildgebung – zunehmend auch der Untersuchung gesunder Menschen zu.
3. Die **Psychophysiologie** („psychophysiology“) als dritte Subdisziplin der Biopsychologie untersucht, wie sich psychische Vorgänge (z. B. kognitive Prozesse) auf der Ebene der physiologischen Prozesse (z. B.

Hirnaktivitäten, Blickbewegungen) abbilden (zur kognitiven Psychophysiologie s. Rösler, 2011). Die Psychophysiologie operiert mit einem Mehr-Ebenen-Ansatz menschlichen Verhaltens (Schandry, 2003). Neben der subjektiv-verbalen Ebene wird der Bereich der körperlichen Prozesse und des offenen motorischen Verhaltens in die Beobachtung mit einbezogen. Die Psychophysiologie sieht ihre wichtigste Forschungsaufgabe darin, immer empfindlichere Indikatoren bereit zu stellen, mit deren Hilfe psychische Prozesse auf nicht-verbalem Wege analysiert werden können. Das Methodenarsenal dieser Forschungsrichtung entstammt v. a. der medizinischen Diagnostik (z. B. Messung von hirnelektrischen Vorgängen oder von Aktivitäten des Herz-Kreislauf-Systems) und wird in den ▶ Abschn. 10.5.2 bis ▶ Abschn. 10.5.6 detaillierter vorgestellt.

Im vorliegenden Kapitel beschränken wir uns auf Datenerhebungsverfahren, die im Rahmen von psychophysiologischen Messungen eingesetzt werden. Diese haben in den Sozialwissenschaften die weiteste Verbreitung erfahren. Denn zum einen können diese Verfahren an gesunden Menschen angewendet werden (es ist z. B. keine Stichprobe von Patienten mit Hirnschädigungen notwendig, wie etwa in der klinischen Neuropsychologie). Zum anderen handelt es sich meist um Verfahren, die auch ohne medizinisch oder tierärztlich geschultes Personal durchgeführt werden können (non-invasive Verfahren). **Psychophysiologische Messungen** haben ihre disziplinären Wurzeln in der Psychophysiologie und befassen sich mit der Erfassung körperlicher Indikatoren und Prozesse im Zusammenhang mit dem subjektiven Erleben und Verhalten. Die Bezeichnung **„physiologische Messung“** wird im Kontext dieses Kapitels weitgehend synonym mit „psychophysiologischer Messung“ verwendet. Sie hat den Vorteil der Kürze und betont außerdem den Einsatz und Nutzen der Methode auch außerhalb der Psychologie (z. B. Kommunikationswissenschaft, Politikwissenschaft, Sexualwissenschaft).

Physiologische Messungen – Die physiologischen Messungen („physiological measurements“) dienen der objektiven Erfassung und Quantifizierung bestimmter Merkmale physiologischer Prozesse in unterschiedlichen Organsystemen des Körpers mittels entsprechender Messgeräte. Die erhobenen Merkmale (z. B. Herzschlagfrequenz) werden als physiologische Indikatoren („physiological indicator“) oder Biosignale („bio signal“) bezeichnet. Meist werden mehrere Biosignale integriert erfasst und ausgewertet (z. B. Hirnaktivität und Blickbewegungen).

Psychophysiologische Messungen – Die psychophysiologischen Messungen („psychophysiological measurements“) entstehen, wenn man physiologische Messungen mit Merkmalen des Verhaltens und Erlebens zeitbasiert verknüpft (z. B. gleichzeitige Messung von Hirnaktivität und Steuerung eines Computerspiels). Erst dadurch werden die physiologischen Messwerte für die sozial- und humanwissenschaftliche Forschung sinnvoll interpretierbar.

Die zunehmende Bedeutung von physiologischen Messungen liegt u. a. darin begründet, dass sie **Einblicke** in die dem menschlichen Erleben und Verhalten zugrunde liegenden somatischen (körperlichen) Prozesse versprechen. Insbesondere Verfahren wie die funktionelle Magnetresonanztomografie (fMRT) , die es ermöglichen, dem arbeitenden Gehirn „zuzuschauen“, lösen eine große Faszination aus (Hussy, Schreier, & Echterhoff, 2010; Schandry, 2003). Diese Aspekte haben dazu geführt, dass psychophysiologische Messungen und insbesondere die bildgebenden Verfahren in den letzten Jahren einen Aufschwung erfahren haben. Die fortschreitende Weiterentwicklung und Verbesserung der technischen Messgeräte haben in diesem Fall dazu geführt, dass immer mehr Studien durchgeführt werden, die psychologische Prozesse live am arbeitenden Gehirn untersuchen. Der „Blick ins Gehirn“ ist fraglos faszinierend, wird jedoch auch zunehmend kritisch gesehen (Jäncke, 2010; Jäncke & Petermann, 2010; Mausfeld, 2010). Dies gilt insbesondere für ein teilweise vorschnelles Vertrauen auf „harte“ Daten, wie sie die Messung physiologischer Indikatoren zu versprechen scheinen und entsprechend weitgreifende Interpretationen menschlichen Erlebens auf deren Grundlage. Wie später noch ausgeführt wird, sind physiologische Daten als Ergebnisse sog. objektiver Messverfahren keinesfalls immer so genau und aussagekräftig, wie es auf den ersten Blick erscheinen mag.

Insgesamt besteht in der Wissenschaft jedoch der Konsens, dass die Kenntnis biologischer Vorgänge ein wichtiger Schlüssel zum Verständnis menschlichen Verhaltens und Erlebens darstellt. So gelten biologische Grundlagen von psychologischen Phänomenen mittlerweile als **zunehmend erforschbar** (Berntson & Cacioppo, 2000). Darauf aufbauend wird die Frage formuliert, inwieweit biologische Funktionen hinreichende Voraussetzungen für das Auftreten von psychischen Phänomenen sind. Besonders wichtig für die Sozial- und Humanwissenschaften ist die Untersuchung der kausalen **Wechselwirkungen zwischen biologischen Faktoren einerseits und psychischen, sozialen sowie kulturellen Faktoren andererseits** (für eine ausführliche Diskussion s. Hussy et al., 2010). So zeigte sich z. B., dass sich durch Psychotherapie störungsspezifische Aktivitätsveränderungen des Gehirns (Überaktivierung bestimmter Regionen und neurochemische Veränderungen etwa bei posttraumatischer Belastungsstörung, Zwangsstörungen oder Depressionen) dauerhaft beeinflussen lassen in Richtung der Hirnaktivität von psychisch unauffälligen Normstichproben (Fuchs, 2009; Schauer et al., 2006).

Objektive Messmethoden – Die objektiven Messmethoden („objective measurement methods/techniques") sind Messverfahren, die nicht auf Selbstauskünften der Untersuchungsobjekte basieren, sondern bei denen die Daten durch Außenstehende oder Apparaturen erfasst werden. Zu den objektiven Messmethoden zählen neben der Fremdbeobachtung und einigen psychologischen Testverfahren vor allem die physiologischen Messungen. Der Hauptvorteil der physiologischen Messung als objektiver Messmethode liegt darin, dass subjektive Verzerrungen durch die Untersuchungsteilnehmenden, wie sie bei Selbstauskunftsdaten vorkommen, weitgehend ausgeschlossen werden können. Ihr Hauptnachteil – neben dem meist hohen untersuchungstechnischen Aufwand – besteht darin, dass von objektiven physiologischen Messdaten kein direkter Rückschluss auf menschliches Erleben und Handeln möglich ist – dazu werden Selbstauskunftsdaten benötigt.

Neben der Biopsychologie beziehen auch andere sozialwissenschaftliche Disziplinen physiologische Messungen **ergänzend** in ihren Kanon der Datenerhebungsmethoden ein. Bei vielen psychologischen Konstrukten spielen Annahmen über somatische Zustände oder Prozesse eine zentrale Rolle. Als Beispiele seien Emotion, Aktivierung oder Stress genannt, die generell unter Einbeziehung physiologischer Prozesse definiert werden. In den psychologischen Anwendungsfeldern sind die klinische Psychologie, die Arbeitspsychologie, die pädagogische Psychologie und die Werbepsychologie zu nennen (Becker-Carus, 1981; Janke & Kallus, 1995; Schandry, 1996, 2003). Des Weiteren werden psychophysiologische Messungen (wie z. B. die Blickbewegungsmessung bzw. das Eye-Tracking) auch vermehrt in der medien- und kommunikationswissenschaftlichen Forschung und nicht zuletzt in der Markt- und Konsumentenforschung eingesetzt (Bente, Eschenburg, & Fürtjes, 2007; Das, Ester, & Kaczmirek, 2011; Meißner & Decker, 2010; Zhou & Jin, 2009).

Ob in einer konkreten Studie physiologische Messungen nützlich bzw. erforderlich sind, ist anhand des Forschungsproblems zu entscheiden. Ist zwischen verschiedenen Datenerhebungsmethoden abzuwägen, so sind folgende **Vorteile physiologischer Messungen** gegenüber den in der Sozialforschung per Interview, Fragebogen oder Persönlichkeitstest erhobenen Selbstauskünften zu berücksichtigen. Diese Vorzüge können dafür sprechen, physiologische Messungen ausschließlich oder – das ist der Regelfall – ergänzend einzusetzen:

1. Physiologische Indikatoren lassen sich durch die Untersuchungspersonen **nicht so direkt und gezielt beeinflussen bzw. verfälschen** wie Selbstauskünfte, welche z. B. bei Bedarf sozial erwünscht gestaltet werden können. In gewissem Rahmen besteht jedoch auch bei physiologischen Messungen die Möglichkeit der Einflussnahme der untersuchten Person auf das Messergebnis (z. B. kann ein Zungenbiss das Ergebnis eines Polygrafen-Tests verfälschen; auch können Probanden im Experiment z. B. die Stimulusvorlage bewusst ignorieren und an etwas anderes denken etc.).
2. **Gedächtnisfehler** sind eine häufige Verfälschungsquelle bei Selbstauskünften (z. B. die Häufigkeit des eigenen Verhaltens wird nicht korrekt erinnert), die bei physiologischen Messungen aufgrund der Objektivität der Methode keine Rolle spielen (z. B. automatische Aufzeichnung der Herzschlagfrequenz per EKG).
3. Mit physiologischen Messmethoden lassen sich Daten über Phänomene erheben, die **nicht bewusst wahrgenommen** werden (Veränderung des Hautwiderstands durch Stress) und/oder **nicht eindeutig verbalisiert** werden können (genauer Anstieg der Herzschlagfrequenz).
4. Physiologische Messungen ermöglichen eine bis zu millisekundengenaue **Erfassung von Daten im Zeitverlauf**. Bei Selbstauskünften sind solche Verlaufsdaten nur mit speziellen, eher selten eingesetzten Methoden erhebbar (z. B. Continous Response Measurement, CRM; ► Abschn. 8.3.2).
5. Physiologische Daten erlauben einen **Einblick in Teilprozesse der Informations- und Emotionsverarbeitung**, den man durch Verhaltensmaße nicht bekommt, denn letztere sind immer nur das Endprodukt der Verarbeitung (Sommer, Ulrich, & Leuthold, 1996; Sternberg, 2001).

Der Einsatz von physiologischen Messungen ist jedoch auch mit **Nachteilen** verbunden:

1. Der forschungspraktische Hauptnachteil besteht darin, dass zur Messung der physiologischen Indikatoren entsprechende **Messgeräte** zur Verfügung stehen müssen. Dies stellt bei der Messung des Blutdrucks mit einer Manschette noch kein großes Problem dar. Die benötigten Geräte für bildgebende Verfahren wie z. B. der funktionellen Magnetresonanztomografie (fMRT) sind jedoch sehr kostspielig. Neben dem Anschaffungspreis sind auch die **Kosten** für jede einzelne Untersuchung sehr hoch.
2. Eine valide und reliable Datenerhebung und -auswertung erfordern bei physiologischen Messungen besonderes **Expertenwissen**, da viele Messgeräte nicht ohne entsprechende Expertise bedient und die Daten auch nicht ohne einschlägiges Hintergrundwissen analysiert und interpretiert werden können. Insgesamt gesehen ist es relativ schwierig, mithilfe einer Messung von physiologischen Indikatoren aussagekräftige Befunde zu erzielen. So können **Interferenzen und Artefakte** die Ergebnisse verfälschen, wenn nicht entsprechende Vorkehrungen getroffen werden.
3. Je nach Messaufbau und Dauer der Messung ist die Ableitung von physiologischen Indikatoren mehr oder weniger **zumutbar für die Untersuchungspersonen**. So können das Anbringen von Elektroden auf der

Der „Lügendetektor" bzw. Polygraf

Der im Volksmund oft als „Lügendetektor" bezeichnete **Polygraf** („polygraph") ist ein von den Psychologen Max Wertheimer und Carl Gustav Jung Anfang des 20. Jahrhunderts entwickeltes spezielles Gerät, das zeitgleich mehrere ausgewählte Indikatoren des vegetativen Nervensystems (v. a. Blutdruck, Herzrate, Schweißabsonderung) misst und darauf abzielt zu unterscheiden, ob ein Mensch gerade die Wahrheit sagt oder lügt (Herbold-Wootten, 1982; Lockhart, 1975). In einigen Staaten der USA wird der Polygraf in Gerichtsverfahren als Beweismittel eingesetzt. In Deutschland ist der Einsatz solcher „Lügendetektoren" im Rahmen von Strafrechtsprozessen jedoch nicht zulässig, obwohl Beschuldigte zuweilen selbst fordern, man möge sie einem Polygrafentest unterziehen, um damit ihre Unschuld zu beweisen. Der Einsatz von Polygrafentests spielt mittlerweile auch in den Medien eine Rolle. So werden in manchen Fernsehshows Lügendetektortests durchgeführt, um z. B. zu prüfen, ob ein der Untreue bezichtigter Partner tatsächlich die Wahrheit sagt, wenn er den Seitensprung abstreitet. In solchen TV-Shows werden die Ergebnisse des Tests nicht hinterfragt, sondern als unumstößlich dargestellt, so dass möglicherweise in der Öffentlichkeit ein falsches Bild der Leistungsfähigkeit des „Lügendetektors" entsteht.

Der umstrittene Polygrafentest läuft darauf hinaus, bei einer Person ausgewählte Indikatoren des vegetativen Nervensystems zu messen, während man sie gleichzeitig in spezifischer Weise befragt (Tatwissenstest oder Kontrollfragentest). Der **Tatwissenstest** kann nur eingesetzt werden, wenn die Ermittlungsbehörden über zuverlässige Detailinformationen zur Tat verfügen, die ansonsten nur Täterin oder Täter kennen. Werden nun eine Reihe von Fragen gestellt, von denen eine einzige tatrelevant und die anderen irrelevant sind („Wurde das Opfer mit einem Schal erdrosselt? Wurde das Opfer mit bloßen Händen erwürgt? Wurde das Opfer mit einem Gürtel erdrosselt? Wurde das Opfer mit einem Kissen erstickt?), so wird eine stärkere Reaktion auf die tatrelevante Frage als Hinweis auf Tatwissen und somit Täterschaft gedeutet. Beim **Kontrollfragentest** werden neben den kritischen Fragen zur Tat („Haben Sie Ihren Mann getötet?"), die das eigentliche Ziel der Lügendetektion darstellen, inhaltlich mit der Tat unverbundene Fragen gestellt, auf die im Sinne sozialer Erwünschtheit typischerweise gelogen wird (z. B. „Haben Sie in den ersten 18 Jahren Ihres Lebens einmal etwas genommen, das Ihnen nicht gehörte?"). Wird die mit dem Polygrafen gemessene physiologische Reaktion bei den tatkritischen Fragen von mehreren Beurteilern als bedeutsam größer eingeschätzt als die Reaktion bei den Kontrollfragen, so folgert man, dass auch bei den tatkritischen Fragen gelogen wurde.

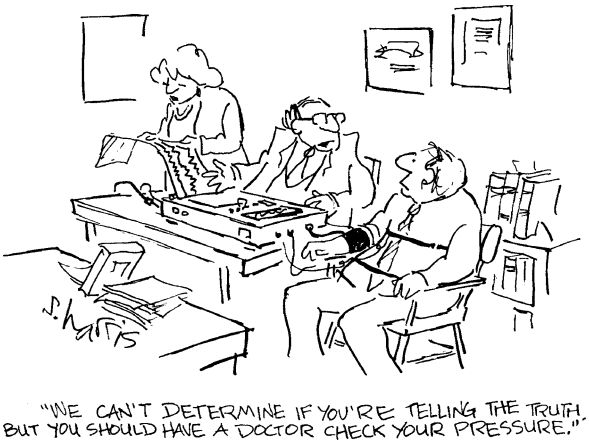
Dieser Rückschluss von einer unspezifischen physiologischen Reaktion auf einen spezifischen Bewusstseinszustand (absichtliche Lüge oder wahrheitsgemäße Aussage) ist jedoch wissenschaftlich nicht haltbar. Denn eine starke physiologische Reaktion auf eine bestimmte tatkritische Frage kann aus vielen und z. T. hochindividuellen psychologischen Gründen erfolgen (z. B. Scham, Ärger, Angst vor falscher Beschuldigung etc.) und stellt somit keinen klaren Beweis für bewusstes Lügen dar. Ausreißerwerte können absichtlich erzeugt werden, etwa indem sich die Testperson unbemerkt in die Zunge beißt. Einen besonders „objektiv" erscheinenden Lügendetektor einzusetzen, der faktisch aber unzuverlässige Ergebnisse liefert (mangelnde Validität), stellt im Rahmen von polizeilichen Ermittlungen und Gerichtsverfahren eine doppelte Gefahr dar: Zum einen können Unschuldige fälschlich verdächtigt und umgekehrt Täter zu Unrecht entlastet werden (zum mangelnden Validitätsnachweis des Polygrafentests s. Fiedler, Schmid, & Stahl, 2002; ◘ Abb. 10.31).

Wie physiologische Prozesse einerseits und Bewusstseinsprozesse andererseits einander bedingen, wird in Philosophie, Kognitionswissenschaft und Psychologie als **Leib-Seele-Problem** diskutiert, für dessen Bearbeitung durch die sich ständig verbessernden physiologischen Messmethoden eine wachsende Datenbasis zur Verfügung steht. So wird heute die Frage diskutiert, ob und wie es mit modernen bildgebenden Verfahren, die einen **„Blick ins arbeitende Gehirn"** gewähren, nicht doch möglich ist, Lügen objektiv zu erkennen und welche ethischen Implikationen dies hat (Racine, Bell, & Illes, 2010).

Haut und das während des Messvorgangs möglichst regungslose Verharren als belastend empfunden werden.

4. Physiologische Messungen stellen **reaktive Methoden** dar (die Untersuchungspersonen wissen, dass sie gerade an einer Studie teilnehmen) und werden typischerweise im **Forschungslabor** durchgeführt. Die Übertragbarkeit der Befunde auf Alltagssituationen ist je nach konkretem Studiendesign zu hinterfragen.
5. Weiterhin muss festgehalten werden, dass physiologische Indikatoren generell **keinen direkten Rückschluss auf das subjektive Erleben und Handeln** ermöglichen.

In ▶ Der „Lügendetektor" bzw. Polygraf wird veranschaulicht, wie irreführend der Glaube ist, durch objektive physiologische Messungen die „wahren" Gedanken und Gefühle von Menschen erfassen zu können.



◘ **Abb. 10.31** Die engen Grenzen der Aussagekraft des „Lügendetektors" sind in der breiten Öffentlichkeit oft nicht bekannt. (© Sidney Harris/Search ID: shrn208, Rights Available from CartoonStock.com)

Abgrenzung wissenschaftlicher und nicht-wissenschaftlicher physiologischer Messungen

Im Alltag begegnen uns Messungen von physiologischen Indikatoren relativ häufig. Wird z. B. der Blutdruck bei der Hausärztin mittels Manschette gemessen, oder beobachten wir unsere Pulsfrequenz beim Joggen mit einer Pulsuhr, so werden bestimmte physiologische Indikatoren erfasst, um Informationen über unseren Körper – genauer: unser Herz-Kreislauf-System – zu erhalten. Es bestehen hier jedoch mehrere wesentliche Unterschiede zum wissenschaftlichen Einsatz physiologischer Messungen. So werden im Alltag physiologische Indikatoren in der Regel einzeln erfasst und auch für sich genommen interpretiert (Blutdruck als Indikator für den Gesundheitszustand eines Menschen; Pulsfrequenz als Indikator für die aktuelle körperliche Belastung beim Sport). In der sozialwissenschaftlichen Forschung werden dagegen meist **mehrere Indikatoren abgeleitet und integriert** ausgewertet.

Die Messung physiologischer Indikatoren erfolgt zwar auch im Alltag durchaus zielgerichtet und regelgeleitet (so werden individuelle Werte mit Normwerten in Verbindung gebracht, die Erfassung erfolgt anhand etablierter Messgeräte und Verfahren, wie z. B. eine Analyse von Blutwerten). Sie dienen jedoch meist praktischen Diagnosezwecken und nicht **allgemeinem Erkenntnisgewinn** bzw. der Überprüfung und Entwicklung von Theorien, wie das in der Wissenschaft der Fall ist (◘ Abb. 10.32).

Schließlich müssen physiologischen Messungen in der Wissenschaft nachweislich die **Gütekriterien** Validität, Reliabilität und Objektivität erfüllen, die auch an andere standardisierte Messverfahren wie psychometrische Skalen und psychologische Tests angelegt werden. Die Auswertung und Interpretation physiologischer Messdaten muss innerhalb der Forschungscommunity intersubjektiv nachvollziehbar sein. Entsprechende Anforderungen werden im Alltag nicht unbedingt gestellt und kaum überprüft. Eine Ausnahme stellt der Einsatz physiologischer Messungen im Rahmen medizinischer Diagnostik dar, die sich ebenfalls an wissenschaftlichen Gütekriterien orientiert.

Methodische Grundlagen physiologischer Messungen

Physiologische Messungen gehören zur Gruppe der objektiven Messmethoden, zu denen z. B. auch Leistungstests wie der Intelligenztest (▶ Abschn. 10.4.3), Fremdbeobachtung (▶ Abschn. 10.1) oder Dokumentenanalyse (▶ Abschn. 10.6) zählen. Von diesen objektiven Messungen abzugrenzen sind Selbstauskunfts- oder Selbstberichtsdaten („self-report data“), die auf subjektiven Selbstbeobachtungen und entsprechenden Angaben der Untersuchungsteilnehmenden basieren (z. B. Interview ▶ Abschn. 10.2 und Fragebogen ▶ Abschn. 10.3).

Bei den meisten physiologischen Indikatoren, die für psychologische Fragestellungen erhoben werden, handelt es sich um elektrische Indikatoren. Die am Körper messbaren Prozesse spiegeln sich in sog. **Biosignalen** („biosignals“) wider, wie z. B. der Hautleitfähigkeit als indirekt elektrisches Signal oder der Herzaktivität und den Gehirnströmen als direkte elektrische Signale. Die Klasse direkter elektrischer Biosignale nennt man auch **Biopotenziale** („biopotentials“). Abgeleitet werden Spannung (mV oder μV), Leitfähigkeit (μSiemens oder μMho) oder Widerstand (kOhm). Nicht-elektrische Signale (z. B. Blutdruck, Temperatur oder Atmung) werden mithilfe von Messfühlern erfasst und durch Messwandler in elektrische Signale transformiert.

Allgemeine Messprinzipien

Eine Anordnung zur Registrierung von Biosignalen besteht prinzipiell immer aus **Elektroden** (bei elektrischen Signalen) bzw. **Messfühlern und Messwandlern** (bei nicht-elektrischen Signalen) sowie aus einer Verstärkungs- und einer Registriereinheit. Die **Verstärkungseinheit** muss aufgrund der geringen Signalamplituden besonders unempfindlich gegen elektrische Störeinflüsse sein (z. B. Netzspannung, elektrostatische Einflüsse). Zusätzlich kommen **Filter** zum Einsatz, die die zu verstärkenden Frequenzen auf den für die Interpretation des jeweiligen Biosignals relevanten Bereich begrenzen. Zur Signalaufzeichnung mittels **Registriereinheit** werden anstelle der früher verwendeten analogen Schreibsysteme (z. B. Tinten- oder Thermoschreiber) heute Registriercomputer verwendet. Auf die so gespeicherten Daten kann jederzeit für spezielle Signalauswertungen zurückgegriffen werden (für weitere Informationen zu Messaufbauten s. Kramme, 2006).

Messprobleme

Bei jeder physiologischen Messanordnung muss damit gerechnet werden, dass die Messwerte des interessierenden Indikators durch andere physiologische Prozesse der Versuchspersonen sowie durch äußere bzw. technische Einflüsse verzerrt werden, man spricht zusammenfassend von Artefakten (für weitere Informationen zur Artefaktbehandlung s. Kramme, 2006):

1. **Physiologische Artefakte** („physiological artifacts“): Die interessierenden Messwerte können durch begleitende physiologische Prozesse verzerrt werden. Eine beschleunigte oder verlangsamte Atmung der Untersuchungsperson schlägt sich z. B. im EKG nieder. Unwillkürliche Augenbewegungen und der Lidschlag beeinflussen das EEG (sog. Augenbewegungs- und Blinkartefakte). Weitere **Bewegungsartefakte** („mo-

vement artifacts“) entstehen durch unruhiges Sitzen, Kauen oder Kopfdrehungen. Um entsprechende Störquellen auszuschließen, werden die Versuchspersonen gebeten, sich während der Untersuchung möglichst nicht zu bewegen. Zudem können Störeinflüsse wie Augenbewegungen oder Atemfrequenz ihrerseits gemessen und durch geeignete Filter oder Algorithmen in der Datenverarbeitung kontrolliert werden.

2. **Technische Artefakte** („non-physiological artifacts“): Neben den Verzerrungen, die durch die Versuchsperson selbst zustande kommen, existieren vielfältige äußere und technische Störeinflüsse. Fehlerhafte Messungen entstehen z. B., wenn Elektroden verrutschen oder einen Wackelkontakt haben. Solche Artefakte lassen sich durch eine optimale Platzierung von gut haftenden funktionsfähigen Elektroden minimieren. Die abgeleiteten Biosignale können auch durch elektrische oder magnetische Felder aus der Umgebung der Messapparatur verfälscht werden (z. B. durch elektrische Leitungen, Mobiltelefone, Radio- und Fernsehsender, Aufzüge etc.). Solche Artefakte sind durch eine möglichst vollständige Abschirmung des Untersuchungsraums, die Entfernung von Geräten, die als Störquelle fungieren, den Einsatz von Vorverstärkern für die interessierenden Signale bzw. von verbesserten elektronischen Komponenten zu vermeiden. Technische Artefakte können auch auftreten, wenn verschiedene physiologische Messungen parallel durchgeführt werden (z. B. können Magnetimpulse der fMRT das EEG beeinflussen).

Neben diesen technischen Störgrößen sind auch einige methodische Grundprobleme zu berücksichtigen, die mit Besonderheiten des „gemessenen“ Individuums verbunden sind. Beispielhaft sollen im Folgenden die Spezifitäts- und Ausgangswertproblematik kurz skizziert werden.

Spezifitätsproblematik („specificity problem“) Der Psychologe John I. Lacey stellte in den 1950er-Jahren in Laboruntersuchungen fest, dass ein Teil seiner Probanden unabhängig von verschiedenen Stresssituationen stets mit einem für sie typischen Reaktionsmuster reagierte (Konzept der autonomen Reaktionsspezifität, „autonomic specificity“; Lacey, Bateman, & Van Lehan, 1953). Dieses Prinzip wird heute unter dem Begriff der **individualspezifischen Reaktion** zusammengefasst (Engel, 1972; Janke, 1976), etwa wenn eine Person auf unterschiedliche Stimuli (Lampenfieber, Ärger, Zeitdruck) stets mit erhöhter Herzfrequenz reagiert. Von der individualspezifischen Reaktion sind stimulusspezifische und motivationsspezifische Reaktionen zu unterscheiden. **Stimulusspezifische Reaktionen** werden von einer bestimmten Umweltbedingung bei allen Individuen in gleicher oder ähnlicher Weise hervorgerufen (z. B. Schreckreaktion). Als **motivationsspezifische Reaktionen** bezeichnet man die durch einen spezifischen Motivationszustand beim Individuum hervorgerufenen Reaktionen (Birbaumer & Schmidt, 2006; Foerster, Schneider, & Walschburger, 1983). Im Idealfall sollte bei biopsychologischen Untersuchungen stets der jeweilige Anteil an individualspezifischen Reaktionen, an stimulusspezifischen und an motivationsspezifischen Reaktionen erfasst werden.

Ausgangswertproblematik („initial value problem“) Die Ausgangswertproblematik betrifft Veränderungsmessungen, d. h. Messungen vor und nach einer bestimmten Intervention oder einem Ereignis. Genauer gesagt spielt die hierbei auftretende Abhängigkeit zwischen Ausgangswert und der Differenz aus Verlaufs- und Ausgangswert (Veränderungswert) eine Rolle. Diese Problematik wurde schon zu Beginn des 20. Jahrhunderts diskutiert und von dem Psychologen Max Wilder (1931) in seinem **Ausgangswertgesetz** („law of initial value“) beschrieben. Demnach ist im Rahmen von psychophysiologischen Messungen die Reaktion auf funktionssteigernde Reize umso geringer, je höher der Ausgangswert (Baseline) des gemessenen physiologischen Indikators ist. Zum Beispiel steigt bei Personen, die bereits einen hohen Ruhepuls haben, bei Stress der Puls geringfügiger an als bei Personen mit sehr niedrigem Ruhepuls. Umgekehrt ist die Reaktion auf funktionshemmende Reize größer, je höher der Ausgangswert ist (für eine ausführliche Diskussion der Ausgangswertproblematik s. Kallus, 1992).

Analyse physiologischer Messwerte

Die Analyse psychophysiologischer Messwerte ist anspruchsvoll: Bereits bei der Untersuchungsplanung und Datenerhebung sollten die Prinzipien der integrierten Datenanalyse beachtet werden, bevor die erhobenen Daten mittels Verlaufsanalysen und Parametrisierung ausgewertet werden.

Integrierte Datenanalyse Mit integrierter Datenanalyse ist gemeint, dass es bei vielen Forschungsproblemen, zu deren Bearbeitung physiologische Messungen herangezogen werden, sinnvoll ist, a) mehrere physiologische Indikatoren zu erfassen, diese b) mit Stimulusbedingungen sowie c) mit Maßen subjektiven Erlebens in Verbindung zu setzen. Für die begleitende Messung von psychologischen Indikatoren bietet sich u. a. das Continuous Response Measurement (CRM) an. Dabei werden subjektive Eindrücke und Bewertungen von Stimulus- und Situationsmerkmalen mittels eines stufenlos verschiebbaren Reglers (z. B. eines Drehknopfs) kontinuierlich über die Zeit hinweg erhoben (► Abschn. 8.3.2).

Mit Blick auf eine integrierte Datenanalyse folgt für die Untersuchungsplanung, dass

- mehrere theoriebasiert ausgewählte physiologische Indikatoren Berücksichtigung finden sollten,
- die psychophysiologische Messdatenerfassung kontinuierlich bzw. in regelmäßigen und möglichst kurzen Zeitintervallen vorgenommen wird,
- die registrierten physiologischen Reaktionen mit den jeweils wirksamen Stimulusbedingungen oder Situationsmerkmalen in einen unmittelbaren Zusammenhang zu bringen und
- die physiologischen Reaktionen zu anderen Verlaufsdaten wie z. B. dem gezeigten Verhalten oder dem per CRM erfassten subjektiven Erleben in Beziehung zu setzen sind (Kempter & Bente, 2004).

Die **zeitsynchrone Dokumentation** von Stimulus- und Reaktionsprozessen ist zur Aufklärung von Wirkzusammenhängen besonders wichtig. Für die theoriegeleitete Interpretation von physiologischen Messdaten ist es von entscheidender Bedeutung, ob die Aktivierung von bestimmten Organsystemen vor, während oder nach dem Auftreten eines Stimulus erfolgt, wie viel Zeit zwischen Stimulus und Reaktion vergeht (Latenz), wie lange die Aktivierung anhält und wie oft sie sich wiederholt.



Abb. 10.32 In der Forschung werden physiologische Messungen theorieprüfend eingesetzt. (© Mike Baldwin/Search ID: mba0369, Rights Available from CartoonStock.com)

Verlaufsanalyse und Parametrisierung Im Zuge der integrierten Datenerhebung und Datenanalyse muss mit einer großen Menge an Daten operiert werden: Psychophysiologische **Messzeitreihen** bzw. eine Fülle an Messdaten für sehr viele zeitlich eng aufeinander folgende Zeitpunkte, Beobachtungsdaten für das offen gezeigte Verhalten und die Dokumentation von sich evtl. fortlaufend ändernden Stimuluskonfigurationen liegen zunächst als **Verlaufsdaten** vor (► Die Integrierte Rezeptions-Prozessanalyse (IRP) in der Medienwirkungsforschung). Die grafische Darstellung und Inspektion (Betrachtung) dieser Kurven steht am Anfang der Auswertung. Aus den kontinuierlichen Messwertveränderungen über die Zeit werden für weitere statistische Analysen (z. B. Mittelwertsvergleiche, Korrelationen) zusätzlich oft **Parameter** gebildet. Mithilfe der **Parametrisierung** können interessierende Strukturmerkmale der verschiedenen kontinuierlichen Messzeitreihen im Sinne von Merkmal und Zeit umfassenden und kontextbezogenen Kenngrößen definiert werden.

Die methodische Grundlage der Parametrisierung liegt im Prinzip der **konstruktbezogenen Messoperation** (Frey & Bente, 1989; Kempter, 1997). Diese Messkonzeption begreift die Parameterbildung als eine Bewertung bzw. Evaluation des inneren und äußeren Verhaltens unter der Perspektive einer bestimmten theoretischen Modellvorstellung (zur Operationalisierung und Messung latenter Konstrukte ► Kap. 8). Um zu einer quantitativen Aussage über interessierende Parameter zu gelangen, sind die Forschenden aufgerufen, ihr Konzept durch definitorische Festlegungen unter folgenden Gesichtspunkten zu explizieren (Kempter & Bente, 2004, S. 288):

- Welche Aspekte des Verhaltens sind in einem bestimmten Kontext zum Zweck der Parameterbestimmung aus den Verlaufsdaten zu selektieren?
- In welcher Weise sind sie untereinander und mit dem Kontext zu verknüpfen?
- In welcher Weise sind die verschiedenen selektierten Datenpunkte numerisch zu gewichten?

In einer Studie zur Aufmerksamkeit gegenüber Werbespots könnten z. B. das Umschaltverhalten, Blickbewegungen (wird hin- oder weggesehen), physiologische Aktivierung (Messung evozierter Potentiale im EEG, Messung der Pulsvolumenamplitude oder der Herzschlagfrequenz) und subjektive Bewertung des Medienangebots untersucht werden. Ein erster Schritt besteht nun in der visuellen Betrachtung der gesamten Verlaufsdaten. Dabei kann überprüft werden, ob bestimmte Parameter möglicherweise sehr konstant bleiben (z. B. die Herzschlagfrequenz). Zeigen sich bei bestimmten Indikatoren kaum Veränderungen, besteht die Möglichkeit, diese aus

Die Integrierte Rezeptions-Prozessanalyse (IRP) in der Medienwirkungsforschung

In der Medienwirkungsforschung werden häufig physiologische Indikatoren ergänzend zu Selbstauskünften über das individuelle Rezeptionserleben erhoben. In den letzten Jahren wurden vermehrt Instrumentarien entwickelt, um die Anforderungen einer integrierten psychophysiologischen Prozessanalyse zu erfüllen, gleichzeitig aber auch eine hohe Forschungsökonomie zu garantieren (Kempter & Bente, 2004, S. 286). Die Integrierte Rezeptions-Prozessanalyse IRP (Bente, 2000) ist eine speziell auf Fernsehwirkungsforschung zugeschnittene Laboreinheit, die die kontinuierliche Erfassung von psychophysiologischen Daten (Pulsvolumenamplitude und Herzschlagfrequenz mittels Plethysmografie), Bewertungsmaßen (mittels Continous Response Measurment, CRM; ▶ Abschn. 8.3.2) und des Umschaltverhaltens von bis zu 60 Probanden zulässt. Als Werkzeug zur Verlaufsanalyse dient das Programm TVLab Pro. In ◘ Abb. 10.33 ist die Benutzungsoberfläche des Programms **TVLab Pro** dargestellt. Das Datenfenster zeigt den z-normierten Verlauf der Pulsvolumenamplitude (oben). Das Ansteigen der mit invertierter Skala dargestellten Werte (negative Werte oben) zeigt Vasokonstriktion an, also hohe Aktivierung. Die untere Kurve stellt die Bewertung mittels eines Schiebereglers im Sinne von Continous Response Measurment (CRM; auf einer Skala von -4 = negativ bis $+4$ = positiv) während der Rezeption dar. Die Messdaten sind dabei mit dem Videobild (mittleres Fenster oben) synchronisiert. Beide Verlaufskurven stellen die gemittelten Daten von $n = 20$ Untersuchungspersonen während der Rezeption einer Quizshow dar. Die grau markierten Bereiche in den Verlaufskurven indizieren die beiden Werbeblöcke. Es zeigt sich insbesondere anhand der Verlaufsdaten im zweiten Werbeblock, dass es zu Beginn des Werbeblocks zu einem deutlichen Nachlassen der zuvor aufgebauten Erregung (Spannung) sowie zu einem Einbruch der Bewertung kommt (markiert durch Pfeile). Die Zuschauer koppeln offensichtlich vom Stimulus ab. Zum Ende des Werbesegments zeigt sich eine Art Reset der Aufmerksamkeit. Erregung und Bewertung erreichen innerhalb kürzester Zeit wieder das Niveau vor dem Einsetzen der Werbung (Kempter & Bente, 2004).

der weiteren Analyse auszuschließen, es sei denn, auch die Konstanz eines physiologischen Indikators bei veränderten Stimulusbedingungen ist von Interesse. Weiterhin ist die Frage zu beantworten, ob sich zu bestimmten Zeitpunkten bzw. an bestimmten Stellen des Werbespots auffällige Veränderungen in den Verlaufsdaten ausmachen lassen. Stehen etwa physiologische Parameter und subjektive Bewertungen des Medienangebots mit dem Umschaltverhalten in einem Zusammenhang?

Ein derartiges Vorgehen zeichnet sich durch die systematische Trennung der Deskriptions- und Evaluationsaspekte der Messoperation aus. Es bietet mehrere Vorteile:

- Das Modell der Forschenden, das sich in der Parameterdefinition niederschlägt, beeinflusst die Gewinnung der Primärdaten nicht, insbesondere dann, wenn im Sinne einer integrierten Datenanalyse physiologische Parameter möglichst umfassend erhoben werden. Dieselben Daten können also auch für andere Fragestellungen und Parameterkonzepte ausgewertet werden, solange die entsprechenden physiologischen Indikatoren erhoben wurden.
- Ein transparentes Parameterkonzept dient der wissenschaftlichen Nachvollziehbarkeit und kann entsprechend in fachlichem Austausch diskutiert und weiterentwickelt werden.
- Die Bewertung von psychophysiologischen Verlaufsdaten aus der Perspektive eines bestimmten Parameterkonzeptes ist aufgrund der Trennung zwischen Deskriptions- und Evaluationsaspekt reversibel. Dies bedeutet, dass die Forschenden keine erneute Datenerhebung durchführen müssen, wenn die für einen Parameter gewählte Definition sich als unhaltbar erweist.
- Zudem können Parameter definiert werden, die zum Zeitpunkt der Datenerhebung evtl. noch gar nicht im Fokus standen. So kann man dieselben Verlaufsdaten immer wieder für neue Fragestellungen auswerten, neue Gesichtspunkte in Betracht ziehen und auch nicht erwartete Effekte entdecken (für weitere Ausführungen zur Parametrisierung s. Kempter & Bente, 2004).

Überblick gängiger Messverfahren für physiologische Indikatoren

Wie sich Aktivitäten des Gehirns, des Herz-Kreislauf-Systems, der Haut, der Muskeln und der Augen messen lassen, wird in den folgenden Abschnitten näher beschrieben. Wir beginnen jeweils mit einer kurzen Skizze der physiologischen Indikatoren, stellen die Messverfahren vor und behandeln danach die interessierenden psychologischen Korrelate. Für vertiefte anatomische Grundlagen und weitere Messverfahren sei auf die biopsychologische und neuroanatomische Fachliteratur verwiesen (Birbaumer & Schmidt, 2006; Cacioppo, Tassinary, & Berntson, 2007; Köhler, 2001; Schandry, 2003). In ◘ Tab. 10.26 sind die in den folgenden Abschnitten behandelten physiologischen Datenerhebungsmethoden nach Organsystemen im Überblick dargestellt.

Bei dieser Auswahl bleiben eine **Reihe von Biosignalen unberücksichtigt**. Dazu zählen z. B. die Körpertemperatur, die Pupillenweite, der Speichelfluss oder Merkmale der Stimme (zur Vertiefung s. Birbaumer & Schmidt, 2006), die in der Forschung eher eine untergeordnete Rolle spielen.

Weiterhin werden **biochemische Indikatoren** („biochemical indicators“), die Laboranalysen erfordern, hier nicht betrachtet (z. B. Hormonwerte in Urinproben), obwohl sie in der Forschung eine wichtige Rolle spielen.

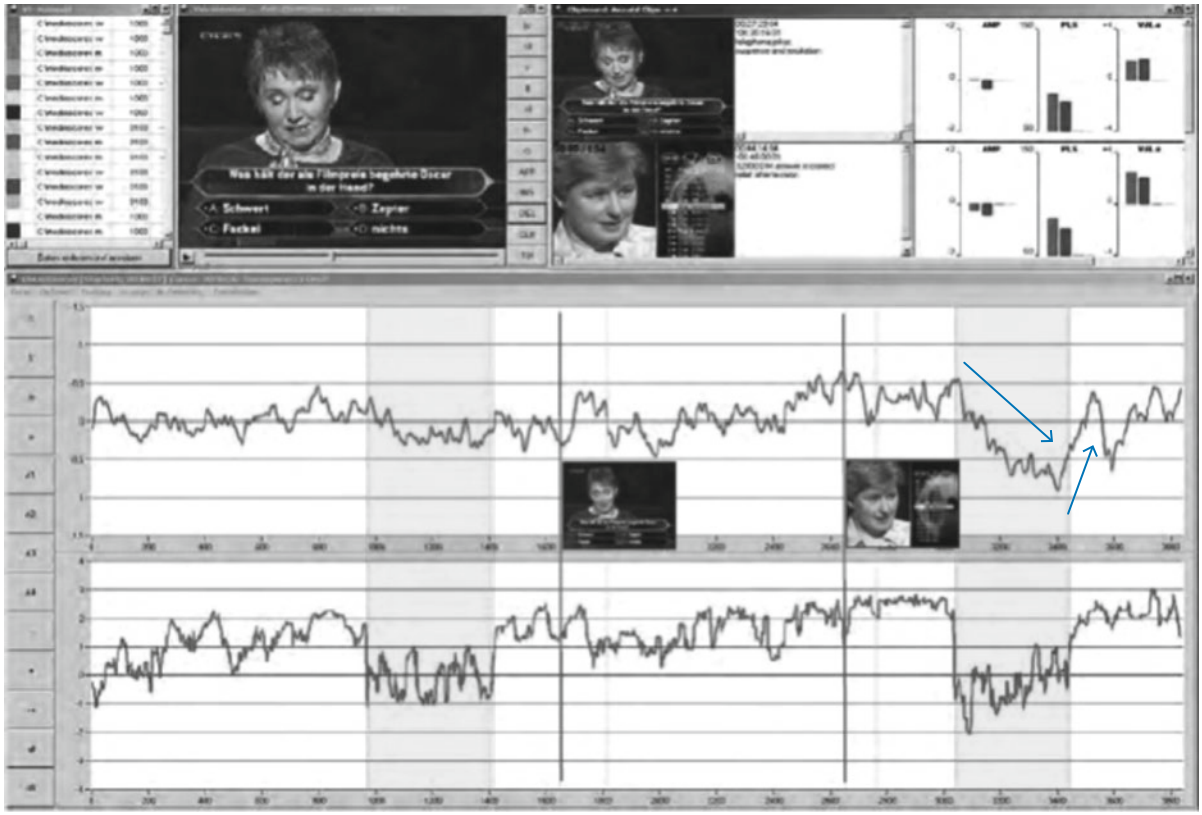


Abb. 10.33 Screenshot der Benutzungsoberfläche des Programms TVLab Pro zur Verlaufsanalyse von Fernsehrezeptionsdaten. (Aus Mangold, R., Bente, G., & Vorderer, P. (Hrsg.) (2004). Lehrbuch der Medienpsychologie. Göttingen: Hogrefe, S. 287. © Hogrefe Verlag GmbH & Co. KG, Göttingen. Mit freundlicher Genehmigung)

Ebenso liegen **invasive Methoden**, die nur von medizinischem Personal durchgeführt werden dürfen und ebenfalls oft Laboranalysen erfordern, außerhalb des Fokus dieses Kapitels (z. B. Hormonwerte in Blutproben; zu biochemischen Indikatoren s. Birbaumer & Schmidt, 2006; Carlson, 2004; Schandry, 2003).

10.5.2 Hirnaktivität

Das Gehirn ist für psychische Prozesse von enormer Bedeutung. Denn es spielt unter allen Organen die Hauptrolle für Erleben und Verhalten, von der Wahrnehmung über Denk- und Urteilsprozesse, Emotion und Motivation bis zur Handlungsplanung und -kontrolle. Im Folgenden werden die Messmethoden Elektroenzephalografie (EEG) und funktionelle Magnetresonanztomografie (fMRT) vorgestellt. Auf die eingangs erwähnte MEG (Magnetoenzephalografie) als Alternative zum EEG und die NIRS (Nahinfrarotspektrografie) als Alternative zur fMRT wird hier nicht näher eingegangen. Ebenso liegen neurowissenschaftliche Untersuchungstechniken wie die transkranielle Magnetstimulation (TMS), mit deren Hilfe über Magnetfelder bestimmte Hirnbereiche zu Forschungszwecken temporär stimuliert oder gehemmt werden können, jenseits des Fokus dieses Kapitels.

Tabelle 10.26 Ausgewählte physiologische Messverfahren im Überblick

Physiologischer Indikator	Messung
Hirnaktivität	– Elektroenzephalografie (EEG) – funktionelle Magnetresonanztomografie (fMRT)
Herz-Kreislauf-Aktivität	– Elektrokardiografie (EKG) – Pulsfrequenzmessung – Blutdruckmessung
Elektrodermale Aktivität	– Hautleitfähigkeitsmessung (EDA)
Muskuläre Aktivität	– Elektromyografie (EMG)
Augenaktivität	– Elektrookulografie (EOG) – Eye-Tracking

Elektroenzephalografie

Die am weitesten verbreitete Methode zur Erfassung der elektrischen Hirnaktivität ist die Elektroenzephalografie (EEG), deren Grundprinzipien von dem Neurologen Hans Berger (1929) an der Universität Jena entwickelt wurden.

Elektroenzephalografie – Die Elektroenzephalografie („electroencephalography", EEG) ist eine Methode zur Messung der elektrischen Hirnaktivität der obersten Schicht der Hirnrinde mittels auf der Kopfhaut angebrachter Elektroden. Sie dient der Messung des Spontan-EEG im wachen oder schlafenden Zustand sowie der Messung von gezielt durch Stimuli, motorische Reaktionen, kognitive und emotionale Prozesse hervorgerufener Hirnaktivität (ereigniskorrelierte Potenziale, EKP; „event related potentials" ERP). Als elektrophysiologisches Verfahren, das direkt die Aktionspotenziale der Nervenzellen der Großhirnrinde misst, hat das EEG eine sehr hohe Zeitauflösung, allerdings eine geringe räumliche Auflösung.

Enzephalogramm – Das Enzephalogramm („electroencephalogram", EEG) ist das Ergebnis einer Elektroenzephalografie und besteht in einer grafischen Darstellung der Potenzialschwankungen (in Mikrovolt: µV) im Zeitverlauf. Als „das EEG" wird üblicherweise nicht nur das Enzephalogramm, sondern auch die Methode der Enzephalografie (eigentlich „die EEG") bezeichnet.

Physiologische Grundlagen des EEG

Die Grundlage der Aktivität des Gehirns sind elektrochemische Prozesse und die daraus resultierenden Potenziale an den Membranen der Nervenzellen (Neurone). Dabei ist zu beachten, dass mit dem EEG lediglich die Aktivität von Neuronen der obersten Schichten der Hirnrinde erfasst werden, und zwar die **postsynaptischen Potenziale**.

Messverfahren beim EEG

Für den wissenschaftlichen Einsatz eines EEG ist eine entsprechende technische Laborinfrastruktur, möglichst mit abgeschirmten Räumen notwendig. Neben dem Registrierungsgerät sind Elektroden, Verbindungskabel und Paste zur Befestigung der Elektroden nötig. Zudem muss die Platzierung der Elektroden anhand eines etablierten Systems erfolgen. Auch die Auswertung der registrierten Daten ist sehr aufwendig und erfordert ein nicht unbeträchtliches Fachwissen.

Zur Ableitung des EEG befestigt man mit einer Klebesubstanz (Kollodium) oder mit selbstklebender Paste Elektroden auf der Schädeloberfläche. Die Platzierung der Elektroden erfolgt in der Regel in Anlehnung an ein internationales Platzierungssystem, das sog. 10/20-System nach Jasper (Schandry, 1996, S. 231 f.), teilweise werden auch mehr Kanäle benutzt. Oft verwendet man Hauben, die die Elektroden an die Schädeloberfläche anpressen. Die abgeleiteten Potenziale sind sehr schwach. Ihre Amplituden liegen im Bereich von 1 bis 200 Mikrovolt (µV), was eine sehr genaue Messtechnik erforderlich macht, um diese Signale artefaktfrei registrieren zu können.

Bei der EEG-Registrierung kommen bipolare Ableitungen (Vergleich zweier aktiver Elektroden) und unipolare Ableitungen (Vergleich einer aktiven mit einer neutralen Referenzelektrode) zum Einsatz. Als Position für die neutrale Referenzelektrode wird bei unipolaren Ableitungen entweder ein Ohrläppchen oder der Knochenvorsprung hinter dem Ohr (Mastoid) gewählt. Bei bipolaren Ableitungen wird eine aktive Elektrode als Referenz verwendet, z. B. die Cz-Elektrode des sog. Vertexpunktes im 10/20-Systems. Meistens platziert man gleichzeitig mehrere Elektroden auf der Schädeloberfläche, deren Spannungsdifferenzen paarweise registriert werden. Alternativ können auch die Potenzialdifferenzen zwischen einer aktiven Elektrode und der mittleren Aktivität aller übrigen bestimmt werden.

Eine bedeutende Artefaktquelle bei der Registrierung elektrophysiologischer Aktivität des Zentralnervensystems sind – neben äußeren Störungen – Lidschlag („blinks") und Augenbewegungen („eye movements") der Probanden, die im EEG-Signal als Potenzialspitzen erscheinen und die deshalb bei Spontan-EEG und ereigniskorrelierten Potenzialen routinemäßig mitregistriert werden (Elektrookulografie, EOG; Schandry, 1996, S. 274ff.). Signalabschnitte mit gleichzeitigem Lidschlag oder Augenbewegungen werden aus der Frequenzanalyse (EEG) oder dem Mittelungsverfahren (bei durch Reizung von Sinnesorganen evozierten = hervorgerufenen Potenzialen, EP) herausgenommen bzw. mittels statistisch-mathematischer Verfahren korrigiert.

Im Ergebnis läuft die Elektroenzephalografie auf eine grafische Darstellung der Potenzialschwankungen über die Zeit hinweg hinaus. Diese Grafik wird **Elektroenzephalogramm** genannt und ebenso wie die Erhebungsmethode mit EEG abgekürzt. An einem EEG interessieren die Schwankungen elektrischer Potenziale (Spannungen) über die Zeit. Dabei spielen die Frequenz (Schwingungshäufigkeit pro Sekunde, gemessen in Hertz, Hz) und die Amplitude (Ausmaß der Differenz, gemessen in µV) der Schwankungen eine Rolle. Da beim EEG die elektrische Aktivität direkt gemessen wird, hat es eine hohe zeitliche Auflösung im Bereich weniger Millisekunden.

Im Wesentlichen lassen sich zwei Arten der Aktivität unterscheiden, das Spontan-EEG (Spontanaktivität) und die evozierte (hervorgerufene) Aktivität bzw. ereigniskorrelierte Potenziale.

Spontan-EEG

Die Spontanaktivität gibt in erster Linie Auskunft über den Zustand von Gehirnfunktionen, die einen Zusammenhang mit Aufmerksamkeit, Bewusstheit, Aktivierung, Wachheit und Schlafstadien aufweisen. Der Hauptfrequenzbereich der EEG-Wellen im Spontan-EEG, das die

Tabelle 10.27 Relevante Frequenzbänder im Elektroenzephalogramm (EEG)

Frequenzband	Frequenzbereich	Amplitudenbereich	Aktivierungszustand
	Hz	µV	
Delta	0.1–4	20–200	Tiefschlaf
Theta	4–8	5–100	Einschlafzustand, Zustand tiefer Entspannung
Alpha	8–13	5–100	Entspannter Wachzustand
Beta	13–30	2–20	Mentale Aktivierung
Gamma	30–100	2–10	Starke Konzentration

kontinuierlich ablaufenden Potenzialschwankungen erfasst, reicht von 0.5–30 Hertz. Gemessen werden die Amplituden oder die relativen Anteile verschiedener Frequenzbereiche der hirnelektrischen Aktivität. Hierzu benötigt man digitale Registriereinheiten, die über numerische Verfahren (Fourier-Analyse) das Rohsignal in seine Bestandteile zerlegen. Bei dieser Technik wird jedes periodische Signal in Sinus- und Kosinus-Schwingungen zerlegt, deren Frequenz und Amplitude registriert werden. Bei modernen Geräten ist die Frequenzanalyse „online" möglich: Das EEG-Signal wird gleich bei seiner Registrierung in vorher festgelegten Zeitintervallen oder kontinuierlich in seine Frequenzbestandteile zerlegt.

Die Fourier-Analyse ermittelt die im EEG-Signal charakteristischen Frequenzbereiche, die man mit den griechischen Buchstaben Delta, Theta, Alpha, Beta und Gamma bezeichnet, sowie deren Amplituden. Wie Tab. 10.27 zu entnehmen ist, dominieren bei unterschiedlichen Aktivierungszuständen des menschlichen Organismus jeweils typische Frequenzbänder und Amplitudenbereiche.

Psychologische Korrelate des Spontan-EGG

Das Spontan-EEG eignet sich insbesondere für die Abbildung von allgemeiner Aktivierung oder Wachheit. Die Schlafforschung ist deshalb ein wichtiges Anwendungsgebiet der Elektroenzephalografie (Birbaumer & Schmidt, 2006, S. 535f.). Die einzelnen **Schlafstadien** sind durch Dominanz unterschiedlicher Frequenzbänder definiert (Tab. 10.27). Eine Sonderstellung nimmt die sog. **REM-Schlafphase** ein („rapid eye movement", REM). In dieser Phase herrschen niederamplitudige Theta-Wellen vor, verbunden mit schnellen Augenbewegungen und kurzen phasischen Muskelaktivitäten bei gleichzeitig niedrigem bis fehlendem Hintergrundtonus der Skelettmuskulatur (Atonie). Die REM-Phasen scheinen mit Traumphasen einherzugehen.

Des Weiteren wird das EEG zur Erforschung von kognitiven Leistungen eingesetzt. Bezüglich der **Intelligenz** bestehen Ansätze, das EEG der rechten und der linken Hirnhemisphäre abzunehmen und die Beziehung der Hirnhälften mit intellektueller Leistung zu untersuchen (Alexander, O'Boyle, & Benbow, 1996; Gasser, Lucadou-Muller, Verleger, & Bacher, 1983; Giannitrapani, 1969). In Bezug auf **Erinnerungsleistungen** werden Alpha-Wellen mit einer Belastung der Merkfähigkeit (De Pascalis & Ray, 1998) und der Schnelligkeit des Informationsabrufs aus dem Gedächtnis in Verbindung gebracht (Klimesch, 1997). Dem Tief- und REM-Schlaf wird eine hohe Bedeutung für die Gedächtniskonsolidierung, also die Überführung gelernter Inhalte vom Kurz- und Arbeitsgedächtnis in das Langzeitgedächtnis, zugeschrieben (Diekelmann & Born, 2010). Außerdem kann die gesamte EEG-Aktivität Hinweise auf die Leistung des räumlichen Arbeitsgedächtnisses geben und zwischen hoch und niedrig leistungsfähigen Individuen unterscheiden (Gevins & Smith, 2000). Nicht zuletzt können EEG-Daten auch in der **Emotionsforschung** genutzt werden (Davidson, Schwartz, Saron, Bennett, & Goleman, 1979; Janke, Schmidt-Daffy, & Debus, 2008).

Ereigniskorrelierte Potenziale im EEG

Beim Spontan-EEG handelt es sich um einen Indikator **tonischer** („tonic") Aktivität des Zentralnervensystems ZNS, der über Niveauveränderungen Auskunft gibt. Im Unterschied hierzu liefern **phasische** („phasic") elektrophysiologische ZNS-Maße Informationen über kurzzeitige Veränderungen, die mit einem externen oder internen Reiz einhergehen. Diese Maße werden wie bereits erwähnt als **ereigniskorrelierte Potenziale (EKP)** bezeichnet (zur Einführung s. Luck, 2005).

EKP können zum einen mit kognitiven Prozessen einhergehen (z. B. Konzentration, Sprachverarbeitung), zum anderen mit der Reizung von Sinnesorganen (z. B.

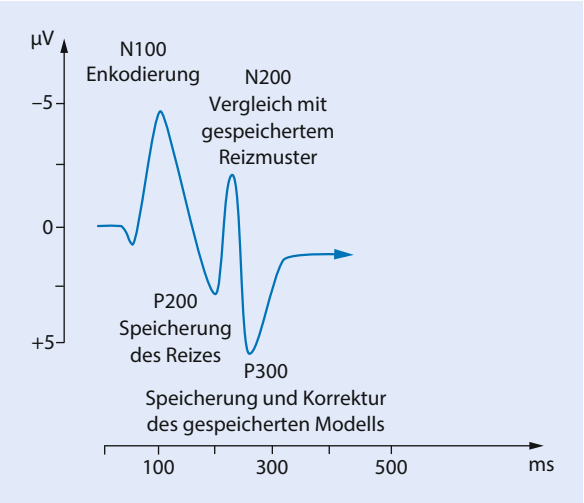


Abb. 10.34 Schematische Darstellung des Verlaufs ereigniskorrelierter Potenziale bei der Aufnahme und Verarbeitung eines visuellen oder auditiven Reizes als Ausschnitt aus einem Enzephalogramm. P positive Amplitude, N negative Amplitude. (Mod. nach Birbaumer & Schmidt, 2006, S. 481)

visuelle oder akustische Stimulation), letztere werden **evozierte Potenziale** („evoked potential", **EP**) genannt. Die am häufigsten untersuchte Gruppe phasischer ZNS-Maße sind sensorisch evozierte Potenziale, die phasische EEG-Veränderungen nach einem sensorischen Reiz beschreiben. Die Amplituden der EP sind um das 5- bis 20-fache kleiner als die Spontan-EEG-Amplituden (Schandry, 1996, S. 241) und deshalb mit dem bloßen Auge meist nicht zu erkennen. Um diese Potenziale sichtbar zu machen, wendet man die **Summations- oder Mittelungstechnik** an, bei der die Reaktionen auf viele aufeinander folgende Darbietungen des interessierenden Reizes zusammengefasst werden. Die bei dieser Technik resultierende Potenzialverlaufskurve stellt eine reliable Abbildung der mittleren evozierten Antwort auf den externen Stimulus dar, da sich die ereignisunabhängigen Anteile des EEG herausmitteln. Eine große Rolle spielen auch Zeit-Frequenz- bzw. **Wavelet-Analysen**, die zeitliche Veränderungen in der Frequenzzusammensetzung hirnelektrischer Aktivität bestimmen.

Ausgewertet werden markante Potenzialkomponenten, die durch ihre Amplituden (P für positiv, N für negativ) sowie ihre mittleren Latenzen (etwa 100 für 100 Millisekunden [ms]) benannt sind. So ist N100 eine nach durchschnittlich 100 ms messbare Verringerung der Spannung um wenige Mikrovolt (µV), sie wird auch als N1 bezeichnet. Demgegenüber ist P200 eine nach durchschnittlich 200 ms auftretende Erhöhung der Spannung (Abb. 10.34). Die Latenzen sind jedoch variabel, so dass z. B. P300 Latenzen zwischen 250 und 700 ms aufweisen kann (Luck, 2005). Einige Potenzialkomponenten werden auch durch Namen bezeichnet (z. B.„error related negativity", ERN; „late positive potential", LPP).

Psychologische Korrelate von EKP im EEG

Die Untersuchung ereigniskorrelierter und insbesondere evozierter Potenziale im EEG nimmt in der psychophysiologischen Forschung sehr großen Raum ein. Die APA-Literaturdatenbank PsycINFO weist allein für die Jahre 2000–2012 rund 13 000 referierte Zeitschriftenartikel mit „evoked potentials" als Keyword aus.

Die Anwendungsfelder sind vielfältig, einige Beispiele seien genannt: Bezüglich **Sinneswahrnehmungen** werden evozierte Potenziale für visuelle und akustische Wahrnehmung, Geruchs- und Geschmackssinn, Schmerz und Körperbeschleunigung abgeleitet (z. B. de Tommasoa et al., 2008; Pilling, 2009). Ereigniskorrelierte Potenziale können auch genutzt werden, um kognitive Prozesse wie z. B. die **Zeitwahrnehmung** zu erforschen (Macar & Vidal, 2004). Des Weiteren spielen ERP bei der Untersuchung von **Aufmerksamkeitsprozessen** wie selektiver Aufmerksamkeit und Ressourcenallokation eine Rolle. Hierbei wird v. a. die Detektion von unregelmäßig und zufällig auftretenden Stimuli untersucht (für einen Überblick Andreassi, 2007). Im Zusammenhang mit **motorischen Reaktionen** wird u. a. erforscht, ob kürzere Reaktionszeiten mit höheren EP-Amplituden oder kürzeren EP-Latenzen zusammenhängen oder welche EP-Unterschiede sich beim Reaktionsverhalten von Profi-Sportlern versus Kontrollpersonen zeigen (z. B. Hung, Spalding, Santa Maria, & Hatfield, 2004).

Bestimmte ereigniskorrelierte Potenzialkomponenten (z. B. P300 und LPP) werden mit der Verarbeitung emotionaler Stimuli und der **Emotionsregulation** in Verbindung gebracht (Hajcak, MacNamara, & Olvet, 2010). Die Ableitung ereigniskorrelierter Potenziale wird im Rahmen der klinischen Neurophysiologie auch genutzt, um **psychische Erkrankungen** besser zu verstehen, etwa Autismus (Jeste & Nelson, 2009) oder die Aufmerksamkeitsdefizit-Hyperaktivitätsstörung ADHS (Barry, Johnstone, & Clarke, 2003). Die **Wirkung von Nikotin** auf die Informationsverarbeitung lässt sich ebenfalls mittels ereigniskorrelierten Potenzialen im EEG erfassen (Pritchard, Sokhadze, & Houlihan, 2004).

Funktionelle Magnetresonanztomografie

Die **Magnetresonanztomografie** (MRT; „magnetic resonance imaging", MRI) – im alltäglichen Sprachgebrauch auch Kernspintomografie oder kurz **Kernspin** genannt – gehört zu den **bildgebenden Verfahren**, die Aufnahmen aus dem Körperinneren bereitstellen. Sie dient der Identifizierung struktureller Merkmale innerer Organe, inklusive des Gehirns. Für die Sozial- und Humanwis-

senschaften ist insbesondere die funktionelle Variante (fMRT) von Bedeutung, da diese Hirnaktivität abbilden und – im Gegensatz zum räumlich schlecht auflösenden EEG – präzise spezifischen Hirnstrukturen zuordnen kann. „Funktionell" meint hier, dass die Hirnaktivität bei bestimmten Aufgaben oder Funktionen gemessen wird. Außerdem ermöglicht es die funktionelle Magnetresonanztomografie, das Gehirn bzgl. seiner Funktion und Struktur in seiner Gesamtheit darzustellen. So wird auch in subkortikale (unter der Großhirnrinde liegende) Bereiche des Gehirns vorgedrungen, die mittels EEG nicht untersucht werden können.

Funktionelle Magnetresonanztomografie (fMRT) – Die funktionelle Magnetresonanztomografie (fMRT; „functional magnetic resonance imaging", fMRI) ist ein bildgebendes Verfahren, das es erlaubt, im Zuge spezifischer Aufgaben oder Tätigkeiten aktivierte Hirnareale zu lokalisieren oder verschaltete Netzwerke zu identifizieren (z. B. sog. Default-Mode oder Resting-State-Netzwerke). Die untersuchte Person wird dabei in eine Untersuchungskammer (Kernspintomograf) geschoben. Dort werden starke Magnetfelder erzeugt, mit deren Hilfe die Anatomie (festes Gewebe vs. Flüssigkeit) sowie der Blutsauerstoffgehalt des Gehirns messbar sind. In stark durchbluteten Arealen ist das Blut sauerstoffreicher, was seine magnetischen Eigenschaften verändert. Steigt bei einem bestimmten Stimulus oder einer bestimmten Aufgabe im Vergleich zum Ausgangszustand der Sauerstoffgehalt im Blut in einer bestimmten Hirnregion statistisch überzufällig an (sog. BOLD-Kontrast, „blood oxygenation level dependence"), so wird das betreffende Hirnareal als aktiviert interpretiert. Aktivierte Areale werden in der fMRT-Aufnahme des Gehirns (**fMRT-Scan**, Hirn-Scan) farblich markiert. Als hämodynamisches Verfahren, das neuronale Aktivität indirekt über den erhöhten Stoffwechselumsatz der lokal aktiven Nervenzellen erfasst, hat die fMRT eine sehr hohe räumliche Auflösung, allerdings eine geringe Zeitauflösung.

Physiologische Grundlagen der fMRT

Hirnaktivität spiegelt sich direkt in der elektrischen Erregung der Nervenzellen wider, die – zumindest in den obersten Schichten der Hirnrinde – mit dem EEG messbar ist. Neuronale Aktivität führt zu einer Erhöhung der lokalen zerebralen Durchblutung: Es wird vermehrt arterielles Blut in die aktiven Hirnregionen geleitet. Dies führt wiederum zu einer Erhöhung der Sauerstoffkonzentration im venösen Blut, da der lokale Sauerstoffverbrauch in der Regel niedriger ist als die zusätzliche Zufuhr. Erhöhte Aktivität in einer bestimmten Gehirnregion („region of interest", ROI) geht also mit erhöhter Durchblutung bzw. erhöhtem Blutsauerstoffgehalt in dieser Region einher (**BOLD-Effekt**; „blood oxygenation level dependence", BOLD)

Messverfahren bei der fMRT

Bei Magnetresonanzverfahren werden starke Magnetfelder zur Bildgebung eingesetzt. Personen, die Metallimplantate im Körper tragen (Herzschrittmacher, künstliche Gelenke) können deswegen nicht untersucht werden. Bei der fMRT fMRT werden Querschnittsbilder des arbeitenden Gehirns erzeugt, so dass im Unterschied zum EEG, das nur die neuronale Aktivität an der Hirnoberfläche erfasst, bei der fMRT fMRT Aktivität im **gesamten Gehirn** messbar wird.

Die untersuchte Person wird meistens auf einer Liege in eine Untersuchungskammer geschoben, welche die Form einer kurzen Röhre hat. Die Methode basiert darauf, dass sich die magnetischen Eigenschaften des sauerstoffbindenden Hämoglobins im Blut von denen des sauerstoffarmen Blutes unterscheiden. Bei Anlegen eines starken pulsierenden Magnetfeldes kann die lokale Erhöhung der Sauerstoffkonzentration bzw. der BOLD-Effekt mittels Hochfrequenzempfängern sichtbar gemacht und dadurch **indirekt auf die Aktivität der umgebenden Neurone geschlossen** werden. Die zeitliche Auflösung ist bei der fMRT fMRT geringer als beim EEG, da der Blutfluss viel langsamer ist als neuronale Prozesse.

Die **räumliche** bzw. **anatomische Auflösung** ist bei der fMRT fMRT indessen sehr hoch, so dass auch wechselnde Aktivitäten anatomischer Strukturen während psychischer Prozesse im fMRT-Scan sichtbar gemacht werden können. Bei der Bilderstellung werden diejenigen Hirnareale farblich hervorgehoben, die in einem bestimmten Moment statistisch überzufällig stark aktiviert sind. Setzt man die Aktivierung zu zeitlich identifizierbaren Reizen in Beziehung, kann auf deren Verarbeitung geschlossen werden.

Wie alle physiologischen Messmethoden können auch fMRT-Aufnahmen durch Artefakte verfälscht werden. Eine wichtige **Artefaktquelle** stellen Bewegungen der untersuchten Person dar. Diese verzerren die Daten auch bei kleinen Bewegungen, wie z. B. beim Sprechen. Eine weitere Herausforderung bei der fMRT stellen die **Aufbereitung, Auswahl und statistische Auswertung der Rohdaten** dar. Ein hochauflösender Scanner produziert alle 2 Sekunden rund 500 000 Messwerte (**Voxel** = Datenpunkte in einer 3D-Grafik). Eine einzelne fMRT-Sitzung mit einer Person generiert nicht selten mehr als **1 Milliarde Messwerte**, von denen nur rund 5 % für die eigentliche Studie relevant sind (Vul & Kanwisher, 2010). Die Messwerte werden nun in vielfacher und unterschiedlicher Weise aufbereitet, relevante Hirnareale werden ausgewählt und die ausgewählten Daten statistisch analysiert. Die einzelnen Transformations- und Berechnungsschritte sind so umfang- und variantenreich, dass sie im Methodenteil von Publikationen meist gar nicht vollständig dargestellt werden (können), was die Möglichkeit einer kritischen Bewertung durch Reviewer sowie Leserschaft einschränkt. Zudem besteht in erhöhtem Maße die Gefahr eines Selektionsbias, der die Voraussetzung statistischer Unabhängigkeit verletzt sowie die Gefahr der Alpha-Fehler-Kumulation durch wiederholte Tests am

selben Datenmaterial. Derartige Effekt können dazu führen,

- dass Zufallsrauschen im Datensatz **fälschlich als signifikanter Befund** interpretiert wird (eine Studie konnte sogar signifikante Effekte mit einem toten Lachs nachweisen: Bennett, Abigail, Michael, & George, 2009 http://prefrontal.org/blog/2009/06/atlantic-salmon-index) oder
- dass Korrelationen zwischen fMRT-Ergebnissen und anderen Konstrukten stark überschätzt werden (sog. „Voodoo-Korrelationen"; Vul, Harris, Winkielman, & Pashler, 2009; http://www.edvul.com/voodoocorr.php).

Auch wenn fMRT-Untersuchungen in der letzten Zeit einen enormen Aufschwung erfahren haben, darf folgendes nicht aus den Augen verloren werden: Die Datenerhebung mittels einer fMRT ist enorm aufwändig und kostenintensiv. Bereits die Kosten für die Anschaffung einer fMRT-Umgebung bewegen sich in Millionenhöhe, Bedienung und Auswertung erfordern umfangreiches Fachwissen bzw. Fachpersonal. Häufig ist eine entsprechende Infrastruktur nur in Kliniken verfügbar, die diese auch zur medizinischen Diagnostik nutzen (es mehren sich allerdings die reinen Forschungsscanner). Auch die Betriebskosten sind sehr hoch, so dass jeder Versuchsdurchgang bzw. jede Versuchsperson schnell Kosten in Höhe von mehreren hundert Euro verursacht. Dies führt dazu, dass die meisten bislang durchgeführten fMRT-Studien sehr niedrige Probandenzahlen (oft im ein- oder niedrigen zweistelligen Bereich) aufweisen.

Psychologische Korrelate der fMRT

Ziel einer fMRT-Untersuchung ist es, hinsichtlich verschiedener psychologischer Prozesse die jeweils aktivierten Hirnareale abzubilden. In der Grundlagenforschung wird z. B. untersucht, welche Hirnregionen bei **Aufmerksamkeitsprozessen**, bei unterschiedlichen **Sinneswahrnehmungen**, **Gedächtnisprozessen**, **Sprache** und **Emotionen** eine wichtige Rolle spielen (für detaillierte Ausführungen s. Andreassi, 2007). Untersuchungen zur **Gehirnplastizität** („brain plasticity") weisen darauf hin, dass bestimmte Hirnregionen bei intensiver Nutzung stärker ausgeprägt sind (z. B. ist die Repräsentation der linken Hand bei Streichmusikern, die mit dieser Hand die Saiten greifen, im Vergleich zu ihrer rechten Hand sowie zur linken Hand von Nichtmusikern stärker ausgeprägt; Elbert & Heim, 2001). Zudem werden fMRT-Messungen auch bei komplexen Konstrukten wie z. B. in der **Aggressionsforschung** eingesetzt. So wurde z. B. die Hirnaktivität mittels fMRT live untersucht, während die Versuchspersonen im Kernspintomografen gewalthaltige Computerspiele spielten (Weber, Ritterfeld, & Mathiak, 2006). Dabei zeigte sich, dass im Zuge gewalthaltiger Handlungen im Computerspiel neuronale Muster auftraten, die typisch für aggressives Denken und Handeln sind. Dieser Befund erlaubt freilich nicht den direkten Rückschluss, dass gewalthaltige Computerspiele Aggressionsverhalten im Alltag verursachen. Eine Anwendung der funktionellen Magnetresonanztomografie in der Gesundheitsforschung zeigt das ▶ Studienbeispiel „fMRT: Sagt neuronale Aktivität beim Betrachten von Gesundheitsanzeigen die Reduktion des Rauchverhaltens vorher?".

Studienbeispiel

fMRT: Sagt neuronale Aktivität beim Betrachten von Gesundheitsanzeigen die Reduktion des Rauchverhaltens vorher?

Fragestellung

Eine gesundheitspsychologische Studie ging der Frage nach, ob die neuronale Aktivität beim Betrachten von Anzeigen einer Anti-Rauchkampagne das spätere Rauchverhalten vorhersagen kann (Falk, Berkman, Whalen, & Lieberman, 2011).

Methodisches Vorgehen

Eine Stichprobe von $n = 28$ Raucherinnen und Rauchern, die mit dem Rauchen aufhören wollten, nahmen an der Studie teil. Es wurde ein Messwiederholungsdesign mit einer Prä-Post-Messung angewendet. Als objektives Maß für das Rauchverhalten wurde der Kohlenmonoxidgehalt des Atems gemessen, der in Abhängigkeit von der Menge der in den vergangenen Stunden gerauchten Zigaretten ansteigt. Die Messung des Rauchverhaltens über den Kohlenmonoxidatemtest erfolgte zum ersten Messzeitpunkt (Baseline) und zum zweiten Messzeitpunkt (einen Monat später). Nach der Baselinemessung betrachteten die Studienteilnehmenden Anzeigen einer Anti-Rauchkampagne, während parallel mittels fMRT die Hirnaktivität im medialen präfrontalen Kortex (ein Areal, das mit Verhaltensänderungen in Verbindung gebracht wird) gemessen wurde.

Ergebnisse

Die Autoren stellten eine überzufällige Korrelation zwischen dem Ausmaß der Aktivierung des Hirnareals beim Betrachten der Anzeige (◘ Abb. 10.35) und der späteren Reduktion des Rauchens fest. Wie erfolgreich die Versuchspersonen bei der angestrebten Reduktion ihres Nikotinkonsums waren, ließ sich deutlich besser vorhersagen, wenn nicht nur ihre subjektive Intention mit dem Rauchen aufzuhören einbezogen, sondern auch der physiologische Indikator der Intensität der Hirnaktivität bei der Verarbeitung von Anti-Rauchbotschaften berücksichtigt wurde.

Schlussfolgerung

Objektive Maße der Hirnaktivität können eine nützliche Ergänzung zu Selbstauskünften darstellen, wenn es um die Vorhersage von Verhaltensänderungen im Gesundheitsbereich geht.

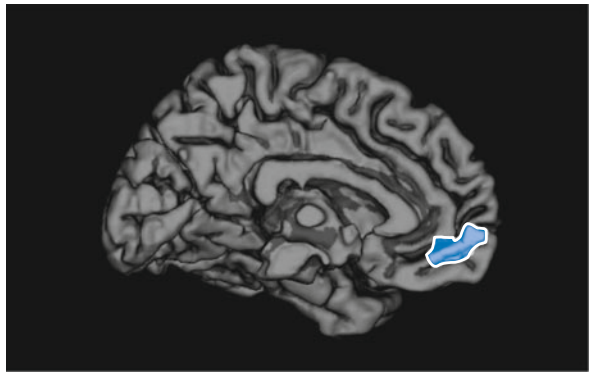


Abb. 10.35 Aktivierung des medialen präfrontalen Kortex (gilt als Indikator für Verhaltensänderungen) während des Betrachtens von Anzeigen einer Anti-Rauch-Kampagne dargestellt in einer fMRT-Aufnahme. (Aus Falk et al., 2011. Copyright © 2011 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA)

10.5.3 Herz-Kreislauf-Aktivität

Eine große Gruppe klassischer physiologischer Indikatoren sind die Maße des kardiovaskulären bzw. Herz-Kreislauf-Systems. Dass man in angespannten Situationen „Herzklopfen" bekommt, der „Puls hochgeht" oder der „Blutdruck steigt" – das sind geradezu sprichwörtliche physiologische Indikatoren psychischen Geschehens.

Physiologische Grundlagen des Herz-Kreislauf-Systems

Im Folgenden werden die physiologischen Grundlagen für die Biosignale beschrieben, die am häufigsten genutzt werden, um psychophysiologische Erkenntnisse zum Herz-Kreislauf-System zu erlangen.

▪ Herzschlagfrequenz

Das kardiovaskuläre oder Herz-Kreislauf-System besteht aus Funktionsorganen, die die ausreichende und adäquate Blutversorgung des Organismus sicherstellen (für eine ausführliche Einführung Birbaumer & Schmidt, 2006). Zentrales Organ ist das **Herz**. Das Herz pumpt das Blut mit seiner linken Hälfte im großen Körperkreislauf und mit seiner rechten Hälfte im kleinen Lungenkreislauf. Es zeigt dabei eine ausgeprägte Autorhythmie: Das Herz schlägt auch bei völliger Isolierung mit einer rhythmischen Schlagfolge weiter, da die Erregungsbildung und Erregungsweiterleitung im Herzen selbst erfolgt. Die Häufigkeit der Herzschläge pro Minute wird als **Herz(schlag)frequenz** („heart beat rate") gemessen (internationale Einheit: „beats per minute", **bpm**; deutsche Einheit: Herzschläge pro Minute, $\mathbf{min}^{-1}$)

Die autorhythmische Erregungsbildung wird normalerweise vom Sinusknoten generiert, einem Muskelgeflecht in der Nähe der oberen Hohlvene. Man spricht beim Sinusknoten auch von einem physiologischen Schrittmacher, dessen Eigenfrequenz ca. 70–80 Entladungen pro Minute beträgt (Silbernagl & Despopoulos, 1991, S. 165) Die **Ruhefrequenz** beim gesunden Herz liegt in der Regel bei ca. **60–70 Herzschlägen pro Minute**, also unterhalb der Sinusknotenautorhythmie. Dies ist darauf zurückzuführen, dass in Ruhe und bei schwacher Belastung über die tonische Aktivität des 10. Hirnnervs, des parasympathischen Vagusnervs, eine Senkung der Herzfrequenz erfolgt. Lediglich bei stärkerer Belastung nimmt der Einfluss des Sympathikus über die Ausschüttung von Adrenalin (aus dem Nebennierenmark) und Noradrenalin (aus den postganglionären sympathischen Nervenfasern) zu, wobei gleichzeitig die Vagusaktivität gedämpft wird. Es kommt zu einem Anstieg der Herzschlagfrequenz bis **maximal 150–180 Schläge pro Minute**. Die Herzschlagfrequenz unterliegt starken intra- und interindividuellen Unterschieden, z. B. in Abhängigkeit vom Lebensalter oder vom Trainingszustand; so weisen Ausdauersportler mit einem stark trainierten Herzmuskel eine deutlich geringere Ruhe-Herzschlagfrequenz auf.

Die Aktivität des Herzens lässt sich in vier Aktionsphasen unterteilen (Silbernagl & Despopoulos, 1991, S. 162f.): die **Systole** („systole") mit der Anspannungs- und Auswurfphase sowie die **Diastole** („diastole") mit der Entspannungs- und Füllungsphase. In der Anspannungsphase steigt der Druck in den Herzkammern aufgrund der Muskelkontraktion bei geschlossenen Herzklappen an. Wenn der Herzinnendruck den Gegendruck der Körper- bzw. Lungenarterie übersteigt, kommt es zur Auswurfphase. Nach der Austreibung des Blutes in die Arterien entspannen sich die Herzkammern (Entspannungsphase) und die Arterienklappen schließen sich wieder. Dies leitet die Füllungsphase ein.

▪ Pulsfrequenz

Im Unterschied zur Herzfrequenz, die als die Anzahl der Herzschläge pro Minute definiert ist, versteht man unter der Pulsfrequenz („pulse rate") die **Anzahl der Impulse pro Minute**, die durch die Herzaktivität verursacht werden und die sich im peripheren Gefäßsystem fortsetzen. Herzfrequenz und Pulsfrequenz stimmen weitgehend überein, allerdings kann es unter bestimmten Umständen zu Abweichungen kommen (sog. Pulsdefizit). Zudem ist zu beachten, dass neben der Pulsfrequenz als quantitativem Maß des Pulses auch qualitative Pulsmerkmale messbar und interpretierbar sind (z. B. Regelmäßigkeit oder Härte des Pulses).

▪ Blutdruck

Eine weitere wichtige Bestimmungsgröße kardiovaskulärer Aktivität ist der Blutdruck („blood pressure"). Hierunter versteht man den Druck des Blutes in den Gefäßen,

der je nach Gefäß und Stelle im Körper variiert. Üblicherweise ist der Druck in den größeren Arterien, insbesondere den Schlagadern auf Herzhöhe gemeint (z. B. Oberarmarterie). Der Blutdruck wird normalerweise in der Druckeinheit mmHg gemessen, was der Anhebung einer normierten Quecksilbersäule in Millimetern entspricht.

Die Regulation des Blutdrucks lässt sich als komplexes Regelkreismodell darstellen. Als direkte Einflussgrößen bestimmen das Herzzeitvolumen (Auswurfvolumen des Herzens pro Zeiteinheit = Herzfrequenz × Schlagvolumen) und der periphere Gesamtwiderstand der Blutgefäße den Blutdruck. Eine Aktivierung des Sympathikus bewirkt über sog. α-Rezeptoren (Wirkungsdominanz des Noradrenalins) eine Verengung der Blutgefäße und über β_1-Rezeptoren (Noradrenalin und Adrenalin) eine Erhöhung der Herzfrequenz sowie indirekt über die Erhöhung der Kontraktionskraft des Herzens auch eine Steigerung des Schlagvolumens. Beides löst jeweils eine Blutdruckerhöhung aus. Die Senkung des Blutdrucks wird durch eine Steigerung der Aktivität des Parasympathikus (Senkung der Herzfrequenz) und ein Aussetzen der Sympathikusaktivität (Erweiterung der peripheren Blutgefäße) hervorgerufen.

Während der Systole wird arterielles Blut aus den Herzkammern herausgeschleudert. Dies ist der Zeitpunkt des höchsten Blutdrucks in den Arterien, des **systolischen Blutdrucks**. Während der Diastole strömt das venöse Blut in die Herzvorhöfe zurück. Dabei sinkt der Blutdruck auf seinen niedrigsten Wert, den **diastolischen Blutdruck**.

Der Blutdruck ist Schwankungen unterworfen. In der Oberarmarterie pendelt der Blutdruck unter Ruhebedingungen zwischen 120–140 mmHg (systolischer Blutdruck) und 80–100 mmHg (diastolischer Blutdruck). Je weiter der Blutdruckmesspunkt vom Herz entfernt ist, desto niedriger wird der Blutdruck und desto geringer sind dessen Schwankungen. In den großen Venen ist keine Blutdruckveränderung mehr messbar. Der Blutdruck liegt dort nur noch bei 1–2 mmHg.

Messverfahren des Herz-Kreislauf-Systems

Die Herzschlagfrequenz wird meist mittels EKG, die Pulsfrequenz mittels Pulsmessgerät und der Blutdruck mittels Manschettenverfahren gemessen.

Elektrokardiografie

Das bekannteste und am weitesten verbreitete Verfahren zur kontinuierlichen Messung der Herzaktivität ist die Elektrokardiografie **(EKG)**. Mit dieser Methode wird der zeitliche Verlauf der summierten Aktionspotenziale der Herzmuskelfasern, deren elektrisches Feld sich durch das leitende Gewebe bis zur Körperoberfläche fortpflanzt, über zwei aktive Elektroden aufgezeichnet. Eine dritte Elektrode dient als Erdung. Es gibt verschiedene Möglichkeiten zur Platzierung der EKG-Elektroden. Neben der Brustwandableitung (je eine Elektrode an den beiden Brustbeinpolen) haben sich auch die drei bipolaren Extremitätenableitungen nach dem Arzt Willem Einthoven (Schandry, 1996, S. 134 f.) etabliert.

Elektrokardiografie (EKG) – Die Elektrokardiografie (EKG; „electrocardiography“, ECG) ist ein physiologisches Messverfahren, das u. a. zur Erfassung der Herzschlagfrequenz dient. Mittels Elektroden auf dem Oberkörper werden die elektrischen Spannungen abgeleitet, die durch die Herzmuskulatur erzeugt werden. Das Ergebnis der Elekrokardiografie ist das **Elektrokardiogramm** (EKG) als grafische Darstellung der Potenzialschwankungen (in Mikrovolt µV) über die Zeit hinweg. Als „das EKG“ wird üblicherweise nicht nur das Elektrokardiogramm, sondern auch die Methode der Elektrokardiografie (eigentlich „die EKG“) bezeichnet.

Bei der sog. II. Ableitung nach Einthoven werden die aktiven Elektroden am Unterarm und am unteren Bein der gegenüberliegenden Seite befestigt, wobei die Erdung über eine Elektrode am zweiten Bein erfolgt. Der Vorteil der Einthoven-Ableitungen liegt darin, dass die Elektroden ohne das Ablegen von Kleidung angebracht werden können. Die Brustwandableitung ist dagegen sehr robust gegenüber Körperbewegungen und eignet sich daher besonders zur EKG-Registrierung während körperlicher Aktivität.

Das EKG-Signal bildet die elektrischen Erregungsprozesse am Herzen kontinuierlich ab und wird grafisch in einem Elektrokardiogramm dargestellt. Es kann in verschiedene Komponenten aufgeteilt werden, wobei die markanteste Komponente die **R-Zacke** ist (Birbaumer & Schmidt, 2006, S. 186ff.). Die Anzahl der R-Zacken bezogen auf ein 1-Minuten-Intervall ergibt die Herzrate der **Herzfrequenz** (ca. 60–70 Schläge pro Minute in Ruhe). Die Veränderung der Herzrate über einen längeren Zeitraum wird als Maß der **tonischen** kardiovaskulären Aktivität herangezogen. Unter **phasischen** Herzratenänderungen versteht man kurzfristige Erhöhungen bzw. Senkungen der Herzfrequenz in Abhängigkeit von Reizen. Zur Ermittlung phasischer Herzratenänderungen verwendet man meist die RR-Abstände (◘ Abb. 10.36).

Als **RR-Abstand** („interbeat intervall“) bezeichnet man das Zeitintervall zwischen zwei R-Zacken im EKG. Unterschiedlich große RR-Abstände sprechen für eine hohe **Herzfrequenzvariabilität** („heart rate variability“, HRV). Im Zuge der Digitalisierung physiologischer Registriertechniken wurden auch Spektralanalysen verschiedener EKG-Indikatoren eingeführt, speziell Spektralanalysen der RR-Abstände. Hierbei werden die Indikatoren mit unterschiedlichen mathematischen Algorithmen nach Zeitverlaufsgesichtspunkten sortiert und in Klassen (Spektren) eingeteilt. Die Spektralanalyse informiert über

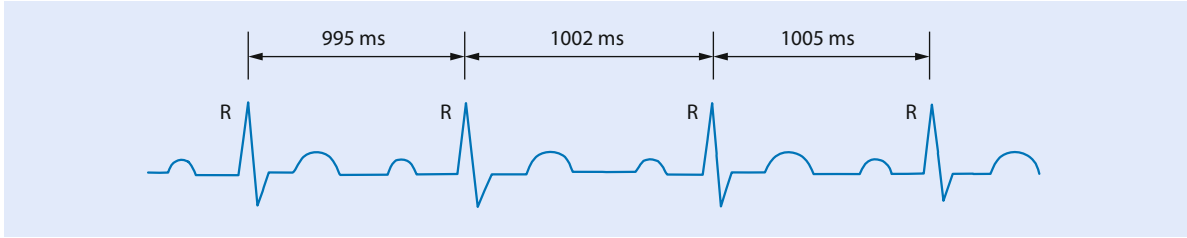


Abb. 10.36 Schematische Darstellung der Herzfrequenzaufzeichnung mit R-Zacken und RR-Abständen als Ausschnitt aus einem Elektrokardiogramm. (Aus RRscope, 2012. Mit freundlicher Genehmigung von Richard Rudnicki)

die Anteile (oder Power) der einzelnen Spektren (ausführlicher hierzu Rösler, 1996, S. 501ff.).

Eine bedeutende Einflussgröße stellt die Atmung bzw. die **Atemfrequenz** („respiratory rate") dar. Sie führt zur sog. **respiratorischen Arhythmie** des Herzschlags, d. h., beim Einatmen erhöht sich die Herzfrequenz, beim Ausatmen sinkt sie ab. Die Atemfrequenz stellt somit eine mögliche Artefaktquelle v. a. bei der Registrierung phasischer Herzratenänderungen dar. Sie sollte daher miterhoben und zur statistischen Bereinigung der EKG-Daten genutzt werden. Die Registrierung erfolgt üblicherweise über einen **Atemgürtel**, der mechanische Signale (die Dehnung des Brustkorbes) in elektrische Signale umwandelt. Alternativ können Gasanalysatoren eingesetzt werden. Dabei werden Atemtiefe, Atemfrequenz und der Gasaustausch über eine Luftentnahme mittels kleinen Plastikröhrchen in einem oder beiden Nasenlöchern registriert (für detailliertere Informationen zur Physiologie der Atmung Birbaumer & Schmidt, 2006). Ferner empfiehlt sich im Zuge der Elektrokardiografie eine Kontrolle des Blutdrucks, der indirekt über den Barorezeptorenreflex die Herzfrequenz ebenfalls beeinflusst.

Eine nahezu EKG-genaue drahtlose Messung der Herzschlagrate ist auch mit einem Brustgurt möglich, in den die Elektroden integriert sind, und der die Daten an eine Pulsuhr sendet. Entsprechende Herzratenmessgeräte werden unter der Bezeichnung **Pulsuhr mit Brustgurt** vermarktet und sind z. B. im Leistungs- und Freizeitsport verbreitet.

Pulsfrequenzmessung

Der arterielle Puls lässt sich z. B. an der Halsschlagader oder am Handgelenk gut tasten, so dass die Pulsfrequenz pro Minute mithilfe einer Uhr einfach ausgezählt werden kann. Neben dieser einfachen Form der Pulsfrequenzmessung („pulse rate measurement") ist eine kontinuierliche Messung mittels **Pulsoxymetrie** bzw. Photoplethysmographie („pulse oximetry") möglich. Dabei wird die Sauerstoffsättigung des pulsierenden Blutes am Finger, Zeh oder Ohrläppchen mit einem Clip-Sensor gemessen. Dieser sendet Infrarotlicht durch das Gewebe und misst auf der anderen Seite die mit dem Puls schwankende Lichtabsorption. Je stärker die Durchblutung und die Sauerstoffsättigung, umso stärker ist die Lichtabsorption durch das Hämoglobin (rote Blutkörperchen).

Pulsfrequenzmessung – Bei der Pulsfrequenzmessung („pulse rate measurement") wird die Pulsrate entweder nicht-apparativ an der Halsschlagader oder am Handgelenk getastet und mit der Uhr ausgezählt oder apparativ durch einen Clip-Sensor am Finger, Ohrläppchen oder Zeh mittels Pulsoxymetrie kontinuierlich gemessen und grafisch in einer **Pulskurve** dargestellt.

Blutdruckmessung

Eine ganz exakte kontinuierliche **Blutdruckmessung** („blood pressure measurement") ist nur **invasiv über eine Kanüle** möglich, die in eine Arterie eingeschoben wird. An die Kanüle wird ein Manometer angeschlossen, das den Arterieninnendruck anzeigt (direkte Druckmessung). Die Belastung der Untersuchungspersonen ist dabei vergleichsweise groß, zudem ist medizinisches Personal für diese invasive Methode notwendig.

Blutdruckmessung – Bei der Blutdruckmessung („blood pressure measurement") wird der Blutdruck entweder direkt – d. h. invasiv mit Kanüle und Manometer – oder indirekt mit einem nicht-invasiven Manschettendruckverfahren erfasst. Blutdruckwerte können grafisch als **Blutdruckkurve** dargestellt werden, indem sie kontinuierlich oder in zeitlichen Abständen abgetragen werden.

Weit verbreitet zur Blutdruckmessung ist deswegen das **nicht-invasive Manschettendruckverfahren** nach dem Mediziner Scipione Riva-Rocci aus dem Jahr 1896, das auch Sphygmomanometrie genannt wird, bzw. die automatische Blutdruckmessung, die ebenfalls nach dem Riva-Rocci-Prinzip arbeitet (indirekte Druckmessung). Bei diesen Geräten wird eine Staumanschette am linken Oberarm angebracht und auf Knopfdruck bis zu einem voreingestellten Wert aufgepumpt. Anschließend wird die Luft wieder langsam abgelassen. Ein Mikrophon in der Manschette registriert dabei das An- und Abschwellen der Geräusche, die durch die abnehmende Stauung des Blutes in der Oberarmarterie entstehen (sog. **Korotkow-Ge-**

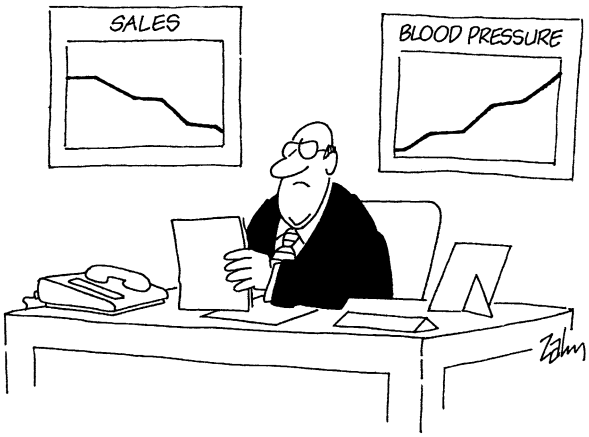


Abb. 10.37 Die Höhe des Blutdrucks korreliert mit dem subjektiven Stresserleben. (© Carroll Zahn/Search ID: cza0598, Rights Available from CartoonStock.com)

räusche). Der systolische und diastolische Blutdruckwert wird dann am Gerät angezeigt. Für kurzfristig wiederholte Blutdruckerfassungen ist das Riva-Rocci-Verfahren nicht geeignet, da jeder Messprozess eine Deformation der Arterie verursacht, die erst nach einer gewissen Zeitspanne zurückgeht.

Psychologische Korrelate von Indikatoren des Herz-Kreislauf-Systems

Ein wichtiges Forschungsfeld ist die Untersuchung der Zusammenhänge von steigender Herzrate und **motorischer Leistung** („motor performance"): Mehrere Studien belegen, dass motorische Aufgaben mit einem Anstieg der Herzrate einhergehen – und zwar über das Maß hinaus, das man von einer einfachen Steigerung des Kreislaufs durch die körperliche Aktivität erwarten würde (für einen Überblick Andreassi, 2007). Zudem werden Veränderungen in der Herzrate mit **Lern- und Problemlösungsprozessen** in Verbindung gebracht (Andreassi & Whalen, 1967; Goldstein, Harman, McGee, & Karasik, 1975). Bezüglich **Aufmerksamkeitsprozessen** wird ein Absinken der Herzrate mit der Wahrnehmung von Stimuli und der Orientierungsreaktion auf Reize verknüpft, während eine erhöhte Herzrate die Stimulusabwehr und eine abwehrende Reaktion begleitet (eine Einführung gibt Andreassi, 2007). In Studien, die **Motivation** untersuchen, führen höhere in Aussicht gestellte Belohnungen zu einer Aktivierung des Herz-Kreislauf-Systems (Elliott, 1974; Fowles, Fischer, & Tranel, 1982). Studien, die eine Veränderung der Aktivität des Herz-Kreislauf-Systems mit hoher zeitlicher Auflösung mit motorischen, kognitiven, emotionalen oder motivationalen Prozessen in Verbindung setzen, greifen meist auf die Herzrate als Indikator und auf das EKG als Messverfahren zurück.

Eine Erhöhung des Blutdrucks ist nachgewiesen durch so unterschiedliche Faktoren wie die **Akkulturation an die westliche Gesellschaft** bei Migranten (Steffen, Smith, Larson, & Butler, 2006), die Konfrontation mit **Rassismus** (Brondolo, Rieppi, Kelly, & Gerin, 2003), den **Koffein-Konsum** (James, 2004) oder auch **chronischen Ärger** (Schum, Jorgensen, Verhaeghen, Sauro, & Thibodeau, 2003; Abb. 10.37). In derartigen Studien ist oft eine weniger feine zeitliche Auflösung notwendig, so dass eine Messwiederholung zu verschiedenen Zeitpunkten mit dem nicht-invasiven Manschettendruckverfahren durchführbar ist. Eine weitere Anwendung zeigt das ▶ Studienbeispiel „Blutdruckmessung: Lässt Emotionsausdruck den Blutdruck steigen oder sinken?".

In der Herzfrequenzvariabilität (HRV) kommt zum Ausdruck, wie flexibel sich die Herzschlagrate den aktuellen Erfordernissen anpassen kann (Malik, Bigger, Camm, & Kleiger, 1996). Eine eingeschränkte HRV tritt u. a. bei **mentaler Beanspruchung** (Nickel & Nachreiner, 2003), **Stress und Sorgen** (Brosschot, van Dijk, & Thayer, 2007) und auch bei **posttraumatischer Belastungsstörung** auf (Hagit, et al., 1998). Mit Biofeedback-Methoden kann die HRV gezielt beeinflusst werden, indem die Probanden lernen, nicht in einem Zustand der Anspannung und konstant erhöhten Herzrate zu verharren, sondern die Herzrate durch Synchronisation mit der Atmung bei Bedarf bewusst zu senken (sog. **Herzkohärenztraining**). Studien zeigen, dass **Yoga**, **Meditation** und bestimmte Atemtechniken zur Senkung der Herzfrequenz – gemessen über die Pulsfrequenz – beitragen (Mohan & Wei, 2002; Sudsuang, Chentanez, & Veluvan, 1991; Telles & Vani, 2002). Diese Beispiele verdeutlichen, dass physiologische Messungen auch im Rahmen von Evaluationsstudien (z. B. Evaluation des Gesundheitsnutzens von Sportkursen) sinnvoll einsetzbar sind.

10.5.4 Elektrodermale Aktivität

Die dritte große Gruppe klassischer physiologischer Indikatoren sind nach der Aktivität des Gehirns und des Herz-Kreislauf-Systems die Maße der elektrodermalen Aktivität, die Leitfähigkeits- oder Potenzialveränderungen der Haut registrieren. Schon seit Ende des 19. Jahrhunderts diskutierte man Zusammenhänge zwischen der elektrodermalen Aktivität und psychischen, insbesondere emotionalen Prozessen (Féré, 1888).

Physiologische Grundlagen der EDA

Unter elektrodermaler Aktivität (**EDA;** „electrodermal activity") werden die Leitfähigkeits- und Potenzialänderungen der Haut verstanden. Diese können im Zusammenhang mit psychischen Prozessen auftreten. Von

Studienbeispiel

Blutdruckmessung: Lässt Emotionsausdruck den Blutdruck steigen oder sinken?

Fragestellung

Der Ausdruck von Emotionen mittels Verhaltensweisen ist häufig, aber nicht immer mit einem Absinken des Blutdrucks verbunden. Eine kulturvergleichende Studie ging der Frage nach, ob der kulturelle Hintergrund von Individuen die Beziehung zwischen emotionaler Expressivität und damit einhergehenden physiologischen Reaktionen (Höhe des Blutdrucks) moderiert (Butler, Lee, & Gross, 2009).

Methodisches Vorgehen

Eine Stichprobe von $n = 32$ Frauen, die entweder einen asiatisch-amerikanischen oder einen europäisch-amerikanischen Hintergrund hatten, nahmen an der Studie teil. Dabei wurden Dyaden mit jeweils 2 Frauen mit demselben kulturellen Hintergrund gebildet. Diese sahen beide gemeinsam eine dreiminütige Naturdokumentation, während der die Baselinemessung des Blutdrucks erfolgte. Danach sahen sie gemeinsam eine elfminütige Kriegsdokumentation. Anschließend diskutierten sie miteinander ihre Gedanken und Gefühle, die Implikationen des Films für die menschliche Natur und dessen Relevanz für ihre religiösen und politischen Überzeugungen. Während der beiden Filme und während der anschließenden Diskussion wurde fortlaufend der Blutdruck gemessen und die Mimik der Probandinnen aufgezeichnet. Die Videos wurden im Nachhinein hinsichtlich emotionaler Gesichtsausdrücke ausgewertet. Zum Abschluss wurden die Versuchsteilnehmerinnen schriftlich zu ihren emotionalen Reaktionen während der Kriegsdokumentation befragt.

Ergebnisse

Bei den europäisch-amerikanischen Frauen ging verstärkter Ausdruck von negativen Emotionen mit einem Absinken des Blutdrucks einher. Bei den asiatisch-amerikanischen Frauen war es genau umgekehrt, je stärker sie ihre negativen Emotionen anlässlich des Kriegsfilms gegenüber der anderen Zuschauerin zum Ausdruck brachten, desto höher war ihr Blutdruck.

Schlussfolgerung

Die Ergebnisse weisen darauf hin, dass die Beziehung zwischen emotionaler Expressivität und begleitenden physiologischen Reaktionen durch den kulturellen Hintergrund moderiert wird. So gilt in asiatischen Kulturen der offene Emotionsausdruck gegenüber fremden Personen als unangemessen. Möglicherweise stieg also bei den asiatisch-amerikanischen Frauen bei offenem Emotionsausdruck der Blutdruck an, weil sie angespannt und über die soziale Wirkung ihres Verhaltens auf das Gegenüber besorgt waren. Umgekehrt ist im westlichen Kulturkreis offener Emotionsausdruck normal und erwünscht, so dass sich die europäisch-amerikanischen Versuchspersonen physiologisch entspannten (Blutdrucksenkung), während sie sich über ihre emotionalen Reaktionen auf die Kriegsdokumentation austauschten.

besonderer Bedeutung für die EDA sind die **Schweißdrüsen**. Diese bestehen aus einem sekretorischen Teil (in dem die eigentliche Schweißproduktion stattfindet) und dem Ausführungsgang. Dieser Gang transportiert den Schweiß zur Hautoberfläche. Die Höhe der Hautleitfähigkeit ist abhängig vom Füllungszustand der Ausführungsgänge mit Schweiß. Je höher dieser gestiegen ist, desto höher ist auch die elektrische Leitfähigkeit der Haut (bzw. desto geringer ist der Hautwiderstand). Die Verteilung der Schweißdrüsen auf der Körperoberfläche ist am dichtesten an der Hand- und der Fußinnenfläche (über 2 000/cm^2), am Rumpf und den Extremitäten wesentlich dünner (100–200/cm^2; Birbaumer & Schmidt, 2006). Neben der Schweißdrüsenaktivität hängt die elektrodermale Aktivität vermutlich auch von der Aktivität einer **elektrisch geladenen Membran** in der Epidermis (Oberhaut) und in den Schweißdrüsengängen ab, die am Entstehen von Hautpotenzialen und an Veränderungen in der Hautleitfähigkeit mitbeteiligt sind. Die Aktivität der Schweißdrüsen wird sympathisch angeregt.

Messverfahren der EDA

Am häufigsten wird in empirischen Untersuchungen die **Hautleitfähigkeit** (bzw. deren Kehrwert: der **Hautwiderstand**) bestimmt. Andere Maße wie das Hautpotenzial (Boucsein, 1988) oder die Hautfeuchte (Köhler, 1992) spielen heute nur noch eine untergeordnete Rolle. Die Hautleitfähigkeit ist eine exosomatische Größe; sie kann daher nur unter Zufuhr von äußerer Energie, meist einer Stromspannung von 0.5 Volt (Edelberg, 1967, 1972), erhoben werden. Die Ableitung der Hautleitfähigkeit ist eine Methode, die sich für den praktischen Einstieg in das Gebiet der physiologischen Messungen gut eignet: Die Anschaffungskosten für die Messgeräte halten sich in Grenzen, so dass sie an Hochschule oft zur Verfügung stehen.

Messung der Hautleitfähigkeit – Die Messung der Hautleitfähigkeit („skin conductance measurement“) bzw. ihres Kehrwertes, des Hautwiderstandes, dient der Erfassung der elektrodermalen Aktivität („electrodermal activity“, EDA) des Organismus. Die Ableitung erfolgt an der Handinnenfläche mittels zweier Elektroden. Als Ergebnis der Messung werden die Veränderungen der Hautleitfähigkeit (in Mikrosiemens µS) über die Zeit hinweg in einer **Hautleitfähigkeitskurve** grafisch dargestellt.

Zur Ableitung der Hautleitfähigkeit werden zwei Elektroden an der Innenseite der beiden mittleren Glieder von Zeige- und Mittelfinger (Venables & Christie, 1980) oder am Daumen- und Kleinfingerballen (Walschburger, 1975)

befestigt. Üblicherweise wird von der Handinnenfläche der nicht-dominanten Hand abgeleitet. Die Maßeinheit der Hautleitfähigkeit ist µSiemens (früher: µMho), wobei die Maßeinheit Siemens für den elektrischen Leitwert genau der Kehrwert des elektrischen Widerstandes (Einheit: Ohm) ist.

Es gibt tonische Hautleitfähigkeitsmaße, die Aussagen über das Niveau der elektrodermalen Aktivität gestatten (das sog. **Hautleitfähigkeitsniveau** oder „skin conductance level", SCL) sowie die Anzahl spontaner Fluktuationen. Neben diesen tonischen Maßen sind auch phasische Maße von Bedeutung, die Reaktionen auf externe Stimuli kennzeichnen (**Hautleitfähigkeitsreaktionen,** „skin conductance response", SCR). Zur Charakterisierung von Hautleitfähigkeitsreaktionen sind die Amplitude und verschiedene Zeitmaße (z. B. Latenzzeit = Zeitdifferenz zwischen Reiz- und Reaktionsbeginn) gebräuchlich.

Bei der Identifikation von SCR in der Hautleitfähigkeitskurve besteht die Gefahr, dass sie mit spontanen Fluktuationen verwechselt werden. Um dieses Problem zu umgehen, werden nur solche Hautleitfähigkeitsreaktionen ausgewertet, deren Amplitudenmaximum in einen festgelegten Zeitbereich nach Beginn des externen Stimulus fällt (z. B. 1.5–6.5 Sekunden; Schandry, 1996, S. 202f.). Liegt das Amplitudenmaximum außerhalb des Zeitfensters, wird der Hautleitfähigkeitsreaktionsamplitude der Wert Null zugewiesen, da die Hautleitfähigkeitsänderung dann mit großer Wahrscheinlichkeit nicht auf den externen Stimulus zurückzuführen ist.

Psychologische Korrelate der EDA

Generell gilt, dass eine Zunahme von Aktiviertheit mit einem Anwachsen der Leitfähigkeit sowie einer Zunahme der Frequenz spontaner Fluktuationen verbunden ist. So weisen Studien darauf hin, dass eine erhöhte Hautleitfähigkeit mit kürzeren **Reaktionszeiten** einhergeht (Andreassi, 1966a; Freeman, 1940). Ebenso stehen **Lernaufgaben** (Yuille & Hare, 1980) und **Orientierungsreaktionen** (Gati, Ben-Shakar, & Avni-Liberty, 1996) mit einer erhöhten elektrodermalen Aktivität in Zusammenhang.

Ein bedeutsames Anwendungsfeld für tonische elektrodermale Maße ist die **Emotionsforschung**. Vor allem die spontane Fluktuation der Hautleitfähigkeit gilt als sensitiver und spezifischer Indikator für emotionale Zustände (Boucsein, 1991) bzw. als **Angstindikator** (zu Grundlagen und Anwendungsfeldern s. Boucsein, 1992; für eine ausführliche Diskussion experimenteller Befunde Boucsein, 1995; Erdmann & Voigt, 1995; Fowles, 1980).

Phasische Indikatoren der elektrodermalen Aktivität sind in der Aufmerksamkeits- und der **Habituationsforschung** von Bedeutung, die sich mit dem Vorgang der Gewöhnung an sich wiederholende Reize befasst. Die Amplituden von physiologischen Reaktionsmaßen werden bei Wiederholung des auslösenden Reizes immer geringer, sie klingen zunächst ab und verschwinden dann meist ganz. Eine Zusammenstellung umfangreicher Ergebnisse zur Habituation elektrodermaler Reaktionen im Zusammenhang mit der sog. **Orientierungsreaktion** liefert Siddle (1983). In der klinischen Psychologie wurden ebenfalls Habituationsuntersuchungen mit Indikatoren des elektrodermalen Systems durchgeführt (Baltissen & Heimann, 1995; ▶ Studienbeispiel „Messung der Hautleitfähigkeit: Wirkt die häufige Nutzung medialer Gewaltdarstellungen abstumpfend?"). Ein bedeutendes Forschungsfeld ist nicht zuletzt die **klassische Konditionierung von elektrodermalen Reaktionen** (electrodermal conditioning; z. B. Lachnit, Kinder, & Reinhard, 2002).

10.5.5 Muskuläre Aktivität

Die vierte große Gruppe physiologischer Indikatoren nach Gehirn, Herz-Kreislauf-System und Haut sind die Maße der muskulären Aktivität.

Physiologische Grundlagen der muskulären Aktivität

Die Aktivität der Skelettmuskulatur wird über das motorische System gesteuert. Das motorische System besteht aus folgenden zentralnervösen Strukturen: Motorische Großhirnrinde, Teile des Thalamus, Kleinhirn, Basalganglien sowie zahlreiche motorische Kerne in Hirnstamm und Rückenmark. Das zentralnervöse motorische System mündet in den Fortsätzen der motorischen Nervenfasern (Axone) und den Skelettmuskelfasern. Als Überträgerstoff (Transmitter) zwischen den Synapsen der motorischen Nervenfasern und den Skelettmuskelfasern wirkt Acetylcholin. Jede motorische Nervenfaser ist mit mehreren **Muskelfasern** verknüpft, die eine **motorische Einheit** bilden. Motorische Einheiten können aus einigen wenigen (z. B. Augenmuskulatur), aber auch aus bis über 1 000 Muskelfasern (z. B. Rückenmuskulatur) bestehen.

Eng verschaltet mit dem motorischen System sind sensorische Einheiten bzw. auch komplexere zentralnervöse Strukturen, so dass psychische Prozesse einen starken Einfluss auf die Steuerung der Skelettmuskulatur ausüben. Im Alltagsverständnis kommt dieser Zusammenhang z. B. in der Redewendung „vor Angst zittern" zum Ausdruck, wobei das Zittern auf Muskelaktivität zurückgeht. Für die biopsychologische Forschung ist die Registrierung elektrischer Muskelaktivität deswegen von großer Bedeutung.

Studienbeispiel

Messung der Hautleitfähigkeit: Wirkt die häufige Nutzung medialer Gewaltdarstellungen abstumpfend?

Fragestellung

Eine medienpsychologische Studie ging der Frage nach, ob und wie die regelmäßige Nutzung gewalthaltiger Medienangebote mit aggressive Kognitionen und aggressivem Verhalten zusammenhängt (Krahé et al., 2011).

Methodisches Vorgehen

Eine Stichprobe von $n = 303$ Studierenden wurden schriftlich zu habitueller Exposition gegenüber gewalthaltigen Mediendarstellungen, Trait-Aggression und Trait-Erregbarkeit sowie normativen Einstellungen gegenüber Aggression befragt. Zwei Wochen später nahmen die Untersuchungspersonen an einem Laborexperiment teil. Alle Teilnehmerinnen und Teilnehmer sahen einen gewalthaltigen Videoclip und jeweils entweder einen lustigen oder einen traurigen Vergleichsfilm. Die Variablen gewalthaltiger Filmclip (ein Clip aus „Casino“ von Martin Scorsese, ein Clip aus „Reservoir Dogs“ von Quentin Tarantino), Vergleichsfilm (lustig vs. traurig) und die Reihenfolge der Filme wurden experimentell manipuliert. Während der Filmrezeption wurde fortlaufend die Hautleitfähigkeit gemessen. Zudem wurden nach jedem Clip schriftliche Ratings zu ängstlicher und angenehmer Erregung durchgeführt. Abschließend bearbeiteten die Versuchspersonen eine lexikalische Entscheidungsaufgabe, um die Zugänglichkeit aggressiver Kognitionen und entsprechende Reaktionszeiten zu messen. Dazu mussten die Teilnehmenden 160 Zeichenketten (40 aggressive Wörter, 40 nichtaggressive Wörter und 80 bedeutungslose Zeichenketten) dahingehend beurteilen, ob diese sinnvolle deutsche Wörter darstellten. Wenn dabei aggressionsbezogene Wörter vergleichsweise schneller erkannt werden, deutet dies auf die stärkere Verfügbarkeit aggressiver Kognitionen hin. Abschließend wurde aggressives Verhalten mit dem Noise-Blast-Paradigm (eine standardmäßige Reaktionszeitaufgabe in der Mediengewaltforschung, bei der ein imaginierter Gegenspieler bei eigenem Gewinn mit einem lauten Geräusch bestraft werden kann) gemessen.

Ergebnisse

Habituelle Exposition gegenüber gewalthaltigen Medieninhalten hing negativ mit der Hautleitfähigkeit während gewalthaltiger Videoclips zusammen, aber positiv mit angenehmer Erregung. Wurden in Pfadanalysen Trait-Aggression, normative Einstellungen und Trait-Erregbarkeit kontrolliert, ergab sich folgender Befund: Regelmäßige Nutzung gewalthaltiger Mediendarstellungen sagte eine schnellere Verfügbarkeit von aggressiven Kognitionen hervor, die teilweise durch eine höhere angenehme Erregung mediiert wurde. Provoziertes aggressives Verhalten in der Reaktionszeitaufgabe wurde aber weder durch habituelle Exposition, noch durch ängstliche oder angenehme Erregung vorhergesagt. Die Hautleitfähigkeit zeigte ebenfalls keine Zusammenhänge mit aggressiven Kognitionen oder Verhaltensweisen.

Schlussfolgerung

Die Studie stützt die Annahme, dass Desensibilierung gegenüber gewalthaltigen Mediendarstellungen zu geringeren physiologischen Reaktionen gegenüber solchen Darstellungen und einer schnelleren Verfügbarkeit von aggressiven Kognitionen führt. Es kann laut Studie von einer Habituierung gegenüber gewalthaltigen Medienstimuli gesprochen werden, auch wenn diese sich nicht zwingend direkt in aggressivem Verhalten niederschlägt.

Messverfahren der muskulären Aktivität

Zur Messung der elektrischen Muskelaktivität wird die **Elektromyografie** (**EMG**) verwendet. Diese Methode registriert die Depolarisationswellen von Muskelaktionspotenzialen, die sich entlang der Zellmembran der Muskelfaser fortpflanzen. Die Elektromyografie kann **invasiv mit Nadelelektroden** durchgeführt werden, was eine Messung der Aktivität einzelner Muskelfasern erlaubt, oder **nicht-invasiv mit Oberflächenelektroden**. Die nichtinvasive Methode ist in der sozialwissenschaftlichen Forschung gebräuchlicher, da kein medizinisches Personal notwendig ist und die Untersuchungspersonen weniger belastet werden. Mit Oberflächenelektroden kann jedoch nur die Aktivität von ganzen Muskeln – manchmal auch nur von mehreren Muskeln zusammen (nicht von einzelnen Fasern) – erfasst werden. Generell werden bei der Elektromyografie zwei bipolare Elektroden benötigt, die im Abstand von ca. 5 cm voneinander in Längsrichtung des Muskels am Muskelbauch (d. h. an der dicksten Stelle des Muskels) anzubringen sind.

Änderungen in der Muskelaktivität sind auf eine erhöhte Entladungsrate der motorischen Einheit zurückzuführen oder auf eine erhöhte Anzahl aktiver motorischer Einheiten, was sich im EMG in höheren Signalamplituden und in höheren Frequenzanteilen ausdrückt. Beim EMG ist die korrekte Platzierung der Elektroden wichtig. Die Auswertung kann eine Herausforderung darstellen, da in der Regel das Zusammenspiel mehrerer Muskelgruppen zu betrachten ist.

Elektromyografie – Die Elektromyografie („electromyography“, EMG) ist ein Messverfahren zur Erhebung der elektrischen Muskelaktivität. Sie kann invasiv mit Nadelelektroden oder nicht-invasiv mittels Oberflächenelektroden durchgeführt werden, die die Aktivationspotenziale der Muskelzellen ableiten. Als Ergebnis der Elektromyografie werden die Potenzialschwankungen an den Muskelzellen (in Mikrovolt, μV) über die Zeit hinweg grafisch in einem **Elektromyogramm** (EMG) dargestellt. Als „das EMG“ wird üblicherweise nicht nur das Elektromyogramm, sondern auch die Methode der Elektromyografie (eigentlich „die EMG“) bezeichnet.

Zur Quantifizierung der gesamten elektrischen Muskelaktivität werden die EMG-Signale zunächst gleichgerichtet, indem man die negativen Potenzialanteile in positive umrechnet. Es folgt eine mathematische Integration des Signals, bei der die Fläche zwischen Nulllinie und dem gleichgerichteten Potenzialverlauf berechnet wird. Neben dem EMG gibt es noch weitere Maße zur Erfassung der

Muskelaktivität, wie z. B. die Bestimmung des Tremors und der Muskelvibration (Fahrenberg, 1983, S. 38f.; Fahrenberg, Leonhard, & Foerster, 2002, Kap. 6.7.2).

Psychologische Korrelate der muskulären Aktivität

Emotionen lassen sich neben der subjektiven Ebene (erlebte Gefühle) und der physiologisch-chemischen Ebene (z. B. Hormonausschüttung) auch auf der Verhaltensebene, nämlich dem mimischen Ausdruck betrachten. Die Erfassung peripherer Muskelaktivität – v. a. im Bereich des Gesichts und des Nackens – ist in der biopsychologischen Emotionsforschung etabliert. Bereits Charles Darwin (1872) erklärte die sichtbaren mimischen Veränderungen mit der Aktivität bestimmter **Gesichtsmuskeln**. Er postulierte, dass sich Emotionen u. a. in der Mimik offenbaren und dass der Ausdruck und die Erkennung von Emotionen einen evolutionären Vorteil darstellen.

Gemäß der **evolutionspsychologischen Emotionstheorie** des Psychologen Paul Ekman (1973) gibt es eine Gruppe von Basisemotionen („basic emotions"), die wichtige Überlebensfunktionen erfüllen. Zunächst postulierte er sechs Basisemotionen (Freude, Ärger, Ekel, Angst, Trauer, Überraschung), später erweiterte er die Liste und elaborierte das theoretische Konzept der „Basisemotion" (Ekman, 1999). Eine Möglichkeit, die Bedeutung von mimischem Emotionsausdruck zu untersuchen, besteht darin, den Untersuchungspersonen entsprechende Fotos vorzulegen und sie zu fragen, was die abgebildete Person wohl gerade empfindet, wobei sich kulturübergreifende Übereinstimmungen finden. Zur Erfassung der Mimik wird auf **Beobachtungsverfahren** zurückgegriffen (z. B. auf das standardisierte Beobachtungssystem FACS, „facial action coding system", von Ekman & Friesen, 1978), mit deren Hilfe einzelne Mimikelemente differenziert protokolliert werden (z. B. Heben der Augenbraue außen, Herabziehen der Mundwinkel, Rümpfen der Nase, Blinzeln etc.). Eine alternative oder ergänzende Methode zur standardisierten Beobachtung der Mimik stellt die physiologische Messung der Gesichtsmuskelaktivität per **Elektromyografie** dar. Wie anhand der Mimik bzw. an kleinsten unwillkürlichen Mimikänderungen (sog. „micro expressions") erkennbar ist, ob eine Person die Wahrheit sagt oder lügt, ist ein aktuelles Forschungsfeld mit vielfältigen Anwendungsmöglichkeiten (Ekman, 2009; Warren, Schertler, & Bull, 2009). Ähnlich wie sich der Polygraf als „Lügendetektor" in der Fernsehunterhaltung großer Beliebtheit erfreut (▶ Abschn. 10.5.1 „Definition und Indikation der physiologischen Messung"), ist auch die **Lügendetektion** anhand von beobachteter Gesichtsmuskelaktivität Gegenstand populärer Unterhaltungsmedien. So greift die US-TV-Serie „Lie to me" Forschungsergebnisse von Paul Ekman (2009) explizit auf (http://www.paulekman.com).

Neben seiner interpersonalen Bedeutung hat der mimische Emotionsausdruck gemäß der **Facial-Feedback-Theorie** auch eine intrapsychische Relevanz: Durch die Aktivität der Gesichtsmuskeln bei unterschiedlicher Mimik können Emotionen verstärkt oder gar ausgelöst werden (Zajonc, Murphy, & Inglehart, 1989), d. h. durch bewusstes Lächeln kann die Stimmung gehoben, durch das Unterdrücken von Schmerz- oder Ekelreaktionen kann das entsprechende Empfinden abgeschwächt werden. Mittels Elektromyografie können nun die Muskelaktivitäten im Gesicht präzise erfasst werden, etwa um Rückschlüsse auf Emotionen zu ziehen. In der **Medienrezeptionsforschung** spielt die Messung der Gesichtsmuskelaktivität eine große Rolle, weil auf Basis dieser fortlaufend registrierten physiologischen Messungen die Emotionen während des Anschauens eines Films oder des Spielens eines digitalen Games detailliert mitverfolgt werden können (▶ Studienbeispiel „EMG: Emotionale Reaktionen auf Tod und Töten in einem Computerspiel"). Experimentelle Studien konnten auch zeigen, dass sich die emotionale Reaktion auf Nachrichtentexte durch begleitende Bilder beeinflussen lässt – sowohl im Hinblick auf subjektives Erleben als auch objektive Gesichtsmuskelaktivität (Kätsyri, Ravaja, & Salminen, 2012).

Vor allem die Aktivität von verschiedenen **spezifischen Gesichtsmuskeln** (z. B. Corrugator supercilii: Zusammenziehen der Augenbrauen; Zygomaticus major: Hochziehen der Mundwinkel) werden als objektive Indikatoren des emotionalen Erlebens vermehrt herangezogen (z. B. Schmidt-Atzert, 1993, 1995). In der klinischen Psychologie wird die Aktivität spezifischer Gesichtsmuskeln z. B. zur Differenzialdiagnose bei **affektiven Störungen** eingesetzt (z. B. Greden, Genero, Price, Feinberg, & Levine, 1986). Die Verwendung des EMG als Indikator für allgemeine Erregung und Spannung ist dagegen eher rückläufig, wie auch die Methode des Muskelbiofeedback zur Behandlung allgemeiner psychischer Spannungszustände.

In der Verhaltensmedizin hat man chronische Spannungskopfschmerzen mit Verspannungen der **Nackenmuskulatur** in Zusammenhang gebracht, die auch elektromyografisch nachgewiesen werden konnten (Gerber, 1986). Auf diesen Befunden basieren **Biofeedbacktrainings** zur Entspannung hypertoner Muskelaktivität (Köhler, 1995, S. 47f.). Auch bei der Diagnose und Therapie von neurologischen Erkrankungen des zentralen motorischen Systems (z. B. Morbus Parkinson oder Dystonien wie die „Schiefhals"-Krankheit Torticollis spasticus) bzw. bei peripheren neuromuskulären Erkrankungen (z. B. Myasthenie) kommt der Erfassung der peripheren Muskelaktivität eine große Bedeutung zu (Poeck, 1990, S. 30ff.;

Studienbeispiel

EMG: Emotionale Reaktionen auf Tod und Töten in einem Computerspiel

Fragestellung

Eine medienpsychologische Studie untersuchte u. a. emotionale psychophysiologische Reaktionen auf unterschiedliche gewalthaltige Episoden im Ego-Shooter-Spiel „James Bond 007: NightFire“ (Ravaja, Turpeinen, Saari, Puttonen, & Keltikangas-Järvinen, 2008).

Methodisches Vorgehen

Eine Stichprobe von $n = 36$ jungen Erwachsenen (25 Männer und 11 Frauen) nahmen an einem Laborexperiment teil. Die Versuchspersonen spielten vier verschiedene Spielepisoden (Verwundung des Gegners, Tod des Gegners, Verwundung des eigenen Charakters, Tod des eigenen Charakters). Während des Spielens wurde ein EMG folgender Gesichtsmuskeln abgeleitet: linker Corrugator supercilii (negative Emotionen), Zygomaticus major (positive Emotionen) und Orbicularis oculi (positive Emotionen, die mit einer starken Erregung verbunden sind). Außerdem wurde die Hautleitfähigkeit aufgezeichnet. Zudem bewerteten die Versuchspersonen ihre emotionale Reaktion auf das Spiel und machten Angaben zu ihren Persönlichkeitseigenschaften (darunter Psychotizismus) auf der revidierten Kurzform des Eysenck-Persönlichkeitsinventars.

Ergebnisse

Verwunden und Töten eines Gegners im Videospiel führten zu einem Anstieg der Hautleitfähigkeit und einem Absinken der Aktivität im Zygomaticus major und dem Orbicularis oculi. Das Absinken der EMG-Aktivität war bei Personen mit hohen Werten auf der Psychotizismus-Skala stärker ausgeprägt. Das Verwunden und der Tod des eigenen Charakters (James Bond) führten zu einem Anstieg der Hautleitfähigkeit sowie der EMG-Aktivität im Zygomaticus major und dem Orbicularis oculi, aber zu einem Absinken der Aktivität im Corrugator supercilii.

Schlussfolgerungen

Anstatt Freude durch einen Sieg und einen damit verbundenen Spielerfolg zeigte die vorliegende Studie, dass das Verwunden und Töten eines Videospielgegners zu einem negativen und mit hoher Erregung verbundenen Affekt (Angst) führen kann. Dies könnte darauf hinweisen, dass die Spielerinnen und Spieler mit einer empathischen Stressreaktion auf das Spielgeschehen reagieren. Personen mit hoch ausgeprägtem Psychotizismus erleben möglicherweise weniger Angst als Personen, die geringere Ausprägungen dieser Persönlichkeitseigenschaft aufweisen. Überraschend erscheint der Befund, dass Verwundung und Tod des eigenen Spiel-Charakters laut Messung der Gesichtsmuskelaktivität mit positiven Emotionen einherging. In der Studie wird dieser Effekt so erklärt, dass mit dem Tod des Spiel-Charakters die Spielenden von den kognitiv und emotional belastenden Aufmerksamkeits- und Verhaltensanforderungen der Spielhandlung entlastet werden. Mögliche Enttäuschung oder Angst angesichts des drohenden oder tatsächlichen Todes des eigenen Spiel-Charakters war bei den Versuchspersonen offenbar nicht so deutlich ausgeprägt wie die Erleichterung durch den Spielabbruch, die ihnen elektromyografisch messbar sozusagen ins Gesicht geschrieben stand.

Schenck, 1992, S. 297ff.). Nicht zuletzt kann das EMG auch mit Bezug zur **Marktforschung** zum Einsatz kommen, etwa um das Kau- und Schluckverhalten beim Verzehr unterschiedlicher Schokoladensorten oder unterschiedlicher Typen von Schokoladenessenden zu vergleichen (vgl. Carvalho-da-Silva, Van Damme, Wolf, & Hort, 2011).

10.5.6 Augenaktivität

Die fünfte an dieser Stelle behandelte Klasse an physiologischen Messverfahren bezieht sich auf **Blickbewegungen** (Okulomotorik, „eye movement“). Blickbewegungen spielen naturgemäß bei Wahrnehmungsprozessen eine große Rolle und werden sowohl in der Grundlagenforschung (z. B. Informationsaufnahme, Erkennung und Unterscheidung von Objekten) als auch in der angewandten Forschung (z. B. Blickbewegungen in der Werbeforschung) untersucht.

Physiologische Grundlagen der Augenaktivität

Augenbewegungen werden durch Bereiche im Frontal- und Okzipitalkortex des Gehirns gesteuert. Die Innervierung erfolgt durch drei Hirnnerven (Kranialnerven). Diese Nerven führen zu den sechs extraokulären Muskeln pro Augapfel, die an der Außenseite befestigt sind und die Feinkontrolle von Blickbewegungen ermöglichen. Die Augen vollziehen drei Typen von Bewegungen:

1. **Vergenzbewegungen** („vergence“) sind kooperative Bewegungen, die beide Augen auf denselben Zielpunkt fixieren. Halten Sie einen Finger senkrecht vor Ihr Gesicht, schauen Sie ihn an und führen Sie ihn dann näher an Ihr Gesicht heran: Jetzt vollziehen Ihre Augen eine Vergenzbewegung in Richtung Ihrer Nase, die Netzhautbilder überlappen sich, ein räumliches Sehen des fixierten Punktes wird möglich.
2. **Sakkadische Bewegungen** („saccadic movement“) verschieben die Blickrichtung ruckartig von einem Punkt zum anderen. Diese treten z. B. beim Lesen auf: Wenn man eine Zeile in einem Buch liest, halten die Augen mehrere Male an, um sich dann zwischen diesen Punkten sehr schnell weiter zu bewegen. Diese Bewegungsgeschwindigkeit lässt sich nicht bewusst steuern.
3. **Folgebewegungen** der Augen („following eye movement“) lassen sich bewusst steuern. Folgen Sie mit Ihrem Blick z. B. einem fahrenden Fahrzeug auf einer Straße, können Sie bewusst steuern, dass sich Ihre Augen der Fahrtgeschwindigkeit anpassen (Carlson, 2004).

Messverfahren der Augenaktivität

Als wichtige Messverfahren für Augenbewegungen sind die Elektrookulografie (EOG) und das Eye-Tracking zu nennen. Das EOG wird v. a. zur Kontrolle von Augenbewegungen in der EEG-Forschung eingesetzt. Eye-Tracker sind in der anwendungsorientierten Forschung verbreitet, etwa wenn es darum geht, Blickbewegungsmuster beim Betrachten einer Werbeanzeige oder Website zu analysieren. Eye-Tracker-Systeme liefern eine Visualisierung der Blickbewegungen und sind für die Untersuchungspersonen weniger störend als das EOG.

■ Elektrookulografie

Die **Elektrookulografie (EOG)** gibt Auskunft darüber, wohin Personen ungefähr wie lange blicken. Sie zeigt typischerweise horizontale und vertikale Bewegungen der Augen an. Bei der Elektrookulografie macht man sich zunutze, dass der Augapfel ein Dipol ist: Der vordere Teil, die Hornhaut, ist gegenüber der dahinter gelegenen Netzhaut geringfügig positiv geladen. Wenn sich das Auge bewegt, führt das zu Potenzialschwankungen in der Umgebung. Diese lassen sich mit Elektroden registrieren, die in der Nähe des Augapfels auf die Haut aufgeklebt werden. Bei einem Lidschlag erhöht sich kurzfristig das Dipolsignal, so dass mit der gleichen Methode auch Lidschlagfrequenzen registriert werden können.

Elektrookulografie – Die Elektrookulografie („electrooculography", EOG) ist ein Verfahren zur Erfassung der elektrischen Aktivität der Augenmuskeln. Diese wird mittels Elektroden abgeleitet, die direkt auf die Haut in der Nähe des Augapfels angebracht werden. Die Elektrookulografie kommt v. a. zum Einsatz, um in der EEG-Forschung Artefakte durch Augenbewegungen zu kontrollieren. Als Ergebnis der Elektrookulografie werden die Potenzialschwankungen am Augapfel (in Mikrovolt, µV) über die Zeit hinweg grafisch in einem **Elektrookulogramm** (EOG) dargestellt. Als „das EOG" wird üblicherweise nicht nur das Elektrookulogramm, sondern auch die Methode der Elektrookulografie (eigentlich „die EOG") bezeichnet.

■ Eye-Tracking

In den letzten Jahren werden verstärkt Eye-Tracker bzw. **Blickbewegungskameras** eingesetzt, um Blickbewegungen präziser als mit der Elektrookulografie zu erfassen. Eye-Tracker nutzen keine elektrischen Signale, sondern arbeiten mit hoch auflösenden Kameras. Sie registrieren die Stellung und Bewegung der Augen, insbesondere Fixationen von bestimmten Punkten und Sakkaden. Entsprechende Kameras werden zum einen **in Brillen integriert** und erlauben den Untersuchungspersonen somit Bewegungsfreiheit. Zum anderen werden die Kameras meist direkt **am Monitor oder Display befestigt**, auf dem das Stimulusmaterial präsentiert wird. Solche modernen Eye-Tracking-Systeme tolerieren bei entsprechend sorgfältiger Kalibrierung sogar größere Kopfbewegungen, die ansonsten die Messergebnisse verfälschen können. Für Blickbewegungskameras samt entsprechender Analysesoftware sind Anschaffungskosten in Höhe von mehreren Tausend Euro zu veranschlagen.

Eye-Tracking – Das Eye-Tracking („eye tracking", Blickbewegungsmessung) ist ein Messverfahren zur Erfassung von Blickbewegungen, das mittels hochauflösender Kameras Stellung und Bewegung der Augen registriert. Entsprechende Blickbewegungskameras werden entweder in Brillen integriert oder an einem Monitor oder Display angebracht, auf dem das Stimulusmaterial präsentiert wird. Im Ergebnis kann der Blickbewegungsverlauf auf dem Stimulusmaterial statisch als **Grafik** oder dynamisch als **Video** visualisiert werden. Zudem können diejenigen Bereiche des Stimulusmaterials, die besonders häufig fixiert wurden, in einer grafischen Darstellung ähnlich einem Wärmebild (sog. „heat maps") farblich markiert werden.

Psychologische Korrelate der Augenaktivität

Blickbewegungsmessungen werden in vielen Forschungsbereichen eingesetzt. Naheliegend ist, dass dieser Indikator im Zusammenhang mit **Leseverhalten** und **Leseschwierigkeiten** untersucht wird. So ließ sich z. B. nachweisen, dass effiziente Leserinnen und Leser weniger und kürzer fixieren und seltener zu bereits fixiertem Material zurückspringen als unerfahrene Lesesende (Starr & Rayner, 2001). Um kognitive Prozesse beim Lesen zu erforschen, ist es auch sinnvoll, Blickbewegungsmessungen mit dem EEG zu kombinieren (Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011).

Ein weiterer wichtiger Forschungsschwerpunkt stellt die Untersuchung der **Aufnahme von visuellen Informationen** dar. Es zeigte sich, dass Menschen bei Bildern diejenigen Bereiche als erstes fixieren, die die meisten Informationen zu enthalten scheinen. Dies ist insbesondere in den ersten Stufen der Betrachtung der Fall. Außerdem steht die Anzahl von Fixierungen mit einer erhöhten Wiedererkennungsleistung in Zusammenhang (Loftus, 1972; Mackworth & Morandi, 1967).

Schließlich werden Blickbewegungen auch mit **psychischen Erkrankungen** wie z. B. Schizophrenie in Zusammenhang gebracht. So ist bei einer Schizophrenie-Erkrankung zu beobachten, dass Betroffene Probleme haben, bewegten Objekten gezielt mit den Augen zu folgen (für einen Überblick Andreassi, 2007). In der angewandten Forschung wird Eye-Tracking regelmäßig in der **Werbewirkungsforschung** eingesetzt, wobei eruiert wird, ob und welche Werbemerkmale Aufmerksamkeit generieren bzw. einen „Blickfang" darstellen (► Studienbeispiel „Eye-Tracking: Wie werden Werbeanzeigen in den Gelben Seiten wahrgenommen?"). Auch im Bereich des **Bildschirm- und Web-Designs** bzw. der Usability-Forschung wird mit Blickbewegungsmessungen gearbeitet, etwa um die optimale Platzierung von Menüleisten oder anderen Elementen der Bildschirmgestaltung zu finden

Studienbeispiel

Eye-Tracking: Wie werden Werbeanzeigen in den Gelben Seiten wahrgenommen?

Fragestellung

Eine werbepsychologische Studie in den USA ging der Frage nach, wie Anzeigen in den Gelben Seiten idealerweise aufgebaut werden müssen, um die Aufmerksamkeit potenzieller Kunden auf sich zu ziehen (Lohse, 1997).

Methodisches Vorgehen

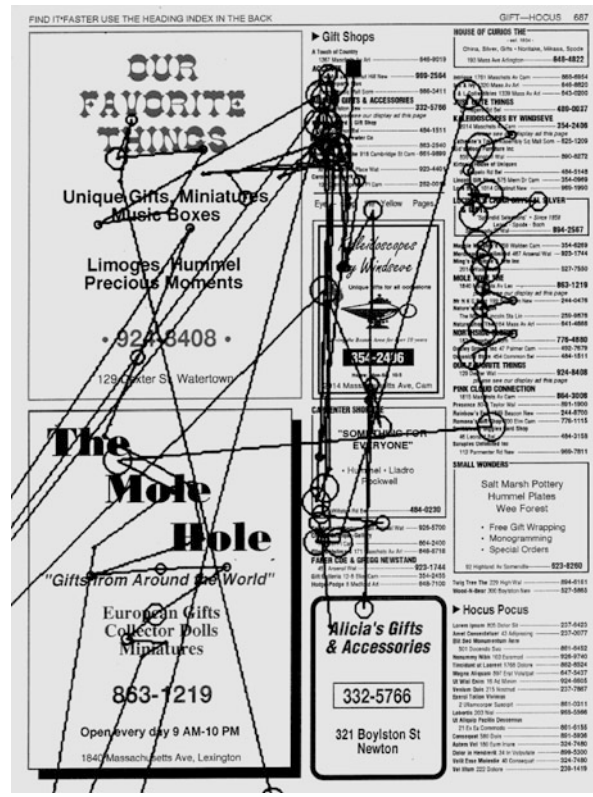
Eine Stichprobe von $n = 32$ Untersuchungspersonen nahmen an einem Laborexperiment teil. Sie erhielten die Aufgabe, gezielt Dienstleister per Anzeige aus eigens erstellten Gelbe-Seiten-Telefonbüchern herauszusuchen, die sie selbst beauftragen würden (z. B. um Blumen zu kaufen oder ihren Wagen reparieren zu lassen). Während dieses Suchprozesses wurden ihre Blickbewegungen per Eye-Tracking aufgezeichnet.

Ergebnisse

Die Versuchspersonen betrachteten die Anzeigeneinträge in alphabetischer Reihenfolge, dieses Scannen war jedoch nicht vollständig. Bevorzugt wurden Anzeigen, die über eine Viertel-Seite gehen gegenüber Anzeigen, die rein textuell in einer Liste aufgeführt sind. Zudem fielen farbige Anzeigen stärker auf, ebenso diejenigen, die in Fettdruck dargestellt sind. Die Versuchsteilnehmenden betrachteten Anzeigen, die sie letztlich für die Inanspruchnahme einer Dienstleistung in Betracht zogen, 54 % länger als Anzeigen, die sie nicht auswählten (◘ Abb. 10.38).

Schlussfolgerung

Die Ergebnisse liefern Hinweise für eine effektive Anzeigengestaltung: Im Vergleich zu den übrigen Einträgen generieren auffällig gestaltete Anzeigen mehr Aufmerksamkeit. Ist die Aufmerksamkeit geweckt worden, wird mehr Zeit für die Betrachtung verwendet, was eine positive Entscheidung des Kunden zugunsten des werbetreibenden Anbieters begünstigen kann.



◘ **Abb. 10.38** Beispiel für eine serielle Betrachtung der alphabetischen Einträge in einem Telefonverzeichnis. (Aus Lohse, 1997. Consumer eye movement patterns on yellow pages advertising. *Journal of Advertising, 26*(1), 61–73. Mit freundlicher Genehmigung von Taylor & Francis Ltd, http://www.tandf.co.uk/journals; http://www.informaworld.com)

(Yom & Wilhelm, 2004). Nicht zuletzt können Blickbewegungsmessungen auch der **Methodenforschung** dienen, etwa wenn es um die Optimierung von Fragebögen geht, die in Online-Umfragen per Bildschirm präsentiert (Galesic, Yan, Das, Ester & Kaczmirek, 2011) oder bei mobilen Befragungen auf sehr kleinen Displays bearbeitet werden sollen (Öquist & Lumsden, 2008).

10.5.7 Messung sexueller Erregung

Während in den vorausgegangenen Abschnitten fünf einzelne Organsysteme (Gehirn, Herz-Kreislauf-System, Haut, Muskulatur, Augen) und die Messung ihrer jeweiligen Aktivitäten separat abgehandelt wurden, soll jetzt ein komplexes **psychophysiologisches Phänomen** im Mittelpunkt stehen, das über unterschiedliche psychologische sowie ganz unterschiedliche physiologische Indikatoren messbar ist. Exemplarisch wird hier als Phänomen die sexuelle Erregung („sexual arousal") herausgegriffen, analog könnte man z. B. auch Stress oder Angst betrachten. Die sexuelle Psychophysiologie (zum Überblick Janssen, 2002) findet Anwendung u. a. in der Sexualforschung, der Sexualmedizin, der Pharmakologie, der Forensik, der Geschlechterforschung sowie der Mediennutzungs- und Medienwirkungsforschung.

Sexuelle Psychophysiologie – Die sexuelle Psychophysiologie („sexual psychophysiology") meint die Anwendung psychophysiologischer Messmethoden zur Erforschung sexueller Erregung. Dabei wird besonderes Gewicht auf die Untersuchung der Wechselwirkungen zwischen einerseits subjektiven und andererseits physiologischen Aspekten der sexuellen Erregung gelegt.

Psychologische Indikatoren sexueller Erregung

Sexuelle Erregung hat eine subjektive Erlebensdimension mit unterschiedlichen kognitiven, emotionalen und motivationalen Aspekten. Diese können als latente Variablen

unterschiedlich operationalisiert werden, um sie über entsprechende Indikatorvariablen zu messen. Beispielsweise kann man Untersuchungspersonen bitten, die Intensität ihrer subjektiven Erregung vor und nach dem Anschauen von sexuell expliziten Fotos oder Filmen auf einer einzelnen Ratingskala (1: gar nicht sexuell erregt bis 7: sehr stark sexuell erregt) anzugeben. Ebenso ist es möglich, die Intensität der subjektiven Erregung kontinuierlich während des Betrachtens eines Films über einen Schieberegler zu erfassen („continuous response measurement", CRM). Um einzelne inhaltliche Aspekte der Erregung (z. B. kognitive, emotionale und motivationale Komponenten) differenzierter zu erheben, können auch standardisierte Fragebogeninstrumente eingesetzt werden wie z. B. „The Sexual Arousal and Desire Inventory" (SADI; Toledano & Pfaus, 2006) oder die Items von Kukkonen, Binik, Amsel & Carrier (2010). In einem qualitativen Untersuchungsdesign können die Untersuchungsteilnehmenden gebeten werden, ihre erregungsbezogenen Gedanken, Gefühle und Handlungstendenzen in eigenen Worten zu schildern.

All diese Datenerhebungsmethoden liefern Selbstauskunftsdaten, bei denen aufgrund der sozial heiklen Thematik – je nach konkreten Untersuchungsbedingungen und Zielgruppen – mit mehr oder minder deutlichen Verzerrungen gerechnet werden muss. So würde man erwarten, dass viele Menschen sich im Sinne sozial erwünschten Antwortens scheuen, sexuelle Erregung in Situationen oder bei Reizvorlagen zuzugeben, die sie als nicht angemessen bewerten, etwa weil sie Geschlechtsrollenerwartungen oder sozialen Normvorstellungen über „richtige" Sexualität widersprechen. Sozial erwünschte Antworten können auch darauf hinauslaufen, dass sexuelle Erregung berichtet wird, obwohl sie nicht oder kaum vorhanden ist. Abgesehen von Antwortverzerrungen aufgrund sozialer und sexueller Normen ist auch zu beachten, dass Menschen sich deutlich darin unterscheiden, wie genau sie ihre sexuelle Erregung wahrnehmen und artikulieren können. So mag z. B. die Abgrenzung zwischen Nervosität angesichts der Untersuchungssituation einerseits und sexueller Erregung angesichts des gezeigten experimentellen Stimulus andererseits nicht allen Versuchspersonen gleichermaßen gut gelingen, was die Validität der Selbstauskunftsdaten einschränken würde.

Physiologische Indikatoren sexueller Erregung

Sexuelle Erregung ist ein psychophysiologisches Phänomen, das sich in der Aktivität aller in diesem Kapitel behandelten Organsysteme widerspiegelt: Eine Aktivierung des Belohnungszentrums im Gehirn, erhöhte Herzschlag- und Atemfrequenz, erhöhte Hautleitfähigkeit und Muskelspannung, gezielte Blickbewegungen etc. sind typisch. Auch weitere – in diesem Kapitel nicht behandelte – Prozesse sind beteiligt (z. B. sekretorische und hormonelle Prozesse). Ein Großteil der physiologischen Reaktionen auf sexuelle Stimuli sind jedoch als **unspezifische physiologische Reaktion** („unspecific physiological response") zu kennzeichnen (Schandry, 1996). Eine solche allgemeine Aktivierung – wie z. B. beschleunigter Herzschlag und Aktivitäten im Belohnungszentrum des Gehirns – würde sich vergleichbar zeigen, wenn Personen mit einem spannenden neuen – aber eben nicht-sexuellen – Stimulus konfrontiert werden.

Als **spezifische physiologische Reaktion** („specific physiological response") bei sexueller Erregung wird v. a. die genitale Reaktion aufgefasst. Ein wesentlicher physiologischer Prozess ist dabei die sowohl bei Frauen als auch bei Männern verstärkte Blutansammlung im Genitalbereich (**Vasokongestion**, „vasocongestion"), die mit einer Volumenveränderung einhergeht. Als Messverfahren für die Volumenveränderung des Genitalbereichs dient die im Zusammenhang mit der Messung der Pulsfrequenz bereits angesprochene **Plethysmografie**, die in verschiedenen Varianten existiert:

- **Penis-Plethysmografie mittels Dehnungsmesseinrichtung** („penile plethysmography [PPG] with mercury-in-rubber or indium/gallium-in-rubber ring strain gauge"): Bei diesem Verfahren wird der mit der Genitaldurchblutung schwankende Umfang des Penis durch eine um den Schaft gelegte elastische Schlaufe erfasst, die einen Messfühler für die Dehnung enthält. Die Schlaufe wird von den Probanden selbst angelegt und ist relativ komfortabel, weshalb sich die Methode durchgesetzt und die in den 1950er-Jahren entwickelten volumetrischen Verfahren (Freund, 1963) weitgehend abgelöst hat.
- **Vaginale Photoplethysmografie** („vaginal plethysmography", VPPG): Zur Messung der Vaginaldurchblutung wird seit Ende der 1960er-Jahre ein tamponförmiges Photoplethysmografie-Messgerät verwendet, das die Probandinnen selbst einführen können (Palti & Bercovici, 1967). Das Messgerät sendet Infrarotlicht aus. Dieses wird vom umliegenden Gewebe umso weniger reflektiert, je stärker das Gewebe durchblutet ist, denn der rote Blutfarbstoff absorbiert das Licht. Die Menge des reflektierten Lichts wird gemessen und als Indikator der Vasokongestion verwendet.
- **Klitorale Photoplethysmografie** („clitoral plethysmography"): Ergänzend zu der seit Jahrzehnten etablierten Messung der vaginalen Durchblutung wird neuerdings auch eine Messung der klitoralen Durchblutung als weiterer – und möglicherweise genauerer – Erregungsindikator bei Frauen vorgeschlagen (Gerritsen, et al., 2009).

Die Messgeräte müssen vor dem Einsatz desinfiziert werden und sind von den Untersuchungspersonen selbst anzulegen. Sie können dann eine ganze Weile relativ komfortabel getragen werden. Die Versuchspersonen werden einzeln untersucht, wobei die gleichgeschlechtliche Zuordnung von Versuchsleitenden empfohlen wird (Schandry, 1996). Da die Genitaldurchblutung als spezifischer physiologischer Erregungsindikator nach Entfernung des Stimulus relativ schnell wieder zur Baseline zurückkehrt, sind pro Person mehrere Untersuchungsdurchgänge mit unterschiedlichen Stimuli (z. B. mehreren Audiofiles, Fotoserien oder Filmclips) möglich (Messwiederholungsdesign, ► Abschn. 7.8).

Studien, in denen eine Messung körperlicher sexueller Erregung vorgenommen wird, sind insofern von Verzerrungen betroffen, als sich für diese Form der Datenerhebung v. a. sexuell besonders aufgeschlossene Personen freiwillig melden (**Freiwilligen-Bias**, „volunteer bias"), was die Generalisierbarkeit der Ergebnisse einschränken kann (Plaud, Gaither, Hegstad, Rowan, & Devitt, 1999). Weniger belastende Messmethoden, die z. B. keinen Hautkontakt mit dem Messgerät erfordern, könnten die Teilnahmebereitschaft größerer Bevölkerungskreise erleichtern. Eine viel versprechende neue Methode, die in bekleidetem Zustand und genau identisch bei Frauen wie Männern einsetzbar ist, stellt die Erfassung der Genitaldurchblutung mittels **Wärmebildkameras** („infrared camera") dar. Diese können Temperaturänderungen mit einer Genauigkeit von Hundertstel Grad Celsius erfassen, sofern das Gerät etwa einen Meter von der Versuchsperson entfernt positioniert wird. Die erhöhte Genitaldurchblutung bei sexueller Erregung schlägt sich in einer lokalen Temperaturerhöhung um ca. 1–3 °C im Genitalbereich nieder (Kukkonen et al., 2010).

Besonders forschungs- und praxisrelevant ist das **ambulatorische Assessment** körperlicher sexueller Erregungsindikatoren, etwa um festzustellen, ob Erektionsstörungen eine körperliche Ursache haben oder nicht. Dazu werden die Probanden bzw. Patienten mit einem ambulanten Plethysmografie-Gerät ausgestattet, das über längere Zeit im Alltag getragen wird und mit dem dann z. B. nächtliche Erektionen aufgezeichnet werden können, deren Existenz darauf hindeutet, dass in sexuellen Situationen erlebte Erektionsprobleme nicht körperlich, sondern psychisch bedingt sind. Darüber hinaus existieren noch diverse weitere, aber weniger verbreitete physiologische Messmethoden zur Erhebung sexueller Erregung (Cacioppo et al., 2007). Der Einsatz **bildgebender Verfahren** („imaging techniques") zur Messung der Hirnaktivität bei sexueller Erregung (z. B. mittels fMRT) ist wissenschaftlich interessant, aber aufgrund der hohen Kosten nur beschränkt möglich.

Verhältnis psychologischer und physiologischer Erregungsindikatoren

Um für das Spannungsverhältnis zwischen psychologischen und physiologischen Indikatoren zu sensibilisieren, werden im Folgenden exemplarisch drei kontroverse Forschungsthemen innerhalb der sexuellen Psychophysiologie angesprochen.

Viagra für Mann und Frau? Seit 1998 sind Phosphodiesterase-V-Hemmer als Wirkstoffe zur Behandlung erektiler Dysfunktion auf dem Markt. Am bekanntesten ist das von der Firma Pfizer unter dem Namen „Viagra" entwickelte Medikament basierend auf dem Wirkstoff Sildenafil. Wenn eine subjektive sexuelle Erregung vorliegt, kann der Wirkstoff die Genitaldurchblutung fördern, sofern diese anderweitig physiologisch blockiert ist. Bei Männern ohne körperlich begründete Erektionsstörung ist Viagra weitgehend wirkungslos, wird aber teilweise als Partydroge genutzt. Empirische Studien zur Wirksamkeit von Viagra bei Männern stützen sich meist auf eine Messung der Genitaldurchblutung mittels Penis-Plethysmografie. Angesichts des großen Erfolgs von Viagra bei Männern wurde die Wirksamkeit auch bei Frauen zur Behandlung von Erregungs- und Orgasmusstörungen überprüft. Die medikamentös nachweislich verstärkte Genitaldurchblutung trug bei Frauen jedoch nicht zu einer stärkeren subjektiven Erregung oder einem befriedigenderen sexuellen Erleben bei. In der Sexualforschung wird die zunehmende **Medikalisierung von Sexualität** kontrovers diskutiert und eine zu starke Fokussierung auf physiologische Prozesse auf Kosten der intensiveren Erforschung der komplexen psychosozialen Dimensionen sexueller Erregung kritisiert (Kaschak & Tiefer, 2002; The New View Campaign: http://www.newviewcampaign.org). Insgesamt wird zur Untersuchung von Sexualität heute ein sog. **biopsychosoziales Modell** zugrunde gelegt, wobei die genaue Definition, Messung und Gewichtung der einzelnen Faktoren sowie ihre Relationen untereinander Gegenstand laufender Forschung sind. Auch aus medizinischer Sicht wird betont, dass sexuelle Funktionsstörungen oft keine physiologischen, sondern psychosoziale Ursachen haben und auf dieser Ebene zu behandeln sind (Weig, 2006).

Sexueller Lügendetektor? Das erste Einsatzfeld der von Kurt Freund (1963) entwickelten volumetrischen Penis-Plethysmografie war die Überprüfung von Rekruten, die behaupteten, sie seien homosexuell und sich davon eine Freistellung vom tschechischen Militärdienst versprachen, zu dem Homosexuellen damals der Zugang verboten war (Wilson & Mathon, 2006). Der physiologische

Erregungsindikator wurde hier als eine Art sexueller Lügendetektortest verwendet: Wer auf Bilder nackter Frauen, nicht aber nackter Männer mit objektiv messbarer Genitaldurchblutung reagierte, wurde nicht als homosexuell anerkannt. Nachdem Homosexualität in den internationalen Klassifikationen DSM („Diagnostic and Statistical Manual of Mental Disorders") der American Psychiatric Association und ICD („International Statistical Classification of Diseases and Related Health Problems") der World Health Organization seit 1973 bzw. 1992 nicht mehr als Krankheit geführt wird, spielt die Diagnose gleichgeschlechtlicher sexueller Orientierung mittels physiologischer Messungen keine Rolle mehr, dafür werden heute Selbstauskunftsdaten verwendet.

Dagegen haben sich physiologische Messverfahren in der klinischen und forensischen Forschung etabliert, wenn es um die Diagnose von sog. **devianter Erregung** geht (Hunter, Goodwin, & Becker, 1994; Seto, Lalumière, & Blanchard, 2000). Damit ist insbesondere eine stärkere genitale Reaktion auf Stimuli mit Kindern als mit Erwachsenen gemeint. Die Penis-Plethysmografie wird in der Literatur teilweise als die beste Methode beschrieben, um das Rückfallrisiko von Sexualstraftätern mit kindlichen Opfern vorherzusagen (Hanson & Bussière, 1998). In den USA werden Sexualstraftäter in der Bewährungszeit regelmäßigen Polygrafentests und Penis-Plethysmografietests unterzogen. Dahinter steht die Intention, möglicherweise geschönte Selbstauskünfte von Sexualstraftätern, die z. B. ein anhaltendes sexuelles Interesse an Kindern leugnen mögen, durch den objektiven Test aufzudecken. Der Penis-Plethysmografietest gilt dabei als relativ täuschungssicher, da es Probanden zwar vereinzelt gelingen kann, genitale sexuelle Reaktion auf deviante Stimuli zu unterdrücken, sie aber dies kaum durchgängig schaffen und v. a. die positive genitale Reaktion auf normkonforme Stimuli kaum aktiv herstellen können. Dennoch gibt es Kritik an dem Verfahren sowie seiner forensischen Einsatzweise in den USA. Neben ethischen und verfassungsrechtlichen Bedenken angsichts des starken Eingriffs in die Privatsphäre wird in methodischer Hinsicht die Validität des Verfahrens infrage gestellt, insbesondere weil überprüfte Normwerte fehlen (Laws, 2003). Da der Schutz vor Sexualstraftaten ein Thema ist, das die breite Öffentlichkeit sehr bewegt, ist auch zu beachten, dass ein vorschnelles Vertrauen in eine einzelne vermeintlich sichere und objektive physiologische Messmethode zur Identifikation von potenziellen Tätern das Risiko von Fehleinschätzungen birgt; etwa wenn dadurch Sexualstraftäter ohne Präferenz für deviante Stimuli aus dem Blick geraten oder deviante Erregung pauschal kriminalisiert wird, auch wenn sie oft gar nicht mit Straftaten einhergeht. Hier besteht sowohl Bedarf nach gesicherten Forschungsergebnissen als auch ihrem Transfer in die Praxis.

Stärkere Reaktion von Frauen als von Männern auf sexuelle Stimuli? In den 1990er-Jahren zeigten Studien aus der Forschungsgruppe um die niederländische Psychologin Ellen Laan erstmals auf, dass die genitale Reaktion von Frauen auf visuelle sexuelle Stimuli nicht etwa geringer, sondern sogar stärker ausfällt als bei Männern: Frauen reagieren binnen Sekunden genital und zwar bei einer viel größeren inhaltlichen Bandbreite an Darstellungen (z. B. hinsichtlich Geschlecht der medial dargestellten Personen, Art der sexuellen Aktivitäten etc.) als Männer. Während bei Männern die genitale Reaktion jedoch sehr stark mit subjektivem Erregungserleben einhergeht (bivariate Korrelation von $r = .66$), ist dieser Zusammenhang bei Frauen schwächer ($r = .26$; Chivers, Seto, Lalumière, Laan, & Grimbos, 2010). Nicht selten berichten Frauen angesichts des präsentierten sexuellen Stimulus keinerlei subjektive sexuelle Erregung oder sogar Ablehnung und Ekel angesichts pornografischer Stimuli, obwohl die vaginale Photoplethysmografie genitale Erregung misst. Diese Inkonsistenz ist offenbar nicht als Methodenartefakt zu erklären, denn die Validität der vaginalen Photoplethysmografie gilt als gesichert (Suschinsky, Lalumière, & Chivers, 2009). Eine Erklärung besteht darin, dass bei Männern subjektive sexuelle Erregung v. a. im Zuge der Wahrnehmung der eigenen genitalen Reaktion entsteht, während bei Frauen die automatische genitale Reaktion ein evolutionärer Schutzmechanismus gegen Verletzungen bei möglichem Geschlechtsverkehr ist und die subjektive Erregung stärker von der Beschaffenheit und Interpretation des jeweiligen sexuellen Stimulus abhängt (Laan & Both, 2008). Angesichts zunehmender Sexualisierung der Medienwelt gewinnt Forschung zu sexuellen und nicht-sexuellen Effekten expliziter medialer Stimuli an Bedeutung. Physiologische Messungen bei Frauen, Männern, Inter- und Transsexuellen liefern dabei auch der Genderforschung Informationen zu Unterschieden und Gemeinsamkeiten zwischen verschiedenen Geschlechtern.

Am Beispiel der physiologischen Messung sexueller Erregung wurden abschließend noch einmal generelle Besonderheiten physiologischer Messverfahren verdeutlicht. Dazu gehört die starke Geräteabhängigkeit und fortwährende Weiterentwicklung der Messgeräte (z. B. neuerdings Erprobung von Wärmebildkameras). Auch die artifizielle Untersuchungssituation ist zu nennen und die relativ starke Beanspruchung der Versuchspersonen, die zu einem Freiwilligen-Bias und somit einer Einschränkung der externen Validität (Generalisierbarkeit auf größere Bevölkerungsgruppen) entsprechender Studi-

en führen kann. Besonders hervorgehoben wurde zudem die Notwendigkeit, physiologische Parameter im Hinblick auf psychische Prozesse stets differenziert zu interpretieren. Das gilt besonders für unspezifische physiologische Reaktionen (z. B. kardiovaskuläre oder elektrodermale Indikatoren), denn beschleunigter Herzschlag oder verstärktes Schwitzen kann mit ganz unterschiedlichen positiven wie negativen Erlebensqualitäten einhergehen. Das gilt aber auch für spezifische physiologische Parameter wie die Genitaldurchblutung, die offenbar nicht mit erlebter sexueller Erregung gleichzusetzen ist.

Lernquiz

1. Warum werden physiologische Messungen auch als „objektive" Datenerhebungsmethoden bezeichnet?
2. Welche Messprobleme treten bei der Erhebung physiologischer Daten auf?
3. Grenzen Sie EEG und fMRT zur Messung von Hirnaktivität hinsichtlich physiologischer Grundlagen, Kosten und Informationsgehalt voneinander ab!
4. In welchen Bereichen bewegt sich üblicherweise die Herzschlagfrequenz beim gesunden Erwachsenen im Ruhe- sowie im maximalen Belastungszustand?
5. Warum wird bei EP im EEG mit der „Mittelungstechnik" gearbeitet?
6. Welche Parallele besteht zwischen Spontan-EEG und Ruhe-EKG?
7. Wie lässt sich ein Ansteigen der Hautleitfähigkeit psychologisch interpretieren?
8. Welche psychologischen Rückschlüsse lassen sich aus der Messung der Aktivität von Gesichtsmuskeln ziehen?
9. Wie werden Blickbewegungen gemessen und in welchen sozialwissenschaftlichen Forschungsfeldern spielen sie eine wichtige Rolle?
10. Intensivierte Genitaldurchblutung ist ein spezifischer physiologischer Indikator sexueller Erregung des Organismus. In welcher Beziehung steht er zu subjektiv erlebter sexueller Erregung?

Übungsaufgaben

1. Messen Sie vor der Betrachtung eines spannenden Films Ihren Ruhepuls manuell mithilfe von Zeige- und Mittelfinger an Ihrer Halsschlagader. Zählen Sie die Pulsschläge für zehn Sekunden und multiplizieren Sie den Wert mit 6. Dokumentieren Sie den Wert. Messen Sie Ihren Puls während des Sehens des Films, und zwar direkt nach einer für Sie sehr spannenden Szene. Vergleichen Sie die beiden Werte.
2. Formulieren Sie eine sozialwissenschaftliche Forschungsfrage, bei der die ergänzende Erhebung physiologischer Indikatoren sinnvoll ist. Wählen Sie aus, welche Indikatoren Sie berücksichtigen wollen und diskutieren Sie die Auswahl in Bezug auf Umsetzbarkeit der Datenerhebung, Datenanalyse und statistische Auswertung der Befunde.
3. Recherchieren Sie eine aktuelle wissenschaftliche Studie zu einem Thema Ihrer Wahl, in der physiologische Indikatoren ergänzend zu Selbstauskünften erhoben wurden. Lesen Sie die Studie und diskutieren Sie die Interpretation der physiologischen Messwerte vor dem theoretischen Hintergrund der Studie kritisch hinsichtlich ihrer Aussagekraft.

Weiterführende Literatur

- Andreassi, J. L. (2007). *Psychophysiology. Human Behavior & Physiological Response*. Mahwah: Taylor & Francis.
- Birbaumer, N. & Schmidt, R. F. (2010). *Biologische Psychologie* (7. Aufl.). Heidelberg: Springer.
- Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). *Handbook of psychophysiology* (3rd ed.). Cambridge: Cambridge University Press.
- Gramann, K. & Schandry, R. (2009). *Psychophysiologie. Körperliche Indikatoren psychischen Geschehens* (4. Aufl.). Weinheim: Beltz.
- Kramme, R. (2006). *Medizintechnik. Verfahren, Systeme, Informationsverarbeitung* (3. Aufl.). Heidelberg: Springer.
- Schandry, R. (2003). *Biologische Psychologie*. Weinheim: Beltz.

Literatur

Alexander, J. E., O'Boyle, M. W., & Benbow, C. P. (1996). Developmentally Advanced EEG Alpha Power in Gifted Male and Female Adolescents. *International Journal of Psychophysiology, 23*(1-2), 25–31

Andreassi, J. L. (1966a). Skin-conductance and reaction-time in a continuous auditory monitoring task. *American Journal of Psychology, 79*(3), 470–474.

Andreassi, J. L. (2007). *Psychophysiology. Human behavior & physiological response*. Mahwah: Taylor & Francis.

Andreassi, J. L. & Whalen, P. M. (1967). Some physiological correlates of learning and overlearning. *Psychophysiology, 3*(4), 406–413.

Baltissen, R. & Heimann, H. (1995). Aktivierung, Orientierung und Habituation bei Gesunden und psychisch Kranken. In G. Debus, G. Erdmann & K. W. Kallus (Hrsg.), *Biopsychologie von Streß und emotionalen Reaktionen* (S. 233–246). Göttingen: Hogrefe.

Barry, R. J., Johnstone, S. J., & Clarke, A. R. (2003). A review of electrophysiology in attention-deficit/hyperactivity disorder: II. Event-related potentials. *Clinical Neurophysiology, 114*(2), 184–198.

Becker-Carus, C. (1981). *Grundriß der physiologischen Psychologie.* heidelberg: quelle & meyer.

Bennett, C. M., Abigail, A. B., Michael, B. M., & George, l. W. (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results, 1*(1), 1–5.

Bente, G. (2000). *Intergrierte Rezeptions-Prozessanalyse, Integration Apparativer Methoden in der angewandten Medienforschung*. Paper presented at the 3. Tagung der Fachgruppe Medienpsychologie. 10.-12. September, 2003, Landau, Germany.

Bente, G., Eschenburg, F., & Fürtjes, M. (2007). Im Auge des Nutzers. Eye-Tracking in der Web-Usability-Forschung. In M. Welker & O. Wenzel (Hrsg.), *Online-Forschung 2007. Grundlagen und Fallstudien* (S. 185–219). Köln: Halem.

Berger, H. (1929). Über das Elektroenzephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten, 87*, 527–570.

Berntson, G. G. & Cacioppo, J. T. (2000). Psychobiology and social psychology: Past, present, and future. *Personality and Social Psychology Review, 4*(1), 3–15.

Birbaumer, N. & Schmidt, R. F. (2006). *Biologische Psychologie* (7. Aufl.). Heidelberg: Springer.

Boucsein, W. (1988). *Elektrodermale Aktivität: Grundlagen, Methoden, Anwendungen.* Heidelberg: Springer.

Boucsein, W. (1991). Arbeitspsychologische Beanspruchungsforschung heute. Eine Herausforderung an die Psychophysiologie. *Psychologische Rundschau, 42*, 129–144.

Boucsein, W. (1992). *Electrodermal Activity*. New York: Plenum Press.

Boucsein, W. (1995). Die elektrodermale Aktivität als Emotionsindikator. In G. Debus, G. Erdmann & K. W. Kallus (Hrsg.), *Biopsychologie von Streß und emotionalen Reaktionen* (S. 143–162). Göttingen: Hogrefe.

Brondolo, E., Rieppi, R., Kelly, K. P., & Gerin, W. (2003). Perceived racism and blood pressure: A review of the literature and conceptual and methodological critique. *Annals of Behavioral Medicine, 25*(1), 55–65.

Brosschot, J. F., van Dijk, E., & Thayer, J. F. (2007). Daily worry is related to low heart rate variability during waking and the subsequent nocturnal sleep period. *International Journal of Psychophysiology, 63*, 39–47.

Butler, E. A., Lee, T. L., & Gross, J. J. (2009). Does expressing your emotions raise or lower your blood pressure? *Journal of Cross-Cultural Psychology, 40*(3), 510–517.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). *Handbook of psychophysiology* (3rd ed.). Cambridge: Cambridge University Press.

Carlson, N. R. (2004). *Physiologische Psychologie* (8. Aufl.). München: Pearson Studium.

Carvalho-da-Silva, A. M., Van Damme, I., Wolf, B., & Hort, J. (2011). Characterisation of chocolate eating behaviour. *Physiology and Behavior, 104*(5), 929–933.

Chivers, M. L., Seto, M. C., Lalumière, M. L., Laan, E., & Grimbos, T. (2010). Agreement of self-reported and genital measures of sexual arousal in men and women: A meta-analysis. *Archives of Sexual Behavior, 39*(1), 5–56.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: Murray.

Das, M., Ester, P., & Kaczmirek, L. (2011). *Social and behavioral research and the internet: advances in applied methods and research strategies*. New York: Routledge.

Davidson, R. J., Schwartz, G. E., Saron, C., Bennett, J., & Goleman, D. J. (1979). Frontal versus parietal EEG asymmetry during positive and negative affect. *Psychophysiology, 16*, 202–203.

De Pascalis, V. & Ray, W. J. (1998). Effects of memory load on event-related patterns of 40 Hz EEG during cognitive and motor tasks. *International Journal of Psychophysiology, 28*(3), 301–316.

De Tommasoa, M., Pecoraroa, C., Sardaroa, M., Serpinoa, C., Lancionib, G., & Livreaa, P. (2008). Influence of aesthetic perception on visual event-related potentials. *Consciousness and Cognition, 17*(3), 933–945.

Diekelmann, S. & Born, J. (2010). Slow-wave sleep takes the leading role in memory reorganization. *Nature Reviews Neuroscience, 11*(128), 114–126.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General, 140*(4), 552–572.

Edelberg, R. (1967). Electrical properties of the skin. In C. C. Brown (Ed.), *Methods in psychophysiology.* (pp. 1–53). Baltimore: Wiliams & Wilkins.

Edelberg, R. (1972). Electrical activity of the skin. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of Psychophysiology* (pp. 367–418). New York: Holt, Rinehart & Winston.

Ekman, P. (1973). *Darwin and facial expression: A century of research in review*. New York: Academic Press.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & T. Power (Eds.), *The handbook of cognition and emotion* (pp. 45–60). Hoboken: Wiley.

Ekman, P. (2009). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York: Norton.

Ekman, P. & Friesen, W. V. (1978). *Facial action coding system. A technique for the measurement of facial movement.* Palo Alto: Consulting Psychologists Press.

Elbert, T. & Heim, S. (2001). A light and a dark side. *Nature, 411*, 139.

Elliott, R. (1974). The motivational significance of heart rate. In P. A. Obrist, A. H. Black, J. Brener & L. V. DiCara (Eds.), *Cardiovascular psychophysiology* (pp. 505–537). Chicago: Aldine.

Engel, B. T. (1972). Response specifity. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 571–576). New York: Holt, Rinehart, & Winston.

Erdmann, G. & Voigt, K. H. (1995). Vegetative und endokrine Reaktionen im Paradigma „Öffentliches Sprechen". Was indizieren sie? In G. Debus, G. Erdmann & K. W. Kallus (Hrsg.), *Biopsychologie von Streß und emotionalen Reaktionen* (S. 113–128). Göttingen: Hogrefe.

Fahrenberg, J. (1983). Psychophysiologische Methodik. In K. J. Groffmann & I. Michel (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich B, Serie 2, Bd. 4, Verhaltensdiagnostik* (S. 1–115). Göttingen: Hogrefe.

Fahrenberg, J., Leonhard, R., & Foerster, F. (2002). *Alltagsnahe Psychologie.* Bern: Huber.

Falk, E. B., Berkman, E. T., Whalen, D., & Lieberman, M. D. (2011). Neural activity during health messaging predicts reductions in smoking above and beyond self-report. *Health Psychology, 30*(2), 177–185.

Féré, C. (1888). Note sur les modifications de la résistance éctrique sous l'influence des excitations sensorielles et des émotions. *Comptes Rendus des Séances de la Société de Biologie, 5*, 217–219.

Fiedler, K., Schmid, J., & Stahl, T. (2002). What is the current truth about polygraph lie detection. *Basic and Applied Social Psychology, 24*(4), 313–324.

Foerster, F., Schneider, H. J., & Walschburger, P. (1983). *Psychophysiologische Reaktionsmuster. Zur Theorie und Methodik der Analyse*

individualspezifischer, stimulusspezifischer und motivationsspezifischer Reaktionsmuster in Aktivierungsprozessen. München: Minerva.

Fowles, D. C. (1980). The three arousal model: Implications of Gray's two-factor learning theory for heart rate, electrodermal activity, and psychopathy. *Psychophysiology, 17*(2), 87–104.

Fowles, D. C., Fischer, A. E., & Tranel, D. T. (1982). The heart beats to reward: The effects of monetary incentives on heart rate. *Psychophysiology, 19*(5), 506–513.

Freeman, G. L. (1940). The relationship between performance level and bodily activity level. *Journal of Experimental Psychology, 26*(6), 602–608.

Freund, K. (1963). Alaboratory method for diagnosing predominance of homo- or hetero-erotic interest in the male. *Behaviour Research and Therapy, 1*(1), 85–93.

Frey, S. & Bente, G. (1989). Mikroanalyse medienvermittelter Informationsprozesse. Zur Anwendung zeitreihen-basierter Notationsprinzipien auf die Untersuchung von Fernsehnachrichten. In M. Kaase & W. Schulz (Hrsg.), *Massenkommunikation. Theorien, Methoden, Befunde. Sonderheft 30 der „Kölner Zeitschrift für Soziologie und Sozialpsychologie"* (S. 508–526). Opladen: Westdeutscher Verlag.

Fuchs, T. (2009). *Das Gehirn. Ein Beziehungsorgan: Eine phänomenologisch-ökologische Konzeption* (2. Auf.). Stuttgart: Kohlhammer.

Galesic, M., Yan, T., Das, M., Ester, P., & Kaczmirek, L. (2011). Use of eye tracking for studying survey response processes. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: advances in applied methods and research strategies* (pp. 349–370). New York: Routledge.

Gasser, T., von Lucadou-Muller, I., Verleger, R., & Bacher, P. (1983). Correlating EEG and IQ: A new look at an old problem using computerized EEG parameters. *Electroencephalography and Clinical Neurophysiology, 55*(5), 493–504.

Gati, I., Ben-Shakar, G., & Avni-Liberty, S. (1996). Stimulus novelty and significance in electrodermal orienting response: The effects of adding versus deleting stimulus components. *Psychophysiology, 33*(6), 637–643.

Gerber, W. D. (1986). Chronischer Kopfschmerz. In W. Miltner, N. Birbaumer & W. D. Gerber (Hrsg.), *Verhaltensmedizin* (S. 135–170). Heidelberg: Springer.

Gerritsen, J., van der Made, F., Bloemers, J., van Ham, D., Kleiverda, G., Everaerd, W., et al. (2009). The clitoral photoplethysmograph: A new way of assessing genital arousal in women. *Journal of Sexual Medicine, 6*(6), 1678–1687.

Gevins, A. S. & Smith, M. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex, 10*(9), 829–839.

Giannitrapani, D. (1969). EEG average frequency and intelligence. *Electroencephalography and Clinical Neurophysiology, 27*(5), 480–486.

Goldstein, J. H., Harman, J., McGee, P. E., & Karasik, R. (1975). Test of an information-processing model of humor: physiological response changes during problem and riddle-solving. *Journal of General Psychology, 92*(1), 59–68.

Greden, J. F., Genero, N., Price, L., Feinberg, M., & Levine, S. (1986). Facial electromyography in depression. *Archives of General Psychiatry, 43*(3), 269–274.

Hagit, C., Kotler, M., Matar, M. A., Kaplan, Z., Loewenthal, U., Miodownik, H., et al. (1998). Analysis of heart rate variability in posttraumatic stress disorder patients in response to a trauma-related reminder. *Biological Psychiatry, 44*(10), 1054–1059.

Hajcak, G., MacNamara, A., & Olvet, D. M. (2010). Event-related potentials, emotion, and emotion regulation: An integrative review. *Developmental Neuropsychology, 35*(2), 129–155.

Hanson, R. K. & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*(2), 348–362.

Herbold-Wootten, H. (1982). The German Tatbestandsdiagnostik; a historical review of the beginnings of scientific lie detection in Germany. *Polygraph, 11*(3), 246–257.

Hung, T.-M., Spalding, T. W., Santa Maria, D. L., & Hatfield, B. D. (2004). Assessment of reactive motor performance with event-related brain potentials: Attention processes in elite table tennis players. *Journal of Sport and Exercise Psychology, 26*(2), 317–337.

Hunter, J. A., Goodwin, D. W., & Becker, J. V. (1994). The relationship between phallometrically measured deviant sexual arousal and clinical characteristics in juvenile sexual offenders. *Behaviour Research and Therapy, 32*(5), 533–538.

Hussy, W., Schreier, M., & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften*. Heidelberg: Springer.

James, J. E. (2004). Critical review of dietary caffeine and blood pressure: A relationship that should be taken more seriously. *Psychosomatic Medicine, 66*(1), 63–71.

Jäncke, L. (2010). Hirnforschung. Sinnvolle Ergänzung oder überflüssiges Anhängsel der Psychologie? *Psychologische Rundschau, 61*(4), 191–198.

Jäncke, L. & Petermann, F. (2010). Zum Verhältnis von Biologie und Psychologie. *Psychologische Rundschau, 61*(4), 175–179.

Janke, W. (1976). Psychophysiologische Grundlagen des Verhaltens. In M. von Kerekjarto (Hrsg.), *Medizinische Psychologie* (S. 1–101). Heidelberg: Springer.

Janke, W. & Kallus, K. W. (1995). Reaktivität. In M. Amelang (Hrsg.), *Verhaltens- und Leistungsunterschiede, Enzyklopädie der Psychologie, Themenbereich C (Theorie und Forschung), Serie VIII (Differentielle Psychologie und Persönlichkeitsforschung), Bd. 2, Verhaltens- und Leistungsunterschiede* (S. 1–89). Göttingen: Hogrefe.

Janke, W., Schmidt-Daffy, M., & Debus, G. (Hrsg.). (2008). *Experimentelle Emotionspsychologie. Methodische Ansätze, Probleme, Ergebnisse*. Lengerich: Pabst.

Janssen, E. (2002). Psychophysiological measurement of sexual arousal. In M. W. Wiederman & B. E. Whitley, Jr. (Eds.), *Handbook for conducting research on human sexuality* (pp. 139–171). Mahwah: Erlbaum.

Jeste, S. S. & Nelson, C. A. (2009). Event related potentials in the understanding of autism spectrum disorders: An analytical review. *Journal of Autism and Developmental Disorders, 39*(3), 495–510.

Kallus, K. W. (1992). *Ausgangszustand und Beanspruchung.* Weinheim: PVU.

Kaschak, E. & Tiefer, L. (2002). *A new view of women's sexual problems*. New York: Harworth.

Kätsyri, J., Ravaja, N., & Salminen, M. (2012). Aesthetic images modulate emotional responses to reading news messages on a small screen: A psychophysiological investigation. *International Journal of Human-Computer Studies, 70*(1), 72–87.

Kempter, G. (1997). *Psychophysiologische Fernsehwirkungsforschung*. Hamburg: Kovac.

Kempter, G. & Bente, G. (2004). Psychophysiologische Wirkungsforschung: Grundlagen und Anwendungen. In R. Mangold, P. Vorderer & G. Bente (Hrsg.), *Lehrbuch der Medienpsychologie* (S. 271–296). Göttingen: Hogrefe.

Klimesch, W. (1997). EEG-Alpha rhythms and memory processes. *International Journal of Psychophysiology, 26*(1–3), 319–340.

Köhler, T. (1992). *Die Zahl aktiver Schweißdrüsen (PSI, Palmar Sweat Index) als Aktivierungsparameter in Labor- und Feldstudien*. Frankfurt: Lang.

Köhler, T. (1995). *Psychosomatische Krankheiten* (3. Aufl.). Stuttgart: Kohlhammer.

Köhler, T. (2001). *Biopsychologie*. Stuttgart: Kohlhammer.

Krahé, B., Möller, I., Huesmann, L. R., Kirwil, L., Felber, J., & Berger, A. (2011). Desensitization to media violence: links with habitual media violence exposure, aggressive cognitions, and aggressive behavior. *Journal of Personality and Social Psychology, 100*(4), 630–646.

Kramme, R. (2006). *Medizintechnik. Verfahren, Systeme, Informationsverarbeitung* (3. Aufl.). Heidelberg: Springer.

Kukkonen, T. M., Binik, Y. M., Amsel, R., & Carrier, S. (2010). An evaluation of the validity of thermography as a physiological measure of sexual arousal in a non-university adult sample. *Archives of Sexual Behavior, 39*(4), 861–873.

Laan, E. & Both, S. (2008). What makes women experience desire? *Feminism and Psychology, 18*(4), 505–514.

Lacey, J. I., Bateman, D. E., & Van Lehan, R. (1953). Autonomic response specifity: An experimental study. *Psychosomatic Medicine, 15*(1), 8–21.

Lachnit, H., Kinder, A., & Reinhard, G. (2002). Are rules applied in pavlovianelectrodermal conditioning with humans general or outcome specific? *Psychophysiology, 39*, 380–387.

Laws, R. (2003). Penile plethysmography: Will we ever get it right? In T. Ward, R. Laws & S. Hudson (Eds.), *Sexual deviance: Issues and controversies* (pp. 82–102). Thousand Oaks: Sage.

Lockhart, R. A. (1975). Jung: A forgotten psychophysiologist remembered. *Polygraph, 4*(1), 18–32.

Loftus, G. R. (1972). Eye fixations and recognition memory for pictures. *Cognitive Psychology, 3*(4), 525–551.

Lohse, G. L. (1997). Consumer eye movement patterns on yellow pages advertising. *Journal of Advertising, 26*(1), 61–73.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge: MIT Press.

Macar, F. & Vidal, F. (2004). Event-related potentials as indices of time processing: A review. *Journal of Psychophysiology, 18*(2–3), 89–104.

Mackworth, N. H. & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Attention, Perception and Psychophysics, 2*(11), 547–552.

Malik, M., Bigger, J., Camm, A., & Kleiger, R. (1996). Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *Circulation, 93*, 1043–1065.

Mausfeld, R. (2010). Psychologie, Biologie, kognitive Neurowissenschaften. Zur gegenwärtigen Dominanz neuroreduktionistischer Positionen und zu ihren stillschweigenden Grundannahmen. *Psychologische Rundschau, 61*(4), 180–190.

Meißner, M. & Decker, R. (2010). Eye-tracking information processing in choice-based conjoint analysis. *International Journal of Market Research, 52*(5), 591–610.

Mohan, S. M. & Wei, L. Y. (2002). Modulation of pulse rate by unilateral nostril breathing. *Journal of Indian Psychology, 20*(1), 32–37.

Nickel, P. & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1–Hz component of heart rate variability as an indicator of mental workload. *Human Factors, 45* (4), 575–590.

Öquist, G. & Lumsden, J. (2008). Three eye movement studies of mobile readability. In J. Lumsden (Ed.), *Handbook of research on user interface design and evaluation for mobile technology* (pp. 945–971). Hershey: IRMA.

Palti, Y. & Bercovici, B. (1967). Photoplethysmographic study of the vaginal blood pulse. *American Journal of Obstetrics and Gynecology, 97*(2), 143–153.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research, 52*(4), 1073–1081.

Plaud, J. J., Gaither, G. A., Hegstad, H. J., Rowan, L., & Devitt, M. K. (1999). Volunteer bias in human psychophysiological sexual arousal research: To whom do our research results apply? *Journal of Sex Research, 36*(2), 171–179.

Poeck, K. (1990). *Neurologie* (7. Aufl.). Heidelberg: Springer.

Pritchard, W., Sokhadze, E., & Houlihan, M. (2004). Effects of nicotine and smoking on event-related potentials: A review. *Nicotine and Tobacco Research, 6*(6), 961–984.

Racine, E., Bell, E., & Illes, J. (2010). Can we read minds? Ethical challenges and responsibilities in the use of neuroimaging research. In J. J. Giordano & B. Gordijn (Eds.), *Scientific and philosophical perspectives in neuroethics* (pp. 244–269). Cambridge: Cambridge University Press.

Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., & Keltikangas-Järvinen, L. (2008). The psychophysiology of James Bond: Phasic emotional responses to violent video game events. *Emotion, 8*(1), 114–120.

Rösler, F. (1996). Methoden der Psychophysiologie. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden.* (S. 490–514). Weinheim: Beltz.

Rösler, F. (2011). *Psychophysiologie der Kognition. Eine Einführung in die Kognitive Neurowissenschaft*. Heidelberg: Spektrum.

RRscope. (2012). Einführung in die Herzfrequenzanalyse. Abgerufen 28. Februar, 2012, unter http://www.rrscope.com/data/Einfuehrung_Herzfrequenzanalyse.shtml

Schandry, R. (1996). *Lehrbuch Psychophysiologie*. Weinheim: Beltz.

Schandry, R. (2003). *Biologische Psychologie*. Weinheim: Beltz.

Schauer, M., Elbert, T., Gotthardt, S., Rockstroh, B., Odenwald, M., & Neuner, F. (2006). Wiedererfahrung durch Psychotherapie modifiziert Geist und Gehirn. *Verhaltenstherapie, 16*, 96–103.

Schenck, E. (1992). *Neurologische Untersuchungsmethoden* (4. Aufl.). Stuttgart: Thieme.

Schmidt-Atzert, L. (1993). *Die Entstehung von Gefühlen. Vom Auslöser zur Mitteilung.* Heidelberg: Springer.

Schmidt-Atzert, L. (1995). Mimik und Emotionen aus psychologischer Sicht. In G. Debus, G. Erdmann & K. W. Kallus (Hrsg.), *Biopsychologie von Streß und emotionalen Reaktionen* (S. 53–66). Göttingen: Hogrefe.

Schum, J. L., Jorgensen, R. S., Verhaeghen, P., Sauro, M., & Thibodeau, R. (2003). Trait anger, anger expression, and ambulatory blood pressure: A meta-analytic review. *Journal of Behavioral Medicine, 26*(5), 395–415.

Seto, M. C., Lalumière, M. L., & Blanchard, R. (2000). The discriminative validity of a phallometric test for pedophilic interests among adolescent sex offenders against children. *Psychological Assessment, 12*(3), 319–327.

Siddle, D. (1983). *Orienting and habituation. Perspectives in human research*. Hoboken: Wiley.

Silbernagl, S. & Despopoulos, A. (1991). *Taschenatlas der Physiologie* (4. Aufl.). Stuttgart: Thieme.

Sommer, W., Ulrich, R., & Leuthold, H. (1996). Das lateralisierte Bereitschaftspotential als psychophysiologischer Zugang bei der Untersuchung kognitiver Prozesse. *Psychologische Rundschau, 47*(1), 1–14.

Starr, M. S. & Rayner, K. (2001). Eye movements during reading: Some current controversies. *Trends in Cognitive Sciences, 5*(4), 156–163.

Steffen, P. R., Smith, T. B., Larson, M., & Butler, L. (2006). Acculturation to western society as a risk factor for high blood pressure: A meta-analytic review. *Psychosomatic Medicine, 68*(3), 386–397.

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica, 106*(1–2), 147–246.

Sudsuang, R., Chentanez, V., & Veluvan, K. (1991). Effect of Buddhist meditation on serum cortisol and total protein levels, blood pressure, pulse rate, lung volume and reaction time. *Physiology and Behavior, 50*(3), 543–548.

Suschinsky, K. D., Lalumière, M. L., & Chivers, M. L. (2009). Sex differences in patterns of genital sexual arousal: Measurement artifacts or true phenomena? *Archives of Sexual Behavior, 38*, 559–573.

Telles, S. & Vani, P. R. (2002). Increase in voluntary pulse rate reduction achieved following yoga training. *International Journal of Stress Management, 9*(3), 236–239.

Toledano, R. & Pfaus, J. (2006). The sexual arousal and desire inventory (SADI): A multidimensional scale to assess subjective sexual arousal and desire. *Journal of Sexual Medicine, 3*(5), 853–877.

Venables, P. H. & Christie, M. (1980). Electrodermal activity. In I. Martin & P. H. Venables (Eds.), *Technics in Psychophysiology*. Hoboken: Wiley.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274–290.

Vul, E. & Kanwisher, N. (2010). Begging the question: The non-independence error in fMRI data analysis. In S. J. Hanson & M. Bunzl (Eds.), *Foundational Issues in human brain mapping* (pp. 71–92). Massachusetts: MIT Press.

Walschburger, P. (1975). Zur Standardisierung und Interpretation elektrodermaler Meßwerte in psychologischen Experimenten. *Zeitschrift für experimentelle und angewandte Psychologie, 22*, 514–533.

Warren, G., Schertler, E., & Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior, 33*(1), 59–69.

Weber, R., Ritterfeld, U., & Mathiak, K. (2006). Does playing violent video games induce aggression? Empirical evidence of a functional magnetic resonance imaging study. *Media Psychology, 8*(1), 39–60.

Weig, W. (2006). Sexuelle Funktionsstörungen aus nervenärztlicher Perspektive. *Nervenarzt, 77*(1), 101–109.

Wilder, M. (1931). Das „Ausgangswertgesetz“, ein unbeachtetes biologisches Gesetz und seine Bedeutung für Forschung und Praxis. *Zeitschrift für Neurologie, 137*(1), 317–338.

Wilson, R. J. & Mathon, H. F. (2006). *Remembering Kurt Freund (1914–1996). ATSA Forum*. Beaverton: Association for the Treatment of Sexual Abusers.

Yom, M. & Wilhelm, T. H. (2004). Methoden und Erkenntnisse der Web-Usability-Forschung. In A. Zerfaß & H. Zimmermann (Hrsg.), *Usability von Internet-Angeboten. Grundlagen und Fallstudien* (S. 25–38). Stuttgart: MFG.

Yuille, J. C. & Hare, R. D. (1980). A psychophysiological investigation of short-term memory. *Psychophysiology, 17*(5), 423–430.

Zajonc, R. B., Murphy, S. T., & Inglehart, M. (1989). Feeling and facial efference: Implications for the vascular theory of emotion. *Psychological Review, 96*(3), 395–416.

Zhou, X.-X. & Jin, Z.-C. (2009). The Influence of involvement on information processing of rational advertising appeals. *Acta Psychologica Sinica, 41*(4), 357–366.

10.6 Dokumentenanalyse

Lernziele

- Wissen, was eine wissenschaftliche Dokumentenanalyse als Datenerhebungsmethode auszeichnet.
- Unterschiede und Gemeinsamkeiten vorgefundener und forschungsgenerierter Dokumente aufzeigen können.
- Verschiedene Zugangswege zu sozialwissenschaftlich aussagekräftigen vorhandenen Dokumenten kennen.
- Verschiedene Ansätze der qualitativen Datenanalyse im Kontext der Dokumentenanalyse im Hinblick auf Vorgehensweise und Anwendungsfelder voneinander abgrenzen können.
- Die Arbeitsschritte einer quantitativen Inhaltsanalyse darstellen und umsetzen können.
- Die Besonderheiten der quantitativen Inhaltsanalyse von Online-Dokumenten und von visuellen Dokumenten kennen.
- Die jeweiligen Vor- und Nachteile qualitativer und quantitativer Typen von Dokumentenanalysen abwägen können.

Um empirische Daten über menschliches Erleben und Verhalten zu gewinnen, können Forschende das Geschehen selbst **beobachten** (▶ Abschn. 10.1) oder die Beteiligten **befragen**, d. h. Selbstauskünfte per **Interview** (▶ Abschn. 10.2) oder **Fragebogen** (▶ Abschn. 10.3) einholen. Zudem können sie **psychologische Testverfahren** (▶ Abschn. 10.4) einsetzen, etwa um Leistungen von Personen objektiv – d. h. jenseits von Selbstauskünften – mittels entsprechender Testaufgaben zu erfassen (z. B. Intelligenztest). Nicht zuletzt lassen sich auch **physiologische Messungen** (▶ Abschn. 10.5) durchführen, um objektive Daten (d. h. wiederum keine Selbstauskünfte) über körperliche Reaktionen zu gewinnen (z. B. Blutdruck) und diese mit Indikatoren des Verhaltens und Erlebens in Zusammenhang zu bringen (z. B. Angst).

Neben diesen bereits behandelten grundlegenden **Methoden der Datenerhebung** stellt die **Dokumentenanalyse** eine weitere eigenständige Verfahrensgruppe dar, um empirische Daten zu gewinnen und auszuwerten. Dabei wird bei einer **genuinen Dokumentenanalyse** auf bereits **vorhandene bzw. vorgefundene Dokumente** („extant documents“) zurückgegriffen, die völlig unabhängig vom Forschungsprozess produziert wurden und als Manifestationen menschlichen Erlebens und Verhaltens angesehen werden können. Dazu zählt z. B. die **Sammlung** von amtlichen Akten, Schulaufsätzen, Briefen, Beiträgen in Onlineforen, Fernsehsendungen, Websites, Kinderzeichnungen, Kontaktanzeigen, Gerichtsurteilen, Sitzungsprotokollen, Urlaubsfotos und so weiter. All diese Dokumente haben qualitativen Charakter, denn es handelt sich nicht um numerisches Ausgangsmaterial, sondern um **verbale bzw. narrative, visuelle, audiovisuelle oder multimediale Dokumente**, deren Bedeutungen es in erster Linie durch eine interpretierend-qualitative Datenanalyse zu rekonstruieren gilt. Aber auch mittels quantitativer Verfahren lassen sich die Ausprägungen einzelner, theoretisch relevanter Merkmale von vorgefundenen Dokumenten erheben. Vorgefundene Offline- und Online-Dokumente können auch als **Verhaltensspuren** aufgefasst werden, denn in ihnen schlagen sich vergangene Verhaltensweisen nieder. Während die Sammlung und Auswertung von relativ einfachen (z. B. Zusammensetzung des Hausmülls; Inschriften auf Toilettenwänden) sowie von quantifizierten Verhaltensspuren (z. B. Logfiles, in denen Internet-Nutzungsverhalten protokolliert ist) der **nonreaktiven Beobachtung** (▶ Abschn. 10.1.3 „Nonreaktive Beobachtung von Verhaltensspuren“) zugerechnet wird, fallen eben Verhaltensspuren in Form von mehr oder minder umfassenden interpretierbaren Dokumenten in den Bereich der Dokumentenanalyse. Da viele Dokumente in Archiven verwaltet werden, spricht man im Zusammenhang mit Dokumentenanalysen auch von **Archivforschung** („archival research“).

Die Auswertung von **forschungsgenerierten Dokumenten** („research generated documents“) ist von der genuinen Dokumentenanalyse als Methode der Sammlung und Analyse vorgefundener Dokumente abzugrenzen. Bei forschungsgenerierten Dokumenten handelt es sich um Dokumente, die während des Forschungsprozesses **durch andere Datenerhebungsmethoden erzeugt** wurden (z. B. Transkripte, die in qualitativen Interviewstudien anfallen; Feldnotizen, die in qualitativen Beobachtungsstudien erzeugt werden; Aufsätze oder offene Kommentare, die im Rahmen von qualitativen Fragebogenstudien generiert werden). Die Arbeit mit forschungsgenerierten Dokumenten wird nicht als Dokumentenanalyse bezeichnet, da die Dokumentenanalyse die Sammlung vorhandener (d. h. unabhängig vom Forschungsprozess produzierter) Dokumente meint. Stattdessen wird bei der Arbeit mit forschungsgenerierten Dokumenten die Methode, durch welche die Dokumente generiert wurden, als die Datenerhebungsmethode der betreffenden Studie betrachtet.

Aus Sicht der **Datenerhebung** unterscheiden sich vorgefundene und forschungsgenerierte Dokumente dahingehend, dass bei vorgefundenen Dokumenten keinerlei Beeinflussung der Dokumente durch Forschungstätigkeit erfolgt (nonreaktive Methode). Der Vorteil der **Nonreaktivität** geht mit zwei Nachteilen einher: 1. Die **Kontextbedingungen der Dokumentenproduktion** sind bei vorgefundenen Dokumente teilweise unbekannt (z. B. fehlen bei vorgefundenen Online-Dokumenten oft genaue Angaben zu den Textproduzierenden, von denen nur ein

Pseudonym bekannt ist). 2. Im Zuge einer genuinen Dokumentenanalyse mit vorgefundenen Dokumenten kann die Dokumentenproduktion nicht im Hinblick auf die größtmögliche **Aussagekraft für das Forschungsproblem** gestaltet werden (z. B. gehen vorgefundene Dokumente oft nur teilweise auf für das Forschungsproblem relevante Themen ein, während mittels qualitativer Befragung umfassende Dokumente erzeugt werden können, in denen alle für das Forschungsproblem relevanten Aspekte angesprochen werden). Umgekehrt sind bei forschungsgenerierten Dokumenten die Produktionsbedingungen exakt bekannt und können durch die Forschenden im Sinne der Erhöhung der Aussagekraft gesteuert werden, dafür sind die resultierenden Dokumente durch den Forschungsprozess künstlich erzeugt und somit potenziell verzerrt, wobei Art und Konsequenzen der Verzerrung im Zusammenhang mit dem Forschungsproblem zu reflektieren sind (z. B. sozial erwünschtes Antworten bei heiklen Themen, um gegenüber dem Interviewer das Gesicht zu wahren).

Wenn es um die **Datenanalyse** von Dokumenten geht, so werden **vorgefundene und forschungsgenerierte Dokumente weitgehend gleich behandelt**. Generell können Dokumente als qualitatives Rohdatenmaterial zum einen direkt qualitativ ausgewertet werden, wozu unterschiedliche Techniken der **qualitativen Datenanalyse** zum Einsatz kommen (z. B. qualitative Inhaltsanalyse, Interpretative Phänomenologische Analyse oder die Kodierung gemäß Grounded-Theory-Methodologie). Zum anderen können die qualitativen Dokumente durch eine quantitative Inhaltsanalyse in Messwerte überführt werden, die anschließend einer **quantitativen Datenanalyse** unterzogen, d. h. statistisch ausgewertet werden (Abb. 10.39).

Bei einer genuinen *Dokumentenanalyse* erfolgt die Datenerhebung durch Sammlung vorgefundener (d. h. nicht durch den Forschungsprozess erzeugter) Dokumente, so dass im Unterschied zu forschungsgenerierten Dokumenten Nonreaktivität gegeben ist. Für die anschließende Datenauswertung können qualitative und quantitative Datenanalysemethoden verwendet werden – und zwar prinzipiell dieselben, die auch auf forschungsgenerierte Dokumente (z. B. Beobachtungsprotokolle, Interview-Transkripte) anwendbar sind.

Die forschungslogische Einordnung der **quantitativen Inhaltsanalyse** sorgt oft für Verwirrung. Sie wird zuweilen als Datenerhebungsmethode oder auch als Datenauswertungsverfahren bezeichnet. Tatsächlich hat sie jedoch eine Zwischenstellung: Sie dient der Quantifizierung qualitativer Dokumente (seien sie vorgefunden oder forschungsgeneriert) und **bereitet eine statistische Analyse** vor. Jede genuine quantitative Dokumentenanalyse (die mit vorgefundenen Dokumenten arbeitet) ist zwingend auf eine quantitative Inhaltsanalyse angewiesen. Aber nicht jede quantitative Inhaltsanalyse findet im Rahmen einer genuinen Dokumentenanalyse statt; sie kann auch im Rahmen einer Beobachtungs- oder Interviewstudie zum Einsatz kommen, um die forschungsgenerierten qualitativen Dokumente für die statistische Auswertung in Messwerte zu transformieren (Abb. 10.39).

Ein **Dokument** („document“) stellt formal betrachtet einen materiellen oder virtuellen **Informationscontainer** (z. B. Schriftstück auf Papier, digitale Datei) samt dessen Inhalten dar. Die Inhalte des Dokuments können dabei in ganz unterschiedlichen **Kommunikationscodes** repräsentiert sein, wobei schriftlicher Text und Bilder von herausragender Bedeutung sind. Aber auch multimediale oder hypermediale Dokumente spielen in Praxis und Forschung eine wichtige Rolle. Ein Dokument muss als Einheit wahrnehmbar, speicherbar bzw. archivierbar und abrufbar sein. Hinsichtlich ihrer **Produktions- und Nutzungskontexte** unterscheiden wir aus dem Blickwinkel der Datenerhebung grob zwischen persönlichen Dokumenten und offiziellen Dokumenten (Scott, 1990):

- **Persönliche Dokumente** („personal documents“) stammen aus dem Bereich der intra- und interpersonalen **informellen Kommunikation** – z. B. Tagebücher, Einkaufszettel, Briefe, Familienfotos, Kontaktanzeigen, Urlaubsvideos, E-Mails etc.
- **Offizielle Dokumente** („official documents“) stammen aus dem Bereich der **formalen Kommunikation** von Organisationen, darunter als Akteure z. B. der Staat oder privatwirtschaftliche Unternehmen – z. B. Dienstvorschriften von Behörden, Sitzungsprotokolle von Firmen, Jahresberichte von Unternehmen, Gerichtsakten, Gesetzestexte. Auch massenmediale Darstellungen (Zeitungsartikel, Radiobeiträge etc.) lassen sich den offiziellen Dokumenten zuordnen, die durch ihre Verfügbarkeit in Archiven meist leichter zu beschaffen sind als persönliche Dokumente.

Methodologische Beiträge zum Stellenwert vorgefundener Dokumente in der empirischen Sozialforschung betonen, dass die Text- oder Bilddokumente nicht ausschließlich als bloße Informationscontainer zu verstehen sind, sondern in ihren Verwendungskontexten sozusagen als **soziale Akteure** fungieren (Prior, 2008): Ebenso wie man in einer Interviewstudie die Äußerungen einer Befragungsperson vor dem Hintergrund ihrer Merkmale sowie der jeweiligen Interviewsituation einordnet, sollten dementsprechend auch bei der wissenschaftlichen Analyse vorgefundener Dokumente deren Entstehungs- und Verwendungszusammenhänge mitberücksichtigt werden (Muckel, 1997).

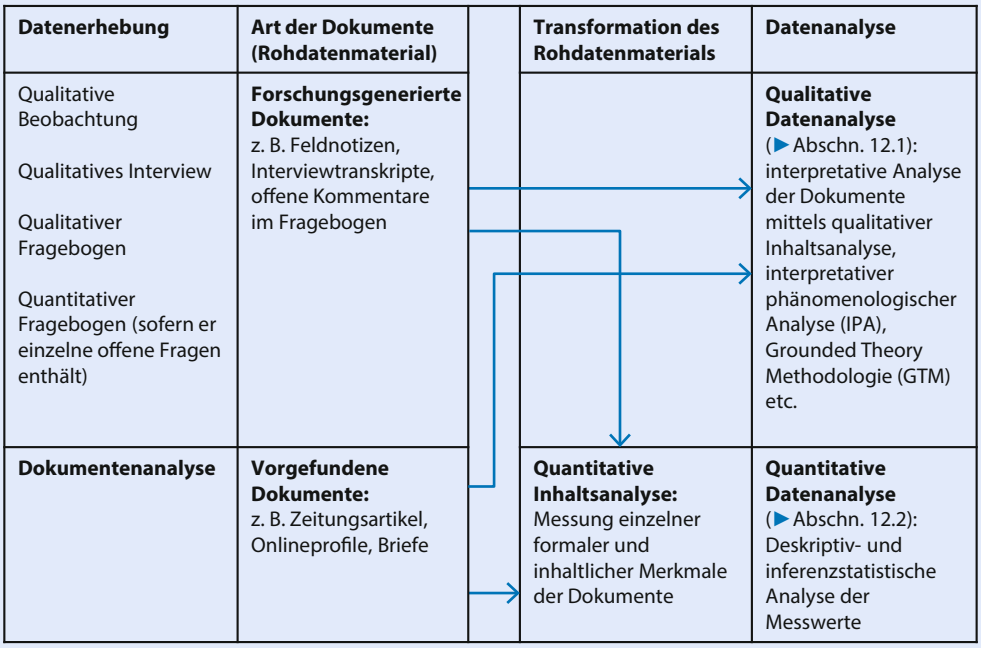


Abb. 10.39 Einordnung der Dokumentenanalyse in den Forschungsprozess

Ein anderes Unterscheidungskriterium betrifft die **Zugänglichkeit der Dokumente**: Ein Teil der Dokumente ist öffentlich zugänglich (z. B. Geschäftsberichte von Aktiengesellschaften, Kontaktanzeigen in der Presse und auf Online-Dating-Portalen), andere haben nicht-öffentlichen Charakter und müssen mit Zustimmung der Produzentinnen oder Besitzer gesammelt werden (z. B. Familienfotos, Sitzungsprotokolle). Damit sind Fragen von Urheberrecht und Datenschutz bzw. von Forschungsethik (▶ Abschn. 4.1) berührt.

Wird eine **qualitative Dokumentenanalyse** durchgeführt, so muss das gesammelte qualitative Ausgangsmaterial einer interpretativen Auswertung unterzogen werden. Dabei kommen die **üblichen Methoden der qualitativen Datenanalyse** (▶ Abschn. 12.1) zum Einsatz, etwa die qualitative Inhaltsanalyse oder die Kodierung gemäß Grounded-Theory-Methodologie. In der Phase der Datenanalyse unterscheidet sich eine qualitative Dokumentenanalyse dementsprechend nicht grundsätzlich von einer qualitativen Interviewstudie, deren Transkripte (als Textdokumente) am Ende interpretativ ausgewertet werden.

Anders ist es bei der **quantitativen Dokumentenanalyse**. Hier muss das qualitative Ausgangsmaterial zunächst quantifiziert werden, bevor es dann einer üblichen quantitativen Datenanalyse mit statistischen Verfahren (▶ Abschn. 12.2) unterzogen werden kann. Als Zwischenschritt hat sich in den Sozialwissenschaften die **quantitative Inhaltsanalyse** etabliert. Sie stellt eine Methode bereit, um die formalen und inhaltlichen Merkmale textueller und/oder nicht-textueller Dokumente zu operationalisieren bzw. zu messen und dadurch quantitative Daten zu generieren.

Während alle anderen in ▶ Kap. 10 behandelten qualitativen Datenerhebungsmethoden relativ aufwändig und anspruchsvoll hinsichtlich der zugrunde liegenden Aktivitäten der Forschenden sind (z. B. notwendige Schulungen für Beobachter/innen oder Interviewer/innen) und bei allen anderen behandelten quantitativen Datenerhebungsmethoden die Erstellung der standardisierten Datenerhebungsinstrumente voraussetzungsvoll ist (z. B. notwendige Arbeitsschritte zur Konstruktion von psychometrischen Fragebögen und psychologischen Tests), stellt die systematische Sammlung von Dokumenten **ein vergleichsweise einfaches Verfahren der Datenerhebung** dar: Weder umfangreiche Schulungen noch standardisierte Instrumente sind dafür notwendig. Je nach Art der Dokumente (persönlich vs. offiziell, öffentlich vs. nicht-öffentlich) sind **unterschiedliche Beschaffungswege** zu beschreiten. Dabei muss der **Stichprobenplan** mit Blick auf das Forschungsproblem gut überlegt werden, wobei alle bekannten qualitativen und quantitativen Stichprobenarten (▶ Kap. 9) einsetzbar sind. Bei der in der Kommunikationswissenschaft einschlägigen Sammlung und

Auswertung von Mediendokumenten stellt der Umgang mit der Periodizität der Berichterstattung ein Problem dar. Hier werden zuweilen Stichproben nach dem Prinzip der natürlichen Woche oder der **künstlichen Woche** gebildet (Früh, 1998; Hester & Dougall, 2007; Riffe, Aust, & Lacy, 1993), um wochentagsbezogene oder saisonale Verzerrungen publizistischer Medien auszuschließen.

Die Besonderheit der Dokumentenanalyse aus Sicht der quantitativen Forschung liegt darin, dass unstrukturiertes qualitativ vorliegendes Material zunächst mittels **quantitativer Inhaltsanalyse** der späteren statistischen Datenanalyse zugänglich gemacht werden muss. Das vorliegende Kapitel widmet sich dementsprechend der quantitativen Inhaltsanalyse von Text-, Online- und Bilddokumenten.

Um die quantitative Inhaltsanalyse als standardisierte Herangehensweise an Dokumente besser zu verstehen, werden im Folgenden auch die **qualitative Inhaltsanalyse** sowie zwei weitere qualitative Auswertungsmethoden für Dokumente näher erläutert. Da es sich bei diesen qualitativen Techniken jedoch gleichzeitig um allgemeine qualitative Auswertungsmethoden handelt, wird hier der qualitativen Datenanalyse (► Abschn. 12.1) vorgegriffen.

Bei einer *qualitativen Dokumentenanalyse* stellt die qualitative Inhaltsanalyse – oder eine andere qualitative Auswertungsmethode – bereits die eigentliche qualitativ-interpretative Datenanalyse dar. Demgegenüber bereitet bei einer rein *quantitativen Dokumentenanalyse* die quantitative Inhaltsanalyse der Dokumente lediglich die daran anschließende quantitativ-statistische Datenanalyse vor, indem sie das qualitative Ausgangsmaterial in quantitative/numerische Daten bzw. Messwerte auf prinzipiell allen vier Skalenniveaus überführt.

Generell wächst die Bedeutung der Sammlung und Auswertung vorgefundener Dokumente in den Sozial-, Human- und Kulturwissenschaften. Denn im Zuge der zunehmenden **Mediatisierung bzw. Digitalisierung sozialen Handelns** in den hochtechnisierten Gesellschaften fallen immer mehr potenziell aussagekräftige Dokumente menschlichen Erlebens und Verhaltens im Alltag an: Menschen präsentieren sich auf Social-Networking-Plattformen massenhaft mit eigenen Online-Profilen, schreiben Internettagebücher und tauschen sich per Online-Chat, Messenger, E-Mail oder in Online-Foren aus, wobei digitale Dokumente der jeweiligen Selbstdarstellungs-, Selbstverständigungs- und interpersonalen sowie Gruppenkommunikationsprozesse entstehen. Darüber hinaus sind offizielle Dokumente (Regierungsberichte, Gesetzestexte, politische Reden etc.) heute ebenfalls über **Online-Archive und Datenbanken** in großem Umfang und in globalem Maßstab öffentlich zugänglich und stehen damit als Datenmaterial für wissenschaftliche Dokumentenanalysen zur Verfügung.

Die große Menge verfügbarer digitaler Dokumente erfordert den Einsatz **computergestützter Recherche-, Archivierungs- und Analysetools**, um das Rohdatenmaterial aus dem Internet für wissenschaftliche Dokumentenanalysen nutzbar zu machen. Die wachsende Bedeutung digitaler Dokumente als Datenmaterial für die Geistes- und Sozialwissenschaften schlägt sich auch in neuen Wissenschaftsdisziplinen nieder wie den **Digital Humanities** (digitale Geisteswissenschaften, z. B. http://www.digitalhumanities.org) und den **Computational Social Sciences** (computergestützte Sozialwissenschaften; Lazer et al., 2009), aus denen heraus entsprechende methodische Innovationen v. a. im Bereich der Dokumentenanalyse stammen, samt der dafür benötigten Computersoftware (z. B. Digital Methods Initiative DMI: http://www.digitalmethods.net).

10.6.1 Wissenschaftliche Dokumentenanalyse

Um zu verstehen, was eine genuine wissenschaftliche Dokumentenanalyse als Datenerhebungsverfahren auszeichnet und worin sie anderen Datenerhebungsmethoden – wie z. B. der wissenschaftlichen Befragung – überlegen ist, wenden wir uns zunächst ihrer **Definition und Indikation** zu. Dabei wird auch erläutert, inwiefern sich die wissenschaftliche Dokumentenanalyse von nichtwissenschaftlichen Dokumentenanalysen unterscheidet. Schließlich werden verschiedene **Typen der Dokumentenanalyse** voneinander abgegrenzt.

Definition und Indikation der wissenschaftlichen Dokumentenanalyse

Das Zusammentragen und Interpretieren von vorhandenen Dokumenten findet auch **außerhalb sozial- und kulturwissenschaftlicher Forschung** statt, etwa im Rahmen von Ausstellungen in Museen, Kunstprojekten und Sammelbänden, in denen Fotos oder Briefe als Alltagsdokumente publiziert werden (z. B. Sammlung von Briefen von Bundeswehrangehörigen aus Afghanistan: Baumann, Langeder, Much, Obermayer, & Storz, 2011). Im Bereich des Bürgerjournalismus haben sich kollaborative Projekte der Dokumentenanalyse bewährt: So stellte die Zeitung *The Guardian* die Spesenabrechnungen der britischen Parlamentsmitglieder – knapp eine halbe Million Seiten – online und bat die Leserschaft, diese auf Ungereimtheiten zu prüfen (http://mps-expenses.guardian.co.uk/). Auch die kollaborative Plagiatssuche in Doktorarbeiten ist ein Beispiel für Dokumentenanalyen.

Im Unterschied zu außerwissenschaftlichen Initiativen der Sammlung und Auswertung von Dokumenten erfolgt die wissenschaftliche Dokumentenanalyse im Kontext eines regelgeleiteten qualitativen und/oder quantitativen Forschungsprozesses mit dem Ziel wissenschaftlichen Erkenntnisgewinns (► Abschn. 1.1). Wichtig ist, dass die Dokumentenanalyse heute zunehmend breiter verstanden wird als eine reine Textanalyse. Denn neben **textuellen Dokumenten** gewinnen sog. **nicht-textuelle Dokumente** stark an gesellschaftlicher Bedeutung. Seit den 1990er-Jahren ist von einer **visuellen Wende** („visual/iconic/pictorial turn“) die Rede, um die wachsende Relevanz und Ausdifferenzierung von Bildern gegenüber Wörtern in unserer Kultur zu betonen („visual culture“) und daraus die Forderung nach einer eigenen Bildwissenschaft abzuleiten. Im Kontext der Dokumentenanalyse wird das Dokument allgemein als Träger von Zeichen oder Informationen verstanden. Häufig haben wir es heute in der Computer- und Internet-Gesellschaft bei vorgefundenen Dokumenten mit **digitalen Dokumenten** zu tun, die jeweils in Form einer Datei auftreten.

Wissenschaftliche Dokumentenanalyse – Unter einer wissenschaftlichen Dokumentenanalyse („scientific document analysis“) als Datenerhebungsmethode verstehen wir die zielgerichtete, systematische und regelgeleitete **Sammlung und Archivierung** von vorhandenen (d. h. unabhängig vom Forschungsprozess produzierten) Dokumenten als Manifestationen menschlichen Erlebens und Verhaltens. Dabei kann es sich inhaltlich um persönliche oder offizielle Dokumente sowie formal um textuelle/verbal-schriftliche sowie um nicht-textuelle (visuelle, auditive, audiovisuelle, multimediale, hypermediale etc.) Dokumente handeln.
An die Sammlung, Archivierung und Aufbereitung des Rohdatenmaterials schließt sich eine **Auswertung der Dokumente** hinsichtlich ihrer inhaltlichen und formalen Merkmale an. Bei der **qualitativen Dokumentenanalyse** stellt die interpretative Auswertung der Dokumente bereits die eigentliche qualitative Datenanalyse dar. Bei der **quantitativen Dokumentenanalyse** wird das qualitative Ausgangsmaterial mittels quantitativer Inhaltsanalyse in Messwerte überführt, die anschließend einer üblichen quantitativ-statistischen Datenanalyse unterzogen werden.

Folgende Umstände sprechen für eine besondere Eignung (Indikation) der Dokumentenanalyse zur Bearbeitung eines Forschungsproblems:

1. Manche **Forschungsthemen** lassen sich mithilfe der Dokumentenanalyse besonders gut bearbeiten, weil die interessierenden Sachverhalte sich systematisch in Dokumenten niederschlagen: So greift z. B. die **medien- und kommunikationswissenschaftliche Produkt- und Medieninhaltsforschung** primär auf Dokumentenanalysen ausgewählter Medienarstellungen zurück. Auch **politikwissenschaftliche Studien** stützen sich oft auf Dokumente wie z. B. Zeitungsartikel über Politikerinnen und Politiker, Gesetzestexte, Gerichtsurteile, Protokolle parlamentarischer Debatten oder Manuskripte politischer Reden. Forschungsprobleme mit **historischem Bezug** sind oft nur anhand von Dokumenten zu bearbeiten (z. B. Tagebücher, Briefe, Berufsordnungen, Studienordnungen, Ratgeberliteratur vergangener Jahrzehnte oder Jahrhunderte), weil zurückliegende Ereignisse nicht mehr beobachtet und verstorbene Zeitzeugen nicht mehr befragt, mit psychologischen Tests untersucht oder physiologischen Messungen unterzogen werden können. Auch wenn größere soziale Gebilde wie **Institutionen** im Fokus des Forschungsinteresses stehen, können Dokumente eine wichtige Datenquelle darstellen: z. B. schulische Curricula, Hausordnungen, Dienstvorschriften, Kranken-, Personal- und Gerichtsakten, Sitzungsprotokolle, Vereinssatzungen etc. Auch Fragestellungen zur **visuellen Kultur** lassen sich mit Dokumentenanalysen bearbeiten (z. B. Sammlung und Auswertung von digitalen Fotos und Videos).
2. Die genuine Dokumentenanalyse zählt zu den **nonreaktiven Verfahren der Datenerhebung**, da auf Dokumente zurückgegriffen wird, die unabhängig vom Forschungsprozess erzeugt wurden. Somit wird Rohdatenmaterial bearbeitet, das hinsichtlich Form und Inhalt nicht durch den Forschungsprozess beeinflusst ist.
3. Je nach Art der untersuchten Dokumente kann es sich bei der Dokumentenanalyse um eine sehr **forschungsökonomische Form der Datenerhebung** handeln. So können z. B. Informationen über ansonsten schwer zugängliche Minoritäten oder unterschiedliche Kulturen anhand von öffentlichen Dokumenten der Onlinekommunikation (z. B. Online-Diskussionsforen, Online-Profile; Online-Videos) in großer Menge und Vielfalt vom Schreibtisch aus gesammelt und ausgewertet werden.

Bei der Dokumentenanalyse sind jedoch auch **Nachteile** zu beachten:

1. Wenn bei einer genuinen Dokumentenanalyse mit vorgefundenen Dokumenten gearbeitet wird, lassen sich keine experimentellen oder quasi-experimentellen Untersuchungsdesigns hinsichtlich der Entstehungsbedingungen der Dokumente realisieren, man ist auf **nicht-experimentelle Studien** festgelegt.
2. Je nach Art und Ursprung der Dokumente stehen mehr oder weniger **Kontextinformationen** zur Verfügung, die Auskunft darüber geben, wer unter welchen Umständen mit welcher Absicht die Dokumente wie erstellt hat und wie sie von welchen Adressaten genutzt und interpretiert werden. Die oft fehlenden Kontextinformationen können dazu führen, dass Dokumente bzw. ihre Bedeutung von den Forschenden falsch eingeschätzt werden. Zuweilen werden Dokumentenanalysen mit Interviews der Textautorinnen

und -autoren kombiniert, um Kontextinformationen zu gewinnen (Dokumentenanalyse von persönlichen Online-Profilen auf Social-Networking-Sites plus Interviews mit den jeweiligen Profilbesitzer/innen).

3. Die Dokumentenanalyse arbeitet typischerweise mit **Stichproben von Dokumenten**. Die Betrachtung eines einzelnen Dokuments als Einzelfallstudie ist selten, ebenso die Betrachtung der gesamten Population an Dokumenten als Vollerhebung. Etablierte qualitative und quantitative Stichprobenverfahren (▶ Kap. 9) sind zwar prinzipiell auf Dokumenten- ebenso wie auf Personenstichproben anwendbar. Praktische und methodische Hinweise für die Planung und Umsetzung aussagekräftiger Dokumentenstichproben sind jedoch weniger stark verbreitet, was die Stichprobenqualität beeinträchtigen kann (zur Stichprobenziehung in der Kommunikationswissenschaft s. Gehrau, Bilandzic, & Woelke, 2005).
4. Inhaltlich relevante Dokumente unterscheiden sich im Grad ihrer Zugänglichkeit sowie hinsichtlich der notwendigen Aktivitäten zur **Beschaffung und Archivierung des Materials**. Dokumente, die bereits digital vorliegen, sind leichter zu verwalten als analoge Dokumente, die erst noch digitalisiert werden müssen (was bei Videos oder Filmen sehr zeit- und ressourcenaufwändig sein kann). Öffentlich zugängliche Dokumente, die über Archive (z. B. von Zeitungen) oder Websites abrufbar sind, erfordern geringeren Beschaffungsaufwand und unterliegen weniger systematischen Verzerrungen als private Dokumente, um deren Herausgabe man die Besitzerinnen und Besitzer einzeln bitten muss.
5. Die **Qualität vorgefundener Dokumente** kann in unterschiedlicher Hinsicht eingeschränkt sein (Scott, 1990): 1. Authentizität (die genaue Urheberschaft oder Herkunft des Dokuments ist unklar, z. B. weiß man bei vielen Online-Diskussionsbeiträgen außer dem selbst gewählten Pseudonym nichts über die Autorinnen und Autoren), 2. Glaubwürdigkeit (es bestehen Zweifel an den im Dokument dargestellten Informationen, z. B. können die in einem Online-Profil abgegebenen Selbstbeschreibungen von der Realität abweichen), 3. Repräsentativität (unabhängig von der Art der Stichprobenziehung ist bereits der zugängliche Dokumentenpool systematisch verzerrt, z. B. weil besser gebildete Bevölkerungsgruppen mehr schriftliche Online-Dokumente erzeugen), 4. Interpretierbarkeit (die Bedeutung der Dokumente kann nicht korrekt entschlüsselt werden, z. B. weil in Dokumenten ein Insider-Jargon verwendet wird, den die Forschenden nicht korrekt dekodieren können).
6. Nicht gänzlich geklärt im Kontext der Dokumentenanalyse sind Fragen der **Forschungsethik**. Dies betrifft z. B. den Umgang mit öffentlichen oder halböffentlichen Dokumenten der Online-Kommunikation im Kontext von Forschungsprojekten. Stehen öffentliche Online-Dokumente (z. B. Beiträge in Online-Foren) für die Forschung frei zur Verfügung? Oder müssen die Autorinnen und Autoren um ihr Einverständnis gebeten werden, wenn ihre Texte für Forschungszwecke genutzt werden? Wie ist mit unterschiedlichen Graden der Öffentlichkeit umzugehen (sind Mailinglisten und Online-Communitys, bei denen man sich anmelden muss, weiterhin öffentlich oder gelten sie als geschützte soziale Räume? Zimmer, 2011). Sollen aus Gründen des Urheberrechts die Autorinnen und Autoren von zitierten Onlinetexten genannt oder aus Gründen der Anonymisierung gerade nicht genannt werden? Pauschale Antworten sind hier oft nicht möglich, stattdessen ist eine differenzierte ethische Reflexion und Diskussion des eigenen Vorgehens im jeweiligen Forschungskontext empfehlenswert (zur Forschungsethik ▶ Abschn. 4.1).

Zur zielgerichteten Beschaffung vorhandener Dokumente stehen unterschiedliche Quellen zur Verfügung (▶ Beschaffung vorhandener Dokumente für eine wissenschaftliche Dokumentenanalyse).

Klassifikationskriterien für wissenschaftliche Dokumentenanalysen

Zur Abgrenzung verschiedener **Typen von Dokumentenanalysen** sind v. a. vier Kriterien ausschlaggebend:

1. **Strukturiertheit des Vorgehens**: Bei der **qualitativen Dokumentenanalyse** werden die Dokumente im Zuge eines qualitativen Forschungsprozesses gesammelt und mit unterschiedlichen qualitativen Datenanalyseverfahren primär induktiv (aus den Daten heraus) sinnrekonstruierend interpretativ ausgewertet. Bei der **quantitativen Dokumentenanalyse** werden die Dokumente im Zuge eines quantitativen Forschungsprozesses gesammelt und mithilfe der quantitativen Inhaltsanalyse ausgewertet. Mittels quantitativer Inhaltsanalyse werden die Ausprägungen einzelner Merkmale der Dokumente in Messwerte überführt, indem ein vorab entwickeltes standardisiertes Kategoriensystem an die Daten angelegt wird (deduktives Vorgehen). Eine Kombination von qualitativen und quantitativen Vorgehensweisen bei einer Dokumentenanalyse ist möglich.
2. **Methode der Datenanalyse:** Bei der qualitativen Dokumentenanalyse kommen verschiedene Methoden der qualitativen Datenanalyse zum Einsatz, wobei die qualitative Inhaltsanalyse, die Interpretative Phänomenologische Analyse und die Kodierung gemäß der Grounded-Theory-Methodologie besonders viel-

Beschaffung vorhandener Dokumente für eine wissenschaftliche Dokumentenanalyse

Die verschiedenen Möglichkeiten, zielgerichtet Dokumente für eine wissenschaftliche genuine Dokumentenanalyse zu beschaffen, sind inhaltlich hinsichtlich der Aussagekraft der beschaffbaren Dokumente für das Forschungsproblem sowie forschungsökonomisch im Hinblick auf Erfolgswahrscheinlichkeit, Zeitaufwand und Kosten zu bewerten.

Zeitungen, Zeitschriften und Bücher

sind über Bibliotheken, über die Online-Archive der jeweiligen Medien (wobei hier teilweise nur die Online-Versionen und nicht die Print-Versionen der Artikel verfügbar sind; beide sind teilweise unterschiedlich formuliert), über kostenpflichtige ePaper-Ausgaben der Printmedien oder über Mediendatenbanken (z. B. LexisNexis) zugänglich. Ältere Zeitschriftenjahrgänge lassen sich teilweise auch über Online-Auktionshäuser wie z. B. Ebay erwerben oder über die Archive der Medienhäuser – meist kostenpflichtig – beschaffen.

Radio- und Fernsehsendungen

sind – meist temporär begrenzt auf die letzten ein bis zwei Wochen – über die Internetmediatheken der Sender bzw. als Internet-Podcasts zugänglich. Zudem können Aufzeichnungen einzelner zurückliegender Sendungen bei den Sendern bestellt werden (oft kostenpflichtig). Aktuelle Programme sind mit entsprechenden Aufzeichnungsgeräten oder mit Internet-Videorekorder-Diensten selbst mitzuschneiden, wobei bei umfangreichen TV-Aufzeichnungen ein hohes Datenvolumen entsteht.

Online-Dokumente

(Websites, Foren, Blogs etc.) sind meist kostenlos zugänglich und über öffentliche Internet-Suchmaschinen oder spezielle Suchprogramme (Webcrawler) recherchierbar. Beim Umgang mit Online-Dokumenten, bei denen es sich um sog. User Generated Content handelt, sind Fragen des Urheberrechts sowie des Datenschutzes und allgemein der Forschungsethik besonders sensibel und bereits bei der Materialsammlung zu beachten (► Abschn. 4.1). Die Annotation von Online-Dokumenten durch Schlagworte – das sog. **Tagging** – ist eine Variante, Online-Dokumente mit **Metadaten** zu versehen und für die automatische Suche und Auswertung vorzubereiten. In Online-Communitys wird diese Verschlagwortung durch die Nutzenden oft selbst vorgenommen („social tagging“) und kann bei wissenschaftlichen Dokumentenanalysen helfen (z. B. können anhand der Tags gezielt themenbezogene Dateien von Fotoplattformen wie Flickr, von Videoplattformen wie YouTube oder aus Weblogs und Microblogs wie Twitter gesammelt werden; zu Details von Online-Dokumentenanalysen siehe Richard, Grünwald, Recht, & Metz, 2010).

Offizielle Dokumente von Organisationen

Sollen Geschäftsberichte oder Informationsbroschüren von Unternehmen, parlamentarische Reden, Vereinssatzungen, Gesetzestexte, Parteiprogramme, ISO-Normen, Hausordnungen, Pressemeldungen von Nichtregierungsorganisationen oder andere offizielle Dokumente gesammelt werden, so sind meist die betreffenden Institutionen bzw. deren Internetpräsenzen direkt zu konsultieren (z. B. http://www.bundestag.de).

Persönliche Dokumente

Sollen persönliche Dokumente wie Familienfotos, Kinderzeichnungen, E-Mails, Einkaufszettel, Kalendereinträge, Briefe etc. gesammelt werden, so müssen die Produzenten bzw. Besitzer der Dokumente um Bereitstellung des Materials für die Forschung gebeten werden. Dies kann z. B. über **öffentliche Aufrufe** in Zeitungen oder im Internet geschehen, in denen die Forschenden um Zusendung entsprechender Dokumente bitten. Zudem können gezielt konkrete Personen angesprochen werden. Historische Dokumente lassen sich teilweise über Flohmärkte erstehen (z. B. Fotoalben früherer Jahrzehnte).

Dokumenten-Archive für die Forschung

Vereinzelt werden inzwischen im Internet Dokumenten-Archive speziell zu Forschungszwecken betrieben. So wurde im Rahmen des Poltext-Projekts ein umfassendes öffentliches Online-Archiv politikwissenschaftlich relevanter Dokumente (z. B. Gesetzestexte, Parteiprogramme, politische Reden) aus Kanada angelegt (http://www.poltext.capp.ulaval.ca/).

seitig einsetzbar und verbreitet sind. Bei der quantitativen Dokumentenanalyse hat sich die quantitative Inhaltsanalyse zur Quantifizierung des Materials etabliert. Bei Mixed-Methods-Designs werden in der Regel qualitative und quantitative Inhaltsanalyse kombiniert.

3. **Art des Materials**: Typischerweise werden im Rahmen der Dokumentenanalyse **textuelle (verbalschriftliche) Dokumente** untersucht. In den letzten Jahren wächst jedoch in den Sozial- und Kulturwissenschaften das Interesse an Dokumentenanalysen mit **visuellen Dokumenten** (z. B. Fotos). Im Rahmen der qualitativen Dokumentenanalyse ist deswegen die Interpretation visueller Dokumente ein eigener Analysezweig. Neben textuellen und visuellen Dokumenten existieren zahlreiche weitere Arten von Dokumenten, etwa auditive (z. B. Aufzeichnungen auf Anrufbeantwortern), audiovisuelle (z. B. Videos), multimediale (z. B. Multimedia-Lernprogramme) oder hypermediale (z. B. Websites, Online-Diskussionsforen). Die verschiedenen Arten von Dokumenten ermöglichen bzw. erfordern es, unterschiedliche Merkmale des Materials zu untersuchen. So können bei hypermedialen bzw. **Online-Dokumenten** neben den Text-, Bild-, Ton- und Videoinhalten von Websites auch interaktive Funktionen (z. B. Online-Formulare) sowie Menge und Art der Verlinkungen innerhalb und zwischen Dokumenten berücksichtigt

Tabelle 10.28 Klassifikation der in diesem Kapitel vorgestellten qualitativen und quantitativen Methoden der Datenanalyse im Kontext der Dokumentenanalyse

Strukturiertheit des Vorgehens	Methodischer Ansatz der Datenanalyse (Auswahl)	Art des Materials
Qualitative Dokumentenanalyse	Qualitative Inhaltsanalyse	Alle Dokumente
	Interpretative Phänomenologische Analyse	Autobiografische Dokumente
	Kodierung gemäß Grounded-Theory-Methodologie	Alle Dokumente
	Qualitative Analyse von visuellen Dokumenten	Visuelle Dokumente
Quantitative Dokumentenanalyse	Quantitative Inhaltsanalyse	Primär Textdokumente
	Quantitative Online-Inhaltsanalyse	Online-Dokumente
	Quantitative Inhaltsanalyse visueller Dokumente	Visuelle Dokumente

werden. Dementsprechend hat sich die quantitative Online-Inhaltsanalyse als eigenes Verfahren etabliert.

4. **Computerunterstützung**: Die Auswertung der gesammelten Dokumente kann **manuell** erfolgen (d. h. ohne Softwareunterstützung). Es kann aber auch **universale Bürosoftware** zum Einsatz kommen (z. B. Textverarbeitungs- oder Tabellenkalkulationsprogramme). Nicht zuletzt existieren eine Reihe von **spezialisierten Computerprogrammen** für die qualitative Datenanalyse wie z. B. NVivo, atlas.ti oder MaxQDA (► Abschn. 12.1.3). Auch für die quantitative Text- bzw. Inhaltsanalyse existieren spezialisierte Auswertungstools v. a. aus dem Bereich der Linguistik (Textanalyse-Software bzw. „software tools for text analysis“ wie z. B. TextQuest). Zudem können die im Zuge einer quantitativen Inhaltsanalyse erzeugten Messwerte auch direkt in ein Statistikprogramm (z. B. SPSS) eingegeben werden (► Abschn. 12.2.3). Abgesehen von der Phase der Datenauswertung kommen Computertools auch verstärkt in der Phase der Datensammlung zum Einsatz, etwa um zielgerichtet bestimmte Online-Dokumente im Web zu suchen und zu archivieren (z. B. mithilfe von Webcrawlern).

Eine Übersicht der in diesem Kapitel vorgestellten methodischen Ansätze der Datenanalyse im Kontext der Dokumentenanalyse bietet Tab. 10.28 (für weitere Verfahren der qualitativen Datenanalyse ► Abschn. 12.1). Die Art der Computerunterstützung wird dabei nicht berücksichtigt. Ebenso ist zu beachten, dass die Einteilung der Verfahren naturgemäß nicht ganz trennscharf ist. So werden die Varianten der qualitativen Datenanalyse v. a. nach dem methodischen Auswertungsansatz differenziert (z. B. Interpretative Phänomenologische Analyse versus Kodierung gemäß Grounded-Theory-Methodologie), die der quantitativen Datenanalyse dagegen nach der Art der Dokumente (z. B. klassische quantitative Inhaltsanalyse für Textdokumente versus quantitative Online-Inhaltsanalyse).

10.6.2 Qualitative Dokumentenanalyse

Typisch für qualitative Dokumentenanalysen ist ein Forschungsproblem, das sich als **offene Forschungsfrage** darstellt und somit eine **explorative** oder **theoriebildende** Funktion hat. Dementsprechend ist das gesamte Vorgehen – trotz einer Anlehnung an den bisherigen Forschungsstand und die Nutzung eines Theorierahmens – eher induktiv, d. h. datengesteuert.

Bei der qualitativen Dokumentenanalyse wird (mit Ausnahme der Technik der qualitativen Inhaltsanalyse) mit meist **sehr kleinen Stichproben** im ein- oder unteren zweistelligen Bereich operiert, deren Struktur entweder im Sinne einer bewussten Auswahl vorab festgelegt wird (z. B. sollen jeweils fünf Online-Tagebücher von Mädchen und von Jungen untersucht werden, die Essstörungen thematisieren) oder im Zuge der Datenerhebung schrittweise aufgebaut wird bis zum Erreichen theoretischer Sättigung (theoretische Stichprobe; ► Abschn. 9.2.1).

Qualitative Dokumentenanalyse – Unter einer qualitativen Dokumentenanalyse („qualitative document analysis“) als Datenerhebungsmethode verstehen wir die zielgerichtete, systematische und regelgeleitete Sammlung (qualitative Stichprobenbildung), Archivierung und Auswertung von vorgefundenen textuellen/verbal-schriftlichen sowie nichttextuellen (visuellen, auditiven, audiovisuellen, multimedialen, hypermedialen etc.) Dokumenten als Manifestationen menschlichen Erlebens und Verhaltens. Die Auswertung der Dokumente erfolgt qualitativ-interpretativ. Dafür stehen zahlreiche qualitative Datenanalysemethoden zur Verfügung, die gleichermaßen für die Auswertung forschungsgenerierter

Dokumente genutzt werden. Die **verschiedenen qualitativen Analysetechniken** unterscheiden sich hinsichtlich ihrer erkenntnistheoretischen Fundierung und inhaltlichen Zielsetzung sowie ihrer Eignung für bestimmte Arten von qualitativem Datenmaterial.

In den meisten Fällen basiert die interpretative Auswertung von Dokumenten auf einem eher induktiven systematischen **Prozess der Kodierung**, bei dem das Material in mehr oder minder kleine Analyseeinheiten segmentiert und die einzelnen Textstellen dann Codes bzw. Kategorien zugewiesen werden, welche der jeweiligen Textstelle Bedeutungen zuschreiben. (Im Unterschied zu dieser **kategorienbildenden Auswertung** existieren noch **sequenzielle Verfahren** wie die z. B. Konversationsanalyse, die Narrationsanalyse oder die objektive Hermeneutik, die nicht mit separaten Codes operieren, sondern Abfolgestrukturen in Dokumenten herausarbeiten). Für die Kodierung vorgefundener (wie auch forschungsgenerierter) Dokumente stehen also sehr unterschiedliche Methoden der qualitativen Datenanalyse („ qualitative data analysis“) zur Verfügung (► Abschn. 12.1).

Auf eine **fallbezogene Auswertung**, bei der jedes Dokument sorgfältig von vorne bis hinten durchgearbeitet, Codes gebildet und jeweils eine zusammenfassende Fallbeschreibung erstellt wird, folgt in der Regel eine vergleichende **fallübergreifende Auswertung**, bei der die in allen Fällen bzw. Dokumenten auftretenden Codes zu **übergeordneten Kategorien** verdichtet werden. Im Ergebnis einer fallübergreifenden Analyse liegen meist entweder die zentralen Themenkategorien des Phänomens vor **(thematische Analyse),** werden einander ähnliche Einzelfälle zu Typen zusammengefasst (**Typenbildung**) oder die Kategorien zu einer Theorie verknüpft (**Theoriebildung**).

Die **verschiedenen qualitativen Analysetechniken** unterscheiden sich hinsichtlich ihrer erkenntnistheoretischen Fundierung und inhaltlichen Zielsetzung sowie ihrer Eignung für bestimmte Arten von qualitativem Datenmaterial. So ist z. B. die **Konversationsanalyse** auf die Auswertung von interpersonaler Kommunikation zugeschnitten (z. B. Sprecherwechsel, Unterbrechungen) und eignet sich u. a. für die Auswertung von Transkripten von Gruppendiskussionen (forschungsgenerierte Dokumente durch mündliche Befragung als Datenerhebungsmethode). Sie ist aber auch geeignet zur Auswertung vorgefundener Dokumente (z. B. computergenerierte Protokolle von schriftlichen Gesprächen in Chat-Rooms, WhatsApp-Gruppen oder Web-Foren). Demgegenüber zielt die **kritische Diskursanalyse** auf gesellschaftliche Machtverhältnisse ab, die sich z. B. darin widerspiegeln, wie bestimmte Personengruppen in den Medien dargestellt oder nicht dargestellt werden. Sie wird typischerweise auf vorgefundene Dokumente aus den Massenmedien angewendet.

Im Folgenden werden drei qualitative Auswertungsmethoden behandelt, die v. a. in der Psychologie breite Anwendung finden und für vorgefundene Dokumente geeignet sind, aber auch zur Auswertung von im Forschungsprozess generierten qualitativen Daten dienen können: Die qualitative Inhaltsanalyse, die Interpretative Phänomenologische Analyse und die Kodierung gemäß Grounded-Theory-Methodologie. Ergänzend wird kurz auf die interpretative Analyse von visuellem Material eingegangen. Die Darstellung kann im Rahmen des vorliegenden Buches jeweils nur kursorisch erfolgen, für methodische und methodologische Details wird jeweils auf vertiefende Literatur verwiesen.

Qualitative Inhaltsanalyse

Die qualitative Inhaltsanalyse hat sich aus der kritischen Auseinandersetzung mit der quantitativen Inhaltsanalyse entwickelt (Kracauer, 1952). Während die **quantitative Inhaltsanalyse** darauf abzielt, anhand eines Forschungsproblems bestimmte formale und inhaltliche Merkmale von Dokumenten mittels eines theoriebasierten vollstandardisierten Kategoriensystems zu messen und dann statistisch auszuwerten (deduktives Vorgehen), will die **qualitative Inhaltsanalyse** anhand eines Forschungsproblems die Bedeutungsgehalte von Dokumenten durch eine datengesteuerte schrittweise Kodierung herausarbeiten (induktives Vorgehen). Die Bedeutungsgehalte der Dokumente werden bei der qualitativen Inhaltsanalyse im Ergebnis durch übergeordnete verbal beschriebene Codes und Kategorien dargestellt. Diese pointierte Gegenüberstellung ist jedoch zu relativieren. Denn eine qualitative Inhaltsanalyse kann deduktive Elemente enthalten (z. B. theoriebasierte Codes berücksichtigen) und eine quantitative Inhaltsanalyse kann induktive Elemente enthalten (z. B. datengenerierte Kategorien; ► Abschn. 10.6.3). Beide Vorgehensweisen haben somit Parallelen und Überschneidungen.

Verglichen mit anderen qualitativen Auswertungsverfahren weist die qualitative Inhaltsanalyse die größte Nähe zur quantitativen Inhaltanalyse auf bzw. nimmt eine **Zwischenposition** zwischen qualitativen und quantitativen Verfahren ein (vgl. Mayring, 2010b, S. 604; Morgan, 1993; Schreier, 2012). Oft kommt eine Kombination von qualitativer und quantitativer Inhaltsanalyse zum Einsatz, um Dokumente auszuwerten (Mayring, 2001). Im Rahmen der qualitativen Inhaltsanalyse werden nicht selten Stichproben von Dokumenten bearbeitet, die hinsichtlich ihres Umfanges weit über die sonst in der qualitativen Forschung üblichen Stichprobengrößen hinausgehen. Dabei wird die Analyse dann üblicherweise computergestützt durchgeführt (zu Computerprogrammen für die qualitative Datenanalyse ► Abschn. 12.1.3).

Qualitative Inhaltsanalyse – Die qualitative Inhaltsanalyse („qualitative content analysis") arbeitet vor allem datengesteuert-induktiv – aber teilweise auch theoriebasiert-deduktiv – durch Kodierung die zentralen manifesten und latenten Bedeutungen von Dokumenten heraus. Sie lässt sich auf sämtliche Arten von vorgefundenen (sowie auch forschungsgenierten) Dokumenten anwenden. Sollen Dokumente im Sinne eines Mixed-Methods-Designs untersucht werden, so ist eine Kombination von qualitativer und quantitativer Inhaltsanalyse üblich.

Es werden aktuell verschiedene Ansätze der qualitativen Inhaltsanalyse verfolgt, etwa die konventionelle, die gerichtete und die summative qualitative Inhaltsanalyse (Hsieh & Shannon, 2005). Im deutschsprachigen Raum hat sich die von dem Psychologen **Philipp Mayring** (Mayring, 2010a, 2010b) entwickelte Form der qualitativen Inhaltsanalyse am stärksten etabliert. Diese beinhaltet wiederum drei verschiedene Vorgehensweisen mit unterschiedlichen Zielsetzungen:

1. Die **zusammenfassende qualitative Inhaltsanalyse** zielt darauf ab, mehr oder minder umfangreiche qualitative Dokumente bzw. qualitatives Datenmaterial schrittweise induktiv auf die manifesten Hauptinhalte zu reduzieren. Sie entspricht der konventionellen qualitativen Inhaltsanalyse (Hsieh & Shannon, 2005). Dazu wird in vier Schritten vorgegangen: 1. **Paraphrasierung**: Inhaltstragende Textstellen werden identifiziert und jeweils in eine Kurzform (Paraphrase) umformuliert. 2. **Generalisierung auf das Abstraktionsniveau**: Die Paraphrasen werden alle auf dasselbe, vorher definierte Abstraktionsniveau gebracht. 3. **Erste Reduktion**: Die relevanten Paraphrasen werden selektiert, indem man bedeutungsgleiche oder unwichtige Paraphrasen streicht. 4. **Zweite Reduktion**: Die verbliebenen Paraphrasen werden fallbezogen gebündelt und integriert, es entstehen somit abschließend neue und komplexere Paraphrasen, die die Hauptinhalte des Falls zusammenfassen (► Studienbeispiel „Qualitative Inhaltsanalyse von vorgefundenen Online-Dokumenten").
2. Die **explizierende qualitative Inhaltsanalyse** zielt darauf ab, unklare Textstellen durch eine Betrachtung des direkten Textumfeldes (enge Kontextanalyse) sowie von Zusatzmaterial, das über den Text hinausgeht (weite Kontextanalyse) verständlich zu machen bzw. zu erläutern (Explikation). Während die zusammenfassende Inhaltsanalyse das Ausgangsmaterial reduziert, wird es im Zuge einer explizierenden Inhaltsanalyse stellenweise erweitert. Denn die betrachteten Textstellen werden durch Kontextinformationen ergänzt, aus denen dann explizierende Paraphrasen abgeleitet werden.
3. Die **strukturierende qualitative Inhaltsanalyse** entspricht im Grunde der quantitativen Inhaltsanalyse (► Abschn. 10.6.3 „Quantitative Inhaltsanalyse"), denn hier wird ein vorher definiertes Kategoriensystem an den Text angelegt und damit numerisches Datenmaterial gewonnen. Nach Mayring (2010a, S. 66) sind vier Varianten der strukturierenden Inhaltsanalyse zu unterscheiden: die formale, die inhaltliche, die typisierende und die skalierende Strukturierung. In der Terminologie der quantitativen Inhaltsanalyse geht es dabei um die Art der Kategorien (formale oder inhaltliche Merkmale des Dokuments) sowie um das Skalenniveau der Kategorien (nominal- oder intervallskalierte Kategorien).

Studienbeispiel

Qualitative Inhaltsanalyse von vorgefundenen Online-Dokumenten

Im Rahmen einer qualitativen Dokumentenanalyse sollte der Frage nachgegangen werden, wie Social-Networking-Sites (z. B. Facebook) von Jugendlichen im Zuge ihres Trauerprozesses genutzt werden, wenn sie im eigenen Freundes- oder Bekanntenkreis mit einem Todesfall konfrontiert werden (Williams & Merten, 2009).

Dazu wurden von 20 plötzlichen Todesfällen (z. B. Unfall, Suizid), die sich zwischen 2005 und 2007 unter 15- bis 19-jährigen Jugendlichen in den USA ereignet hatten (14 Jungen, 6 Mädchen), die **öffentlichen Online-Profile der Verstorbenen** herausgesucht. Als Dokumente wurden die auf den Online-Profilen der Verstorbenen nach deren Tod **von ihren Freunden und Bekannten** hinterlassenen **Postings** ($n = 4\,780$) ausgewertet. Das Verfahren entspricht der Logik einer zusammenfassenden qualitativen Inhaltsanalyse. Die den Nachrichten zu entnehmenden Bewältigungsformen von Trauer wurden zunächst zu Codes paraphrasiert und diese Zusammenfassungen dann zu neun übergeordneten Kategorien verknüpft. Für die Verbreitung der Kategorien werden zudem Häufigkeiten angegeben, womit die Brücke zur quantitativen Inhaltsanalyse geschlagen wird. ◘ Abb. 10.40 ist zu entnehmen, dass in allen untersuchten Nachrichten der Hinterbliebenen die verstorbene Person direkt angesprochen und der Trauer Ausdruck verliehen wurde. Zudem enthielten 45 % aller Postings Hinweise auf emotionale und kognitive Bewältigungsstrategien („indicators of emotional and cognitve coping strategies"): Die einzelnen Coping-Strategien sind als 8 Codes nachvollziehbar etikettiert und mit Textzitaten unterlegt.

Im Rahmen von Dokumentenanalysen werden meist zusammenfassende qualitative Inhaltsanalysen eingesetzt, etwa um Publikumszeitschriften dahingehend auszuwerten, welche Informationen sie zu gesunder Ernährung oder zu sexuell übertragbaren Krankheiten vermitteln (Clarke, 2010; Geyzen, 2011). Neben vorgefundenen Dokumenten wie z. B. Zeitungsartikeln wird die qualitative Inhaltsanalyse auch oft auf im Forschungsprozess generierte Dokumente angewendet, etwa auf Transkripte von Leitfadeninterviews oder Gruppendiskussionen. Auch nicht-textuelle Dokumente wie Videos lassen sich einer

Textstellen aus den Postings Hinterbliebener auf den Online-Profilen Verstorbener	Zusammenfassungen (Codes)	Übergeordnete Themen-Kategorien	Absolute Häufigkeiten	Relative Häufigkeiten
Ok so it has almost been a year, and this is my first time up here. I never thought it would take me this long to get up here. For [an assignment] I wrote about when [I found out] you had died. I don't know why I chose to write about this, I guess it was because I had been thinking about you. Well, it was HARD to write. I cried so much.	writing about bereavement	1. Comments directed towards deceased (i.e. talking to the deceased)	4780	100%
I went to see you yesterday. . . . I bought roses before I got on the metro to get home and a man inside asked me for a dollar so he could make it to where he needed to get to, I gave him one and then he asked me who the flowers were for. I told them they were for a friend, and he was like "I hope they're ok." I kind of just looked at him and said that I was going to visit your grave. I think it was then that it really, really hit me that you were gone for good.	discussing a friend's death with a stranger	2. Comments with memorial sentiment (e.g., RIP: rest in peace)	4780	100%
I just wanted to say that I am a mess right now. . I thought that I was here but I'm NOT ~ Everything is hitting me so hard and no one can see it or care to notice that I'm falling through the cracks and living on the edge.	depression	**3. Indicators of emotional or cognitive coping strategies**	**2167**	**45%**
hate the fact that u did this to yourself. I hate it that u didn't tell anyone. But there's nothing I can do. I can be mad all I want. But nothing will change, I wish u never would've made such a permanent decision.	anger	4. Comments about current events	933	19%
So essentially I'm feeling very guilty right now that I have some amount of happiness. How can I be happy when you're dead? I feel like just the idea of being happy is a betrayal to you. Being so much closer to all your friends and being with J—- now makes me so happy.	guilt	5. Comments regarding the act of posting comments to the deceased	407	8%
One month today. god this is so hard for me. I try to be strong. I do. But life just sucks so bad without you. I never imagined living life without you by my side. No more growing old together. No more kids together . . . no more kisses. No more smiles. No more laughs. No more firsts.	lost opportunity	6. Comments about the cause of death	311	6%
If I said that I cried over this tragedy after Friday and before your wake, I would be lying. From Saturday to Tuesday, I watched video after video on every news site I could find, read every story in every newspaper, saw every picture of everyone and everything. I went to [address], touched the scar on the tree. Prayed. I didn't cry once.	information seeking	7. Reminiscing	141	3%
I know that your lookin down and thankin god that you don't have to sit through 90 minutes of Trigonometry! Lol Well, hold it down up there. .and don't watch me in the shower or anything! Lol. JK!	humor	8. Comments from distal or unknown peers	120	2%
		9. Comment about the funeral, seeing the body, and the afterlife	101	2%

Ergebnisse der qualitativen Inhaltsanalyse von n = 4780 Postings, die auf den öffentlichen Social-Networking-Profilen von 20 plötzlich verstorbenen Jugendlichen nach deren Tod von Peers hinterlassen wurden (Daten aus Williams & Merten, 2009)

Abb. 10.40 Exemplarische Kodierung von Textstellen aus den Nachrichten von Hinterbliebenen auf Online-Profilen Verstorbener im Rahmen einer qualitativen Inhaltsanalyse. (Aus Williams, A. & Merten, M. (2009). Adolescents' online social networking following the death of a peer. *Journal of Adolescent Research, 24*(1), 67–90. © 2013 by SAGE. Reprinted by permission of SAGE Publications)

qualitativen Inhaltsanalyse unterziehen. Generell befasst sich die qualitative Inhaltsanalyse eher mit den **manifesten Inhalten** (unmittelbare Wort-/Bildbedeutungen), weniger mit den **latenten Inhalten** (tieferen Bedeutungsebenen), obwohl diese teilweise im Rahmen einer explizierenden qualitativen Inhaltsanalyse herausgearbeitet werden können. Allerdings wird die qualitativen Inhaltsanalyse in viel stärkerem Maße als die quantitative Inhaltsanalyse für geeignet gehalten, mit Bezug zum Forschungsproblem auch komplexere und latente Inhalte von Dokumenten aufzudecken (Schreier, 2012).

Interpretative Phänomenologische Analyse

Die **Interpretative Phänomenologische Analyse** (IPA) ist – ebenso wie die qualitative Inhaltsanalyse – eine in der Psychologie verwurzelte qualitative Datenanalysemethode, allerdings hat sie einen sehr viel engeren Anwendungsbereich. Die IPA wurde maßgeblich von dem britischen Psychologen **Jonathan Smith** entwickelt (Colaizzi, 1978; Smith, 1996; Smith, Flowers, & Larkin, 2009; http://www.ipa.bbk.ac.uk) und eignet sich besonders für **autobiografische Textdokumente**, in denen Personen ihr Erleben bedeutsamer bzw. außeralltäglicher Ereignisse schildern („personal lived experience"). Die subjektiven Sichtweisen und Sinngebungen der Individuen zu rekonstruieren, ist Ziel der Interpretativen Phänomenologischen Analyse, die erkenntnistheoretisch auf philosophischen Ansätzen der **Phänomenologie** basiert (Moran, 2000), maßgeblich auf den Werken des deutschen Philosophen Edmund Husserl (1913). Geeignet für eine Auswertung gemäß IPA sind dementsprechend persönliche Dokumente wie Briefe, Tagebücher, Weblogs etc. Ungeeignet für die Methode sind dagegen Dokumente, die keine autobiografischen Inhalte haben (z. B. Zeitungsartikel, Sitzungsprotokolle). Die IPA wird nicht nur für Analysen vorgefundener autobiografischer Dokumente, sondern primär im Rahmen von qualitativen Befragungsstudien eingesetzt, um die im Forschungsprozess erzeugten Transkripte von narrativen Interviews, biografischen Leitfadeninterviews oder Fokusgruppendiskussionen (Tomkins & Eatough, 2010) auszuwerten.

Interpretative Phänomenologische Analyse – Die Interpretative Phänomenologische Analyse („interpretative phenomenological analysis", IPA) arbeitet induktiv die zentralen Bedeutungen autobiografischer Dokumente heraus. Sie wird auf vorgefundene autobiografische Texte (z. B. Briefe, Tagebücher) sowie auf im Forschungsprozess generierte Dokumente (v. a. Transkripte von narrativen bzw. biografischen Interviews) angewendet. Zielsetzung ist es, das subjektive Erleben bestimmter autobiografischer Erfahrungen aus den Dokumenteninhalten zu rekonstruieren.

Anwendungsgebiete der IPA sind Gesundheits-, klinische und Sozialpsychologie, aber auch Management-, Beratungs-, Kriminal- und Medizinforschung (Reid, Flowers, & Larkin, 2005). Aktuelle qualitative Interviewstudien, deren Transkripte als **forschungsgenerierte Dokumente** mittels IPA ausgewertet wurden, befassen sich z. B. mit

- den Erfahrungen von vier Psychotherapeuten mit Patienten, die aus ihrer Sicht in Selbsttäuschung verhaftet sind (Westland & Shinebourne, 2009),
- dem Erleben des Eintritts in das Rentenalter durch acht Senioren (Pettican & Prior, 2011),
- den Erfahrungen von 17 homosexuellen Jugendlichen mit Diskriminierung in der Schule (sechs Einzelinterviews, drei Fokusgruppendiskussionen; Robinson, 2010),
- den Erfahrungen von Online-Gamern mit virtuellen Tabuverletzungen (Whitty, Young, & Goodings, 2011) oder
- den Erfahrungen von Soldaten, die Väter werden (▶ Studienbeispiel „Interpretative Phänomenologische Analyse von forschungsgenerierten Dokumenten").

Qualitative Dokumentenanalysen, bei denen **vorgefundene Dokumente** mittels IPA ausgewertet wurden, thematisieren z. B.

- die Bedürfnisse, die in Entschuldigungsbriefen von Tätern an Opfer artikuliert werden (Duff, 2010) oder
- den Umgang mit Essstörungen in den Postings von Pro-Ana-(Pro Anorexie-)Foren im Internet (Mulveen & Hepworth, 2006).

Eine Interpretative Phänomenologische Analyse beinhaltet in der Regel folgende **Arbeitsschritte** (Colaizzi, 1978; Reid et al., 2005) und wird meist computergestützt durchgeführt (für entsprechende Computerprogramme ▶ Abschn. 12.1.3):

1. Jedes Dokument (z. B. Tagebucheintrag, Interview-Transkript) wird sorgfältig gelesen, um ein **Globalverständnis** des untersuchten Phänomens aus Sicht der jeweiligen Untersuchungsperson zu entwickeln.
2. Bei jedem Dokument werden die im Hinblick auf das Forschungsproblem **aussagekräftigsten Stellen** bzw. **Zitate** markiert bzw. extrahiert.
3. Für jedes dieser Zitate wird vor dem Hintergrundverständnis der Sichtweisen der einzelnen Textautor/innen jeweils die **zentrale Bedeutung** zusammengefasst. Die Zuordnung eines Zitats aus dem Dokument zu einer von den Forschenden formulierten Zusammenfassung der Bedeutung entspricht der Tätigkeit des interpretierenden Kodierens. Die Forschenden nehmen dabei sozusagen eine phänomenologische oder **Insider-Position** ein und versuchen die Bedeutung aus Sicht der Autorinnen und Autoren der

Studienbeispiel

Interpretative Phänomenologische Analyse von forschungsgenerierten Dokumenten

Eine qualitative Interviewstudie ging der Frage nach, wie die Geburt des ersten Kindes von jungen Vätern erlebt wird, die zum Zeitpunkt der Geburt als Soldaten an einem militärischen Auslandseinsatz teilnehmen (Schachman, 2010). Rund einstündige narrative Interviews (▶ Abschn. 10.2.2) mit 17 amerikanischen Soldaten wurden zwei bis sechs Monate nach der Geburt der Kinder durchgeführt. Der Erzählanstoß lautete: „What is it like to become a father while deployed overseas to a combat region?"

Da heute die Anwesenheit der Väter bei der Geburt ihrer Kinder der Normalfall ist und in den USA bei über 90 % der Geburten vorkommt, stellt die Abwesenheit ein außeralltägliches Ereignis dar, wobei der Einsatz in einer Kriegsregion als eine zusätzliche außergewöhnliche Belastung für die Familien anzusehen ist. Im Zuge der Interpretativen Phänomenologischen Analyse der **17 Interview-Transkripte (forschungsgenerierte Textdokumente)** extrahierte die Autorin fünf Themencluster, die sie zu zwei Hauptthemen bündelte:

1. Hauptthema: Unterbrechung der Beschützer- und Versorgerrolle
   - Themencluster: „Sorge: Eine traumatische und einsame Geburt"
   - Themencluster: „Bedauern: Verpasste Gelegenheit"
   - Themencluster: „Schuldgefühle: Ein abwesender Vater"
   - Themencluster: „Angst: Tod oder Verwundung"
2. Hauptthema: Wiederherstellung der Beschützer- und Versorgerrolle
   - Themencluster: „Kommunikation: Kanäle, die verbinden"

Die befragten Soldaten erlebten die Abwesenheit bei der Geburt ihrer Kinder als Unterbrechung der von ihnen als zentral betrachteten Beschützer- und Versorgerrolle, mit der sie sich identifizierten. Zur Wiederherstellung dieser Rolle wurde die kommunikative Verbindung zu Partnerin und Kind per Telefon sowie per Internet (Messenger, Skype) genutzt und als hilfreich eingeschätzt (◘ Tab. 10.29).

Das Beispiel bezieht sich auf die Analyse von im Forschungsprozess generierten Dokumenten (Interviewtransskripte), kann aber auf vorgefundene Dokumente (z. B. Postings in Online-Diskussionsforen, Briefe) übertragen werden.

autobiografischen Dokumente möglichst genau auf den Punkt zu bringen.

4. Inhaltsähnliche Bedeutungen bzw. Paraphrasen werden zu **Themenclustern** bzw. Codes gebündelt und mit einem entsprechenden Namen versehen. Dabei werden in der Regel vier bis zehn Themencluster gebildet. Bei diesem Schritt – sowie bei den folgenden Interpretationsschritten – nehmen die Forschenden dann eine **Outsider-Position** ein und bündeln das Material mit Blick auf die Beantwortung der Forschungsfrage.
5. Inhaltlich zusammenhängende Themencluster werden auf höchster Abstraktionsebene bei Bedarf noch zu **Hauptthemen** bzw. Themenkategorien zusammengefasst. Die Themencluster sind dann die Subdimensionen dieser Kategorien.
6. Die grundlegende **Struktur** des untersuchten Phänomens wird durch die (oft tabellarische) Darstellung aller Hauptthemen, aller untergeordneten Themencluster sowie exemplarisch der zugeordneten Codes und Zitate dargestellt.
7. Die detaillierte **Beschreibung** des untersuchten Phänomens erfolgt anhand der Gedanken, Gefühle, Handlungsweisen der Untersuchungspersonen im Hinblick auf die einzelnen Themencluster und wird durch Zitate illustriert. Dabei werden fallübergreifende Gemeinsamkeiten betont, es wird aber zuweilen auch auf individuelle Besonderheiten einzelner Fälle hingewiesen.
8. Die **Gültigkeit des Analyseergebnisses** wird dann als gut gesichert angesehen, wenn die Vorgehensweise transparent ist (z. B. exemplarische Erläuterung der Bildung eines Themenclusters anhand von Codes und Zitaten), wenn der Interpretationsprozess als solcher ausdrücklich reflektiert erfolgt (z. B. Selbstreflexion der Rolle der Forschenden), wenn das Analyseergebnis mit den Textautor/innen und/oder mit Fachkolleg/innen abgestimmt wurde und wenn das Ergebnis für die Leserschaft des Berichts plausibel wirkt (zu Gütekriterien qualitativer Forschung ▶ Abschn. 3.3).

Kodierung gemäß Grounded-Theory-Methodologie

Die Grounded-Theory-Methodologie (GTM) gehört zu den etabliertesten qualitativen Forschungsansätzen. Sie wurde in den 1960er-Jahren von den **amerikanischen Soziologen Anselm Strauss** (1916–1996) und **Barney Glaser** entwickelt im Rahmen medizinsoziologischer Studien (Glaser & Strauss, 1967; Strauss & Corbin, 1994). Der Ansatz wurde in den vergangenen Dekaden von Strauss und Glaser sowie deren Schülerinnen (insbesondere Kathy Charmaz, Adele Clarke) unabhängig voneinander weiterentwickelt und verfeinert (für aktuelle Darstellungen der Methode s. Bryant & Charmaz, 2010; Breuer, 1996, 2010; Mey & Mruck, 2007, 2011).

Es handelt sich bei der GTM um eine **Forschungsstrategie**, die nicht auf die bloße Beschreibung eines Phänomens, sondern ausdrücklich auf die Bildung einer **gegenstandsverankerten Theorie** („grounded theory") mit Erklärungskraft hinausläuft. Dementsprechend wurde die GTM bereits in ▶ Abschn. 6.3.5 im Zusammenhang mit Theoriebildung angesprochen. Eine gegenstandsverankerte Theorie besteht aus meist ein bis fünf **Katego-**

Tabelle 10.29 Illustration der Bildung eines Themenclusters im Rahmen einer Interpretativen Phänomenologischen Analyse. (Aus Schachman, K. A. (2010). Online Fathering: The Experience of First-Time Fatherhood in Combat-Deployed Troops. *Nursing Research, 59*(1), 11–17, S. 13)

Aussagekräftige Zitate aus den Interview-Transkripten („significant statements“)	Bedeutung („formulated meaning“)	Themencluster („theme cluster“)
I remember when [fellow Marine] died. I starting thinking, “Holy cow, that could have been me.“ I didn’t sleep all night, thinking about that, worrying. I kept picturing the Chaplain going to our house, telling [my wife] that I was dead. Her and the baby – alone.	When his friend was killed, he thought about his own mortality. He worried about how his wife would receive the news and the prospect of her and the baby being alone.	Fear of death and dismemberment: Who will be the father?
I worried every day. I was afraid I would not make it back. Sure, money-wise they would be okay. I have plenty of [military life insurance]. But who would raise him? Make sure he stayed out of trouble? I am his father that is MY job. I need to get home, I need to get home that is all I could think about.	Although his death would leave them financially secure, he worried that others would assume his role as father.	
Dying would be bad but – worse than that – what if you got totally [maimed]? Like I’m supposed to throw a football to my son with no arms? And then someone has to take care of ME for the rest of my life like I’m a baby.	He felt that a serious injury would be worse than death, and that he would be unable to engage in physical activities with his child, and would be dependent on others for his physical care.	

rien, die miteinander verknüpft und in Subkategorien ausdifferenziert sind. Jede Kategorie und Subkategorie kann dabei noch hinsichtlich ihrer qualitativen (z. B. positiv – negativ) und quantitativen (z. B. selten – häufig) Dimensionen gekennzeichnet sein. Mithilfe der Kategorien kann ein theoretisches Modell formuliert werden, das unter Verwendung eines **Kodierparadigmas** (Strauss & Corbin, 1996) die konkreten Ursachen, intervenierenden Bedingungen und Konsequenzen sowie Kontext und Strategien im Zusammenhang mit dem untersuchten Phänomen herausarbeitet (Mey & Mruck, 2009, S. 136). Oft wird eine einzelne **Schlüsselkategorie** („core category“) als Zentrum der Theorie angegeben.

Kodierung gemäß Grounded-Theory-Methodologie – Die Grounded-Theory-Methodologie („grounded theory methodology“, GTM) ist eine qualitative Forschungsstrategie, die auf theoretischer Stichprobenziehung („theoretical sampling“), der **Kodiermethode des permanenten Vergleichs** („constant comparison coding/analysis“) sowie dem systematischen Verfassen von Memos basiert („memoing, memowriting“). Im Ergebnis läuft die Methode auf eine gegenstandsverankerte Theorie („grounded theory“) hinaus. Die Kodierung gemäß GTM kann auf vorgefundene ebenso wie auf forschungsgenerierte Dokumente angewendet werden, wobei der gesamte Forschungsprozess an der Grounded-Theory-Methodologie auszurichten ist.
Im Unterschied zur qualitativen Inhaltsanalyse und zur Interpretativen Phänomenologischen Analyse, deren Kategorien eher zusammenfassend-deskriptiven Charakter haben, werden mit der Kodiermethode des permanenten Vergleichs im Rahmen der GTM Kategorien gebildet, die einen erklärenden Charakter aufweisen. Die Kodiermethode des permanenten Vergleichs beinhaltet dabei das offene Kodieren, das axiale Kodieren und das selektive Kodieren.

Forschungsprojekte, die der GTM folgen, zeichnen sich – bei allem Variantenreichtum der Ansätze – durch drei „Essentials“ aus (Mey & Mruck, 2009):

1. Die zugrunde gelegten qualitativen Daten (z. B. vorgefundene Dokumente bei der Dokumentenanalyse; forschungsgenerierte Dokumente bei anderen Datenerhebungsverfahren) werden nach dem Prinzip der **Theoretischen Stichprobenziehung** („theoretical sampling“, ▶ Abschn. 9.2.1) sukzessive gesammelt bis bei der Auswertung theoretische Sättigung eingetreten ist. Bei dieser Form der Stichprobenziehung werden die verschiedenen Fälle permanent miteinander verglichen und gezielt maximal aussagekräftige weitere Fälle in das Sample aufgenommen.
2. Die vorliegenden Daten werden einer speziellen Kodiermethode, nämlich der **Methode des permanenten Vergleichs** („constant comparison coding/analysis“) unterzogen. Diese Kodiermethode umfasst in der insbesondere in der hiesigen Forschungslandschaft populären Fassung von Strauss und Corbin (1996) drei Kodiertechniken: das offene Kodieren, das axiale Kodieren und das selektive Kodieren:
   Beim **offenen Kodieren** werden die zu analysierenden Dokumente in kleine Sinneinheiten segmentiert. An die Sinneinheiten werden dann W-Fragen herangetragen (Was, Wer, Wie, Wann, Wie lange, Wo, Warum, Womit, Wozu?). Auf diese Weise können aus einem Datensegment unterschiedliche Codes/theoretische Konzepte generiert werden. Jeder Code wird durch eine **Codenotiz** erläutert. Weitergehende Ideen werden in Memos festgehalten. Das offene Kodieren des Materials läuft auf eine lange **Codeliste** hinaus, wobei manche Codes bereits provisorisch zu übergeordneten bzw. abstrakteren **Kategorien** gebündelt werden können.

Die so gebildeten Kategorien werden beim **axialen Kodieren** fallübergreifend in Relation zueinander gesetzt, v. a. wird versucht, eine Ordnung in der Weise herzustellen, dass hinsichtlich des untersuchten Phänomens zwischen Ursachen, intervenierenden Bedingungen und Konsequenzen, Handlungskontext und Strategien/Techniken unterschieden wird (sog. **Kodierparadigma** nach Strauss, 1991; im Ansatz von Glaser, 1998 wird stattdessen mit so genannten **Kodierfamilien** gearbeitet). Beim axialen Kodieren wird das Kodierparadigma mit entsprechenden Kategorien gefüllt, für die umfangreiche Theorie-Memos zu erstellen sind.

Beim **selektiven Kodieren** schließlich konzentriert man sich auf die ausgewählte **Kernkategorie**: Ihre Subkategorien, Dimensionen und Interrelationen werden detailliert am Material herausgearbeitet. Das im Zuge des Kodierens erstellte theoretische Modell wird auf diese Weise finalisiert, mit klarem roten Faden argumentativ in einer Geschichte entfaltet und zudem oft grafisch dargestellt (► Abschn. 12.1.2).

3. Der gesamte Forschungs- und Kodierprozess wird durch das **Schreiben von Memos** („memoing", „memowriting") begleitet (Groenewald, 2008). Memos sind schriftliche Notizen, in denen die Forschenden ihre Ideen und Zwischenergebnisse fortlaufend protokollieren. Nach ihren Inhalten können Planungs-, Methoden-, Auswertungs- und Theorememos unterschieden werden. Das Schreiben der Memos hilft den Forschenden bei der Strukturierung und Dokumentation ihrer Gedanken. Jedes Memo sollte mit Datum, Autor, Memotyp, Datenbezug und Thema gekennzeichnet werden. Zudem muss das Memo nach Bedarf mit anderen Memos verknüpft werden. Jedes Memo sollte ein einzelnes Thema fokussieren, für weitere Themen werden neue Memos erstellt. Memos stellen sicher, dass keine zwischenzeitlichen Erkenntnisse durch Vergessen verloren gehen und der Prozess der Theoriebildung nachvollziehbar ist. Sie bilden auch erste Bausteine für den späteren Forschungsbericht. Beim Memoschreiben können Forschende durchaus ihren individuellen Stil entwickeln, wobei jedoch eine Abstimmung im Forschungsteam sinnvoll ist. Im Verlauf eines Forschungsprojektes werden die Memos in der Regel immer differenzierter.

Die APA-Literaturdatenbank PsycInfo weist Tausende referierter Zeitschriftenartikel aus, die sich dem Grounded-Theory-Ansatz zuordnen. Dabei ist jedoch zu beachten, dass Studien nicht selten mit dem „guten Namen" der Grounded-Theory-Methodologie operieren, ohne die oben genannten Kriterien vollständig zu erfüllen. So wird die Bedingung einer Theoretischen Stichprobenziehung, dass nämlich in Abhängigkeit von Zwischenergebnissen sukzessive und theoretisch reflektiert immer wieder neue Fälle in das Sample aufgenommen werden, oft nicht erfüllt, sondern die Stichprobe in einem einzigen Schritt gezogen. Auch die Details der Datenanalyse – z. B. das axiale Kodieren nach dem Kodierparadigma oder das konsequente Schreiben von Memos – werden oft vernachlässigt. Im Ergebnis resultieren dann nicht selten eher beschreibende Kategorien als eine gegenstandsverankerte Theorie. Kenntnisse der GTM sind deswegen wichtig, um vorliegende Studien kritisch einschätzen zu können. Ebenso sollte man selbst bei eigenen Studien immer deutlich machen, ob man bestimmte **Elemente** der GTM (z. B. Datenanalyse gemäß der Kodiermethode des permanenten Vergleichs) aufgegriffen oder tatsächlich den **gesamten Forschungsprozess** gemäß der Logik dieses Ansatzes gestaltet hat.

Die Grounded-Theory-Methodologie wird hier im Zusammenhang mit der Dokumentenanalyse vorgestellt, weil im Rahmen dieses Forschungsansatzes tatsächlich eine besonders differenzierte Form der Sammlung und Analyse von Dokumenten entwickelt wurde. Wie die Textbeispiele im folgenden Studienbeispiel zur GTM zeigen, ist der Auswertungsprozess sehr viel kleinteiliger und anspruchsvoller als bei der qualitativen Inhaltsanalyse und der Interpretativen Phänomenologischen Analyse. Umso wichtiger ist ein sorgfältiges Memoschreiben, um die Übersicht zu behalten. Zudem kann eine geschickte Theoretische Stichprobenziehung (► Abschn. 9.2.1), die sich gezielt und reflektiert auf aussagekräftige Fälle beschränkt, den Arbeitsaufwand reduzieren. Nicht zuletzt profitiert man hier besonders von **Forschungsgruppen bzw. Forschungswerkstätten**, in denen sich ein fester Kreis von Forschenden regelmäßig (online oder offline) über Fortgang und Zwischenergebnisse der eigenen GTM-Projekte austauscht: Durch das Argumentieren werden Konzepte geschärft, Interpretationen im Dialog validiert, subjektive Standpunkte reflektiert, neue Ideen generiert und Durststrecken im Forschungsprozess durch wechselseitige Unterstützung besser überstanden (Mey & Mruck, 2009, S. 143).

Die Forschungsfelder, in denen die GTM und ihre Kodiermethode zum Einsatz kommt, umfassen u. a. Freizeit- (Piggott, 2010), Management- (Kempster & Parry, 2011), Familien- (Dieris, 2006), Gesundheits- (Neill, 2010), Migrations- (Yakushko, 2010) und Sportforschung (Holt & Tamminen, 2010). Es geht z. B. um das Ernährungsverhalten im Urlaub (Kim, Eves, & Scarles, 2009) oder den Umgang mit gefundenen Gegenständen (Camic, 2010). Es liegen auch Vorschläge zur Auswertung audio-visueller Dokumente mittels GTM vor (Figueroa, 2008). Das ► Studienbeispiel „Kodierung gemäß Grounded-Theory-Methodologie von forschungsgenerierten Dokumenten" illustriert das Vorgehen.

Studienbeispiel

Kodierung gemäß Grounded-Theory-Methodologie von forschungsgenerierten Dokumenten

Eine Studie des Entwicklungspsychologen Günter Mey (1999; 2007) ging der Frage nach, wie Jugendliche selbst **Jugend bzw. jugendlichen Identität** verstehen. Im Zuge des Forschungsprozesses wurde eine Theoretische Stichprobe von 16 Jugendlichen im Alter zwischen 16 und 20 Jahren zusammengestellt, die jeweils zwei Mal (im Abstand von etwa einem Jahr) an einem rund zweistündigen narrativen Interview teilnahmen. Als Erzählanstoß wurden die Befragten aufgefordert, auf ihr bisheriges Leben zurückzublicken, sich an den Moment zu erinnern, an dem sie „kein Kind mehr waren" und von diesem Moment an ausführlich weiterzuerzählen. Die Beispielstudie bezieht sich zwar auf Interview-Transkripte und somit auf forschungsgenerierte Dokumente, das Vorgehen ist jedoch auf die Auswertung vorgefundener Dokumente im Rahmen einer genuinen Dokumentenanalyse direkt übertragbar (für eine detaillierte Darstellung des methodischen Vorgehens anhand von Datenbeispielen der genannten Studie s. Mey & Mruck, 2009). Das Beispiel wurde gewählt, da hier die Vorgehensweise beim Kodieren besonders detailliert beschrieben wird; in vielen Studien mit vorgefundenen oder forschungsgenerierten Dokumenten wird nur sehr kursorisch auf den Kodierprozess verwiesen.

Um das Vorgehen beim Kodieren nach der **Methode des permanenten Vergleichs** zu veranschaulichen, werden die drei Schritte – offenes Kodieren, axiales Kodieren und selektives Kodieren (Strauss & Corbin, 1996) – behandelt und Theoriememos präsentiert.

Offenes Kodieren

Für das offene Kodieren wird das zu kodierende Dokument in einzelne Analyseeinheiten segmentiert. Dann werden **generative W-Fragen** an das Material herangetragen und Codes gebildet. Zu beachten ist, dass die Codebildung im GTM-Ansatz eine theoretische Abstraktions- und Reflexionsarbeit beinhaltet und sich nicht auf das bloße Paraphrasieren der Inhalte beschränkt. Das Vorgehen ist dabei in der Regel computergestützt.

Textpassage: Eine 19-jährige Interviewpartnerin der Studie leitete ihre autobiografische Erzählung folgendermaßen ein: „Na gut, also, ich glaub' erstmal es ist unheimlich schwer zu sagen, wann, . . . also selber so zu sagen, wann man kein Kind mehr war, wann man erwachsen war, weil teilweise wünsche ich mir immer noch Kind zu sein, so als Flucht. . ."

Die im Folgenden dargestellte offene Kodierung von zwei Textsegmenten aus dieser Textpassage ist Mey & Mruck (2009, S. 122f.) entnommen.

Textsegment Nr. 1: **„Na gut also ich glaub' erstmal es ist unheimlich schwer zu sagen"**

» Wieso „ich glaub'"? Ist „ich glaub'" Ausdruck einer Überzeugung oder Ausdruck von Unsicherheit und damit Gegenpart von Gewissheit? Was kann man glauben, was kann man wissen? Wer ist Experte/Expertin für was? Wieso „erstmal"? Meint das auf den ersten Blick, auf die Schnelle? Was ändert sich bei genauerer Betrachtung? Wenn „es" schwer zu erzählen ist, muss es dann nicht auch Anderes geben, was leicht zu erzählen ist? Was wäre leicht zu erzählen? Und ist alles von „es" schwer erzählbar oder nur „etwas von es"? Oder wäre „es" von Anderen leicht zu erzählen? Ist also zu unterscheiden zwischen schwer erzählbaren und leicht erzählbaren Geschichten? Könnte ebenso zwischen leichten und schweren Fragen unterschieden werden? Was macht eine Frage zur leichten, was zu einer schweren Aufgabe?
→ **Code: Unklare Zuständigkeiten für Lebensberichte.**
→ **Code: Geschichten vom eigenen Leben** (Dimension: Erfolgsgeschichte – Misserfolgsgeschichte; Erzählenswertes – Nicht-Erzählenswertes, Aufregendes – Banales, Richtiges – Falsches).

Textsegment Nr. 2: **„wann, [. . .] also selber so zu sagen, wann man kein Kind mehr war, wann man erwachsen war"**

» Wieso „selber" – wenn nicht „sie selber", wer dann? Könnten Andere das besser/richtiger? Gibt es die eine richtige Darstellung? Gibt es andere Zuständigkeit oder gar Instanzen dafür? Kann die Frage vielleicht besser „von außen" als „von innen" beantwortet werden? „wann man" – wieso „man" und nicht ich? Ist es ein Zuschreibungsprozess von außen? Gibt es gar nur die einzige Darstellung? Eine Norm, die für alle gilt? Und gibt es Abweichungen von der Norm? Wie werden diese Abweichungen erlebt (beängstigend, Widerstand erzeugend)?
„Kein Kind mehr" und „wann man erwachsen war" – Jugend wird nicht benannt. Ist Jugend nicht benennbar? Nicht definierbar? Hat keine Jugend stattgefunden?
→ **Code: Zuschreibungsprozesse von Jugend** (Dimension: Eigenzuschreibung– Fremdzuschreibung).
→ **Code: Aussparung von Jugend.**

Zu den sich als relevant erweisenden Codes werden **Codenotizen** (Bedeutung des Codes, Regeln der Zuordnung) erstellt, die zusammengenommen ein Codebuch ergeben. Zudem werden fortlaufend **Memos** verfasst, in denen Ideen zur Theoriebildung (z. B. Verknüpfung von Codes, dahinterliegende Prozesse), zum weiteren methodischen Vorgehen (z. B. Ideen für weitere in das theoretische Sample aufzunehmende Fälle) etc. festgehalten werden. Im Zuge des offenen Kodierens entsteht eine lange **Codeliste**. Inhaltlich ähnliche bzw. miteinander verknüpfte Codes werden zu übergeordneten theoretischen **Kategorien** gebündelt (Abb. 10.41). Auf der linken Seite in Abb. 10.41 sind die Codes alphabetisch geordnet aufgelistet. Die *blau unterlegten Codes* wurden zu einer übergeordneten Kategorie „Selbstkonstruktion zwischen Frühreife und Versagen" zusammengefasst. Die Gemeinsamkeit der gebündelten Codes wird in einem dahinterliegenden psychologischen Prozess vermutet, demgemäß die 19-jährige Interviewpartnerin sehr hohe Ansprüche an sich selbst als erwachsene Frau stellt, was gleichzeitig mit einem Erleben von Überforderung und Versagen einhergeht, wobei das Aussparen der Jugendphase ein auffälliges Merkmal ist (Mey & Mruck, 2009, S. 127f.).

Im Sinne der Methode des permanenten Vergleichs werden die übergeordneten Kategorien gebildet, indem die gebündelten Codes miteinander sowie abgrenzend mit den nicht einbezogenen Codes verglichen werden. Zudem findet auch immer wieder ein Abgleich mit dem Datenmaterial statt. Nicht zuletzt werden auch Vergleiche zwischen den Fällen angestellt. Diese permanenten Vergleiche gehen mit einer ständigen Überarbeitung und Verfeinerung der Codenotizen sowie der Theoriememos einher.

Axiales Kodieren

Beim axialen Kodieren werden die aus den Codes gebildeten Kategorien durch vergleichende Analyse verschiedener Textstellen eines Falles (sowie meist auch durch Vergleiche zwischen Fällen) weiter elaboriert und schließlich im Sinne des **Kodierparadigmas** in ein theoretisches Modell integriert (Abb. 10.42). Im hier betrachteten Einzelfall wird das Identitätsverständnis der 19-jährigen Interviewpartnerin, das sich durch eine „Selbstkonstruktion zwischen Frühreife und Versagen" auszeichnet, als erklärungsbedürftiges Phänomen in den Mittelpunkt gerückt (Abb. 10.42).

Ergänzend zur grafischen Darstellung des Theoriemodells wird ein erläuterndes Theoriememo erstellt. Auch hier sei noch einmal auf den hohen Komplexitätsgrad der Konzept- und Theoriebildung im GTM-Ansatz hingewiesen im Vergleich zu den paraphrasierend-beschreibenden Kategorienbildungen bei der Methode der qualitativen Inhaltsanalyse oder der Interpretativen Phänomenologischen Analyse.

» **Theoriememo zum Modell**: „Extreme, einander gegenüberstehende Pole und Widersprüche sind Kennzeichen der berichteten Biografie. Die Interviewte begnügt sich nicht mit dem, was sie kann und ist, sondern sie beurteilt sich bereits in der Gegenwart nach dem, was sie glaubt, sein bzw. können zu müssen [Phänomen]. Soziale Vergleiche, die die Interviewte kontinuierlich zur Bewertung ihres eigenen Handelns unternimmt/anstellt [Strategie] und mit denen sie konfrontiert wird bzw. die sie antizipiert [Handlungskontext], sind getragen von einem grundsätzlichen Zerrissensein zwischen einem (als mangelhaft dargestellten) Ich und den (als kritisch, sie potenziell abwertend dargestellten) Anderen. Dieses Zerrissensein nährt sich biografisch aus ihrer Orientierung an starken Frauen/Mutter oder schwachen Männern/Vater [Ursächliche Bedingungen] und wird aufgeladen durch Anforderungen an das ‚moderne' Subjekt im Allgemeinen und an ‚die moderne Frau' im Besonderen [Intervenierende Bedingungen]: Es ist die überfordernde Vision der Einheit von Schönheit (hier als Schlanksein) und Erfolg (intelligent zu sein) in Familie und Beruf und von der Doppelaufgabe in der Partnerschaft, sowohl emotionale Stütze als auch gleichberechtigte Gesprächspartnerin sein zu wollen. Von hier ausgehend, schwankt die Interviewte in ihrer Darstellung (und vermutlich auch in ihrem [Er-]Leben) zwischen einer ausgeprägten Selbstabwertung [Strategie] und diese begleitenden Versuchen der Selbst-(Wieder-)Errichtung bzw. Selbst-Behauptung [Strategien]. Allerdings: Wenn (Selbst-)Abwertung in Katastrophen mündet, reichen Strategien von Verharmlosung/Relativierung nicht mehr aus, sondern – nicht nur die Identität bedrohende, sondern die leib-seelische Unversehrtheit affizierende – Ereignisse (Suizidversuch, Klaustrophobie, Magersucht, Fehlgeburt) werden über eine ‚Verinselung' [Strategie] separiert; sie stehen in dem Lebens- und Erzählfluss nur mehr eruptiv und ohne Bezüge. Vor diesem Hintergrund muss eine Versöhnung von Stärke/Kompetenzen und Schwäche/Makeln, von Kindheitswünschen und Erwachsenenpflichten in eine ferne Zukunft projiziert werden [Konsequenz]." (Mey & Mruck, 2009, S. 133; Ergänzungen in eckigen Klammern hinzugefügt durch die Autorin).

Betrachtet man dieses Theoriememo, so wird auch deutlich, dass eine gegenstandsverankerte Theoriebildung keinesfalls rein induktiv, d. h. ohne jegliches theoretisches Vorverständnis erfolgen kann. Im Theoriememo angeführte theoretische Konzepte (z. B. Anforderungen an das „moderne Subjekt") wurden nicht direkt von der Befragungsperson genannt, sondern stehen in Bezug zu voraussetzungsvollen wissenschaftlichen Diskursen (Modernisierungstheorien, Subjekttheorien etc.). In Abgrenzung von einem „induktiven Selbstmissverständnis der GTM", demgemäß keinerlei theoretisches Vorwissen in die Analyse einfließen dürfe, wird heute stärker betont, dass Vorwissen auf Seiten der Forschenden nicht nur nicht ausgeblendet, sondern für eine gegenstandsverankerte Theoriebildung sogar fruchtbar gemacht werden kann, wenn es den Blick auf das Phänomen nicht verengt, sondern erweitert und schärft (Mey & Mruck, 2009, S. 105ff.).

Selektives Kodieren

Bislang wurde anhand der Interviewstudie zu jugendlichen Identitätskonstruktionen mit den Datenbeispielen v. a. auf eine einzelne 19-jährige Interviewpartnerin Bezug genommen, deren „Selbstkonstruktion zwischen Frühreife und Versagen" angesiedelt ist, weil sie sich keine Jugendphase zugesteht, sondern mit 19 Jahren bereits sehr hohe Anforderung an sich als erwachsene Frau stellt, an diesen dann scheitert und sich immer wieder wünscht „noch Kind zu sein". Diese spezielle Form der Selbstkonstruktion wird in der Interviewstudie mit den ebenso gründlich herausgearbeiteten Selbstkonstruktionen von 15 anderen Jugendlichen verglichen. Als **Schlüsselkategorie** der Studie dient somit der **„Modus der Selbstkonstruktion"**. Die Vorstellung der betrachteten Interviewpartnerin, mit dem Ende der Kindheit unmittelbar in das Erwachsenenleben eintreten zu müssen (geringe Kontinuität) und dabei zahlreichen äußeren Anforderungen und Normen genügen zu müssen (heteronome Orientierung), ordnet sich als „Sprunghaftigkeit" in ein Vierfelderschema der Selbstkonstruktionen (Abb. 10.43) ein. Dementsprechend lassen sich im Ergebnis der Studie **vier Typen von Identitätskonstruktionen von Jugendlichen** unterscheiden (Abb. 10.43).

Gemäß Kodierparadigma ließe sich auch für die Schlüsselkategorie „Modus der Selbstkonstruktion" auf abstrakterer Ebene fallübergreifend ein Theoriemodell erstellen. Im Unterschied zu dem in Abb. 10.42 dargestellten Theoriemodell für das Phänomen „Selbstkonstruktion zwischen Frühreife und Versagen" der einzelnen Interviewpartnerin, in dem ein Ursachenfaktor wie „starke Frauen als Vorbild" auftaucht, würde für das Theoriemodell der Schlüsselkategorie „Modus der Selbstkonstruktion", das für das gesamte Sample gilt, dann z. B. abstrakter „Rollenvorbilder" als Ursachenfaktor angeführt werden. (Für ein weiteres Studienbeispiel aus dem GTM-Ansatz zur fallübergreifenden qualitativen Datenanalyse ▶ Abschn. 12.1.2).

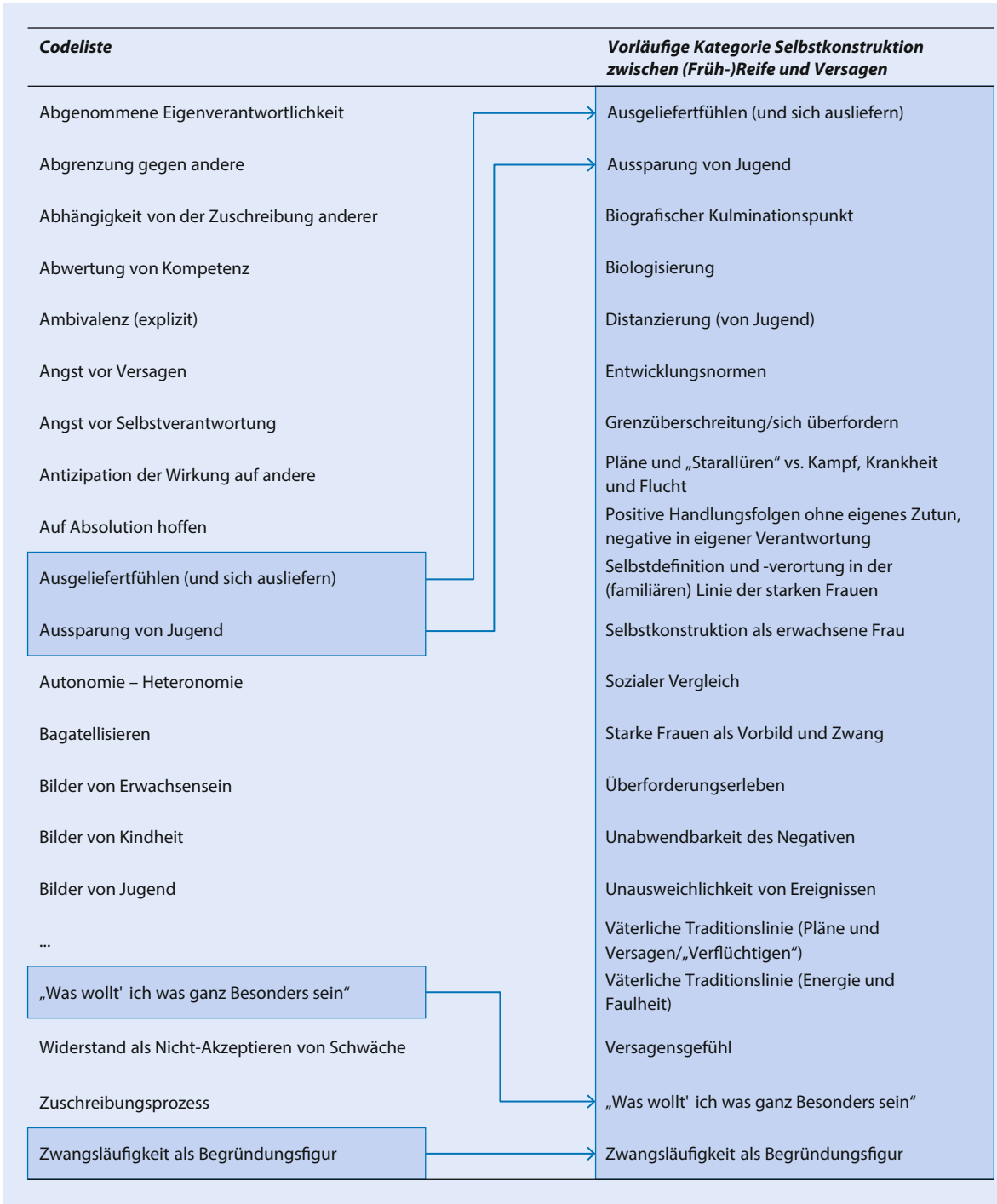


Abb. 10.41 Codeliste (Ausschnitt aus der Liste mit weit über 100 Codes) des Transkripts eines über zweistündigen narrativen Interviews zu jugendlicher Identität mit einer 19-jährigen Frau. (Aus Mey & Mruck, 2009, S. 128. Mit freundlicher Genehmigung von Regener, Berlin). In der linken Spalte sind die Codes blau hinterlegt, die in der rechten Spalte in einer vorläufigen Kategorie zusammengefasst wurden

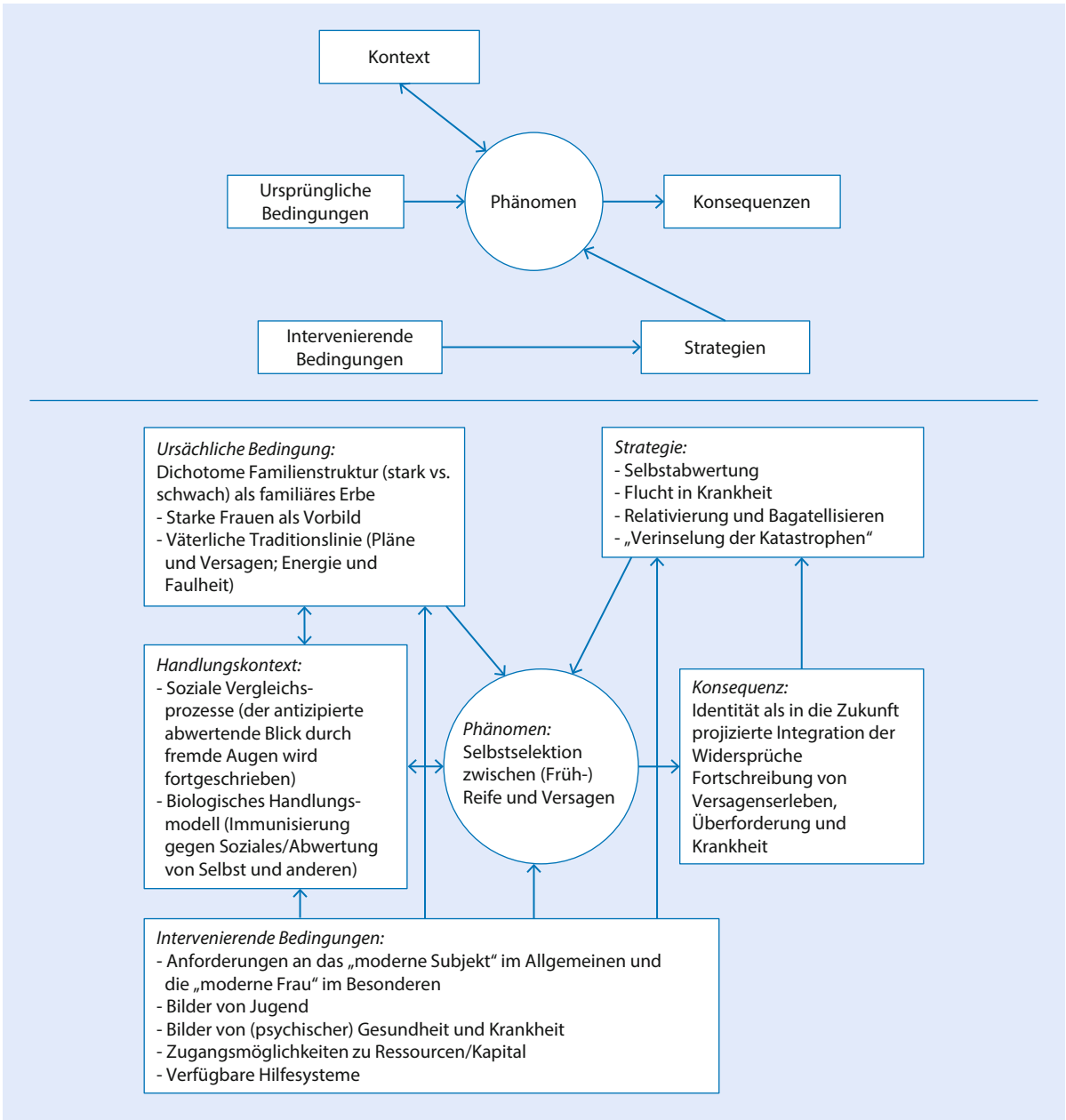


Abb. 10.42 *Oben*: Kodierparadigma nach Strauss. (Aus Strauss, 1991. Copyright © 1987 by Cambridge University Press; zitiert nach Mey & Mruck, 2009). *Unten*: Anwendung des paradigmatischen Modells nach Strauss (1991) am Beispiel der Ausdifferenzierung des Identitätsverständnisses der 19-jährigen Interviewpartnerin. (Aus Mey & Mruck, 2009, S. 131)

Qualitative Analyse von visuellen Dokumenten

Auf die wachsende Bedeutung visueller bzw. nicht-textueller Dokumente in unserer Kultur (Fotos, Videos, hypermediale Online-Dokumente etc.) wurde bereits hingewiesen. Dementsprechend wächst auch die Bedeutung **visueller Methoden der Datenerhebung** in der Sozialforschung, insbesondere in der qualitativen Forschung (zum Überblick Spencer, 2011). Unter Stichworten wie **visuelle Anthropologie** („visual anthropology“), **visuelle Ethnografie** („visual ethnography“), **visuelle Geschichte** („visual history“) und **visuelle Soziologie** („visual sociology“) werden Zugänge zu visuellen Dokumenten in etablierten Wissenschaftsdisziplinen zusammengefasst (Pauwels, 2010). Mit der **Bildwissenschaft**, die das Bild zu ihrem Untersuchungsgegenstand erklärt (z. B. Gesellschaft für interdisziplinäre Bildwissenschaft GIB: http://www.bildwissenschaft.org), und den repräsentationskritischen **Visual Studies**, die die Herstellung gesellschaftlicher Macht- und Herrschaftsverhältnisse durch bildliche Repräsentationen untersuchen (Elkins, 2003), haben sich

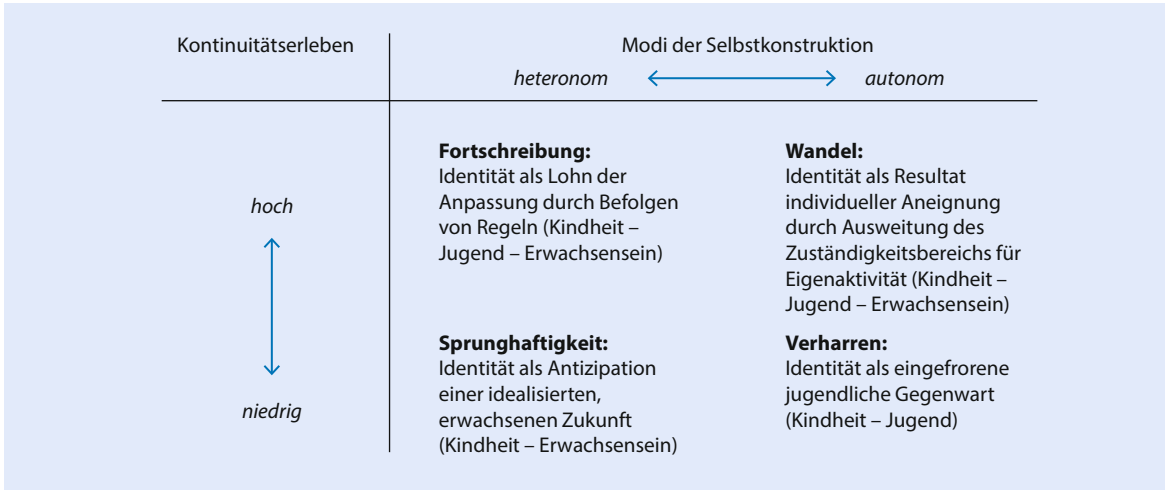


Abb. 10.43 Die Schlüsselkategorie „Modus der Selbstkonstruktion" lässt sich anhand von zwei Dimensionen (Modi der Selbstkonstruktion: heteronom/autonom; Kontinuitätserleben: hoch/gering) in ein Vierfelderschema der Identitätskonstruktion von Jugendlichen einteilen. (Aus Mey & Mruck, 2009, S. 135. Mit freundlicher Genehmigung von Regener, Berlin)

10

neue Forschungsfelder entwickelt, die auf den Trend zur visuellen Kultur reagieren.

Die genuine visuelle Dokumentenanalyse umfasst dabei die systematische Sammlung und Auswertung **vorgefundener visueller Dokumente**, („extant documents") die unabhängig vom Forschungsprozess im Alltag produziert und genutzt werden (z. B. Kinderzeichnungen, Familienfotos, Urlaubsvideos, Online-Fotoalben). Die Menge und Vielfalt entsprechenden visuellen Materials hat sich mit der allgegenwärtigen Verfügbarkeit von Foto- und Videokameras in Mobiltelefonen in jüngster Zeit immens vergrößert. Einer qualitativen Datenanalyse unterzogen werden können aber auch forschungsgenerierte visuelle Dokumente („research-generated documents"), seien sie durch die Forschenden erzeugt (z. B. Foto- und Videodokumentation einer teilnehmenden Feldbeobachtung) oder im Auftrag der Forschenden durch die Untersuchungsteilnehmenden hergestellt worden (z. B. Produktion von Zeichnungen, Collagen, Videoclips zu vorgegebenen Themen durch die Untersuchungsteilnehmenden).

Das Spektrum der qualitativen Auswertungsmethoden für visuelle Dokumente ist breit und kann hier nicht erschöpfend dargestellt werden (Van Leeuwen & Jewitt, 2001). Manche Verfahren sind nach dem Untersuchungsgegenstand benannt, wie z. B. die **Foto-** und die **Videoanalyse** (Knoblauch, Baer, Laurier, Petschke, & Schnettler, 2008) oder die **Filmanalyse** (Faulstich, 2008). Andere basieren auf bestimmten methodologischen Voraussetzungen wie z. B. die **semiotische** oder die **ikonografische Analyse** (Van Leeuwen & Jewitt, 2001), die neben den manifesten Bildinhalten v. a. auch eine Interpretation tieferliegender Bedeutungsebenen anstreben.

Aktuelle qualitative Dokumentenanalysen mit vorgefundenem visuellen Material beschäftigen sich z. B. mit der unterschiedlichen Darstellung von Hillary Clinton und Barrack Obama in Pressefotos während des US-Präsidentschaftswahlkampfs (Goodnow, 2010), mit der visuellen Medienrepräsentation von Opfern des Hurricane Katrina (Faux & Kim, 2006), mit der Auswertung von Bildern aus einem kommunalen Kunstprojekt (Mason, 2005) oder mit der Angst vor dem Zahnarzt, wie sie in selbstproduzierten Online-Videos dargestellt wird (Gao, Hamzah, Yiu, McGrath, & King, 2013).

10.6.3 Quantitative Dokumentenanalyse

Sollen im Rahmen einer quantitativen Dokumentenanalyse vorgefundene Dokumente einer quantitativ-statistischen Analyse zugänglich gemacht werden, so muss zunächst eine **quantitative Inhaltsanalyse** durchgeführt werden, um Messwerte zu erzeugen. Im Unterschied zur qualitativen Dokumentenanalyse, die wenige Dokumente detailliert interpretiert, wird bei der quantitativen Dokumentenanalyse mit sehr viel **größeren und teilweise auch probabilistischen Stichproben von Dokumenten** gearbeitet (▶ Abschn. 9.3.2). Dabei werden die Dokumente vor dem Hintergrund des jeweiligen Forschungsproblems hinsichtlich einzelner, theoretisch relevanter quantitativer Merkmale ausgewertet. Im Folgenden werden drei Varianten der quantitativen Inhaltsanalyse vorgestellt, die klassische quantitative Inhaltsanalyse, die primär für Textdokumente entwickelt wurde, die quantitative Online-Inhaltsanalyse, die auf Online-Dokumente

spezialisiert ist und die quantitative Inhaltsanalyse von visuellen Dokumenten, die für Fotos oder Videos geeignet ist.

Quantitative Dokumentenanalyse – Unter einer quantitativen Dokumentenanalyse („quantitative document analysis") als Datenerhebungsmethode verstehen wir die zielgerichtete, systematische und regelgeleitete Sammlung (quantitative Stichprobenziehung) und Archivierung von vorgefundenen textuellen/verbal-schriftlichen sowie nicht-textuellen (visuellen, auditiven, audio-visuellen, multimedialen, hypermedialen etc.) Dokumenten als Manifestationen menschlichen Erlebens und Verhaltens. Die Auswertung der Dokumente erfolgt **quantitativ-statistisch**. Um aus dem qualitativen Ausgangsmaterial Messwerte zu gewinnen, werden die Dokumente im ersten Schritt einer **quantitativen Inhaltsanalyse** unterzogen und mithilfe eines vollstandardisierten Kategoriensystems kodiert. Im zweiten Schritt werden die mit der quantitativen Inhaltsanalyse gewonnenen Messwerte dann mit den üblichen deskriptiv- und inferenzstatistischen Methoden der Datenanalyse ausgewertet.

Quantitative Inhaltsanalyse

Während bei der qualitativen Dokumentenanalyse eine Fülle von – teilweise erkenntnistheoretisch sehr unterschiedlich begründeten – qualitativen Datenanalyseverfahren zum Einsatz kommen, ist das Vorgehen bei der quantitativen Dokumentenanalyse mittels quantitativer Inhaltsanalyse sehr viel einheitlicher und strukturierter. Bei der quantitativen Inhaltsanalyse geht es darum, die formalen und inhaltlichen Merkmale von Dokumenten systematisch herauszuarbeiten und zu quantifizieren. Bei den **formalen Merkmalen** handelt es sich um äußerliche Dokumenteneigenschaften wie die Länge (z. B. Seiten- oder Zeichenzahl eines Zeitungsartikels), die Platzierung (z. B. Platzierung eines Zeitungsartikels auf der Titelseite bzw. nicht auf der Titelseite) oder die Bebilderung (z. B. wie viele Fotos enthält der Zeitungsartikel). Bei den **inhaltlichen Merkmalen** geht es um Aspekte dessen, was im Dokument inhaltlich ausgesagt wird (z. B. kommen in dem Artikel Experten zu Wort oder nicht; thematisiert der Artikel eher Vorteile, eher Nachteile oder Vor- und Nachteile einer Maßnahme).

Quantitative Inhaltsanalyse – Die quantitative Inhaltsanalyse („quantitative content analysis") erfasst deduktiv anhand eines vollstandardisierten Kategoriensystems die Ausprägungen unterschiedlicher formaler und inhaltlicher Merkmale von Dokumenten. Die Merkmalsausprägungen der Dokumente werden dabei quantifiziert, wobei je nach Definition der einzelnen Kategorien Messwerte auf allen vier Skalenniveaus generiert werden. Die per Inhaltsanalyse gewonnenen Messwerte werden anschließend statistisch ausgewertet.

Inhaltsanalytisches Kategoriensystem – Das inhaltsanalytische Kategoriensystem („content analysis coding system", „content analytic coding system", „content analytic coding scheme") als vollstandardisiertes Erhebungsinstrument operationalisiert theoriebasiert immer genau jene formalen und inhaltlichen Merkmale von Dokumenten, die für das jeweilige Forschungsproblem relevant sind. Damit das Kategoriensystem intersubjektiv nachvollziehbar ist, wird es mit genauen Definitionen aller Kategorien und Kodieranweisungen versehen, dadurch entsteht ein **Codebuch**. Die Kodierung der Dokumente anhand des Codebuches erfolgt durch **geschulte Kodierende**. Die psychometrischen Gütekriterien des Kategoriensystems, insbesondere **Validität** und **Reliabilität** sind zu prüfen und ihre mindestens ausreichende Qualität ist zu belegen (argumentativer Rückgriff auf Theorien und Forschungsstand zur Definition und Inhaltsvalidierung der Kategorien; Validitäts-Beurteilung der Kategorien bzw. des Kategoriensystems durch Fachexperten; empirische Reliabilitätsprüfung des Kategoriensystems).

Für das Grundverständnis der Methode der quantitativen Inhaltsanalyse und ihrer Fachbegriffe ist ein Vergleich zwischen quantitativer Inhaltsanalyse und quantitativer Befragung nützlich (Tab. 10.30).

Die quantitative Inhaltsanalyse ist historisch eng mit der **Untersuchung von massenmedialen Botschaften** verknüpft. So wurde im 18. Jahrhundert in Schweden die Häufigkeit religiöser Schlüsselbegriffe in lutherischen und pietistischen Texten verglichen, um deren Rechtgläubigkeit zu prüfen. Der wichtigste Anwendungsbereich der quantitativen Inhaltsanalyse war im 19. Jahrhundert die Zeitungsanalyse, im 20. Jahrhundert kamen Hörfunk- und Fernsehsendungen und neuerdings Online-Medienpublikationen als Untersuchungsobjekte hinzu. Inhaltlich stand bei der inhaltsanalytischen Medienauswertung die Frage nach propagandistischen und ideologischen Gehalten häufig im Mittelpunkt.

Als sozialwissenschaftliche Methode wurde die quantitative Inhaltsanalyse in den 1920er- und 1930er-Jahren ausgearbeitet. Erstmals in einem Lehrbuch beschäftigte sich der amerikanischen **Sozial- und Kommunikationswissenschaftler Bernard Berelson** mit der Inhaltsanalyse und definierte diese folgendermaßen

> » Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication. (Berelson, 1952, S. 18)

Der Politikwissenschaftler Ole R. Holsti definierte die quantitative Inhaltsanalyse als

> » any technique for making inferences by objectively and systematically identifying specified characteristics of messages. (Holsti, 1969, S. 14)

Unter den sozialwissenschaftlichen Disziplinen stützt sich v. a. die **Kommunikationswissenschaft** mit ihrer Medieninhaltsforschung auf die Methode der Inhaltsanalyse und bietet entsprechende Einführungen (z. B. Früh, 2007; Krippendorff, 2003; Merten, 1995; Neuendorf, 2001). Der Kommunikationswissenschaftler Werner Früh definiert:

> » Die Inhaltsanalyse ist eine empirische Methode zur systematischen, intersubjektiv nachvollziehbaren

Tabelle 10.30 Gegenüberstellung der Datenerhebung mittels quantitativer Inhaltsanalyse und quantitativer Befragung. (Mod. nach Scheufele & Engelmann, 2009)

	Quantitative Inhaltsanalyse	Quantitative Befragung
Datenerhebungs-instrument	Standardisiertes Kategoriensystem (bzw. Codebuch mit Kategoriensystem plus Kodieranweisungen)	Standardisierter Fragebogen
Merkmalsträger	Dokumente (z. B. Zeitungsartikel, Websites, Fotos)	Personen
Merkmale	Inhaltliche und formale Merkmale von Dokumenten (z. B. Anzahl der Wörter, Größe der Bilder, Hauptthemen, Bewertungen etc.)	Einstellungen, Eigenschaften, Verhaltensweisen etc. von Personen
Operationalisierung der Merkmale	Inhaltsanalytische Kategorien plus deren Ausprägungen	Statements/Fragen plus deren Antwortformate (= Fragebogen-Items)
Erhebungsform	Kodierung der Dokumente durch geschulte Kodierende	Selbstadministration eines Fragebogens (schriftliche Befragung) oder Interieweradministration eines Fragebogens (d. h. mündliches oder telefonisches Interview durch geschulte Interviewende)

> Beschreibung inhaltlicher und formaler Merkmale von Mitteilungen, meist mit dem Ziel einer darauf gestützten interpretativen Inferenz auf mitteilungsexterne Sachverhalte. (Früh, 2007, S. 27)

Die quantitative Inhaltsanalyse misst formale und inhaltliche Merkmale von Kommunikationsdokumenten bzw. Mitteilungen. Im Ergebnis können die Dokumente exakt beschrieben, verglichen und bewertet werden. Darüber hinaus werden auf der Basis inhaltsanalytischer Befunde auch oft Rückschlüsse auf Sachverhalte außerhalb der Dokumenteninhalte gezogen, z. B. **Rückschlüsse auf die Kommunikatoren** bzw. die Textproduzenten (z. B. politische Ausrichtung der Redaktion einer Tageszeitung, deren Artikel man analysiert hat) oder **Rückschlüsse auf die Rezipienten** bzw. das Publikum (z. B. verzerrte Wirklichkeitsvorstellungen durch realitätsabweichende Mediendarstellungen, die mittels Medieninhaltsanalyse aufgedeckt wurden). Auch **Rückschlüsse auf soziale, politische oder historische Situationen** können angestrebt werden, etwa durch einen systematischen inhaltsanalytischen Vergleich entsprechender Mediendokumente (Rössler, 2005). So lässt sich die These eines angeblichen „Sprachverfalls" bei Jugendlichen durch den ständigen Gebrauch neuer Medien (z. B. mailen, chatten, simsen etc.) anhand einer Inhaltsanalyse von **Schulaufsätzen** der letzten Dekaden empirisch überprüfen. Hierbei zeigt sich, dass die Schulaufsätze heute wesentlich lebendiger und individueller geschrieben sind, was eher auf eine Zunahme der Schreibkompetenz durch ein medienbedingt vermehrtes Schreiben im Alltag hindeutet (Lunsford et al., 2012; Stanford Study of Writing: http://ssw.stanford.edu). Rückschlüsse auf Produktions-, Rezeptions- oder Wirkungsprozesse, die über die Dokumenteninhalte hinausgehen, können jedoch schnell spekulativ geraten und müssen in der Regel durch Erhebungen mit anderen Methoden abgesichert werden (z. B. Interviews mit Journalisten, Beobachtung von Mediennutzern; zur Methodenkombination von quantitativer Inhaltsanalyse und Rezipientenbefragung in der Medienwirkungsforschung s. Wolling, 2002).

Im Zuge der Kritik an quantitativen Methoden in den 1970er-Jahren wurde verstärkt auf die Einseitigkeiten der – zuweilen als **„Discontent Analysis"** apostrophierten – quantitativen Inhaltsanalyse hingewiesen. Kritiker der quantitativen Inhaltsanalyse bezweifelten die Annahme, die Häufigkeiten bestimmter Begriffe oder Sprachformen seien indikativ für den Aussagegehalt eines Textes. Stumpfsinnige „Wortzählerei" könne den komplexen Bedeutungsgehalt von Textdokumenten nicht erfassen. Doch die quantitative Inhaltsanalyse beschränkt sich keineswegs auf das Auszählen einzelner Wörter, sondern kann – je nach Beschaffenheit des zugrunde liegenden Kategoriensystems – sehr viel komplexere Bedeutungsgehalte herausarbeiten. Zudem werden quantitative und qualitative Methoden auch bei der Inhaltsanalyse heute nicht mehr in einem Konkurrenz-, sondern in einem Ergänzungsverhältnis betrachtet und oft im Rahmen von Mixed-Methods-Designs kombiniert.

Das Spektrum der Forschungsthemen, die mit der quantitativen Inhaltsanalyse bearbeitet werden, ist breit gefächert, wobei traditionell die Auswertung von massenmedialen Dokumenten besonders verbreitet ist (z. B. Darstellung von Ostdeutschland im deutschen Fernsehen: Früh, Hasebrink, Krotz, Kuhlmann, & Stiehler, 1999; Darstellung von Hautkrebs in chinesischen Zeitungen: Liu, Liu, Xiao, Cai, & Xu, 2010). Inhaltsanalysen werden aber auch auf andere Arten von vorgefundenen Mate-

rialien angewendet, etwa auf Dokumente von sozialen Bewegungen (z. B. Anti-Glücksspiel-Bewegung in den USA: Bernhard, Futrell, & Harper, 2010) auf wissenschaftliche Konferenzprogramme (Helwig & Schmidt, 2011) und Fachzeitschriften (Wong, Steinfeldt, Speight, & Hickman, 2010) sowie auf Tagebücher (Gackenbach, Sample, Mandel, & Tomashewsky, 2011). Neben Dokumenten werden auch Artefakte wie z. B. Merchandising-Artikel inhaltsanalytisch ausgewertet (Hirdes, Woods, & Badzinski, 2009). In der Psychologie wird die quantitative Inhaltsanalyse u. a. genutzt, um anhand von Kontaktanzeigen („personal ads") das Partnerwahlverhalten zu studieren (Lance, 1998; Sakai & Johnson, 1997; Salska et al., 2008). Auch Aspekte der geschlechts- und kulturspezifischen Darstellung von Personen werden mit quantitativen Inhaltsanalysen untersucht. So zeigen Inhaltsanalysen von Dokumenten aus Massenmedien sowie auch von selbstproduzierten Dokumenten (z. B. Fotos auf Social Networking Sites wie Facebook) teilweise einen sog. Facial-Prominence-Effekt in dem Sinne, dass bei Männern häufiger nur das Gesicht („face-isms), bei Frauen dagegen häufiger der ganze Körper („body-ims") abgebildet wird, was wiederum Einfluss auf die Personenwahrnehmung haben kann (Archer, Iritani, Kimes, & Barrios, 1983; Copeland, 1989; Zuckerman & Kieffer, 1994).

Ein Sonderfall der Inhaltsanalyse – der auch im Rahmen von Evaluationsstudien zum Einsatz kommt (zur Evaluationsforschung ▶ Kap. 18) – ist die **Qualitätsanalyse von Dokumenten**. Hierbei wird ein Kategoriensystem zugrunde gelegt, das theoretisch fundierte inhaltliche Qualitätskriterien operationalisiert. So lässt sich z. B. anhand eines Kategoriensystems journalistischer Qualitätskriterien überprüfen, ob und inwiefern speziell in der Kriegsberichterstattung journalistische Standards verletzt werden (Gleich, 2003). Oder es kann die Qualität von Online-Gesundheitsinformationen wissenschaftlich überprüft werden anhand einer quantitativen Inhaltsanalyse mit einem Kategoriensystem, das Kriterien für „evidenzbasierte Patienteninformationen" operationalisiert (Mühlhauser & Oser, 2008).

Ein spezielles Anwendungsfeld der Inhaltsanalyse sind schließlich **Input-Output-Analysen**, bei denen z. B. der Input in ein Mediensystem in Form politischer Pressemitteilungen mit dem medialen Output hinsichtlich der Quantität und Qualität von Argumenten in Presseartikeln verglichen wird (Kuhlmann, 1999).

Die Untersuchungsobjekte/Untersuchungseinheiten einer Inhaltsanalyse werden **Analyseeinheiten** genannt. Wird z. B. eine Stichprobe von Zeitungsartikeln ausgewertet, so sind die Zeitungsartikel die Analyseeinheiten. Soll untersucht werden, wie stark in der Presseberichterstattung auf wissenschaftliche Quellen zurückgegriffen wird, so wird die Kategorie „wissenschaftliche Quelle/n genannt" mit den Ausprägungen „ja/nein" pro Artikel kodiert. Da die Kategorie genau einmal auf jeden Artikel angewendet wird, stellt er gleichzeitig auch die Kodiereinheit dar. Als Ergebnis kann dann angegeben werden, welcher Prozentsatz der untersuchten Artikel wissenschaftliche Quellen nennt.

In manchen Fällen sind Analyseeinheit und **Kodiereinheit** jedoch nicht identisch (Früh, 2007, S. 95). Das ist immer dann der Fall, wenn eine Kategorie auf mehrere Elemente einer Analyseeinheit anzuwenden ist. Will man etwa genauer untersuchen, wie Zeitungsartikel visuell aufbereitet sind, so wären Merkmale der visuellen Aufbereitung nicht auf den Artikel insgesamt, sondern auf jede einzelne Abbildung im Artikel anzuwenden (Größe, Farbigkeit, Motive der Abbildung etc.), so dass die Abbildungen somit die Kodiereinheiten darstellen. Bei der Ergebnisdarstellung einer quantitativen Inhaltsanalyse ist dann zu beachten, dass Aussagen über Kodiereinheiten (z. B. Prozentsatz aller Abbildungen, die Personen zeigen) in Aussagen über Analyseeinheiten umgerechnet werden müssen (z. B. Prozentsatz aller Zeitungsartikel, die Abbildungen mit Personen enthalten; Lauf & Berens, 2003). Eine besonders komplexe Kodiereinheit stellen oft **Akteure** dar, seien sie in textueller und/oder bildlicher Form in Dokumenten repräsentiert. Akteure können nämlich hinsichtlich sehr vieler verschiedener Kategorien eingeordnet werden, angefangen von einfachen soziodemografischen Merkmalen bis hin zu komplexen sozialen Rollen. Im Rahmen der Umweltberichterstattung können Personen z. B. in der Rolle von Stromverbrauchern, Wahlbürgern, politischen Aktivisten, Verkehrsteilnehmern, Fachexperten etc. auftauchen. Hier muss für die Bearbeitung eines entsprechenden Forschungsproblems jede relevante Person als Kodiereinheit sorgfältig mit verschiedenen (jeweils eindeutig definierten, trennscharfen und erschöpfenden) Kategorien beschrieben werden.

Im Folgenden wird das Vorgehen bei einer quantitativen Inhaltsanalyse vorgestellt und anschließend anhand eines authentischen Studienbeispiels (Darstellung der UN-Klimakonferenz 2009 in Kopenhagen in der deutschen Presse) schrittweise illustriert. Zudem wird die Reliabilitätsprüfung eines inhaltsanalytischen Kategoriensystems behandelt.

■ Arbeitsschritte einer quantitativen Inhaltsanalyse

Die Durchführung einer quantitativen Dokumentenanalyse mittels quantitativer Inhaltsanalyse kann in zwölf Arbeitsschritte aufgeteilt werden: 1.) Formulierung der Forschungsfragen bzw. Forschungshypothesen, 2.) Untersuchungsplanung, 3.) Populationsdefinition und Stichprobenziehung, 4.) Archivierung und Aufbereitung des Untersuchungsmaterials, 5.) Deduktiver Entwurf des Kategoriensystems sowie des Codebuches, 6.) Induktive

Überarbeitung des Kategoriensystems sowie des Codebuches, 7.) Pretest und Revision des Kategoriensystems sowie des Codebuches, 8.) Kodiererschulung, 9.) Reliabilitätsanalyse und Finalisierung des Kategoriensystems sowie des Codebuches, 10.) Datenerhebung durch Kodierung der gesamten Stichprobe, 11.) Statistische Datenanalyse der per Inhaltsanalyse generierten Messwerte und 12.) Interpretation der statistischen Ergebnisse der Inhaltsanalyse.

1. **Formulierung der Forschungsfragen bzw. Forschungshypothesen**: Auf der Basis einer gründlichen Recherche einschlägiger Theorien sowie des aktuellen Forschungsstandes (▶ Kap. 6) werden konkrete Forschungsfragen bzw. Forschungshypothesen formuliert (▶ Kap. 5). Inhaltsanalytische Studien dienen oft der Beantwortung von zwei bis vier **offenen Forschungsfragen**, in denen es relativ allgemein um die Art und Weise der Repräsentation bestimmter Themen oder Phänomene in bestimmten Typen von Dokumenten geht. **Gerichtete Hypothesen** sind v. a. sinnvoll, wenn Dokumente unterschiedlicher Urheber oder Mediengattungen (z. B. Radio- versus TV-Nachrichtensendungen; TV-Nachrichten im öffentlich-rechtlichen versus privatwirtschaftlichen Fernsehen), unterschiedlicher Kulturkreise (z. B. TV-Nachrichten in Deutschland versus in der Schweiz) oder verschiedener Zeitperioden (z. B. TV-Nachrichten in den 1960er- versus 1990er-Jahren) verglichen werden sollen.
2. **Untersuchungsplanung**: Auf der Basis der Forschungsfragen bzw. -hypothesen sind Überlegungen zur Untersuchungsplanung anzustellen. Dies betrifft das **Untersuchungsdesign** (▶ Kap. 7; z. B. nicht-experimenteller querschnittlicher Vergleich zwischen verschiedenen Mediengattungen; Längsschnittstudie über zwei Dekaden für ein ausgewähltes Medium etc.) sowie die **Definition der Grundgesamtheit** und die **Methode der Stichprobenziehung der Dokumente** (▶ Kap. 9) inklusive Vorüberlegungen zum **optimalen Stichprobenumfang** für ausreichende statistische Teststärke (▶ Kap. 14). Diese methodischen Überlegungen sind zu verknüpfen mit praktischen Planungen der Dokumentenbeschaffung (▶ Beschaffung vorhandener Dokumente für eine wissenschaftliche Dokumentenanalyse). Entscheidungen hinsichtlich geeigneter **Software-Tools** für die Archivierung und Analyse des Materials sind zu treffen. Es ist empfehlenswert, unabhängig von der späteren Stichprobe der Haupterhebung bereits im Zuge der Untersuchungsplanung probehalber erstes Material aus der avisierten Grundgesamtheit zu sammeln, um dessen Beschaffenheit bei der Entwicklung des Kategoriensystems vor Augen zu haben. Bei umfangreichen Studien sollte vorab geplant werden, wie viele Kodierer/innen wie zu rekrutieren und zu entlohnen sind. Im Zuge der Untersuchungsplanung sind auch die Zeitabläufe zu klären. Sollen z. B. Dokumente über einen längeren Zeitraum hinweg gesammelt werden, so muss die Stichprobenziehung teilweise parallel zur Entwicklung des Kategoriensystems erfolgen.
3. **Populationsdefinition und Stichprobenziehung**: Die Zielpopulation der Dokumente, über die Aussagen getroffen werden sollen, ist zu definieren. Bei vorgefundenen Mediendokumenten müssen z. B. Eingrenzungen hinsichtlich Mediengattung (z. B. regionale und überregionale Tageszeitungen), Medienangebot (z. B. Zeitungsartikel aus dem Politik- und/oder Sportressort), inhaltlicher Ausrichtung (z. B. Zeitungsartikel zum Thema Klimawandel), Kulturkreis (z. B. deutsche Zeitungen) und Zeitperiode (z. B. Beiträge aus dem Jahr 2011) getroffen werden. Aus dem auf diese Weise zielgerichtet gesammelten Material kann dann z. B. eine einfache Zufallsstichprobe gezogen werden. Aus der Population aller Ausgaben einer Zeitung oder Zeitschrift eines Jahres kann eine systematische Auswahl getroffen werden im Sinne einer **künstlichen Woche**, indem über das Jahr verteilt die verschiedenen Wochentage pro Monat oder pro Woche gewählt werden (z. B. erster Montag im Januar, erster Dienstag im Februar etc.). Die Wahl eines Wochentages kann dabei als Klumpen verstanden werden, wenn aus der Zeitungsausgabe bzw. dem TV-Programm des betreffenden Tages dann noch einzelne Artikel bzw. Sendungen ausgewählt werden (**mehrstufige Zufallsstichproben** ▶ Abschn. 9.3.2). Dabei ist allerdings zu beachten, dass die meist geringe Zahl der Klumpen pro Jahr (z. B. $n = 14$ bei zwei künstlichen Wochen pro Jahr) zu erheblichen **Klumpungsfehlern** führen kann: Wenn z. B. der Anteil des Themas Religion an der Medienberichterstattung analysiert wird, hat es entscheidenden Einfluss auf die Ergebnisse, ob zufällig einer der Stichprobentage auf einen Papstbesuch oder Weihnachten entfällt (Jandura, Jandura, & Kuhlmann, 2005). Weniger fehleranfällig (wenn auch organisatorisch oft wesentlich aufwändiger) wären Zufallsstichproben mit kleineren Erhebungseinheiten und größeren Fallzahlen: Statt 14 Fernsehsendetagen 24 Stunden könnten dann z. B. 336 ($= 14 \times 24$) per Zufall ausgewählte Sendestunden erhoben werden, was zumindest bei computergestützter Aufzeichnungstechnik auch nicht sehr viel mehr Aufwand bedeuten würde. Bei gedruckten Tageszeitungen wären analog anstelle von kompletten Ausgaben jeweils einzelne Artikel in größerer Zahl per Zufall zu ziehen. Der zu betreibende Aufwand hängt auch hier wieder von der Zielsetzung der Studie sowie vom disziplinären Umfeld ab: In der auf Me-

dieninhaltsanalysen spezialisierten Kommunikationswissenschaft werden diesbezüglich z. B. meist höhere Anforderungen gestellt als in anderen sozialwissenschaftlichen Disziplinen.

4. **Archivierung und Aufbereitung des Untersuchungsmaterials**: Wurden die Dokumente beschafft, so müssen diese systematisch archiviert und aufbereitet werden. Eine Sammlung von Printartikeln muss z. B. eingescannt und unter sprechenden Dateinamen abgespeichert werden. Werden private Dokumente gesammelt und archiviert, ist im Zuge der Datenaufbereitung für eine sorgfältige Anonymisierung zu sorgen (► Kap. 11).
5. **Deduktiver Entwurf des Kategoriensystems sowie des Codebuches**: Anhand der Forschungsfragen bzw. Forschungshypothesen müssen die zentralen formalen und inhaltlichen Merkmale herausgearbeitet werden, hinsichtlich derer die untersuchten Dokumente analysiert werden sollen. Bei der Auswahl, Definition und Ausdifferenzierung der relevanten Konzepte im Sinne einer **Konzeptspezifikation** (► Abschn. 8.2) wird auf Vorläuferstudien sowie auf thematisch einschlägige wissenschaftliche Theorien zurückgegriffen. Oftmals kann hierarchisch vorgegangen werden, indem zunächst grobe Themenblöcke bzw. übergeordnete Konzepte festgelegt werden, die dann im Sinne einer **dimensionalen Analyse** in Unteraspekte ausdifferenziert werden. Da es sich bei dem Kategoriensystem einer quantitativen Inhaltsanalyse um ein vollstandardisiertes Datenerhebungsinstrument handelt, müssen im Zuge der Ausformulierung der einzelnen Kategorien auch die Merkmalsausprägungen präzise festgelegt werden. Die Ausprägungen einer nominalen Kategorie müssen dabei die **Kriterien der Genauigkeit, Exklusivität und Exhaustivität** erfüllen (► Abschn. 8.4.2). Den Kategorien-Ausprägungen werden dann im Zuge der Kodierung und anschließenden **Messung** numerische Werte zugeordnet. Die in einem Kategoriensystem untergebrachten Kategorien können dabei alle vier Skalenniveaus umfassen.

   Um eine intersubjektiv nachvollziehbare und einheitliche spätere Anwendung des Kategoriensystems zu gewährleisten, ist parallel zur Entwicklung des Kategoriensystems am **Codebuch** zu arbeiten.
6. **Induktive Überarbeitung des Kategoriensystems sowie des Codebuches**: Der deduktiv entwickelte Entwurf des Kategoriensystems wird an Beispielmaterial (das nicht zur späteren Stichprobe gehört) erprobt. Dabei können dann die **Kategorienausprägungen** gegenstandsangemessen festgelegt werden. Ebenso können die **Kodieranweisungen im Codebuch** präzisiert und mit Beispielen unterlegt werden. Nicht zuletzt stößt man beim Abgleich zwischen Kategoriensystem und Beispieldokumenten im Material auch immer wieder auf formale und inhaltliche Dokumenten-Merkmale, die für das Thema relevant sind, aber bislang im Kategoriensystem fehlten. Die induktive Überarbeitung beinhaltet also auch das **Hinzunehmen neuer Kategorien** sowie ggf. die **Streichung von unbrauchbaren Kategorien**, die im Material nur extrem selten vorkommen oder zu schwer kodierbar sind. Durch die induktive – datenbasierte – Überarbeitung des Kategoriensystems soll dessen inhaltliche Passung sowie Anwendbarkeit optimiert und zudem das Codebuch elaboriert werden. Die induktive Überarbeitung des Kategoriensystems sollte nicht anhand von Daten aus der späteren Stichprobe erfolgen, um zu vermeiden, dass das Erhebungsinstrument genau in der Weise auf die Daten abgestimmt werden kann, dass Wunschergebnisse entstehen, was als Manipulation und Verletzung der Regeln guter wissenschaftlicher Praxis anzusehen wäre (zur Wissenschaftsethik ► Abschn. 4.2). Bei der theoriebasiert-deduktiven und datenbasiert-induktiven Konstruktion des Kategoriensystems kommt es vor allem darauf an, die **Validität** des entwickelten Datenerhebungsinstruments zu sichern, damit es tatsächlich diejenigen Merkmale der zu untersuchenden Dokumente erfasst, die für das Forschungsproblem relevant und für die Studie angezielt sind. Für Prüfung und Sicherstellung der Validität der einzelnen Kategorien sowie des gesamten inhaltsanalytischen Kategoriensystems sind neben einer sorgfältigen theoretischen Ableitung und Begründung der Kategorien auch Beurteilungen durch Fachexperten empfehlenswert.
7. **Pretest und Revision des Kategoriensystems sowie des Codebuches**: Das deduktiv und induktiv entwickelte Kategoriensystem wird anhand einer Stichprobe von Beispieldokumenten einem Pretest durch mehrere Kodierer unterzogen, wobei es sich meist um Mitglieder aus dem Forschungsteam handelt, die mit dem Kategoriensystem vertraut sind. Für den Pretest können auch weitere Kodierer eingesetzt werden, die jedoch zuvor eine Schulung zur korrekten Anwendung des Kategoriensystems durchlaufen sollten (► Punkt 8). Anhand der Rückmeldungen und Ergebnisse des Pretests (z. B. sehr stark abweichende Kodierungen, unverständliche Kodieranweisungen, sehr großer Zeitaufwand zur Anwendung einer Kategorie etc.) werden die Kategorien und ihre Ausprägungen sowie die Erläuterungen im Codebuch überarbeitet (Revision des Kategoriensystems). An die finale Fassung von Kategoriensystem und Codebuch sind die Anforderungen der klaren Verständlichkeit sowie der komfortablen Anwendung zu stellen. Das Codebuch enthält auch einen Codebogen bzw. eine Datenmas-

ke, in welche die im Zuge der Kodierung gewonnenen Messwerte eingetragen werden können.

8. **Kodiererschulung**: Bei Studien mit sehr hohem Kodieraufwand werden mehrere Kodierer eingesetzt. Der gesamte **Kodieraufwand** berechnet sich über die Zeit für die Kodierung eines Dokuments multipliziert mit dem Stichprobenumfang. Werden z. B. für die Kodierung eines einzelnen Zeitungsartikels als Analyseeinheit auf der Basis des vorliegenden Kategoriensystems ca. 30 Minuten benötigt, entsteht für die Kodierung eines Samples von $n = 400$ Artikeln ein Zeitaufwand von rund $0.5 \times 400 = 200$ Stunden. Bei einer 40-Stunden-Arbeitswoche würde ein einzelner Kodierer also fünf Wochen lang an der Kodierung der Dokumentenstichprobe arbeiten. Somit kann es sinnvoll sein, die Arbeit auf mehrere Kodierende zu verteilen, die z. B. als studentische Hilfskräfte beschäftigt werden oder im Rahmen eines Forschungsseminars am Projekt mitarbeiten. Es ist jedoch zu beachten, dass eine größere Zahl an Kodierenden wiederum auch einen höheren Organisationsaufwand (Kodiererschulung) und neue Fehlerquellen (Integration der verschiedenen Datensätze etc.) mit sich bringt. Bei Qualifikationsarbeiten wird die Kodierung meist von den jeweiligen Studierenden vollständig selbst übernommen.

   Da die Kodierenden zusammen mit dem Kategoriensystem als Messinstrumente fungieren, müssen sie vor der Haupterhebung gemeinsam gründlich geschult werden, um sicherzustellen, dass sie in vergleichbarer Weise kodieren. Im Rahmen der Kodiererschulung wird allen Kodierenden die Zielsetzung der Studie erläutert sowie das Kategoriensystem und Codebuch samt Codebogen vorgestellt. Anschließend werden Beispieldokumente in Einzelarbeit kodiert und die Kodiererfahrungen (z. B. Unklarheiten im Codebuch) sowie die Kodierergebnisse (z. B. abweichende und übereinstimmende Kodierungen) im Rahmen einer gemeinsamen Diskussion ausgetauscht. Als Maßstab für die gültige (valide) Kodierung dienen im Zweifelsfall die Vorgaben des Forschungsteams. Unklarheiten und Abweichungen werden analysiert und schlagen sich notfalls in einer weiteren Überarbeitung (Ergänzung, Präzisierung) von Kategoriensystem und Codebuch nieder.

9. **Reliabilitätsanalyse und Finalisierung des Kategoriensystems sowie des Codebuches**: Jede einzelne Kategorie im inhaltsanalytischen Kategoriensystem muss einer Reliabilitätsanalyse unterzogen werden. Dazu wird entweder eine Pretest-Stichprobe von Dokumenten (die aus der Zielpopulation stammen, aber nicht zum Datensatz der Haupterhebung gehören) oder ein Ausschnitt von konventionell ca. 10–20 % des Datenmaterials der Hauptuntersuchung von mindestens zwei unabhängigen geschulten Kodierern ausgewertet und die **Kodiererübereinstimmung** bzw. **Inter-Kodierer-Reliabilität** („inter-coder agreement/reliability“) pro Kategorie bestimmt. Weniger überzeugend als eine Reliabilitätsüberprüfung mittels Inter-Kodierer-Übereinstimmung ist die **Intra-Kodierer-Reliabilität** („intra-coder agreement/reliability“), bei der dieselbe Person in zeitlichem Abstand dieselben Dokumente zweimal kodiert und die Übereinstimmung zwischen den beiden Messzeitpunkten pro Kategorie berechnet wird. Welcher statistische Übereinstimmungskoeffizient dabei zu berechnen ist, hängt vom Skalenniveau der jeweiligen Kategorie ab (die Kodiererübereinstimmung als Reliabilitätsmaß für Kategorien in einem inhaltsanalytischen Kategoriensystem entspricht der Beobachterübereinstimmung als Reliabilitätsmaß für Kategorien in einem standardisierten Beobachtungssytem; ► Abschn. 10.1.3 „Beobachterübereinstimmung“).

   Die Reliabilitätskoeffizienten sollten bei einfachen manifesten Kategorien sehr nahe bei 1 bzw. 100 % liegen. Bei Kategorien, deren Kodierung mehr Interpretationsaufwand erfordert, sind Übereinstimmungen unter 100 % typisch, sie sollten aber je nach verwendetem Übereinstimmungskoeffizienten bestimmte Mindeststandards nicht unterschreiten, da die erhobenen Daten sonst aufgrund der hohen Belastung durch Messfehler kaum interpretierbar sind. Bei der Beurteilung von Reliabilitätskoeffizienten ist immer zu beachten, dass die kursierenden **Reliabilitätsstandards** („benchmarks“), die angeben, ab welchem Wert von geringer, mittelmäßiger, hoher oder ausgezeichneter Reliabilität auszugehen ist, nur grobe Anhaltspunkte liefern können, da hier jeweils auch inhaltliche Aspekte (Komplexität der betreffenden Kategorie) und empirische Faktoren (z. B. Anzahl und Besetzung der Kategorienausprägungen) eine Rolle spielen. Bei Kategorien mit unzureichender Reliabilität müssen die Ursachen eruiert werden. Notfalls muss abermals an der Präzisierung von Codebuches und Kategoriensystems gearbeitet werden. Ist eine Erhöhung der Reliabilität inhaltlich oder aus forschungsökonomischen Gründen nicht möglich, so sind die messungenauen Kategorien als solche zu kennzeichnen und dürfen nur mit großer Vorsicht interpretiert werden.

   Nicht zuletzt sei erwähnt, dass zur Reliabilitätsabschätzung jeweils eine Stichprobe **ausreichender Mindestgröße** bzw. eine Stichprobe mit **optimalem Stichprobenumfang** (zum Konzept des „optimalen Stichprobenumfanges“ ► Abschn. 14.3) zu wählen ist, damit der jeweilige Übereinstimmungskoeffizient als aussagekräftig gelten kann. Eine Reliabilitätsschät-

zung auf der Basis einer stark verzerrten sowie zu kleinen Stichprobe ist fragwürdig (zu Stichprobenanforderungen bei der Reliabilitätsanalyse z. B. Gwet, 2012 sowie ► Abschn. 10.1.3 „Beobachterübereinstimmung“).

10. **Datenerhebung durch Kodierung der gesamten Stichprobe**: Liegt ein finales Kategoriensystem samt Codebuch vor, dessen Kategorien sich alle als valide und ausreichend reliabel erwiesen haben, so können die geschulten Kodierenden mit der Kodierung der ihnen zugeteilten Dokumente beginnen. Kodierung und Messung erfolgen dabei anhand von **Codebögen** bzw. durch direkte Eingabe der Messwerte in ein **Statistikprogramm** wie SPSS. Dabei müssen alle Kodierenden mit derselben Datenmaske arbeiten. Die Datenfiles der einzelnen Kodierenden sind am Ende in ein Gesamtdatenfile zu integrieren und z. B. um Tippfehler zu bereinigen (zur Datenaufbereitung ► Abschn. 11.4). Wird in der Hauptuntersuchung mit mehreren Kodierenden gearbeitet, so empfiehlt sich eine erneute Prüfung der Urteiler-Übereinstimmung durch Parallelkodierungen, um ggf. Kodierende zu identifizieren, deren Urteile systematisch von denen der anderen Kodierenden abweichen.
11. **Statistische Datenanalyse der per Inhaltsanalyse generierten Messwerte**: Der aufbereitete Gesamtdatensatz wird einer deskriptiv- und inferenzstatistischen Analyse gemäß den Forschungsfragen bzw. Forschungshypothesen unterzogen, wobei z. B. Unterschiede, Zusammenhänge oder Veränderungen geprüft werden (zur quantitativen Datenanalyse ► Abschn. 12.2). Die Ergebnisdarstellung erfolgt durch Kennwerte im Fließtext sowie – bei wichtigen Befunden – in Form von Tabellen und Grafiken. Da viele inhaltsanalytische Kategoriensysteme mehrheitlich nominalskalierte Kategorien enthalten, spielt die Darstellung und Analyse der Häufigkeiten der Ausprägungen von nominalen Kategorien eine wichtige Rolle (**Frequenzanalyse**, z. B. mittels eindimensionalem Chi-Quadrat-Test). Werden mehrere nominalskalierte Kategorien in ihrem Zusammenspiel betrachtet, handelt es sich um eine **Kontingenzanalyse** (z. B. Auswertung von Kreuztabellen über den zweidimensionalen bzw. $k \times l$-Chi-Quadrat-Test). Werden kardinalskalierte Kategorien ausgewertet – etwa über Mittelswertsunterschiede oder Korrelationen, so spricht man im Kontext der quantitativen Inhaltsanalyse von **Intensitätsanalysen**.
12. **Interpretation der statistischen Ergebnisse der Inhaltsanalyse**: Die Ergebnisinterpretation einer quantitativen Inhaltsanalyse sollte sich zunächst auf die Inhalte der Dokumente konzentrieren. Rückschlüsse von den Dokumenten-Inhalten auf die möglichen Umstände ihrer Produktion, die Effekte ihrer Rezeption oder sonstige soziale und historische Rahmenbedingungen sollten erst auf dieser Basis und immer mit großer Vorsicht erfolgen. Ein typischer Fehler inhaltsanalytischer Studien ist die Überinterpretation der Dokumenten-Inhalte im Hinblick auf die Wirkungen beim Publikum. Da eine inhaltsanalytische Studie weder Rezipierende noch Rezeptionsprozesse erfasst, bergen direkte Rückschlüsse vom Material auf Publikumswirkungen die Gefahr, mediendeterministischer Vereinfachungen und Spekulationen. Denn ein aktives Publikum hat immer auch die Möglichkeit, bestimmte Medieninhalte **nicht** oder sehr **selektiv** zu nutzen, umzudeuten und „gegen den Strich“ zu lesen (zum Inferenzproblem der Inhaltsanalyse z. B. Früh, 2007; Merten, 1995).

■ Studienbeispiel einer quantitativen Inhaltsanalyse mit vorgefundenen Dokumenten

Wie wurde die **UN-Klimakonferenz 2009 in Kopenhagen** in der **deutschen Presse** dargestellt? Welche Konflikte und Lösungsansätze der unterschiedlichen politischen Akteure wurden thematisiert? Um diese Forschungsfragen zu beantworten, war eine quantitative Inhaltsanalyse von Zeitungsartikeln (also vorgefundenen Dokumenten) indiziert, die im Rahmen eines Forschungsseminars an der Technischen Universität Ilmenau durchgeführt wurde (Arlt & Wolling, 2012). Präzise inhaltsanalytische Kenntnisse über die Umweltberichterstattung in der Presse sind u. a. für die Journalismusforschung, die Politikwissenschaft oder auch die Umweltpsychologie (z. B. Rückschlüsse auf Umweltwissen der Bevölkerung) relevant.

Stichprobe Das Untersuchungsmaterial (Presseberichterstattung in Deutschland über die UN-Klimakonferenz 2009) wurde zunächst in verschiedener Hinsicht eingegrenzt:

1. **Zeitraum:** Die Klimakonferenz fand vom 7.–18. Dezember 2009 in Kopenhagen statt. Um nicht nur die Berichterstattung *während*, sondern auch *vor* und *nach* dem Ereignis zu erfassen, wurde ein 7-wöchiger Erhebungszeitraum vom 16.11.2009 bis 31.12.2009 festgelegt. Damit lässt sich der Erscheinungszeitpunkt der Artikel als unabhängige nominalskalierte Variable mit den Ausprägungen vor, während und nach der Konferenz nutzen.
2. **Kulturkreis:** Die Studie bezieht sich allein auf die deutsche Presse. Ländervergleiche waren nicht vorgesehen.
3. **Mediengattung:** Da nicht die gesamte deutsche Presselandschaft untersucht werden konnte, wurden wichtige **regionale Tageszeitungen** aus Thüringen *(Thüringer Allgemeine, Freies Wort),* **überregionale Ta-**

Tabelle 10.31 Zusammensetzung der Grundgesamtheit und der Stichprobe der Presseartikel, die in den sieben ausgewählten Medien im Zeitraum 16.11.2009 bis 31.12.2009 über die UN-Klimakonferenz 2009 berichteten. (Aus Arlt & Wolling, 2011. Mit freundlicher Genehmigung der Autoren)

Medium	Artikel in der Grundgesamtheit ($N = 599$)	Artikel in der Stichprobe ($n = 394$)
Süddeutsche Zeitung	196	98
Frankfurter Allgemeine Zeitung	214	107
Thüringer Allgemeine	91	91
Freies Wort	56	56
Die Zeit	24	24
Focus	9	9
Der Spiegel	9	9

10

geszeitungen *(Süddeutsche Zeitung, Frankfurter Allgemeine)* sowie **Wochenzeitungen bzw. Magazine** ausgewählt *(Die Zeit, Der Spiegel, Focus)*. Insgesamt wurden sieben Printmedien einbezogen, die zum Teil auf regionaler, zum Teil auf nationaler Ebene als besonders einflussreiche Vertreter der Presselandschaft gelten können. Somit lässt sich die Mediengattung, aus der ein Artikel stammt, als nominalskalierte unabhängige Variable betrachten, mit den Ausprägungen regionale Tageszeitung, überregionale Tageszeitung und Wochenzeitung/Magazin.

4. **Inhaltliche Selektion:** Für die Studie relevant waren nur jene Artikel, die im definierten Untersuchungszeitraum in den genannten sieben Presseorganen erschienen und sich inhaltlich mit der Weltklimakonferenz 2009 befassten. Diese Presseartikel wurden durch eine komplette manuelle Sichtung aller Ausgaben der betreffenden Zeitungen und Magazine extrahiert und digital archiviert. Auf diese Weise wurde ein Materialkorpus von $N = 599$ Dokumenten (= Zeitungsartikeln) zusammengestellt. Da eine vollständige Analyse aller Artikel zu zeitaufwändig gewesen wäre, wurde aus der Grundgesamtheit eine Stichprobe von $n = 394$ Artikeln gezogen: Da die überregionalen Zeitungen SZ und FAZ sehr viel häufiger über die Klimakonferenz berichtet hatten als die anderen Medien, wurden aus ihren Artikeln in der Grundgesamheit jeweils die Hälfte per Zufall ausgewählt, bei allen anderen Medien wurden sämtliche Populationselemente in die Stichprobe übernommen (Tab. 10.31).

Kategoriensystem Ebenso wie bei einem standardisierten Fragebogen die einzelnen Fragebogen-Items in thematisch zusammengehörende Blöcke zu gliedern sind (▶ Abschn. 10.3.3), werden auch bei einem standardisierten Kategoriensystem die einzelnen Kategorien in **thematische Blöcke** bzw. **Module** gebündelt (Tab. 10.32). Hinter den Kategorien in den einzelnen Modulen stehen dabei jeweils die zugrunde gelegten Theorien. In der hier vorgestellten Studie kommen Kategorien zum Einsatz, die aus insgesamt sechs verschiedenen kommunikationswissenschaftlichen Theorien bzw. Ansätzen stammen (journalistische Qualitätskriterien, Nachrichtenwerttheorie, Agenda-Setting, Framing-Ansatz, Umweltkommunikation, Medienereignisse/Ereignismanagement).

Kategorien Das **Skalenniveau der Kategorien** entscheidet darüber, welche Methoden der quantitativen Datenanalyse später sinnvoll anwendbar und interpretierbar sind. Im Folgenden werden Beispiele für Kategorien aller vier Skalenniveaus präsentiert.

Für jede Kategorie im Kategoriensystem als einem vollstandardisierten Datenerhebungsinstrument werden alle möglichen Ausprägungen vorgegeben. Bei nominalskalierten Kategorien müssen die Ausprägungen die Kriterien der Genauigkeit/Eindeutigkeit, Exklusivität/Trennschärfe und Exhaustivität/Vollständigkeit erfüllen (▶ Abschn. 8.4.2). So werden für die inhaltsanalytisch Kategorie „Anthropogene Ursachen: Thematisiert der Artikel anthropogene Ursachen für Klimawandel – wenn ja, welche?“ folgende vier Ausprägungen auf **Nominalskalenniveau** definiert:

a. keine anthropogenen Ursachen werden genannt;
b. Schadstoffausstoß durch Industrie, Verkehr, Privathaushalte wird genannt;
c. Waldrodung und/oder Massentierhaltung/Ernährungsgewohnheiten wird genannt;
d. Schadstoffausstoß (b) sowie Waldrodung und/oder Massentierhaltung/Ernährungsgewohnheiten (c) werden genannt.

Tabelle 10.32 Struktur des standardisierten inhaltsanalytischen Kategoriensystems zur Presseberichterstattung über die UN-Klimakonferenz 2009 mit insgesamt 56 Kategorien, gebündelt in 6 theoretisch fundierte inhaltliche Module und 1 formales Modul. (Mod. nach Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Retrieved November 26, 2013, from http://www.db-thueringen.de/servlets/DocumentServlet?id=20414. Mit freundlicher Genehmigung der Autoren)

Thematische Module im Kategoriensystem (1 + 6 Module)	Theoretische Bezüge der Module (6 Theorien)	Anzahl der Kategorien (54 Kategorien)	Inhaltsanalytische Kategorien (Beispiele)
Formale Kategorien	–	11	Medium: In welcher Zeitung ist der Artikel erschienen? Seitenzahl: Auf welcher Seite ist der Artikel erschienen? Fotos: Wie viele Fotos enthält der Artikel? Darstellungsform: Welche Darstellungsform hat der Artikel (z. B. Nachricht, Kommentar, Reportage etc.)?
Inhaltliche Kategorien			
Eigenschaften der Berichterstattung	Journalistische Qualitätskriterien Nachrichtenwerttheorie	5	Faktenvermittlung: In welchem Ausmaß vermittelt der Artikel Fakten? Informationsquellen: Welche Informationsquellen werden im Artikel genannt?
Energie- und klimapolitische Ziele	Agenda-Setting Framing Umweltkommunikation	5	Maximale Erderwärmungsgrenze: Thematisiert der Artikel das Gradziel? Festlegung von Emissionsgrenzen: Thematisiert der Artikel CO_2-Grenzen?
Die Weltklimakonferenz	Agenda-Setting Framing Medienereignisse/ Ereignismanagement	13	Bewertung des Konferenzergebnisses: Wie positiv/negativ bewertet der Artikel das Konferenzergebnis oder findet keine Bewertung statt? Entwicklungsländer: Thematisiert der Artikel Konflikte um Finanzhilfen für Entwicklungsländer? Wenn ja, im Zusammenhang mit welchen Akteuren?
Inszenierte Nebenereignisse im Konferenzkontext	Medienereignisse/ Ereignismanagement	3	Nebenereignisse: Thematisiert der Artikel die Gewalthaltigkeit von Nebenereignissen wie Demonstrationen?
Maßnahmen zum Klimaschutz	Agenda-Setting Framing Umweltkommunikation	6	Maßnahmen zum Klimaschutz: Welche Maßnahmen zum Klimaschutz thematisiert der Artikel? Maßnahmenbewertung: Welche und durch wen getroffene Bewertungen der genannten Maßnahmen zum Klimaschutz thematisiert der Artikel?
Klimawandel	AgendaSetting Framing Umweltkommunikation	11	Anthropogene Ursachen: Thematisiert der Artikel anthropogene Ursachen für Klimawandel – wenn ja welche? Wetterextreme: Thematisiert der Artikel Wetterextreme als Folgen des Klimawandels?

Das **Problem der Mehrfachkodierung** – die Kategorienausprägungen b und c können beide im selben Zeitungsartikel auftreten – wird hier durch die zusätzliche Kategorienausprägung d abgedeckt. Treten Mehrfachkodierungen im größeren Umfang auf, so ist dieser Lösungsansatz nicht sinnvoll. Stattdessen müssen die einzelnen Subkategorien jeweils separat als Variablen erfasst werden (z. B. „Anthropogene Ursache Schadstoffausstoß“: ja/nein; „Anthropogene Ursache Waldrodung“: ja/nein etc.)

Während die Kategorie mit ihren vier Kategorienausprägungen Teil des **Kategoriensystems** ist, werden im **Codebuch** zusätzlich eine Definition der Kategorie, erläuternde Hinweise zur Kodierung sowie die zuzuordnenden Messwerte angegeben (Abb. 10.44).

Die inhaltsanalytische Kategorie **„Inszenierte Nebenereignisse“** wurde laut Kategorienausprägungen auf **Ordinalskalenniveau** operationalisiert, weil die Intensität ihrer Thematisierung in drei Rangstufen eingeteilt wird (Abb. 10.45).

Die inhaltsanalytische Kategorie **„Faktenvermittlung“** wurde mit einer 4-stufigen Ratingskala gemessen, die als **Intervallskala** an der Grenze zur Ordinalskala zu interpretieren ist (Abb. 10.46; zum Skalenniveau von Ratingskalen ▶ Abschn. 8.4.4). Eindeutiger wäre der

Anthropogene Ursachen des Klimawandels *werden sowohl bei expliziter als auch bei impliziter Nennung im Pressebeitrag kodiert. Implizit werden sie dann genannt, wenn sie nicht direkt als Ursache genannt werden, sondern als Aspekte thematisiert werden, die verändert werden müssen (z.B.: „Gesetz zur CO_2-Reduktion“→ CO_2-Ausstoß wird implizit als Ursache des Klimawandels thematisiert.*

Code	Ausprägung
0	**Keine anthropogenen Ursachen werden genannt**
1	**Schadstoffausstoß durch Industrie, Verkehr, Privathaushalte** *(Ausstoß von Treibhausgasen, CO_2-Ausstoß, Kohlendioxid-Emissionen)*
2	**Waldrodung und/oder Massentierhaltung/Ernährungsgewohnheiten** *(Nur bei expliziter Nennung, sonst wird 1 kodiert)*
3	**Schadstoffausstoß, Waldrodung und/oder Massentierhaltung/Ernährungsgewohnheiten** *(wenn beides als Ursache genannt wird, dann wird 3 kodiert)*

Abb. 10.44 Die nominalskalierte inhaltsanalytische Kategorie „Anthropogene Ursachen des Klimawandels“. (Aus Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.db-thueringen.de/. Mit freundlicher Genehmigung der Autoren)

10

Inszenierte Nebenereignisse *sind: Demonstrationen, Kundgebungen, Internetappelle, E-Mail-Veröffentlichung von Klimawissenschaftlern, Zeitungsappelle etc., die im Kontext der Verhandlungen und der Konferenz stattfinden.*

Code	Ausprägung
0	Inszenierte Ereignisse werden nicht thematisiert
1	Inszenierte Ereignisse werden im Beitrag nur beiläufig erwähnt
2	Inszenierte Ereignisse stehen im Mittelpunkt des Beitrags

Abb. 10.45 Die ordinalskalierte inhaltsanalytische Kategorie „Inszenierte Nebenereignisse“. (Aus Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.db-thueringen.de/. Mit freundlicher Genehmigung der Autoren)

Faktenvermittlung: *Es soll ermittelt werden, ob durch die Berichterstattung im Presseartikel quantitative Fakten vermittelt werden, oder ob Ereignisse ohne die Vermittlung harter Fakten dargestellt werden. Häufig bedeutet der Verzicht auf Faktenvermittlung, dass das Erzählen einer Geschichte im Vordergrund der Berichterstattung steht. Indikatoren für eine faktenbetonte Berichterstattung sind die Nennung von Prozentzahlen, (z.B.: Reduktionsziele), Mengenangaben, Wiedergabe von Expertenmeinungen (nur wissenschaftliche Experten), Nennung genauer Daten, konkrete Zeitangaben, Jahreszahlen etc.*

Beispiele für Fakten: EU will CO_2-Emmissionen bis 2020 um mindestens 20 Prozent senken. Entwicklungsländer erhalten von 2010 bis 2012 7.2 Milliarden Euro.

Code	Ausprägung
0	**keine** Faktenvermittlung
1	Faktenvermittlung ist im Beitrag von eher **geringer** Bedeutung
2	Faktenvermittlung ist im Beitrag von **großer** Bedeutung
3	nur / fast **nur** Faktenvermittlung

Abb. 10.46 Die intervallskalierte inhaltsanalytische Kategorie „Faktenvermittlung“. (Aus Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.db-thueringen.de/. Mit freundlicher Genehmigung der Autoren)

Anzahl der Fotos: *Eingetragen wird die Anzahl der Fotos, die eindeutig zu einem Presseartikel gehören.*	
Code	**Ausprägung**
0	kein Foto
1	1 Foto
2	2 Fotos
3	3 Fotos
...	...
x	x Fotos

Abb. 10.47 Die verhältnisskalierte inhaltsanalytische Kategorie „Fotos". (Aus Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.db-thueringen.de/. Mit freundlicher Genehmigung der Autoren; Darstellungsweise soll den Charakter der Verhältnisskala verdeutlichen)

Intervallskalen-Charakter, wenn der Umfang der Faktenvermittlung von den Kodierenden auf einer 5-, 6- oder 7-stufigen Ratingskala eingeschätzt worden wäre, sofern diese genaue Abstufung aus Kodierersicht möglich ist und zu reliablen Werten führt. Zu beachten ist bei der Definition der Kategorie im Codebuch auch die konkrete Angabe von Indikatoren für Faktenvermittlung (Prozentzahlen, Jahreszahlen etc.).

Die formale inhaltsanalytische Kategorie **„Anzahl der Fotos"** wird auf **Verhältnisskalenniveau** operationalisiert, denn die Zahl der Fotos ist gleichabständig und besitzt einen absoluten Nullpunkt (Abb. 10.47).

Die quantitative Inhaltsanalyse misst manifeste formale und inhaltliche Merkmale von Dokumenten in intersubjektiv nachvollziehbarer Weise. Dementsprechend sollten Kategorien so konstruiert werden, dass **subjektive Interpretationen** der Kodierenden möglichst nicht unkontrolliert einfließen. Manche Kategorien sind so eindeutig, dass ohne Probleme 100 %ige Urteiler- bzw. Kodierer-Übereinstimmung erzielt werden kann (z. B. Anzahl der Fotos in einem Artikel). Andere Kategorien erfordern größeren **Interpretationsaufwand**. Bei diesen Kategorien müssen sehr **detaillierte Kodieranweisungen** gegeben werden (am besten mit Beispielen aus dem Material), um einheitliche Interpretationen bzw. hohe Reliabilität sicherzustellen. Dies trifft z. B. auf die relativ komplexe Kategorie „Beurteilung des Konferenzergebnisses" zu (Abb. 10.48).

Eine besondere Herausforderung für die Entwicklung eines Kategoriensystems stellen jene Kodiereinheiten dar, die sehr viele verschiedene Merkmale besitzen, wie z. B. Akteure: Im Umfeld der UN-Klimakonferenz 2009 und ihren Nebenereignissen tauchen buchstäblich Dutzende von **Akteurinnen und Akteuren** in der Presse auf: Etwa die Autofahrer, Klimaforscher, der Energiekonzern EON, Greenpeace, das Bundesamt für Naturschutz, der Bundesverkehrsminister, China, die dänische Polizei, der australische Regierungschef, der UN-Generalsekretär etc. Die Kodierung der Akteure erfolgt über **Akteurslisten**, wobei zu beachten ist, dass zur Strukturierung der Akteursgruppen in der Regel mehrere inhaltliche Kategorien notwendig sind, die jeweils die Kriterien der Eindeutigkeit, Trennschärfe und Vollständigkeit erfüllen müssen. So lassen sich Akteure z. B. auf einer geografischen Dimension einordnen (transnational – national – regional), unterschiedlichen Gesellschaftssystemen zuordnen (z. B. Medien, Politik, Wissenschaft, Wirtschaft, Kunst, Zivilgesellschaft) und innerhalb der Gesellschaftssysteme noch nach Funktionen differenzieren (z. B. Funktionen von Akteuren in der Zivilgesellschaft: z. B. Kirchen, NGOs, Vereine etc.). In Akteurslisten können aber auch einzelne Personen aufgenommen werden (z. B. Spitzenpolitiker/innen), deren Auftreten in den untersuchten Artikeln dann ausgezählt wird.

Codebuch Das Codebuch enthält Angaben zum Ziel der Studie, zum Datenmaterial sowie allgemeine Kodieranweisungen (z. B. den Hinweis, jeden zu kodierenden Artikel zunächst zweimal gründlich durchzulesen und dann eine Kurzzusammenfassung zu schreiben, bevor mit seiner Kodierung begonnen wird). Es folgen daraufhin im Codebuch die einzelnen Kategorien des Kategoriensystems, geordnet nach den thematischen Modulen. Jede Kategorie wird dabei exakt definiert, ggf. werden auch den einzelnen Kategorienausprägungen noch Kodierhinweise hinzugefügt. Die Frage nach der Validität des Kategoriensystems, d. h. ob es tatsächlich die inhaltlich-theoretisch interessierenden Konstrukte angemessen erfasst, ist anhand der Definition der Kategorien sowie der Kodieranweisungen und deren jeweiligen Theoriebezügen zu diskutieren. Zudem wird angegeben, welcher Messwert welcher Kategorienausprägung zuzuordnen ist. Schließlich gehört auch der Codebogen zum Kategoriensystem. Bei diesem handelt es sich heute meist um eine

Beurteilung der Konferenzergebnisse: *Kategorie kann nur bei Artikeln/Beiträgen zutreffen, die am letzten Tag der Konferenz (Veröffentlichungsdatum 20.12. 2009) oder danach erscheinen.*	
Code	**Ausprägung**
0	**Ergebnisse der Konferenz werden nicht thematisiert**
1	**Eindeutig negative Bewertung der Konferenzergebnisse** – *Konferenz [war] kein Erfolg, Ziel wurde nicht erreicht* – *völliges Versagen der Politiker* – *Konferenz ist gescheitert* – *Notwendige Vereinbarungen wurde nicht erreicht*
2	**Überwiegend negative Bewertung der Konferenzergebnisse** – *Es werden positive Aspekte gesehen, das Gesamtergebnis ist aber negativ* – *Insgesamt ist die Konferenz nicht zufriedenstellend trotz einzelner Erfolge* – *Einigungen/Zielvereinbarungen sind enttäuschend* – *Die Ergebnisse sind zu unverbindlich*
3	**Überwiegend positive Bewertung der Konferenzergebnisse** – *erste Schritte in die richtige Richtung, aber noch nicht genug* – *kleiner Erfolg, aber noch nicht ausreichend* – *nicht langfristig genug* – *Hoffnungsvolles Signal*
4	**Eindeutig positive Bewertung der Konferenzergebnisse** – *Konferenz war ein Erfolg* – *verbindliches Abkommen wurde erreicht* – *wichtiger Beitrag zum Klimaschutz*

Abb. 10.48 Die komplexe inhaltsanalytische Kategorie „Beurteilung der Konferenzergebnisse". (Aus Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.db-thueringen.de/. Mit freundlicher Genehmigung der Autoren)

Thematisierung der Weltklimakonferenz: *Es wird kodiert, ob die Weltklimakonferenz das eigentliche zentrale Thema des Beitrags ist und im Mittelpunkt steht oder ob sie eher als Aufhänger des Beitrags dient bzw. als Anlass genommen wird, um einen Beitrag über Klimawandel, Klimaschutz oder Klimapolitik zu schreiben. Nebenereignisse der Konferenz werden nicht als Teil der Konferenz betrachtet.*	
Code	**Ausprägung**
1	Weltklimakonferenz wird nur beiläufig erwähnt (Nebenthema)
2	Weltklimakonferenz wird gleichrangig mit anderen Themen (z.B. Klimawandel) behandelt
3	Weltklimakonferenz steht im Mittelpunkt des Beitrags (Hauptthema)

Abb. 10.49 Die inhaltsanalytische Kategorie „Thematisierung der Weltklimakonferenz" (**erste Fassung** vor Pretest). (Aus Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.db-thueringen.de/. Mit freundlicher Genehmigung der Autoren)

Excel- oder SPSS-Dateivorlage, in der die einzelnen Variablen bereits angelegt und die Wertelabel definiert sind, so dass die Kodierenden direkt die Messwerte eintragen können.

Um die Bedeutung gründlicher Pretests für die Verfeinerung des Codebuches zu unterstreichen, wird die Überarbeitung einer exemplarischen Kategorie hier präsentiert. So stellte sich im Pretest heraus, dass es für die Kodierenden schwierig war, den Ausprägungsgrad der Kategorie „Thematisierung der Weltklimakonferenz in einem Pressertikel" eindeutig quantitativ festzulegen, so dass in der überarbeiteten Darstellung der Kategorie im Codebuch deutlich umfangreichere und detailliertere Kodierhinweise gegeben werden (Wolling & Arlt, 2012; Abb. 10.49, Abb. 10.50).

Ergebnisse Anhand der Ergebnisse der quantitativen Inhaltsanalyse lassen sich Unterschiede in der Berichterstattung zwischen den Mediengattungen zeigen (z. B. regionale versus überregionale Tagespresse) sowie Differenzen in der Berichterstattung vor, während und nach der Konferenz belegen.

Thematisierung der Weltklimakonferenz: *Als Teil der Konferenz wird nur kodiert: Berichte über Verhandlungen, Aussagen über Strategien während der Konferenz, Berichte über die Teilnehmer der Konferenz. Es wird kodiert, ob die Weltklimakonferenz das eigentliche Thema des Beitrags ist und im Mittelpunkt steht oder ob*

a) es um inhaltliche Frage geht, d.h. Klimawandel, Klimaschutz oder Klimapolitik,
b) die Konferenz nur als Anlass genommen wird, um über andere Themen zu schreiben, die nur indirekt mit der Konferenz zu tun haben (z.B. internationale Politik allgemein).
c) inszenierte Nebenereignisse im Kontext der Konferenz dargestellt werden (siehe Modul Nebenereignisse)
d) über vorbereitende oder begleitende oder nachfolgende Treffen berichtet wird. Bsp.: „Eine Woche vor der Weltklimakonferenz haben sich die Entwicklungsländer getroffen, um ihren Klima-Kurs abzustimmen".

Code	Ausprägung
1	Weltklimakonferenz wird nur beiläufig erwähnt. Die anderen oben erwähnten Themen (a–d) stehen eindeutig im Mittelpunkt.
2	Weltklimakonferenz spielt eine geringere Rolle als andere Themen (a–d)
3	Weltklimakonferenz spielt eine größere Rolle als andere Themen (a–d)
4	Weltklimakonferenz steht im Mittelpunkt des Beitrags (Hauptthema). Es geht nur um die Abläufe und Verhandlungen auf der Konferenz.

Abb. 10.50 Die inhaltsanalytische Kategorie „Thematisierung der Weltklimakonferenz" (**überarbeitete Fassung** nach Pretest). (Aus Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.db-thueringen.de/. Mit freundlicher Genehmigung der Autoren)

Die Auswertung der Kategorie **„Anthropogene Ursachen für Klimawandel"** führte z. B. zu dem in Tab. 10.33 dargestellten deskriptiv-statistischen Ergebnis.

Die Häufigkeitsunterschiede in der Thematisierung verschiedener anthropogener Ursachen waren laut eindimensionalem Chi-Quadrat-Test statistisch signifikant ($p < .001$).

Tabelle 10.33 Auswertung der nominalskalierten inhaltsanalytischen Kategorie „Anthropogene Ursachen für Klimawandel": Deskriptiv-statistische Ergebnisse (Frequenzanalyse). (Mod. nach Arlt & Wolling, 2011)

Anthropogene Ursachen: Thematisiert der Artikel anthropogene Ursachen für Klimawandel – wenn ja, welche?	Anzahl der Artikel ($n = 394$)	Anteil der Artikel (gesamt = 100 %)
		[%]
Keine anthropogenen Ursachen genannt	135	34
Schadstoffausstoß durch Industrie, Verkehr, Privathaushalte genannt	235	60
Waldrodung und/oder Massentierhaltung/Ernährung genannt	10	2.5
Schadstoffausstoß sowie Waldrodung und/oder Massentierhaltung/Ernährung genannt	14	3.5

Schlussfolgerungen Eine quantitative Inhaltsanalyse der Presseberichterstattung über die UN-Klimakonferenz 2009 liefert über die konkreten Ergebnisse zur Beschaffenheit der Berichterstattung hinaus u. a. folgenden Erkenntnisgewinn und praktischen Nutzen:

- Rückschlüsse darauf, was die **Bundesbürger/innen über das Thema Klima wissen bzw. nicht wissen können.** Da Klimapolitik im Alltag meist nicht direkt erlebt wird, sind Medien bei diesem Thema die Hauptinformationsquelle.
- Hinweise darauf, in welcher Weise das Umweltwissen der Bevölkerung gefördert oder beeinflusst werden sollte (z. B. Konzeption von **Kampagnen zu Umweltwissen/Umweltschutz**, in denen Aspekte thematisiert werden, die in der Presse vernachlässigt wurden, z. B. Einfluss von Waldrodung/Massentierhaltung/Ernährung auf den Klimawandel. Solches Wissen ist auch potenziell handlungsrelevant im Sinne konkreten Umweltverhaltens etwa durch eine Ernährungsumstellung).
- Hinweise darauf, wie positiv/negativ verschiedene Akteure im Zusammenhang mit Klimapolitik dargestellt werden. Dies kann Akteur/innen in der Politik Anhaltspunkte liefern für die **Konzeption von Image- oder Wahlkampfkampagnen**.

Reliabilitätsprüfung eines inhaltsanalytischen Kategoriensystems

Zu den Arbeitsschritten einer quantitativen Inhaltsanalyse gehört nach der Sicherstellung der Validität immer auch die Reliabilitätsüberprüfung des finalen Kategoriensystems. Dabei muss jede einzelne Kategorie im Kategoriensystem einer Reliabilitätsanalyse unterzogen werden. Bei der Reliabilitätsprüfung sind fünf Arbeitsschritte durchzuführen: 1.) Auswahl des Materials zur Reliabilitätsbestimmung, 2.) Festlegung der Anzahl der Kodierenden für die Reliabilitätsbestimmung, 3.) Wahl der geeigneten Reliabilitätskoeffizienten passend zum Skalenniveau der Kategorien, 4.) Interpretation der Reliabilitätskoeffizienten anhand von Reliabilitätsstandards sowie inhaltlichen Besonderheiten der Kategorien und 5.) Darstellung der Reliabilitätsmaße im Ergebnisbericht.

Auswahl des Materials zur Reliabilitätsbestimmung Zur Reliabilitätsprüfung wird entweder eine separate Pretest-Stichprobe oder ein Ausschnitt des Materials der Haupterhebung verwendet. **Konventionell** werden hier beispielsweise 10–20 % des Materials oder ca. 20–25 Dokumente herangezogen. Bei Kategorien, mit sehr selten auftretenden Ausprägungen ist ein größerer oder ein vorselektierter Materialausschnitt zugrunde zu legen, damit auch für seltene Ausprägungen genügend Daten zur Reliabilitätsschätzung zur Verfügung stehen. Anstelle einer Orientierung an Konventionen ist für die Reliabilitätsbestimmung die statistisch begründete Wahl einer ausreichenden bzw. **optimalen Stichprobengröße** vorzuziehen (Gwet, 2012). Ausschlaggend ist dabei diejenige Kategorie im Kategoriensystem, deren Reliabilitätsschätzung für angemessene Schätzgenauigkeit den höchsten Stichprobenumfang verlangt.

Festlegung der Anzahl der Kodierenden für die Reliabilitätsbestimmung Die Reliabilitätsbestimmung kann durch einen, zwei oder mehr Kodierende erfolgen:

- **Ein Kodierer**: Die einfachste Form der Reliabilitätsprüfung erfolgt mittels **Intra**-Kodierer-Übereinstimmung („intra-coder reliability"), bei der eine Person einen (ausreichend großen) Ausschnitt des Materials in zeitlichem Abstand (z. B. von einer Woche) zweimal kodiert und die Übereinstimmung beider Messzeitpunkte berechnet wird.
- **Zwei Kodierer**: Die typische Form der Reliabilitätprüfung erfolgt mittels **Inter**-Kodierer-Übereinstimmung („inter-coder reliability") bei zwei Kodierern, die denselben (ausreichend großen) Ausschnitt des Materials anhand des Kategoriensystems kodieren, so dass dann für jede Kategorie ein Reliabilitätskoeffizient berechnet werden kann.

Tabelle 10.34 Maße der Kodiererübereinstimmung (Reliabilität) für nominal-, ordinal- und kardinalskalierte Kategorien in einem standardisierten inhaltsanalytischen Kategoriensystem

Skalenniveau der inhaltsanalytischen Kategorie	Kodiererübereinstimmungs-Koeffizient
Nominalskalierte Kategorien	Prozentuale Übereinstimmung („Holsti-Formel")
	Cohens Kappa (Cohens κ)
Ordinalskalierte Kategorien	Cohens Weighted Kappa (Cohens Weighted κ)
	Spearmans Rho (Spearmans ρ)
	Kendalls Tau (Kendalls τ)
Kardinalskalierte (intervall- oder verhältnisskalierte) Kategorien	Bivariate Korrelation *r*
	Intra-Class-Correlation ICC

- **Mehr als zwei Kodierer**: Soll die Übereinstimmung zwischen mehr als zwei Kodierern bestimmt werden, so sind zunächst die paarweisen Übereinstimmungen der Kodierenden bei der betreffenden Kategorie zu berechnen. Diese sind dann zu einer Gesamtreliabilität pro Kategorie zu mitteln. Dieser Zusatzaufwand ist z. B. sinnvoll, wenn in einer Studie mehrere Kodierende eingesetzt werden und differenziert werden soll, ob auftretende Reliabilitätsprobleme auf das Kategoriensystem/Codebuch oder auf einzelne Kodierer zurückzuführen sind. Sollte sich herausstellen, dass ein bestimmter Kodierer stets deutlich von den anderen Kodierern abweicht und somit messfehlerbehaftete Daten erzeugt, so sollte er nachgeschult oder ersetzt werden. Bei groß angelegten Studien beschränkt man sich zuweilen nicht darauf, die Reliabilität der Kodierer nur am Anfang zu prüfen: Ergänzend wird zwischendurch eine erneute (unangekündigte) Prüfung anhand des eingereichten Datenmaterials vorgenommen, um die Qualität der laufenden Arbeit sicherzustellen.

Wahl der geeigneten Reliabilitätskoeffizienten passend zum Skalenniveau der Kategorien Ausschlaggebend für die Wahl des Reliabilitätskoeffizienten ist das Skalenniveau der jeweiligen Kategorie. Bei der Reliabilitätsprüfung eines inhaltsanalytischen Kategoriensystems muss also pro Kategorie der jeweils passende Reliabilitäts- bzw. Übereinstimmungskoeffizient ausgewählt werden (Tab. 10.34; die Maße der Kodiererübereinstimmung zur Reliabilitätsprüfung inhaltsanalytischer Kategorien entsprechen denen der Beobachtereinstimmung für Kategori-

Tabelle 10.35 Zwei Kodierende haben 18 Zeitungsartikel hinsichtlich der inhaltsanalytischen Kategorie „Anthropogene Ursachen für Klimawandel" auf vier nominalskalierten Ausprägungen (0–3) kodiert

Ausprägungen: 0 (nicht genannt) 1 (CO_2-Ausstoß) 2 (Waldrodung/Ernährung) 3 (CO_2-Ausstoß & Waldrodung/Ernährung)
Artikel: $n = 18$
Kodierer: K1: Kodierer 1, K2: Kodierer 2

	Artikel																	
Kodierer	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
K1	1	1	**1**	1	2	0	0	**0**	1	0	2	0	0	1	1	**1**	1	0
K2	1	1	**0**	1	2	0	0	**1**	1	0	2	0	0	1	1	**2**	1	0

en in standardisierten Beobachtungssystemen ▶ Abschn. 10.1.3 „Beobachterübereinstimmung").

Prozentuale Übereinstimmung versus Cohens Kappa-Koeffizient:

Nominalskalierte Kategorien sind in inhaltsanalytischen Kategoriensystemen am häufigsten vertreten (Popping, 2010). Die Verwendung der **prozentualen Übereinstimmung** als Reliabilitätsmaß hat den Vorteil, dass sie sehr leicht zu berechnen und intuitiv zu verstehen ist. Beispiel: Zwei Kodierer (K1, K2) kodieren bei $n = 18$ Zeitungsartikeln das nominalskalierte Merkmal „Anthropogene Ursachen für Klimawandel", das vier Ausprägungen hat (Tab. 10.35).

Bei diesen 18 Artikeln urteilten die beiden Kodierer in 15 Fällen gleich (absolute Urteilerübereinstimmung) und in drei Fällen *(Artikel 3, 8 und 16; fett gedruckt)* unterschiedlich. Die prozentuale Übereinstimmung berechnet sich nun einfach, indem die Anzahl der übereinstimmenden Kodierungen (15) durch die Anzahl aller kodierten Analyseeinheiten (18) dividiert wird. Der Kodiererübereinstimmungs-Koeffizient bzw. Reliabilitäts-Koeffizient der Kategorie liegt also bei $\mathbf{Rel = 15/18 = .83 = 83\,\%}$.

Der Nachteil der prozentualen Übereinstimmung als Reliabilitätsmaß besteht jedoch darin, dass dieses Maß die echte Urteilerübereinstimmung **überschätzt**. Wissenschaftlich ist diese Verzerrung problematisch, da eine Überschätzung der Reliabilität die tatsächlichen Schwächen der Datenerhebung (d. h. ihre Belastung durch Messfehler) kaschiert. Soll aus Gründen der Einfachheit dennoch die prozentuale Übereinstimmung als Reliabilitätsmaß genutzt werden, so muss die Problematik der Überschätzung der Messgenauigkeit zumindest ausdrücklich thematisiert werden.

Ein besseres Reliabilitätsmaß als die einfache prozentuale Übereinstimmmung ist **Cohens Kappa-Koeffizient**, der die prozentuale Übereinstimmung **um Zufallsübereinstimmungen bereinigt**. Dahinter steht die Überlegung, dass bei einer rein zufälligen (z. B. per Los erfolgten) Zuordnung von 18 Artikeln zu vier Kategorien eine nennenswerte Zahl an Zufallsübereinstimmungen entsteht, die jedoch inhaltlich nichts mit der Messgenauigkeit der Kategorie zu tun haben. Das von dem Statistiker Jacob Cohen (1960) entwickelte Reliabilitätsmaß – Cohens Kappa-Koeffizient – zieht von der einfachen Urteilerübereinstimmung deswegen alle Zufallsübereinstimmungen ab. Die Zufallsverfälschung ist umso größer, je weniger Ausprägungen eine Kategorie hat. Aufgesplittet nach den einzelnen Kategorienausprägungen lässt sich das Datenbeispiel aus Tab. 10.35 auch wie in Abb. 10.51 darstellen.

Die Abbildung stellt die **beobachteten Häufigkeiten** (f_b) der Kodierungen dar: In der Diagonale der Kreuztabelle *(fett gedruckt)* stehen die insgesamt 15 beobachteten Urteilerüberstimmungen ($6 + 7 + 2 + 0$) für die einzelnen vier Kategorienausprägungen. Außerhalb der Diagonale finden sich *(blau unterlegt)* die drei abweichenden Kodierungen der beiden Urteiler. Die **beobachtete Urteilerübereinstimmung** beträgt also $\boldsymbol{p = (6 + 7 + 2 + 0)/18 = .83 = 83\,\%}$. Dieser Wert entspricht der bereits oben bestimmten prozentualen Übereinstimmung.

Der Kappa-Koeffizient von Cohen folgt der Logik aller statistischen Chi-Quadrat-Verfahren, bei denen die beobachteten Häufigkeiten (f_b) mit den unter der Nullhypothese **erwarteten Häufigkeiten** (f_e) verglichen werden. Dementsprechend stellt sich die Frage, welche Urteile beider Kodierer zu erwarten sind, wenn die Kodierung rein zufällig erfolgt (z. B. per Los), also – gemäß Nullhypothese – keine systematische Urteilerübereinstimmung vorliegt. Die erwarteten Häufigkeiten (f_e) berechnen sich dabei aus den Randsummen nach der Formel: (Zeilensumme × Spaltensumme)/n (Abb. 10.52).

Die Abbildung stellt nur die **erwarteten Häufigkeiten** (f_e) für die Diagonalzellen dar. Diese Häufigkeiten entsprechen den absoluten Zufallsübereinstimmungen. Die **prozentuale Zufallsübereinstimmung** (p_e) berechnet sich als Summe der erwarteten Häufigkeiten in der Diagonale dividiert durch n:

Beobachtete Häufigkeiten f_b		Kodierer 1				Zeilen-summe
		0	1	2	3	
Kodierer 2	0	**6**	1	0	0	7
	1	1	**7**	0	0	8
	2	0	1	**2**	0	4
	3	0	0	0	**0**	0
Spaltensumme		7	9	2	0	$n = 18$

Abb. 10.51 Die inhaltsanalytische Kategorie „Anthropogene Ursachen für Klimawandel" mit vier nominalskalierten Ausprägungen (0–3). *Beobachtete Häufigkeiten (f_b)*

Erwartete Häufigkeiten f_e = (Zeilensumme × Spalten-summe) / n		Kodierer 1				Zeilen-summe
		0	1	2	3	
Kodierer 2	0	7×7/18 = **2.72**				7
	1		8×9/18 = **4.00**			8
	2			4×2/18 = **0.30**		4
	3				0×0/18 = **0.00**	0
Spaltensumme		7	9	2	0	$n = 18$

Abb. 10.52 Die inhaltsanalytische Kategorie: „Anthropogene Ursachen für Klimawandel" mit vier nominalskalierten Ausprägungen (0–3). *Erwartete Häufigkeit (f_e)*

$\boldsymbol{p_e = (2.72 + 4 + 0.30 + 0)/18 = .39 = 39\,\%}$. Man sieht also, dass fast 40 % Urteilerübereinstimmung von zwei Kodierern bei einer Kategorie mit vier Ausprägungen und 18 kodierten Artikeln bereits durch puren Zufall zustande kommen.

Der Kappa-Koeffizient bereinigt nun die einfache Urteilerübereinstimmung (p) nach folgender Formel um die prozentuale Zufallsübereinstimmung (p_e; Gl. 10.17):

$$\kappa = \frac{p - p_e}{1 - p_e} \text{ mit } p = \frac{\sum_{j=1}^{k} f_{bjj}}{n};\ p_e = \frac{\sum_{j=1}^{k} f_{ejj}}{n} \tag{10.17}$$

Für unser Beispiel ergibt sich

$$\kappa = \frac{p - p_e}{1 - p_e} = \frac{.83 - .39}{1 - .39} = \frac{.44}{.61} = .72 = 72\,\%$$

Im Unterschied zur einfachen prozentualen Übereinstimmung von Rel = 83 % beträgt die um Zufallsübereinstimmungen bereinigte Reliabilitätsschätzung also nur $\kappa = 72\,\%$ (SPSS → Analysieren → Deskriptive Statistiken → Kreuztabellen → Statistiken: Kappa).

Bivariate Korrelation versus Intra-Klassen-Korrelation:

Will man die Urteiler- bzw. Kodiererübereinstimmung bei zwei Kodierern und einer kardinalskalierten (d. h. intervall- oder verhältnisskalierten) Kategorie berechnen, deren Ausprägungen z. B. Anzahlen (Anzahl der Fotos pro Artikel) oder Ratingskalen (Umfang der Faktenvermittlung von 0: keine Faktenvermittlung bis 3: fast nur Faktenvermittlung) darstellen, so scheint zunächst der **bivariate Korrelationskoeffizient *r*** (▶ Abschn. 12.5.2) indiziert. Dieser ist jedoch als Reliabilitätsmaß ungeeignet, da er die wahre Kodiererübereinstimmung **überschätzt**. Im Extremfall kann es sogar vorkommen, dass eine perfekte bivariate Korrelation von $r = 1.00$ berechnet wird, auch wenn zwei Kodierer bei keiner einzigen Analyseeinheit übereinstimmend kodiert haben. Der bivariate Korrelationskoeffizient ist nämlich für Verschiebungen im Niveau der Werte unsensibel, während die **Intra-**

Kodierer 1	Kodierer 2
3	2
2	1
2	1
2	1
4	3
1	0
1	0
1	0
2	1
1	0
3	2
3	2
3	2
3	2
4	3
1	0
2	1
3	2

Abb. 10.53 Zwei Kodierer haben 18 Zeitungsartikel hinsichtlich des kardinalskalierten Merkmals „Umfang der Faktenvermittlung" auf einer Ratingskala mit fünf Ausprägungen (0–4) kodiert

Klassen-Korrelation diese Information verarbeitet und somit das genauere Reliabilitätsmaß darstellt. Beispiel: Zwei Kodierer haben 18 Zeitungsartikel hinsichtlich einer kardinalskalierten Kategorie mit 5 Ausprägungen (Ratingskala 0–4) kodiert. Dabei liegen die Messwerte des zweiten Kodierers durchgängig einen Skalenpunkt unter denen des ersten Kodierers (Abb. 10.53).

Bei dieser Datenlage wird ein bivariater Korrelationskoeffizient von $r = 1.00$ berechnet, der fälschlich eine perfekte Urteilerübereinstimmung suggerieren würde. Die Berechnung des als Übereinstimmungs- bzw. Reliabilitätsmaß bei kardinalskalierten Kategorien angemessenen Intra-Klassen-Korrelationskoeffizienten ergibt dagegen eine realistische Urteilerübereinstimmung von **ICC = .81** (SPSS → Analysieren → Skalierung → Reliabilitätsanalyse → Statistiken → Modell: zweifach gemischt; Typ: absolute Übereinstimmung). Hier ist allerdings inhaltlich zu reflektieren, ob die Urteilsabweichungen wirklich dem Kategoriensystem zuzuschreiben sind (z. B. Kodieranweisung ist noch immer zu unklar) oder ob sie auf individuelle Urteilstendenzen der Kodiererenden zurückzuführen sind. Im letztgenannten Fall würde man als Übereinstimmungsmaß nicht die unjustierte Intra-Klassen-Korrelation (Typ: absolute Übereinstimmung), sondern die weniger strenge justierte Intra-Klassen-Korrelation (Typ: Konsistenz) verwenden, die für das Beispiel einen Wert von $ICC_{just} = 1.0$ ergibt.

Interpretation der Reliabilitätskoeffizienten anhand von Reliabilitätsstandards sowie inhaltlichen Besonderheiten der Kategorien Reliabilitätskoeffizienten schätzen die Messgenauigkeit der inhaltsanalytischen Kategorien. Ihre Ausprägung hängt von der Art des erhobenen Merkmals, vom verwendeten Übereinstimmungskoeffizienten und von den Umständen der Reliabilitätsprüfung ab (Anzahl und Besetzung der Ausprägungen einer Kategorie; Anzahl der Kodierenden etc.). Reliabilitätsstandards („benchmarks"), die angeben, ab wann von guter oder zumindest ausreichender Reliabilität auszugehen ist, sind deswegen nur **grobe Richtlinien**. Sie müssen **inhaltlich interpretiert** und diskutiert werden. So kann im Einzelfall auch ein numerisch hoher Reliabilitätskoeffizient problematisch sein (wenn er bei einer sehr einfachen Kategorie vorkommt, die eigentlich perfekte Übereinstimmung verlangt), während ein vergleichsweise niedriger Reliabilitätskoeffizient als ausreichend akzeptiert wird (wenn er z. B. bei einer sehr komplexen Kategorie auftritt; vgl. Gwet, 2012).

Die **prozentuale Übereinstimmung** als Reliabilitätsmaß für nominalskalierte inhaltsanalytische Kategorien (die einen Wertebereich von 0–100 % annehmen kann) sollte nahe 100 % liegen, um von guter Reliabilität auszugehen (Wirtz & Caspar, 2002). Gegenüber der einfachen prozentualen Übereinstimmung ist der um Zufallsübereinstimmungen korrigierte **Cohens Kappa-Koeffizient** in der Regel vorzuziehen. Sein Wertebereich liegt zwischen 0 und theoretisch maximal 1 (wenn keinerlei Zufallsüberstimmungen angenommen werden). Da jedoch in Abhängigkeit von der Zahl der Kategorienausprägungen von mehr oder minder vielen Zufallsübereinstimmungen auszugehen ist (bei zwei Kodierenden und einer binären Kategorie z. B. 50 % Zufallsübereinstimmungen), ist der maximal erreichbare Kappa-Koeffizient oft deutlich geringer als der Wert 1. Dementsprechend müssen Bewertungsstandards, denen gemäß Kappa-Werte über .75 als sehr gut, zwischen .75 und .60 als gut und zwischen .60 und .40 als ausreichend eingestuft werden, inhaltlich jeweils relativiert werden (Gwet, 2012, Kap. 6).

Die **Intra-Class-Korrelation** als Reliabilitätsmaß für intervall- und verhältnisskalierte Kategorien weist einen Wertebereich von −1 bis +1 auf. Ein Wert von 0 sowie ein negativer Wert bedeuten, dass keine Kodiererübereinstimmung vorliegt und die entsprechende Kategorie somit nicht reliabel ist. Je stärker sich der Wert der Intra-Klassen-Korrelation dem Wert +1 nähert, desto besser ist die

Tabelle 10.36 Gegenüberstellung von klassischer quantitativer Inhaltsanalyse und quantitativer Online-Inhaltsanalyse. (Eigene Darstellung nach Welker, et al., 2010)

Besonderheiten von Mediendokumenten	Klassische Inhaltsanalyse von Massenmedien	Online-Inhaltsanalyse von Internet-Content
Innovativität	Bekannte Textgattungen	Innovative Textgattungen
Multimedialität	Stark textorientiert	Multimedial, multimodal
Hypertextualität	Lineare Inhalte	Hypertextuelle, hypermediale Inhalte
Interaktivität	Nicht-interaktive Inhalte	Interaktive Inhalte
Kommunikatoren	Medienunternehmen	Medienunternehmen, Organisationen, Einzelpersonen: identifizierbare sowie anonyme Akteure
Flüchtigkeit/Transitorik	Periodische Erscheinungsweise und vollständige Archivierung	Unregelmäßige und unvorhersehbare Produktions- und Veränderungszeitpunkte, keine vollständige Archivierung
Dynamik	Statische Inhalte	Dynamische Inhalte, die fortwährend geändert, gekürzt, erweitert, aktualisiert etc. werden
Digitalisierung	Oftmals nicht-digitale Dokumente (z. B. Printmedien)	Ausschließlich digitale Dokumente
Quantität	Große Menge an Dokumenten	Riesige und exponentiell wachsende Menge an Dokumenten im globalen Internet
Reaktivität	Nichtreaktive, universale Inhalte	Teilweise reaktive/personalisierte Inhalte auf der Basis individueller Nutzereingaben/Nutzerprofile

Übereinstimmung und entsprechend auch die Reliabilität der untersuchten Kategorie. ICC-Werte über .70 werden konventionell als gute Reliabilität gewertet (Wirtz & Caspar, 2002). Weitere Hinweise zur Berechnung und Einordnung von Urteilerübereinstimmungen sind ▶ Abschn. 10.1.3 zur Beobachterübereinstimmung zu entnehmen.

Darstellung der Reliabilitätsmaße im Ergebnisbericht Im späteren Untersuchungsbericht sind die Reliabilitätsmaße aller Kategorien separat anzugeben (z. B. als Tabelle im Anhang). Zur zusammenfassenden Kennzeichnung des gesamten Kategoriensystems können der Range oder der Durchschnitts- bzw. Medianwert der einzelnen Reliabilitätskoeffizienten berichtet werden. Bei der Einordnung der Reliabilitätsmaße ist einerseits auf die vorgenannten etablierten Standards zu verweisen, andererseits aber auch inhaltlich anhand der Komplexität der jeweiligen Kategorien zu urteilen.

Quantitative Online-Inhaltsanalyse

Die quantitative Online-Inhaltsanalyse unterscheidet sich von der klassischen quantitativen Inhaltsanalyse, die primär Offline-Textmaterial (z. B. Zeitungsartikel) untersucht, hinsichtlich ihres Gegenstandes. Online-Dokumente weisen eine Reihe von medialen Besonderheiten auf, die bei der inhaltsanalytischen Arbeit zu beachten sind (Welker, et al., 2010, Tab. 10.36).

Eine quantitative Inhaltsanalyse von Online-Dokumenten kann der klassischen Inhaltsanalyse sehr ähnlich sein, etwa wenn Artikel einer Online-Zeitung oder Kontaktanzeigen eines Dating-Portals thematisch ausgewertet werden, ohne dass medienspezifische Merkmale (z. B. Verlinkungen, Leserkommentare, Online-Banner etc.) berücksichtigt werden. Allerdings kann eine quantitative Online-Inhaltsanalyse sich auch deutlich von der klassischen Inhaltsanalyse massenmedialer Dokumente unterscheiden, etwa wenn Weblogs von Privatpersonen ausgewertet werden, wobei deren Veränderungen über die Zeit ebenso berücksichtigt werden wie Verlinkungen und Leserkommentare. Bei einer solchen Weblog-Studie schlagen sich die medialen Besonderheiten bei der Stichprobenziehung, der fortlaufenden Archivierung des Materials über die Zeit hinweg, der Bildung inhaltsanalytischer Kategorien zu Verlinkungen und Kommentaren sowie der Verknüpfung der unterschiedlichen Analyseeinheiten nieder. Bei der quantitativen Online-Inhaltsanalyse stellen sich Fragen der Validität und Reliabilität des Erhebungsinstruments in neuer Weise (Herbers & Friedemann, 2010). Ebenso wird der Einsatz computergestützter Selektions-, Archivierungs- und v. a. Auswertungstools – etwa Text-Mining – vordringlicher (Scharkow, 2010).

Studienbeispiel

Quantitative Online-Inhaltsanalyse von vorgefundenen Online-Dokumenten

Wichtige politische Ereignisse wie z. B. die UN-Klimakonferenzen werden in den klassischen Massenmedien, aber auch im Internet thematisiert. Somit lässt sich z. B. fragen, welche klimabezogenen Themen **während** und **im Nachgang** der **UN-Klimakonferenz 2007 auf Bali** im WWW thematisiert wurden. Außerdem lässt sich fragen, welche Themen in Verbindung mit welchen Spitzenpolitiker/innen im Web dargestellt wurden. Für die Beantwortung dieser Forschungsfragen wurde eine Online-Inhaltsanalyse durchgeführt (Arens, Böcking, Kummer, & Rüf, 2010).

Materialsammlung

Um die Repräsentation der UN-Klimakonferenz 2007 im deutschsprachigen Web zu analysieren, wurden drei Typen von Webdokumenten unterschieden:

1. **Online-News**, die primär von Medienunternehmen stammen (zugänglich über Google News);
2. **Weblogs**, die v. a. von Bürgerinnen und Bürgern stammen (zugänglich über Google Blogs);
3. **Allgemeine Websites** als „Sammelbecken" unterschiedlichster Quellen (zugänglich über die Suchmaschine Google).

Vier zentrale deutschsprachige **Suchbegriffe** wurden – neben „UN-Klimakonferenz" – verwendet: „Klimawandel", „Klimaschutz", „globale Erwärmung", „Treibhauseffekt". Mindestens einer dieser Suchbegriffe musste auf der Webpage vorkommen, damit diese zur Population gezählt und in eine Datenbank heruntergeladen wurde.

Die deutschsprachigen Online-Dokumente wurden in zwei Zeitperioden gesammelt: zum einen *während* der UN-Klimakonferenz (3.–15. Dezember 2007; $N = 4\,108$ **Webpages**), zum anderen während der sechs Monate *nach* der Konferenz (16. Dezember bis 15. Juni 2008; $N = 46\,444$ **Webpages**).

Kategoriensystem

Das Kategoriensystem wurde deduktiv entwickelt und bestand aus folgenden Themenkomplexen bzw. Modulen:

1. Klimawandel (mit Unterthemen wie Treibhausgase, CO_2-Emissionen);
2. Klimaschutz (mit Unterthemen wie Kyoto-Protokoll, Emissionshandel);
3. Energie (mit Unterthemen wie erneuerbare Energiequellen, Energieeinsparmöglichkeiten);
4. Akteursgruppen (Umweltverbände, politische Parteien, Energieunternehmen, staatliche und nicht-staatliche Institutionen wie z. B. Verbraucherzentralen);
5. Einzelakteure: Spitzenpolitiker/innen wie Bundeskanzlerin Angela Merkel (CDU) und Umweltminister Sigmar Gabriel (SPD);
6. Aktion „Licht aus!" am Internationalen Tag des Klimaschutzes;
7. Entwicklungs- und Schwellenländer.

Als Analyseeinheit fungierte jeweils eine Webpage, wobei die Nennung der einzelnen Themen anhand der Suchbegriffe (und ihrer Flexionsformen) gezählt wurde. Die Analyse erfolgte computergestützt mithilfe eines proprietären Auswertungs-Tools (des SINDBAD Knowledge Generators).

Onlinespezifisch ist in der hier betrachteten Studie die Materialauswahl (Online-News, Weblogs, Websites). Das Kategoriensystem beschränkt sich auf inhaltliche Kategorien und deren Messung auf Nominalskalenniveau (Nennung von Themen und Unterthemen: Ja/Nein). Typische Gestaltungsmerkmale von Online-Dokumenten, etwa deren Verlinkungen (analysierbar mittels **Linkanalyse**) oder deren strukturbildende Elemente wie z. B. Texte, Bilder, Formulare, Download-Optionen, E-Mail-Adressen etc. (auswertbar mittels **Strukturanalyse**) werden hier indessen nicht näher betrachtet (Schweitzer, 2010).

Ergebnisse

Während der UN-Klimakonferenz 2007 fanden sich themenbezogene Online-Dokumente relativ ausgewogen in den Bereichen Online-News (39 %), Blogs (35 %) und allgemeine Websites (26 %). Als Topthemen im deutschsprachigen Web kristallisierten sich rund um die UN-Klimakonferenz 2007 das Kyoto-Protokoll, Treibhausgase und Entwicklungs-/Schwellenländer heraus. Der Umweltminister wurde im Zusammenhang mit allen Topthemen genannt, die Bundeskanzlerin nur im Zusammenhang mit zwei der drei Themen (Koyoto-Protokoll und Entwicklungs-/Schwellenländer).

Schlussfolgerungen

Eine Analyse des Meinungsklimas im Web kann Politiker/innen Informationen darüber liefern, „wie sie sich im Rahmen bestimmter Themen im Internet profilieren und wie gut sie in die Diskussion der einzelnen Themenaspekte eingebunden sind" (Arens et al., 2010, S. 445).

Quantitative Online-Inhaltsanalyse – Die quantitative Online-Inhaltsanalyse („quantitative online content analysis") erfasst anhand eines vollstandardisierten Kategoriensystems die Ausprägungen unterschiedlicher formaler und inhaltlicher Merkmale von Online-Dokumenten. Dabei ist die inhaltsanalytische Vorgehensweise den Besonderheiten der Online-Dokumente anzupassen; dies betrifft z. B. die Festlegung der Analyse- und Kodiereinheiten, die Stichprobenziehung, die Archivierung des Materials, die Medienspezifik der betrachteten formalen und inhaltlichen Kategorien und nicht zuletzt die Darstellung von Materialbeispielen in Publikationen.

Aktuelle quantitative Online-Inhaltsanalysen befassen sich z. B. mit Online-Werbespielen (Lee, Choi, Quilliam, & Cole, 2009), mit Online-Communitys für Senioren (Nimrod, 2010), mit Unternehmensblogs (Cho & Huh, 2010), mit Online-Restaurantbewertungen (Pantelidis, 2010), mit Online-Selbsthilfegruppen (Orizio, Schulz, Gasparotti, Caimi, & Gelatti, 2010), mit Pro-Essstörungs-Websites (Borzekowski, Schenk, Wilson, & Peebles, 2010) oder mit Politik im Web (► Studienbeispiel „Quantitative Online-Inhaltsanalyse von vorgefundenen Online-Dokumenten").

Quantitative Inhaltsanalyse von visuellen Dokumenten

Die Methode der quantitativen Inhaltsanalyse ist auch auf visuelle Dokumente anwendbar (Niesyto & Marotzki, 2005). Dabei wird in der Praxis oft mit relativ einfachen inhaltlichen Kategorien gearbeitet und z. B. Geschlecht, Alter oder Hautfarbe der visuell dargestellten

Studienbeispiel

Quantitative Inhaltsanalyse von vorgefundenen visuellen Dokumenten

„Wie häufig und in welcher Weise wurden Psychotherapie, Psychotherapeuten und verwandte Berufsgruppen in der Nachkriegszeit auf den Titelbildern und in den Schlagzeilen der wichtigsten überregionalen deutschen Zeitschriften und Illustrierten dargestellt?“ Um diese Forschungsfrage zu beantworten, wurde eine quantitative Inhaltsanalyse von Zeitschriften-Titelbildern durchgeführt (von Sydow, Weber & Reimer, 1998).

Materialsammlung

Zunächst wurden die zum Untersuchungszeitpunkt auflagenstärksten überregionalen deutschen Zeitschriften identifiziert (Stern, Spiegel, Focus, Bunte etc.) und von diesen im Zeitraum zwischen 1947–1995 alle Titelseiten gesammelt (Vollerhebung), sofern sie über Datenbanken sowie die Bildredaktionen der jeweiligen Medien zugänglich gemacht werden konnten ($N = 9\,771$).

Stichprobe

Es wurden dann diejenigen Titelseiten aus dem Materialkorpus bzw. der Titelseitenpopulation herausgesucht, die das Thema Psychotherapie bzw. Psychotherapeuten visuell dargestellten ($n = 14$; Abb. 10.54).

Kategoriensystem

Diese visuellen Darstellungen von Psychotherapie wurden anhand eines standardisierten Kategoriensystems ausgewertet. Das Kategoriensystem wurde deduktiv auf der Basis früherer Studien entwickelt und induktiv ergänzt. Es enthält z. B. Kategorien und Ausprägungen wie die folgenden, die sich als Kodiereinheit z. B. auf die dargestellten Psychotherapeuten beziehen:

- Geschlecht: männlich, weiblich;
- Geschätztes Alter: < 40 Jahre / 40–60 Jahre / > 60 Jahre;
- Attribute des Aussehens: Brille: ja/nein; graue/weiße Haare: ja/nein; Arztkittel: ja/nein;
- Accessoires: Stift/Block: ja/nein; Pfeife: ja/nein;
- Räumliche Umgebung: Schreibtisch: ja/nein; Couch: ja/nein; Orientteppich: ja/nein.

Ergebnisse

Es zeigte sich, dass Psychotherapie nur sehr selten visuell auf den Titelseiten der auflagenstarken Zeitschriften auftauchte (14 von 9 771 Titeln), wobei die Thematisierungshäufigkeit anstieg (1947–1979: 1 Titelseite versus 1999–1995: 8 Titelseiten mit visueller Psychotherapiedarstellung). Wenn Psychotherapeuten visuell auf den Titelseiten dargestellt wurden, handelte es sich mit einer Ausnahme um Männer, entweder um Sigmund Freud persönlich oder um „freudianisch“ aussehende und gekleidete Männer mittleren Alters (weiße Haare, Brille, Bart). Meist wurde als Setting der Therapieraum mit Schreibtisch und Couch gezeigt.

Schlussfolgerungen

Der Beitrag diskutiert, ob die einseitig „freudianisch“ geprägte visuelle Darstellung von Psychotherapie möglicherweise beim Publikum falsche Vorstellungen erzeugt und z. B. Skepsis gegenüber Fachleuten schüren könnte, die diesem Image nicht entsprechen. Aus heutiger Sicht ist dazu relativierend einzuwenden, dass die mediale Repräsentation von Psychotherapie und Psychotherapeuten weiter stark zugenommen hat, insbesondere auch in Film und Fernsehen, und somit möglicherweise eine Diversifizierung des Bildes von Psychotherapie bzw. typischen Psychotherapeut/innen in der Medienöffentlichkeit eingetreten ist (Abele, 1990; Jaeggi & Möller, 1997; Klonoff, 1983; Schorr, 1994; Walter, 1992).

Personen quantifiziert (▶ Studienbeispiel „Quantitative Inhaltsanalyse von vorgefundenen visuellen Dokumenten“). **Visuelle Zeichen** sind generell sehr viel uneindeutiger bzw. bedeutungsreicher (sog. **Polysemie**) als sprachliche Zeichen. Während sich geschulte Kodierende derselben Sprachgemeinschaft z. B. mit recht hoher Übereinstimmung (Reliabilität) darauf einigen können, ob eine Textbotschaft einen negativen, neutralen oder positiven Tenor hat, ist es sehr viel schwieriger, Bildbotschaften hinsichtlich abstrakter Merkmale zuverlässig zu kodieren. Dementsprechend müssen Kategoriensystem und Codebuch für die quantitative Analyse visueller Dokumente besonders gründlich geprüft und überarbeitet werden, um valide und messgenaue Daten zu erhalten. Für eine tiefergehende Analyse der Inhalte und Bedeutungen von visuellem Material sind qualitative Verfahren indiziert (▶ Abschn. 10.6.2 „Qualitative Analyse von visuellen Dokumenten“).

Quantitative Inhaltsanalyse von visuellen Dokumenten – Die quantitative Inhaltsanalyse von visuellen Dokumenten („quantitative content analysis of visual documents“) erfasst anhand eines vollstandardisierten Kategoriensystems die Ausprägungen unterschiedlicher formaler und inhaltlicher Merkmale von visuellen Dokumenten (Fotos, Grafiken, Cartoons, Videos, Animationen etc.). Dabei ist die inhaltsanalytische Vorgehensweisen den Besonderheiten der visuellen Dokumente anzupassen, dies betrifft z. B. die Festlegung der Analyse- und Kodiereinheiten, die Stichprobenziehung, die Archivierung des Materials, die Medienspezifik der betrachteten formalen und inhaltlichen Kategorien und nicht zuletzt die Darstellung von Materialbeispielen in Publikationen.

Aktuelle quantitative Inhaltsanalysen vorgefundener visueller Dokumente befassen sich unter unterschiedlichen Fragestellungen z. B. mit politischen Cartoons in Zeitungen (Zurbriggen & Sherman, 2010), Charakteren in Computerspielen (Martins, Williams, Ratan, & Harrison, 2011), TV-Werbespots (Guran, et al., 2010), TV-Serien (Jensen & Jensen, 2007), Fotos in Publikumszeitschriften (Fabrianesi, Jones, & Reid, 2008; Krassas, Blauwkamp, & Wesselink, 2003), Fotos auf Online-Profilen (Siibak, 2010) oder mit YouTube-Videos (Hussin, Frazier, & Thompson, 2011; Paek, Kim, & Hove, 2010; Siibak, 2010). Bei diesen Beispielen wird deutlich, dass eine quantitative Inhaltsanalyse von *visuellen* Dokumenten heutzutage nicht selten gleichzeitig eine quantitative **Online**-Inhaltsanalyse ist, weil das Bildmaterial aus dem Internet stammt (z. B. Analyse von YouTube-Videos samt entsprechenden Kommentaren zu den Videos und Profil-



Abb. 10.54 Visuelle Darstellung von Psychotherapie/Psychotherapeuten auf Zeitschriften-Titelseiten von 1945–1995. (Aus von Sydow et al., 1998. Psychologie Heute, 1976/9, Titel Nr. 14 sowie Psychologie Heute 1989/6, Titel Nr. 18. Mit freundlicher Genehmigung von Psychologie Heute; Der Spiegel 1994/30, Titel Nr. 6. Mit freundlicher Genehmigung von Spiegel; Stern 1995/27, Titel Nr. 9. Mit freundlicher Genehmigung)

Informationen der Videoanbieter sowie der Kommentatoren).

Bei der Ergebnisdarstellung von inhaltsanalytischen Studien mit visuellem Material ergibt sich oft das Problem, dass in gedruckte Zeitschriftenartikel in der Regel keine Farbfotos oder Filmausschnitte eingebunden werden können. Dementsprechend leidet teilweise die intersubjektive Nachvollziehbarkeit im Vergleich zur Analyse von Textdokumenten, bei denen die Ergebnispräsentation durch Textzitate illustriert werden kann. **Wissenschaftliche Online-Fachzeitschriften**, die die Einbindung von multimedialem Datenmaterial ermöglichen, können hier Abhilfe schaffen. Ein weiteres Problem im Umgang mit der Sammlung, Archivierung und Analyse von Bildmaterial stellen bei offiziellen Dokumenten die Abdruckrechte und bei persönlichen Dokumenten die Anonymisierung dar, etwa wenn mit privaten Fotos gearbeitet wird (zur Datenaufbereitung ► Kap. 11).

Lernquiz

1. Wodurch kann die Qualität von vorgefundenen Dokumenten eingeschränkt sein?
2. Welche Arbeitsschritte gehören zur zusammenfassenden qualitativen Inhaltsanalyse nach Mayring?
3. Auf welche Dokumententypen ist die Interpretative Phänomenologische Analyse (IPA) nicht anwendbar?
4. Wie erfolgt eine Kodierung gemäß Grounded-Theory-Methodologie?
5. Welche Arbeitsschritte sind im Kontext der quantitativen Inhaltsanalyse notwendig für die Konstruktion eines standardisierten Kategoriensystems?
6. Die Darstellung von Wissenschaftlerinnen und Wissenschaftlern in den Edutainment-TV-Formaten „Galileo“, „Nano“ und „Wunder Welt Wissen“ soll untersucht werden. Nennen Sie
   a) eine sinnvolle ordinalskalierte formale Kategorie mit ihren Ausprägungen;
   b) eine sinnvolle kardinalskalierte formale Kategorie mit ihren Ausprägungen;
   c) eine sinnvolle nominalskalierte inhaltliche Kategorie mit ihren Ausprägungen und
   d) eine sinnvolle kardinalskalierte inhaltliche Kategorie mit ihren Ausprägungen!
7. Wozu dient Cohens Kappa im Rahmen der Dokumentenanalyse und was zeichnet diesen Koeffizienten aus?
8. Im Rahmen einer quantitativen Dokumentenanalyse werden 50 Zeitungsartikel von 2 Kodierern mit einem aus 30 dichotomen Kategorien bestehenden inhaltsanalytischen Kategoriensystem kodiert. Wie viele Reliabilitätskoeffizienten sind zur Beurteilung der Messgenauigkeit des Kategoriensystems zu bestimmen?
9. Wozu dient im Kontext der quantitativen Inhaltsanalyse die Intra-Klassen-Korrelation (ICC)?
10. Was versteht man unter der Transitorik von Dokumenten und welche Bedeutung hat sie für die Online-Inhaltsanalyse?

Übungsaufgaben

1. Erarbeiten Sie fünf formale und fünf inhaltliche Kategorien inklusive Kategorienausprägungen für eine quantitative Inhaltsanalyse von Online-Gesundheitsportalen (z. B. http://www.gesundheit.de, http://www.netdoktor.de, http://www.vitanet.de)! Geben Sie auch an, wie Sie die Analyseeinheiten definieren.

2. Recherchieren Sie für eine Dokumentenanalyse die Parteiprogramme (Grundsatzprogramme) aller Parteien des deutschen Bundestages und archivieren Sie diese! Wie würden Sie vorgehen, um im Zuge einer systematischen wissenschaftlichen Dokumentenanalyse die in den Grundsatzprogrammen der Parteien artikulierten Positionen zu Wissenschaft(spolitik) herauszuarbeiten und zu vergleichen?
3. Angenommen, Sie interessieren sich dafür, wie Jugendliche versus Erwachsene sich in ihren fotografischen Selbstportraits auf Social-Networking- und Foto-Plattformen (z. B. Facebook, Flickr, Instragram) darstellen. Wie würden Sie die Foto-Stichprobe zusammenstellen? Würden Sie eine qualitative und/oder eine quantitative Auswertungsstrategie für die visuellen Dokumente (Fotos) wählen? Begründen Sie Ihre Entscheidungen und schätzen Sie den Arbeitsaufwand für die Umsetzung Ihrer Entscheidung ab!

Weiterführende Literatur

- Früh, W. (2007). *Inhaltsanalyse: Theorie und Praxis* (5. Aufl.). Stuttgart: UTB.
- Gwet, K. L. (2012). *Handbook of inter-rater reliability* (3rd ed.). Gaithersburg: Advanced Analytics, LLC.
- Mayring, P. (2010). Qualitative Inhaltsanalyse (11. Aufl.). Weinheim: Beltz.
- Mey, G. & Mruck, K. (Hrsg.) (2011). Grounded theory reader (2. Aufl.). Wiesbaden: VS Verlag.
- Neuendorf, K. A. (2001). *The content analysis guidebook online*. Thousand Oaks: Sage. Buch-Website: http://academic.csuohio.edu/neuendorf_ka/content/.
- Schreier, M. (2012). *Qualitative content analysis in practice*. London: Sage.
- Van Leeuwen, T. & Jewitt, C. (2001). *Handbook of visual analysis*. Thousand Oaks: Sage.
- Welker, M. & Wünsch, C. (Hrsg.) (2010). *Online-Inhaltsanalyse*. Köln: Halem.

Literatur

Abele, A. (1990). Psychologie in den Medien: Anmerkungen zu einem diffizilen Thema. *Psychologische Rundschau, 41*(1), 37–45.

Archer, D., Iritani, B., Kimes, D. D., & Barrios, M. (1983). Face-ism: Five studies of sex difference in facial prominence. *Journal of Personality and Social Psychology, 45*(4), 725–735.

Arens, M., Böcking, S., Kummer, S., & Rüf, F. (2010). Das Meinungsklima zur Klimakonferenz: Wie sich Themenkarrieren im Internet entwickeln und mit welchen Themen sich Politiker positionieren. In M. Welker & C. Wünsch (Hrsg.), *Online-Inhaltsanalyse* (S. 427–447). Köln: Halem.

Arlt, D. & Wolling, J. (2011). *Event-Berichterstattung oder Themen-Berichterstattung? Potenzial und Realität der Medienberichterstattung über die Weltklimakonferenz in Kopenhagen.* Unveröffentlichtes Manuskript. Institut für Medien und Kommunikationswissenschaft, TU Ilmenau.

Arlt, D. & Wolling, J. (2012). Die Presseberichterstattung über die Weltklimakonferenz in Kopenhagen. Normative Anforderungen und empirische Befunde. *SCM Studies in Communication I Media, 1*(3), 283–297.

Baumann, M., Langeder, M., Much, M., Obermayer, B., & Storz, F. (Hrsg.). (2011). *Feldpost: Briefe deutscher Soldaten aus Afghanistan*. Hamburg: Rowohlt.

Berelson, B. (1952). *Content analysis in communication research*. New York: Hafner.

Bernhard, B. J., Futrell, R., & Harper, A. (2010). „Shots from the pulpit": An ethnographic content analysis of United States anti-gambling social movement documents. *UNLV Gaming Research and Review Journal, 14*(2), 15–32.

Borzekowski, D. L. G., Schenk, S., Wilson, J. L., & Peebles, R. (2010). E-Ana and e-Mia: A content analysis of pro-eating disorder web sites. *American Journal of Public Health, 100*(8), 1526–1534.

Breuer, F. (1996). *Qualitative Psychologie. Grundlagen, Methoden und Anwendungen eines Forschungsstils.* Opladen: Westdeutscher Verlag.

Breuer, F. (unter Mitarbeit von Dieris, B. & Lettau, A.). (2010). *Reflexive Grounded Theory. Eine Einführung für die Forschungspraxis* (2. Aufl.). Wiesbaden: VS.

Bryant, A. & Charmaz, K. (Eds.). (2010). *The SAGE handbook of grounded theory*. Thousand Oaks: Sage.

Camic, P. M. (2010). From trashed to treasured: A grounded theory analysis of the found object. *Psychology of Aesthetics, Creativity, and the Arts, 4*(2), 81–92.

Cho, S. & Huh, J. (2010). Content analysis of corporate blogs as a relationship management tool. *Corporate Communications, 15*(1), 30–48.

Clarke, J. N. (2010). The paradoxical portrayal of the risk of sexually transmitted infections and sexuality in US magazines glamour and cosmopolitan 2000–2007. *Health, Risk and Society, 12*(6), 560–574.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Colaizzi, P. (1978). Psychological research as the phenomenologist views it. In R. Valle & M. King (Eds.), *Existential-phenomenological alternatives for psychology* (pp. 48–71). Oxford: Oxford University Press.

Copeland, G. A. (1989). Face-ism and primetime television. *Journal of Broadcasting and Electronic Media, 33*(2), 209–214.

Dieris, B. (2006). "Och Mutter, was ist aus dir geworden?!" Eine Grounded-Theory-Studie über die Neupositionierung in der Beziehung zwischen alternden Eltern und ihren erwachsenen, sich kümmernden Kindern. *Forum Qualitative Social Research, 7*(3), Art. 25. Abgerufen 26. November, 2013, unter http://nbn-resolving.de/urn:nbn:de:0114-fqs0603253.

Duff, S. (2010). Exploring criminogenic need through victim apology letters: An interpretative phenomenological analysis. *Journal of Aggression, Conflict and Peace Research, 2*(2), 33–43.

Elkins, J. (2003). *Visual studies: A skeptical introduction*. London: Routledge Chapman & Hall.

Fabrianesi, B., Jones, S. C., & Reid, A. (2008). Are pre-adolescent girls' magazines providing age-appropriate role models? *Health Education, 108*(6), 437–449.

Faulstich, W. (2008). *Grundkurs Filmanalyse*. Stuttgart: UTB.

Faux, W. V. & Kim, H. (2006). Visual representation of the victims of Hurricane Katrina: A dialectical approach to content analysis and discourse. *Space and Culture, 9*(1), 55–59.

Figueroa, S. K. (2008). The grounded theory and the analysis of audio-visual texts. *International Journal of Social Research Methodology, 11*(1), 1–12.

Früh, W. (1998). *Inhaltsanalyse* (4. Aufl.). Konstanz: UVK.

Früh, W. (2007). *Inhaltsanalyse: Theorie und Praxis* (6. Aufl.). Stuttgart: UTB.

Früh, W., Hasebrink, U., Krotz, F., Kuhlmann, C. & Stiehler, H.-J. (1999). *Ostdeutschland im Fernsehen*. München: kopaed.

Gackenbach, J., Sample, T., Mandel, G., & Tomashewsky, M. (2011). Dream and blog content analysis of a long term diary of a video game player with obsessive compulsive disorder. *Dreaming, 21*(2), 124–147.

Gao, X., Hamzah, S. H., Yiu, Y. C. K., McGrath, C., & King, M. N. (2013). Dental fear and anxiety in children and adolescents: Qualitative study using YouTube. *Journal of Medical Internet Research, 15*(2), e29. Retrieved November 26, 2013, from http://www.jmir.org/2013/2/e29/.

Gehrau, V., Bilandzic, H., & Woelke, J. (Hrsg.). (2005). *Rezeptionsstrategien und Rezeptionsmodalitäten*. München: Fischer.

Geyzen, A. (2011). Popular discourse on nutrition, Health and indulgence in Flanders, 1945–1960. *Appetite, 56*(2), 278–283.

Glaser, B. G. (1998). *Doing grounded theory: Issues & Discussion*. Mill Valley: Sociology Press.

Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory.* Chicago: Aldine.

Gleich, U. (2003). Qualität im Journalismus am Beispiel der Kriegsberichterstattung. Forschungsbeiträge zur Qualitätsdebatte. *Media Perspektiven, 3*, 139–148.

Goodnow, T. (2010). Visual bias in time's „The Great Divide": A semiotic analysis of Clinton and Obama photographs. *American Behavioral Scientist, 54*(4), 406–416.

Groenewald, T. (2008). Memos and memoing. In L. M. Given (Ed.), *The SAGE encyclopedia of qualitative research methods* (Vol. 2, pp. 505–506). Thousand Oaks: Sage.

Guran, T., Turan, S., Akcay, T., Degirmenci, F., Avci, O., Asan, A., et al. (2010). Content analysis of food advertising in Turkish television. *Journal of Paediatrics and Child Health, 46*(7–8), 427–430.

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability* (3rd ed.). Gaithersburg: Advanced Analytics, LLC.

Helwig, A. A. & Schmidt, L. L. L. (2011). Content analysis of 32 years of American Counseling Association convention programs. *Journal of Counseling & Development, 89*(2), 148–154.

Herbers, M. & Friedemann, A. (2010). Spezielle Fragen der Reliabilität und Validität bei Online-Inhaltsanalysen. In M. Welker & C. Wünsch (Hrsg.), *Die Online-Inhaltsanalyse. Forschungsobjekt Internet* (S. 240–266). Köln: Halem.

Hester, J.–B. & Dougall, E. (2007). the effectiveness and efficiency of constructed week sampling for content analysis of online news. *Journalism and Mass Communication Quarterly, 84*(4), 811–824.

Hirdes, W., Woods, R., & Badzinski, D. M. (2009). A content analysis of Jesus merchandise. *Journal of Media and Religion, 8*(3), 141–157.

Holsti, O. R. (1969). *Content analysis for the social sciences*. Reading: Addison-Wesley.

Holt, N. L. & Tamminen, K. A. (2010). Moving forward with grounded theory in sport and exercise psychology. *Psychology of Sport and Exercise, 11*(6), 419–422.

Hsieh, H.-F. & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288.

Husserl, E. (1913). *Ideen zu einer Phänomenologie und phänomenologischen Philosophie. Erstes Buch: Allgemeine Einführung in die reine Phänomenologie*. Halle (Saale): Niemeyer.

Hussin, M., Frazier, S., & Thompson, J. K. (2011). Fat stigmatization on YouTube: A content analysis. *Body Image, 8*(1), 90–92.

Jaeggi, E. & Möller, H. (1997). Die Rolle der PsychologInnen in den Medien. *Journal für Psychologie, 5*(1), 59–64.

Jandura, G., Jandura, O., & Kuhlmann, C. (2005). Stichprobenziehung in der Inhaltsanalyse. Gegen den Mythos der künstlichen Woche. In V. Gehrau, B. Fretwurst, B. Krause & G. Daschmann (Hrsg.), *Auswahlverfahren in der Kommunikationswissenschaft* (S. 71–116). Köln: Halem.

Jensen, R. E. & Jensen, J. D. (2007). Entertainment media and sexual health: a content analysis of sexual talk, behavior, and risks in a popular television series. *Sex Roles, 56*(5–6), 275–284.

Kempster, S. & Parry, K. W. (2011). Grounded theory and leadership research: A critical realist perspective. *The Leadership Quarterly, 22*(1), 106–120.

Kim, Y. G., Eves, A., & Scarles, C. (2009). Building a model of local food consumption on trips and holidays: A grounded theory approach. *International Journal of Hospitality Management, 28*(3), 423–431.

Klonoff, E. A. (1983). A star is born: Psychologists and the media. *Professional Psychology: Research and Practice, 14*(6), 847–854.

Knoblauch, H., Baer, A., Laurier, E., Petschke, S., & Schnettler, B. (2008). Visual analysis. New developments in the interpretative analysis of video and photography. *Forum Qualitative Social Research, 9*(3), Art. 14. Retrieved November 26, 2013, from http://www.qualitativeresearch.net/index.php/fqs/article/view/1170

Kracauer, S. (1952). The challenge of qualitative content analysis. *Public Opinion Quarterly, 16*(4), 631–642.

Krassas, N. R., Blauwkamp, J. M., & Wesselink, P. (2003). „Master your Johnson": Sexual rhetoric in maxim and stuff magazines. *Sexuality and Culture: An Interdisciplinary Quarterly, 7*(3), 98–119.

Krippendorff, K. H. (2003). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks: Sage.

Kuhlmann, C. (1999). *Die öffentliche Begründung politschen Handelns. Zur Argumentationsrationalität in der politischen Massenkommunikation*. Opladen: Westdeutscher Verlag.

Lance, L. M. (1998). Gender differences in heterosexual dating: A content analysis of personal ads. *The Journal of Men's Studies, 6*(3), 297–305.

Lauf, E. & Berens, H. (2003). Das Risiko ökologischer und individualistischer Fehlschlüsse in der Inhaltsanalyse. *Publizistik: Vierteljahreshefte für Kommunikationsforschung, 48*(4), 454–469.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., et al. (2009). „Computational Social Science". *Science, 323*(5915), 721–723.

Lee, M., Choi, Y., Quilliam, E. T., & Cole, R. T. (2009). Playing with food: Content analysis of food advergames. *Journal of Consumer Affairs, 43*(1), 129–154.

Liu, Y., Liu, M., Xiao, H., Cai, J., & Xu, W. (2010). A content analysis of news coverage of skin cancer in China newspapers. *Health Communication, 25*(8), 647–649.

Lunsford, A. A., Stapleton, L., Fishman, J., Krampetz, E., Rogers, P. M., Diogenes, M., et al. (2012). *The Stanford study of writing.* Retrieved February 28, 2012, from http://ssw.stanford.edu/

Martins, N., Williams, D. C., Ratan, R. A., & Harrison, K. (2011). Virtual muscularity: A content analysis of male video game characters. *Body Image, 8*(1), 43–51.

Mason, P. (2005). Visual data in applied qualitative research – Lessons from experience. *Qualitative Research, 5*(3), 325–346.

Mayring, P. (2001). Combination and Integration of Qualitative and Quantitative Analysis. *Forum Qualitative Social Research, 2*(1), Art. 6. Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/967/2110

Mayring, P. (2010a). *Qualitative Inhaltsanalyse* (11. Aufl.). Weinheim: Beltz.

Mayring, P. (2010b). Qualitative Inhaltsanalyse. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie*. Wiesbaden: VS.

Merten, K. (1995). *Inhaltsanalyse*. Opladen: Westdeutscher Verlag.

Mey, G. (1999). *Adoleszens, Identität, Erzählung. Theoretische, methodologische und empirische Erkundungen*. Berlin: Köster.

Mey, G. (2007). Qualitative research on „adolescence, identity, narration": Programmatic and empirical examples. In M. Watzlawik & A. Born (Eds.), *Capturing identity. Quantitative and qualitative methods* (pp. 53–69). Lanham: University Press of America.

Mey, G. & Mruck, K. (2009). Methodologie und Methodik der Grounded Theory. In W. Kempf & M. Kiefer (Hrsg.), *Forschungsmethoden der Psychologie. Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik. Band 3.* (S. 100–152). Berlin: Regener.

Mey, G. & Mruck, K. (Hrsg.). (2007). *Grounded theory reader (HSR Supplement 19)*. Köln: ZHSF.

Mey, G. & Mruck, K. (Hrsg.). (2011). *Grounded theory reader* (2. Aufl.). Wiesbaden: VS Verlag.

Moran, D. (2000). *Introduction to phenomenology*. Abingdon: Routledge.

Morgan, D. L. (1993). Qualitative content analysis: A guide to paths not taken. *Qualitative Health Research, 3*(1), 112–121.

Muckel, P. (1997). *Der Alltag mit Akten – psychologische Rekonstruktionen bürokratischer Phänomene.* Eine empirische Untersuchung in verschiedenen Institutionen auf der Grundlage der Grounded Theory. Aachen: Shaker.

Mühlhauser, I. & Oser, F. (2008). Sind medizinische und Gesundheitsinformationen auf den Internetseiten von Wikipedia evidenzbasiert? Eine Inhaltsanalyse. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 102*(7), 441–448.

Mulveen, R. & Hepworth, J. (2006). An interpretative phenomenological analysis of participation in a pro-anorexia internet site and its relationship with disordered eating. *Journal of Health Psychology, 11*(2), 283–296.

Neill, S. J. (2010). Containing acute childhood illness within family life: A substantive grounded theory. *Journal of Child Health Care, 14*(4), 327–344.

Neuendorf, K. A. (2001). *The content analysis guidebook*. Thousand Oaks: Sage.

Niesyto, H. & Marotzki, W. (Hrsg.). (2005). *Visuelle Methoden in der Forschung*. Zürich: MedienPädagogik.

Nimrod, G. (2010). Seniors' online communities: A quantitative content analysis. *The Gerontologist, 50*(3), 382–392.

Orizio, G., Schulz, P., Gasparotti, C., Caimi, L., & Gelatti, U. (2010). The world of e-patients: A content analysis of online social networks focusing on diseases. *Telemedicine and e-Health, 16*(10), 1060–1066.

Paek, H.-J., Kim, K., & Hove, T. (2010). Content analysis of antismoking videos on YouTube: Message sensation value, message appeals, and their relationships with viewer responses. *Health Education Research, 25*(6), 1085–1099.

Pantelidis, I. S. (2010). Electronic meal experience: A content analysis of online restaurant comments. *Cornell Hospitality Quarterly, 51*(4), 483–491.

Pauwels, L. (2010). Visual sociology reframed: An analytical synthesis and discussion of visual methods in social and cultural research. *Sociological Methods & Research, 38*(4), 545–581.

Pettican, A. & Prior, S. (2011). 'It's a new way of life': An exploration of the occupational transition of retirement. *The British Journal of Occupational Therapy, 74*(1), 12–19.

Piggott, D. (2010). Listening to young people in leisure research: The critical application of grounded theory. *Leisure Studies, 29*(4), 415–433.

Popping, R. (2010). Some views on agreement to be used in content analysis studies. *Quality and Quantity: International Journal of Methodology, 44*(6), 1067–1078.

Prior, L. (2008). Repositioning documents in social research. *Sociology, 42*(5), 821–836.

Reid, K., Flowers, P., & Larkin, M. (2005). Exploring lived experience. *The Psychologist, 18*, 20–23.

Richard, B., Grünwald, J., Recht, M. & Metz, N. (2010). *Flickernde Jugend – rauschende Bilder: Netzkulturen im Web 2.0*. Frankfurt: Campus.

Riffe, D., Aust, C. F., & Lacy, S. R. (1993). The effectiveness of random, consecutive day and constructed week sampling in newspaper content analysis. *Journalism Quarterly, 70*(1), 133–139.

Robinson, K. (2010). A study of young lesbian and gay people's school experiences. *Educational Psychology in Practice, 26*(4), 331–351.

Rössler, P. (2005). *Inhaltsanalyse*. Konstanz: UVK.

Sakai, D. K. & Johnson, R. C. (1997). Active phenotypic assortment in mate selection: Self-descriptions and sought-for-attributes of mates in dating advertisements. *Biodemography and Social Biology, 44*(3–4), 258–264.

Salska, I., Frederick, D. A., Pawlowski, B., Reilly, A. H., Laird, K. T., & Rudd, N. A. (2008). Conditional mate preferences: Factors influencing preferences for height. *Personality and Individual Differences, 44*(1), 203–215.

Schachman, K. A. (2010). Online fathering: The experience of first-time fatherhood in combat-deployed troops. *Nursing Research, 59*(1), 11–17.

Scharkow, M. (2010). Lesen und lesen lassen – State of the Art automatischer Textanalyse. In M. Welker & C. Wünsch (Hrsg.), *Online-Inhaltsanalyse* (S. 340–364). Köln: Halem.

Scheufele, B. & Engelmann, I. (2009). *Empirische Kommunikationsforschung*. Stuttgart: UTB.

Schorr, A. (1994). Die Psychologie in den Printmedien. Eine Inhaltsanalyse auflagenstarker Tageszeitungen und Zeitschriften. *Medienpsychologie, 6*(3), 199–219.

Schreier, M. (2012). *Qualitative content analysis in practice.* London: Sage.

Schweitzer, E. J. (2010). Normalisierung 2.0. Die Online-Wahlkämpfe deutscher Parteien zu den Bundestagswahlen 2002–2009. In C. Holtz-Bacha (Hrsg.), *Die Massenmedien im Wahlkampf: Die Bundestagswahl 2005* (S. 189–244). Wiesbaden: VS Verlag.

Scott, J. (1990). *A matter of record*. Cambridge: Poilty.

Siibak, A. (2010). Constructing masculinity on a social networking site: The case-study of visual self-presentations of young men on the profile images of SNS rate. *Young, 18*(4), 403–425.

Smith, J. A. (1996). Beyond the divide between cognition and discourse: Using interpretative phenomenological analysis in health psychology. *Psychology and Health, 11*(2), 261–271.

Smith, J. A., Flowers, P., & Larkin, M. (2009). *Interpretative phenomenological analysis: Theory method and research*. Thousand Oaks: Sage.

Spencer, S. (2011). *Visual research methods in the social sciences. Awakening visions*. Oxon: Routledge.

Strauss, A. L. (1991). *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.

Strauss, A. L. & Corbin, J. M. (1994). Grounded theory methodology: An overview. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273–285). Thousand Oaks: Sage.

Strauss, A. L. & Corbin, J. M. (1996). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks: Sage.

Tomkins, L. & Eatough, V. (2010). Reflecting on the use of IPA with focus groups: Pitfalls and potentials. *Qualitative Research in Psychology, 7*(3), 244–262.

Van Leeuwen, T. & Jewitt, C. (Eds.). (2001). *Handbook of visual analysis*. Thousand Oaks: Sage.

Von Sydow, K., Weber, A., & Reimer, C. (1998). „Psychos" in den Medien. Eine Inhaltsanalyse der Titelbilder von acht deutschen Zeitschriften aus dem Zeitraum von 1947 bis 1995. *Psychotherapeut, 43*(2), 80–91.

Walter, G. (1992). The psychiatrist in American cartoons, 1941–1990. *Acta Psychiatrica Scandinavia, 85*(2), 167–172.

Welker, M., Wünsch, C., Böcking, S., Bock, A., Friedemann, A., Herbers, M., et al. (2010). Die Online-Inhaltsanalyse: methodische Herausforderung, aber ohne Alternative. In M. Welker & C. Wünsch (Hrsg.), *Die Online-Inhaltsanalyse. Forschungsobjekt Internet* (S. 9–30). Köln: Herbert von Halem Verlag.

Westland, S. & Shinebourne, P. (2009). Self-deception and the therapist: An interpretative phenomenological analysis of the experiences and understandings of therapists working with clients they describe as self-deceptive. *Psychology and Psychotherapy: Theory, Research and Practice, 82*(4), 385–401.

Whitty, M. T., Young, G., & Goodings, L. (2011). What I won't do in pixel: Examining the limits of taboo violation in MMORPGs. *Computers in Human Behavior, 27*(1), 268–275.

Williams, A. & Merten, M. (2009). Adolescents' online social networking following the death of a peer. *Journal of Adolescent Research, 24*(1), 67–90.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wolling, J. (2002). Methodenkombination in der Medienwirkungsforschung. Der Entscheidungsprozess bei der Verknüpfung von Umfrage- und Inhaltsanalysedaten. *ZUMA-Nachrichten, 26*(50), 54–85.

Wolling, J. & Arlt, D. (2012). Codebuch zur inhaltsanalytischen Untersuchung der Weltklimakonferenz 2009 in den Medien. Abgerufen 26. November, 2013, unter http://www.dbthueringen.de/servlets/DocumentServlet?id=20414

Wong, Y. J., Steinfeldt, J. A., Speight, Q. L., & Hickman, S. J. (2010). Content analysis of psychology of men & masculinity (2000–2008). *Psychology of Men and Masculinity, 11*(3), 170–181.

Yakushko, O. (2010). Stress and coping strategies in the lives of recent immigrants: A grounded theory model. *International Journal for the Advancement of Conselling, 32*(4), 256–273.

Zimmer, M. (2011). „But the data is already public": On the ethics of research in facebook. *Ethics and Information Technology, 12*(4), 313–325.

Zuckerman, M. & Kieffer, S. C. (1994). Race differences in face-ism: Does facial prominence imply dominance? *Journal of Personality and Social Psychology, 66*(1), 86–92.

Zurbriggen, E. L. & Sherman, A. M. (2010). Race and gender in the 2008 U.S. presidential election: A content analysis of editorial cartoons. *Analyses of Social Issues and Public Policy, 10*(1), 223–247.

Datenaufbereitung

Nicola Döring, Jürgen Bortz

11.1 Definition der Datenaufbereitung – 580

11.2 Funktionen und Bedeutung der Datenaufbereitung – 581

11.3 Aufbereitung qualitativer Daten – 582

11.4 Aufbereitung quantitativer Daten – 584

Literatur – 594

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_11, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Wissen, was man unter Datenaufbereitung versteht.
- Die Funktionen der Datenaufbereitung kennen.
- Zentrale Arbeitsschritte der Aufbereitung qualitativer Daten beschreiben und durchführen können.
- Wichtige Arbeitsschritte der Aufbereitung quantitativer Daten erläutern und umsetzen können.

Nach der Datenerhebung und vor der Datenanalyse erfolgt die **Datenaufbereitung**. Welche Funktionen sie erfüllt und welche Schritte für die Aufbereitung qualitativer und quantitativer Daten empfehlenswert sind, wird im Folgenden dargestellt.



Abb. 11.1 Rohdaten müssen vor der Datenanalyse gründlich aufbereitet werden. (© Harley Schwadron/Search ID: hsc4679, Rights Available from CartoonStock.com)

11.1 Definition der Datenaufbereitung

Nach Abschluss der Datenerhebung liegen z. B. ausgefüllte Fragebögen, Audioaufzeichnungen von Leitfadeninterviews, handschriftliche Feldnotizen und/oder physiologische Messwertreihen vor. Das **Rohdatenmaterial** einer empirischen Studie existiert oft in verschiedenen Formen und Formaten, enthält Lücken und Inkonsistenzen und ist in seiner Gesamtheit schwer zu überblicken. Es muss zunächst sorgfältig sortiert, kommentiert, formatiert, anonymisiert, bereinigt und oft auch transformiert werden, um eine systematische Datenanalyse überhaupt zu ermöglichen (Abb. 11.1). Am Ende der Datenaufbereitung stehen in der Regel mit Metainformationen bzw. Kommentierungen versehene, fehlerbereinigte und anonymisierte elektronische **Datensätze**, die dann direkt mit entsprechender qualitativer oder quantitativer Analysesoftware auswertbar sind.

Zu den wichtigsten, allerdings nicht immer trennscharf abgrenzbaren und oft auch unterschiedlich bezeichneten Schritten der Datenaufbereitung gehören:

1. **Erstellung der Datensätze**: Sortierung, Zuordnung, Digitalisierung, Formatierung des Rohdatenmaterials zu strukturierten Datensätzen. Bei der qualitativen Forschung gehört hierzu auch die Verschriftlichung (Transkription) von Audio- und Videoaufzeichnungen.
2. **Kommentierung der Datensätze**: Ergänzung der qualitativen und quantitativen Datensätze um Metainformationen wie das Erhebungsdatum (im Sinne einer Methodendokumentation im Datensatz). Zudem werden in quantitativen Datensätzen die Bedeutungen der Variablen und der vergebenen numerischen Werte festgehalten (sog. Labeling, d. h. Vergabe von Variablen- und Wertelabels).
3. **Anonymisierung der Datensätze**: Entfernung/Ersetzung von Informationen, die zur Identifizierung von Untersuchungsteilnehmenden führen könnten.
4. **Datenbereinigung**: Bei qualitativen Daten werden zuweilen für die jeweilige Auswertung unwichtige Details entfernt (etwa „Glättung“ von Transkripten um Versprecher, verschluckte Silben, Dialekt), sofern dies die Arbeit mit den Daten erleichtert. Bei quantitativen Daten geht es um die Identifikation und Korrektur unplausibler und fehlerhafter numerischer Angaben (z. B. Werte außerhalb des zulässigen Wertebereichs), Entfernung doppelter Datensätze, sinnvolle Behandlung von Lücken im Datenmaterial (z. B. Nacherhebung, Kennzeichnung fehlender Werte, Ersetzung fehlender Werte durch Schätzwerte gemäß entsprechender statistischer Verfahren).
5. **Datentransformation**: Umpolung und Zusammenfassung von quantitativen Variablen, ggf. Gewichtung von komplexen probabilistischen Stichproben, um deren Repräsentativität zu erhöhen.

Datenaufbereitung – Zur Datenaufbereitung („data preparation“) gehören all jene begründeten und dokumentierten Bearbeitungen bzw. Veränderungen des Rohdatenmaterials, welche die Aussagekraft und (Wieder-)Verwendbarkeit der Daten steigern und die inhaltliche Datenanalyse vorbereiten. Dazu zählen v. a. die Erstellung strukturierter Datensätze aus dem Rohdatenmaterial, die Kommentierung und die Anonymisierung der Datensätze sowie Datenbereinigung und Datentransformation. Ziel der Datenaufbereitung ist eine Steigerung der Datenqualität.

Die Begrifflichkeiten im Feld der wissenschaftlichen Datenaufbereitung sind noch uneinheitlich. Meist werden die unbehandelten Daten als „**Rohdaten**“ oder „Rohdatenmaterial“ bezeichnet („raw data“) und die aufbereiteten Daten als „Datensätze“. Dabei ist zu beachten, dass die Bezeichnung „**Datensatz**“ oft in unterschiedlichen Bedeutungen verwendet wird:

- Mit dem Datensatz können zum einen alle in einer Datei zusammengefassten **Daten einer Stichprobe** gemeint sein (z. B. sämtliche Fragebogenantworten aller $n = 500$ Befragungspersonen, die in einer SPSS-Datei zusammengeführt sind).
- Mit einem Datensatz können aber auch nur die **Daten einer einzelnen Untersuchungseinheit** gemeint sein (z. B. sämtliche Fragebogenantworten von Befragungsperson 17).

Manche Forschungsprojekte operieren nur mit **einem** Datensatz im erstgenannten Sinne (z. B. einem einzigen SPSS-Datenfile), andere mit **mehreren** verschiedenen Datensätzen, weil sie verschiedene Stichproben untersucht und/oder verschiedene Datenerhebungsmethoden eingesetzt haben. Insbesondere bei Mixed-Methods-Studien ist es von Belang, ob und wie die verschiedenen qualitativen und quantitativen Datensätze direkt aufeinander bezogen bzw. miteinander verknüpft werden. Je komplexer Studien sind, umso wichtiger ist es, Fragen der Sicherung der Datenqualität bereits bei der Untersuchungsplanung zu berücksichtigen (z. B. Vergabe von Identifizierungscodes, um die Daten einer Person über verschiedene Teilerhebungen hinweg einander anonymisiert zuordnen zu können).

11.2 Funktionen und Bedeutung der Datenaufbereitung

In vielen Modellen des empirischen Forschungsprozesses wird die Datenaufbereitung nicht gesondert als eigene Phase aufgeführt. Ein Grund mag sein, dass **Datenqualität** („data quality“) zu oft als selbstverständlich betrachtet wird und nicht als etwas, das anhand von Kriterien geprüft, belegt bzw. aktiv hergestellt werden sollte. Geringe Datenqualität führt zu hohen Folgekosten, etwa wenn auf dieser Basis betriebswirtschaftlich ungünstige Entscheidungen getroffen werden. In der Wissenschaft führt das Arbeiten mit Daten fragwürdiger Qualität zu unzuverlässigen Ergebnissen, die nicht nur den Erkenntnisgewinn behindern, sondern auch unerwünschte praktische Folgen haben können. Eine gründliche Datenaufbereitung inklusive Datenbereinigung kann im Einzelfall ebenso viel oder sogar mehr Zeitaufwand und Expertise erfordern als die nachgelagerte Datenanalyse (Schendera, 2007).

Die Datenaufbereitung erfüllt mehrere **Funktionen**:

1. **Vermeidung fehlerhafter Ergebnisse:** Eine Auswertung unvollständiger oder fehlerhafter Daten führt unweigerlich zu fehlerhaften Ergebnissen (sog. Garbage-in-Garbage-out-Prinzip). Die Analyse eines unbereinigten Datensatzes, in dem sich z. B. Dopplungen, Tippfehler oder unplausible Antworten befinden, kann das gesamte Ergebnis verzerren und zu falschen inhaltlichen Schlussfolgerungen führen. Derartige Probleme sind vermeidbar, wenn **von Anfang an** die Datenqualität geprüft und sichergestellt wird. Das Fatale an suboptimaler Datenqualität ist, dass sie erst in dem Moment erkannt wird, wenn sie geprüft wird. Wird in einem Forschungsprozess erst spät – etwa erst während oder am Ende der Datenanalyse – mangelhafte Datenqualität festgestellt, sind oft alle zuvor ausgeführten Analyseschritte umsonst gewesen.
2. **Vermeidung von Schwierigkeiten und Verzögerungen bei der Daten(re)analyse:** Angestrebt wird eine reibungslose aktuelle und spätere Datenanalyse durch die Forschenden selbst, durch Forschungspartner oder andere Forschungsteams, die die Daten einer **Re- bzw. Sekundäranalyse** unterziehen möchten. Dies setzt v. a. eine stringente Organisation und ausreichende Kommentierung der Datensätze durch **Metadaten** voraus (z. B. genaue Angaben, wann, wo und durch wen die Daten erhoben wurden, was Variablennamen und Messwerte bedeuten etc.). Zudem sind für Austausch und Auswertung der Daten einheitliche Darstellungsformen und kompatible Formatierungen notwendig. Ein schlecht aufbereiteter Datensatz kann unbrauchbar werden, etwa wenn mangelnde Kommentierung bzw. ein fehlender Codeplan dazu führen, dass zu einem späteren Zeitpunkt nicht – oder nur mit großem Aufwand – rekonstruierbar ist, was bestimmte Messwerte eigentlich inhaltlich bedeuten.
3. **Vermeidung ethischer Probleme:** Insbesondere in der qualitativen Forschung können Versäumnisse bei der Anonymisierung des Rohdatenmaterials die teilnehmenden Personen identifizierbar machen. Identifizierbarkeit ist – sofern kein ausdrückliches Einverständnis der Untersuchungsteilnehmenden vorliegt (z. B. erlaubte namentliche Nennung bei einem Experteninterview) – nicht nur ein Verstoß gegen die Forschungsethik, sondern auch gegen Datenschutzgesetze. Dies gilt unabhängig davon, ob durch die Identifizierbarkeit tatsächlich Beeinträchtigungen für eine einzelne Untersuchungsteilnehmerin oder einen einzelnen Untersuchungsteilnehmer spürbar werden (z. B. soziale Stigmatisierung angesichts von Interviewäußerungen zu heiklen Themen).

Da die Datenaufbereitung einen Eingriff in die Daten darstellt, der prinzipiell auch zu **Manipulationszwecken** missbraucht werden kann (z. B. hypothesenkonträre Fälle werden als „Ausreißer“ ausgeschlossen), ist das Vorgehen bei der Datenaufbereitung gründlich zu reflektieren und im Ergebnisbericht entsprechend zu **begründen** und zu **dokumentieren**. Manipulationen am Datenmaterial, die darauf hinauslaufen, ein bestimmtes „Wunschergebnis“

bewusst herzustellen, stellen eine gravierende Verletzung der Wissenschaftsethik dar (▶ Abschn. 4.2.2).

Die **praktische Bedeutung** der Datenaufbereitung ist in den letzten Jahren deutlich gestiegen. Dies hat v. a. folgende Gründe:

- Die **computergestützte Datenanalyse** ist bei qualitativen wie quantitativen Studien heute der Regelfall (zu entsprechender Software ▶ Abschn. 12.1.3 und ▶ Abschn. 12.2.3). Sie setzt eine einheitliche Formatierung und digitale Aufbereitung des Materials voraus.
- Die steigende Zahl an **kooperativen Forschungsprojekten**, in denen mehrere geografisch verstreute Forschende oder Forschungsteams in wechselseitiger Abstimmung Daten erheben, untereinander austauschen und gemeinsam auswerten, verlangt ein professionelles Datenmanagement.
- Der Trend zur **Archivierung von Primärdatensätzen** in Online-Archiven stellt ebenfalls erhöhte Anforderungen an eine vereinheitliche und gleichzeitig strikt anonymisierte Datenaufbereitung.
- Die **Masse an digitalen Daten**, die in hochtechnisierten Gesellschaften in allen Lebensbereichen automatisch generiert werden („big data"), soll verstärkt durch sozialwissenschaftliche Forschung erschlossen werden. Riesige Datenmengen erfordern eine systematische Datenaufbereitung, da es den Forschenden prinzipiell gar nicht mehr möglich ist, einen vollständigen Datensatz mit eigenen Augen zu sichten (Dateninspektion), um auf diese Weise Irregularitäten zu erkennen.

Die wachsende praktische Bedeutung der Datenaufbereitung schlägt sich auch in der **Methodenforschung** nieder: Die Zahl an wissenschaftlichen Beiträgen, die unterschiedliche Verfahren und Aspekte der Datenaufbereitung sowie deren Implikationen behandeln, nimmt zu. So befassten sich mehrere Ausgaben der referierten Open-Access-Zeitschrift *Forum Qualitative Sozialforschung* (FQS) im Zusammenhang mit der Archivierung und Reanalyse qualitativer Daten ausführlich mit ethischen und methodologischen Fragen der Datenaufbereitung (Corti, Witzel, & Bishop, 2005; Eberle & Bergman, 2005). Im Zusammenhang mit der Aufbereitung quantitativer Daten haben sich zu einzelnen Problemen – z. B. adäquater Umgang mit fehlenden Werten im Datensatz – eigene traditionsreiche Forschungsfelder entwickelt (Enders, 2010). Auch zu Spezialfragen wie etwa der Aufbereitung von Scannerdaten (d. h. Daten, die beim Erfassen der Produkt-Barcodes durch Scannerkassen entstehen) für die Konsumentenforschung liegen Methodenvergleichsstudien vor (z. B. Andrews & Currim, 2005).

11.3 Aufbereitung qualitativer Daten

Welche Arbeitsschritte bei der Aufbereitung des Rohdatenmaterials einer qualitativen Studie im Einzelnen durchzuführen sind, hängt u. a. von Inhalt und Art der erhobenen Daten sowie vom Umfang des Forschungsprojekts ab. Nicht zuletzt ergeben sich unterschiedliche Konsequenzen, je nachdem, ob die Daten für eine manuelle Analyse oder – was heute eher der Normalfall ist – für eine elektronische bzw. computergestützte Datenauswertung (▶ Abschn. 12.1.3) präpariert werden sollen. Aus den verschiedenen Empfehlungen, die in der Literatur genannt werden, müssen also jeweils die für die eigene Datenlage passenden herausgesucht werden.

11.3.1 Erstellung und Kommentierung qualitativer Datensätze

Ein großer Aufgabenkomplex bei der Aufbereitung qualitativer Daten betrifft die Erstellung strukturierter Datensätze aus dem oft umfangreichen und heterogenen qualitativen Rohdatenmaterial (LeCompte, 2000, S. 148f.). Dieser Arbeitsschritt ist in der qualitativen Forschung oft aufwändiger als bei quantitativen Studien. So werden z. B. bei einer standardisierten Onlineumfrage alle quantitativen Daten bereits automatisch in ein einheitlich formatiertes Datenfile eingelesen. Demgegenüber produziert eine über mehrere Wochen oder Monate im Team durchgeführte ethnografische Feldstudie (▶ Abschn. 10.1.2) eine Fülle unterschiedlichen Materials (von Feldnotizen über Interviewaufzeichnungen bis zu Fotos), dessen Zusammensetzung und Menge im Vorfeld nicht exakt planbar ist. Das gesammelte qualitative Rohdatenmaterial sollte im ersten Schritt **kopiert** werden. Die weitere Datenaufbereitung anhand der Kopien bzw. Scans (z. B. von handschriftlichen Feldnotizen) bewahrt vor Datenverlust, weil zur Sicherheit die Originale noch vorliegen.

Das kopierte Material muss zunächst sorgfältig **sortiert** werden (z. B. chronologisch, nach Erhebungsorten), sofern dies nicht unmittelbar bei der Datenerhebung geschehen ist. Hierzu wird das Material physisch in Stapel, Boxen oder Ordnern organisiert. Zudem muss sichergestellt werden, dass alle Rohmaterialien vollständig mit **Metadaten** versehen bzw. kommentiert werden (z. B. bei jeder Feldnotiz und jedem Interview sind zusätzliche Informationen über Ort und Zeit der Datenerhebung sowie die an der Erhebung beteiligten Mitglieder des Forschungsteams mitzuliefern). Bei Interviews müssen die **Einwilligungsformulare** der Teilnehmerinnen und Teilnehmer vorliegen. Sofern sich gravierende **Lücken im**

Datenmaterial finden, muss bei Bedarf weiteres Material nacherhoben werden.

Teil des Datenmanagements ist heute meist eine vollständige **Digitalisierung** des Materials (handschriftliche Notizen oder Skizzen werden eingescannt oder digital neu erfasst). Ein Großteil des Materials dürfte ohnehin bereits elektronisch vorliegen (z. B. digitale Fotos von Schauplätzen, digitale Audioaufzeichnungen von Interviews). Im Idealfall sollten die zur Studie gehörenden Dateien in eine möglichst einheitliche Form gebracht werden. So erleichtert es das Arbeiten mit Beobachtungsprotokollen oder Interviewtranskripten, wenn der Kopf des Dokuments jeweils in **Formularform** gestaltet ist und alle wesentlichen Metadaten in wohlstrukturierter Weise enthält. Auch die digitalen Dokumente gilt es wiederum zu sortieren und mit entsprechend **aussagekräftigen Dateinamen** zu versehen (eine entsprechend detaillierte Anleitung mit Beispielen findet sich z. B. in McLellan, MacQueen, & Neidig, 2003). Das Gesamtmaterial sollte schließlich in einer kommentierten **Tabelle** oder **Liste** aller Dateititel bzw. Einzeldokumente zusammengefasst werden.

Die qualitative Datenauswertung kann manuell, mit Bürosoftware oder mit speziellen qualitativen Datenanalyseprogrammen (QDA-Software) durchgeführt werden (▶ Abschn. 12.1.3). Je nach geplanter Vorgehensweise müssen die Datensätze unterschiedlich aufbereitet werden:

- **Vorbereitung für die manuelle Auswertung**: Soll das Material manuell ausgewertet werden, so sind mehrere sortierte und kommentierte Arbeitskopien aller Dokumente auf Papier anzufertigen, damit dann im Zuge der Auswertung großzügig mit Unterstreichungen, Ausschneiden, Umgruppieren etc. gearbeitet werden kann. Diese Form der analogen Auswertung ist jedoch aufgrund ihrer Umständlichkeit weniger empfehlenswert.
- **Vorbereitung für die Auswertung mit Bürosoftware**: Zuweilen wird argumentiert, dass bei kleineren Forschungsprojekten Anschaffung von und Einarbeitung in qualitative Datenanalyse-Tools nicht lohnen, so dass zur Auswertung auf allgemeine Bürosoftware (z. B. MS Word, MS Excel) zurückgegriffen wird (Nideröst, 2002). Dementsprechend müssen die aufbereiteten qualitativen Daten z. B. in eine Word-Datei zusammenkopiert werden. Bei dieser Vorgehensweise verzichtet man auf zahlreiche nützliche Optionen der Datenaufbereitung und -auswertung, wie sie professionelle Analysetools bieten.
- **Vorbereitung für die Auswertung mit QDA-Software**: Der übliche Weg ist die Nutzung von QDA-Software für die qualitative Datenanalyse. Im Zuge der Datenaufbereitung muss also im Vorfeld das geeignete Programm ausgewählt und beschafft werden (zur Programmauswahl ▶ Abschn. 12.1.3). Zudem müssen die aufbereiteten Daten in das Programm eingelesen werden. Der aufbereitete qualitative Datensatz besteht dann z. B. in einer Atlas.ti- oder MaxQDA-Datei, die alle Interviewtranskripte sowie die Angaben aller Befragten aus dem begleitenden Kurzfragebogen enthält.

11.3.2 Umfang, Art und Überprüfung der Transkription

Ein wichtiges Thema der Aufbereitung qualitativer Daten ist die Frage nach Umfang und Art der Verschriftlichung von Video- und Audioaufzeichnungen, die bei Beobachtungen sowie bei Einzelinterviews und Gruppendiskussionen anfallen (Dresing & Pehl, 2010). Eine **Volltranskription** („full transcription"), d. h. eine vollständige Verschriftlichung aller Audio- bzw. Videoaufzeichnungen kann insbesondere im Kontext größerer Feldstudien oder bei einer hohen Zahl an Interviews sehr zeit- und kostenintensiv sein und somit nicht in angemessenem Verhältnis zum Erkenntniswert stehen. Mit entsprechender Begründung ist deswegen auch eine **Teiltranskription** („partial transcription", „semi-transcription") des Materials möglich, d. h. es werden nur die für das Forschungsproblem besonders relevanten Interview- oder Diskussionspassagen wortwörtlich verschriftet und die sonstigen Gesprächspassagen von den Forschenden summarisch in eigenen Worten zusammengefasst. Ein Problem dieser Vorgehensweise besteht jedoch darin, dass oft erst nach der Auswertung klar ist, ob eine Passage für die Forschungsfrage ergiebig oder eher vernachlässigbar ist. Welches Vorgehen gewählt wird, hängt nicht nur von theoretischen und methodologischen Erwägungen ab, sondern auch von den forschungspraktischen Umständen eines Projekts: Bei einer Qualifikationsarbeit, bei der Transkription und Datenanalyse in einer Hand liegen und somit guter Überblick über das Material besteht, kann eine begründete Teiltranskription eine ggf. notwendige Zeitersparnis bieten. Bei arbeitsteiligen Großprojekten wird man eher eine Volltranskription durch externe Transkripteure in Auftrag geben, weil die Selektion der Teilpassagen und das Hineinhören in die Aufzeichnungen für die Forschenden wiederum Mehraufwand generieren würden.

Ist der Umfang der Transkription (also Voll- oder Teiltranskription) entschieden, muss noch das passende **Transkriptionssystem** gewählt und durchgängig umgesetzt werden. Auch hier ist die Entscheidung am Forschungsproblem auszurichten sowie im Einklang mit den zur Verfügung stehenden zeitlichen und finanziellen Ressourcen zu treffen. Eine minutiöse Feinprotokollierung der Dauer von Sprechpausen oder der Stimmmodulation

der Interviewten mag nur in bestimmten Fällen inhaltlich wichtig und später tatsächlich auch Teil der Datenanalyse sein (zur Transkription von qualitativen Interviews ► Abschn. 10.2.2).

Es ist zu beachten, dass mit der Transkription von Audio- oder Videoaufzeichnungen (z. B. von Interviews oder Gruppendiskussionen als dem Rohdatenmaterial der Studie) bereits eine **Datentransformation** vollzogen wird, indem je nach Umfang der Transkription nur Teile des Materials verschriftet und je nach Transkriptionssystem bestimmte Details vernachlässigt oder hervorgehoben werden.

Im Zuge der **Datenbereinigung** von Transkripten müssen Transkriptionsfehler identifiziert und korrigiert werden. Fehler und Ungenauigkeiten in Transkripten werden wesentlich durch die Merkmale der Aufzeichnungen (z. B. schlechte Tonqualität) sowie durch Auswahl, Schulung und Arbeitsbedingungen der Transkripteure beeinflusst. Zur Überprüfung kann man alle oder ausgewählte Transkripte mit der Tonaufzeichnung vergleichen und Abweichungen korrigieren. QDA-Software ermöglicht es, das Transkript synchron zum Ton darstellen zu lassen, so dass Vergleich und Verbesserung komfortabel möglich sind. Eine weitere Methode, um sicherzustellen, dass das aufbereitete Datenmaterial tatsächlich die Erfahrungswirklichkeit widerspiegelt, ist eine Überprüfung durch die Studienteilnehmenden (z. B. Interviewpartner erhalten die Transkripte ihrer Interviews und können Korrekturen und Kommentare anbringen, etwa um Missverständnisse aufzuklären). Hierbei ist jedoch zu beachten, dass Interviewte, die mit Transkriptionssystemen nicht vertraut sind, zuweilen irritiert auf die Datensätze reagieren und der Meinung sind, so würden sie ja gar nicht sprechen. Die Einbeziehung der Untersuchungsteilnehmenden in die Datenaufbereitung bedarf also entsprechender Vorbereitung und Begleitung.

11.3.3 Anonymisierung qualitativer Datensätze

Sehr viel komplexer als in der quantitativen Forschung ist in der qualitativen Forschung die Frage der angemessenen **Anonymisierung** des Rohdatenmaterials. Das Ankreuzmuster in einem standardisierten Fragebogen ist als solches kaum auf eine konkrete Person zurückführbar: Tausende von Menschen können hier ganz ähnliche Antwortmuster aufweisen. Ausführliche Schilderungen im Interview sind dagegen hochgradig individuell. Die geschilderten Ereignisse und ihre Umstände, Wortwahl und Dialekt der Interviewpartnerin, ihr Beruf und Wohnort können zusammengenommen eine Identifizierung sehr leicht machen. Dies ist besonders problematisch, wenn der Primärdatensatz über ein **Datenarchiv** auch Dritten zugänglich gemacht werden soll, so dass hier besondere Standards einzuhalten sind (Corti, Day, & Backhouse, 2000; Corti et al., 2005; Schmidt, 2005).

Identifizierende Informationen vollständig zu entfernen ist schwerer, als es klingt. Zum einen sind damit – je nach Umfang des Materials – teilweise ein immenser **Aufwand** bzw. hohe Kosten verbunden. Zum anderen würde eine konsequente Eliminierung oder Veränderung aller Kontextinformationen (z. B. „Berlin“ wird durch „Großstadt“ oder „München“ ersetzt) die Aussagekraft des Datensatzes reduzieren, denn in der qualitativen Forschung kommt es auf die kontextuelle Einbettung der Geschehnisse an. Möglicherweise ist eben genau der Berlin-Bezug inhaltlich-theoretisch wichtig. Eine zu rigorose Entfernung von Kontextmerkmalen ist also kein gangbarer Weg. Auf der anderen Seite ist schwer einschätzbar, wo die Grenzen zwischen zu viel Anonymisierung **(Verlust wichtiger Informationen)** und zu wenig Anonymisierung (ungewollte Identifizierbarkeit der Interviewpartner) im Einzelfall genau liegt. Welche Details zur Identifizierbarkeit führen könnten, mag im Einzelfall vorab schwer einschätzbar sein. Etwas reduziert wird diese Problematik, wenn die Anonymisierung erst nach der Datenauswertung erfolgt.

Neben dem sorgfältigen und kritischen Umgang mit potenziell identifizierenden Angaben in den Transkripten und sonstigen Dokumenten gehört zum Datenschutz auch die **sichere Verwahrung** (abschließbare Schränke und Räume; passwortgeschützte Computer) sowie die **Vernichtung** des Rohdatenmaterials, da dieses oft hochgradig identifizierbar ist (Audio- und Videoaufzeichnungen). Besondere Vorkehrungen müssen auch getroffen werden, damit die identifizierenden **Einwilligungserklärungen** zwar archiviert, aber eben nicht mit dem Datenmaterial direkt verknüpft werden, da damit Identifizierbarkeit gegeben wäre (z. B. Aufbewahrung in separaten Räumen).

11.4 Aufbereitung quantitativer Daten

Während bei der Aufbereitung qualitativer Daten Transkription und Anonymisierung ein großes Thema sind, stehen bei der Aufbereitung quantitativer Daten die Kodierung, Bereinigung und Transformation der numerischen Variablenwerte im Zentrum. Hinzukommen können Spezialtechniken wie das Auffüllen fehlender Werte im Datensatz durch statistische Imputationsmethoden oder die Angleichung der Zusammensetzung probabilistischer Stichproben an die Population durch statistische Gewichtungsverfahren. All diese Verfahren erfordern eine elektronische Datenverarbeitung.

Während für einfache Auswertungen zuweilen auf **Tabellenkalkulationsprogramme** (z. B. Microsoft Excel) zurückgegriffen wird, werden in der professionellen wissenschaftlichen Forschung **Statistikprogramme** bevorzugt, wobei u. a. SPSS besonders populär ist. Die folgende Darstellung fokussiert deshalb **SPSS**. Die Arbeitsschritte sind jedoch mit anderer Statistiksoftware (z. B. SAS, R, Stata, Statistica, Systat; zur Programmauswahl ▶ Abschn. 12.2.3) meist ganz ähnlich durchführbar. Es ist empfehlenswert, jeweils die neueste Softwareversion zu nutzen, da sich gerade der **Leistungsumfang von Datenqualitätsmaßnahmen** stetig weiterentwickelt (zur Einführung in das Statistikprogrammpaket SPSS Bühl, 2009; Janssen & Laatz, 2009; speziell zur Datenaufbereitung mit SPSS Schendera, 2007).

Aus Gründen der Veranschaulichung wird die Aufbereitung quantitativer Daten im Folgenden v. a. am Beispiel von Daten erläutert, die im Rahmen einer **standardisierten Fragebogenerhebung** gewonnen wurden (▶ Abschn. 10.3.3). Die Prinzipien sind auf alle anderen vollstrukturierten Datenerhebungsverfahren und daraus resultierenden numerischen Rohdaten übertragbar.

Generell gilt: Der **Aufwand für die Aufbereitung der Rohdaten** wächst mit der Menge der Daten (z. B. $n = 200$ Befragungspersonen $\times$ 100 Variablen $= 20\,000$ Datenzellen vs. $n = 2\,000$ Befragungspersonen $\times$ 300 Variablen $= 600\,000$ Datenzellen) sowie mit der Heterogenität des Datenmaterials (z. B. ein Fragebogen vs. Fragebogen plus Beobachtungsprotokoll). Insbesondere hängt der Aufbereitungsaufwand davon ab, wie stark die Daten „verschmutzt" sind.

An dieser Stelle sei noch einmal an die große Bedeutung einer sorgfältigen Untersuchungsplanung im Vorfeld erinnert, die gründliche empirische Pretests einschließen muss, damit sich Probleme bei der Operationalisierung und Datenerhebung eben nicht erst in der Haupterhebung zeigen und in Qualitätsproblemen beim Rohdatenmaterial niederschlagen. Schwächen bei der Instrumentenentwicklung (z. B. schlechte Operationalisierung durch unverständliche Fragebogenitems) sowie Probleme bei der Datenerhebung (z. B. viele unvollständig ausgefüllte Fragebögen aufgrund mangelnder Motivierung oder zeitlich und räumlich ungünstiger Befragungssituation) können beträchtlichen Zusatzaufwand bei der Datenaufbereitung erzeugen und oft im Nachhinein nicht mehr kompensiert werden. Der Aufwand für die Datenaufbereitung reduziert sich teilweise, wenn Daten nicht analog (z. B. Paper-Pencil-Fragebogen, handschriftliches Beobachtungsprotokoll), sondern elektronisch erfasst wurden (z. B. Online-Fragebogen), da hier automatische Plausibilitätschecks eingebaut und z. B. Eingaben außerhalb des zulässigen Wertebereiches verhindert werden können. Die elektronische Datenerfassung kann indessen auch wieder neue Fehlerquellen bergen (für ein Beispiel einer Studienkritik u. a. wegen mangelnder Datenqualität s. Schendera, 2006).

Das Ziel der Datenaufbereitung ist stets die Sicherstellung hoher **Datenqualität**. Diese kann bei quantitativen Daten über eine Reihe von Kriterien spezifiziert werden, dazu zählen v. a. (Schendera, 2007):

1. Vollständigkeit,
2. Einheitlichkeit (z. B. von Datums- und Währungsangaben, Verwendung von Akronymen),
3. Ausschluss doppelter Werte/mehrfacher Datenzeilen,
4. sachgerechte Behandlung von fehlenden Werten,
5. Erkennung und Behandlung von Ausreißerwerten, wobei diese oft auch erst im Rahmen der Datenanalyse erfolgt, sowie
6. Plausibilität der Antwortmuster.

Im Folgenden werden einige praxisorientierte Hinweise zur Aufbereitung quantitativer Daten mit SPSS gegeben, zudem wird auf fortgeschrittene Verfahren und vertiefende Literatur verwiesen.

11.4.1 Erstellung und Kommentierung quantitativer Datensätze

Quantitative Datensätze werden in der Form einer **Datenmatrix** in der Weise dargestellt, dass die einzelnen **Untersuchungseinheiten die Zeilen** und die **Variablen die Spalten** darstellen. Eine Studie mit $n = 500$ Personen, an denen per Fragebogen jeweils 100 Variablen erhoben wurden – was z. B. auch für eine studentische Abschlussarbeit eine realistische Dimension sein kann –, ergibt also eine Datenmatrix mit 500 Zeilen $\times$ 100 Spalten $= 50\,000$ Zellen. In größer angelegten Projekten können Tausende von Untersuchungseinheiten und Hunderte von Variablen in einer einzigen elektronischen Datenmatrix zusammengeführt werden.

In quantitativen Studien bestehen die Daten definitionsgemäß aus **numerischen Messwerten** (meist ergänzt um kurze Texteingaben zu offenen Fragen), die als Ergebnis einer Messung auf verschiedenen Skalenniveaus die unterschiedlichen **Ausprägungen der Variablen** bei den Untersuchungseinheiten repräsentieren sollen (zur Operationalisierung, ▶ Kap. 8). Dabei kann der Messwert „2" z. B. die verschiedensten Bedeutungen haben:

- Er kann als Ausprägung der Variable **Versuchspersonennummer** besagen, dass es sich um die 2. Versuchsperson im Sample handelt.
- Er kann als Ausprägung der **experimentellen unabhängigen Variable** besagen, dass die betreffende Person der 2. Untersuchungsgruppe angehörte und z. B. einen lustigen TV-Clip gesehen hat.

Nehmen Sie bitte zu den folgenden menschlichen Verhaltensweisen Stellung. Was würden Sie sagen:
Inwieweit trifft der jeweilige Satz auf Sie persönlich zu?

Item	Antwortskala und Messwerte							Polung
	trifft überhaupt nicht zu	trifft nicht zu	trifft teilweise nicht zu	teils/ teils	trifft teilweise zu	trifft zu	trifft voll zu	
1. Mein erster Eindruck von Menschen stellt sich gewöhnlich als richtig heraus.	1	2	3	4	5	6	7	P
2. Ich bin mir oft unsicher in meinem Urteil.	1	2	3	4	5	6	7	N
3. Ich weiß immer genau, wieso ich etwas mag.	1	2	3	4	5	6	7	P
4. Ich habe schon mal zuviel Wechselgeld zurückbekommen und nichts gesagt.	1	2	3	4	5	6	7	N
5. Ich bin immer ehrlich zu anderen.	1	2	3	4	5	6	7	P
6. Ich habe gelegentlich mal jemanden ausgenutzt.	1	2	3	4	5	6	7	N

Abb. 11.2 Codeplan eines Fragebogens zur Messung sozialer Erwünschtheit (deutsche Kurzskala auf Basis des „Balanced Inventory of Desirable Responding" [BIDR] von Paulhus, 1991. Copyright 1991, with permission from Elsevier). (Aus Winkler, Kroh, & Spiess (2006). Entwicklung einer deutschen Kurzskala zur zweidimensionalen Messung von sozialer Erwünschtheit. DIW Discussion Papers, 579, 1–37, Abgerufen 4. Oktober 2013 von http://www.diw.de/documents/publikationen/73/diw_01.c.44281.de/dp579.pdf. Mit freundlicher Genehmigung des DIW)

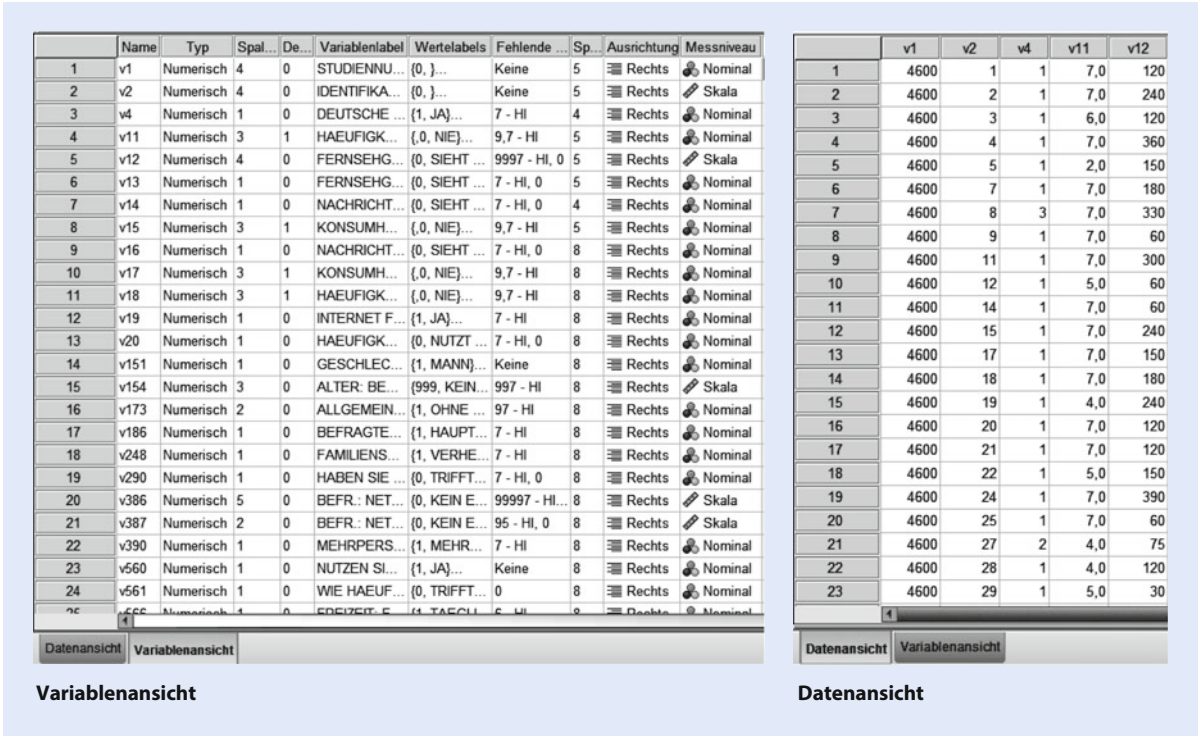
- Er kann als Messwert der **abhängigen Variable** „depressive Stimmung" besagen, dass die Person sich nach dem Anschauen des Filmclips „sehr wenig depressiv" fühlte.
- Er kann als Ausprägung der **soziodemografischen Variable** Kinderzahl besagen, dass die Person 2 Kinder hat.

Welche Bedeutung ein einzelner Messwert „2" einer Variable hat, muss in einem **Codeplan** festgelegt werden. Erfolgte bereits die Datenerhebung elektronisch (z. B. über einen Online-Fragebogen), so wurde der Codeplan schon vor der Datenerhebung erstellt und im elektronischen Fragebogen verankert (z. B. wenn die Befragungsperson bei Item 47 „stimmt gar nicht" anklickt, wird der Wert „4" vergeben), so dass die Daten bereits vercodet vorliegen.

Bei einer nicht-elektronischen Datenerhebung (z. B. Paper-Pencil-Fragebogen) wird der Codeplan zuweilen erst nach der Datenerhebung im Zuge der Datenaufbereitung erstellt. Dazu kann man z. B. in einem Musterfragebogen bei den verschiedenen Antwortoptionen die zu vergebenden Messwerte eintragen. Diese sollten so vergeben werden, dass eine zügige Dateneingabe gefördert wird und man möglichst wenig umdenken muss. Beispielsweise werden Antwortalternativen **in Leserichtung** von links nach rechts fortlaufend ab 1 (oder ab 0, sofern man mit einem Ziffernblock auf der Tastatur arbeitet und die 0 gut erreichbar ist) **durchnummeriert** (Abb. 11.2). Wenn dadurch Werte kontraintuitiv vergeben werden (z. B. stärkere Merkmalsausprägung erhält niedrigeren Zahlenwert), dann ist dies später durch **Umpolung** zu korrigieren.

Der Codeplan enthält alle Variablen mit allen Ausprägungen und den zugeordneten Messwerten. Dabei ist zu beachten, dass im Codeplan einer Fragebogenstudie zusätzlich auch Variablen auftauchen, die keine Fragebogen-Items sind. Dazu gehört die **Fall-** bzw. **Fragebogennummer** durch die der Originalfragebogen mit dem zu generierenden elektronischen Datensatz verknüpft ist. Zudem können Variablen des Untersuchungsdesigns wie z. B. Untersuchungsgruppe, Untersuchungsort, Untersuchungszeit vercodet werden, die sich aus den Umständen des Austeilens des Fragebogens ergeben und nicht direkt von den Untersuchungspersonen abgefragt werden.

Bei elektronischer Datenerhebung per Online-Fragebogen erfolgt die **Vercodung** automatisch. Ebenso bei



	Name	Typ	Spal...	De...	Variablenlabel	Wertelabels	Fehlende ...	Sp...	Ausrichtung	Messniveau
1	v1	Numerisch	4	0	STUDIENNU...	{0, }...	Keine	5	Rechts	Nominal
2	v2	Numerisch	4	0	IDENTIFIKA...	{0, }...	Keine	5	Rechts	Skala
3	v4	Numerisch	1	0	DEUTSCHE ...	{1, JA}...	7 - HI	4	Rechts	Nominal
4	v11	Numerisch	3	1	HAEUFIGK...	{,0, NIE}...	9,7 - HI	5	Rechts	Nominal
5	v12	Numerisch	4	0	FERNSEHG...	{0, SIEHT ...	9997 - HI, 0	5	Rechts	Skala
6	v13	Numerisch	1	0	FERNSEHG...	{0, SIEHT ...	7 - HI, 0	5	Rechts	Nominal
7	v14	Numerisch	1	0	NACHRICHT...	{0, SIEHT ...	7 - HI, 0	4	Rechts	Nominal
8	v15	Numerisch	3	1	KONSUMH...	{,0, NIE}...	9,7 - HI	5	Rechts	Nominal
9	v16	Numerisch	1	0	NACHRICHT...	{0, SIEHT ...	7 - HI, 0	8	Rechts	Nominal
10	v17	Numerisch	3	1	KONSUMH...	{,0, NIE}...	9,7 - HI	8	Rechts	Nominal
11	v18	Numerisch	3	1	HAEUFIGK...	{,0, NIE}...	9,7 - HI	8	Rechts	Nominal
12	v19	Numerisch	1	0	INTERNET F...	{1, JA}...	7 - HI	8	Rechts	Nominal
13	v20	Numerisch	1	0	HAEUFIGK...	{0, NUTZT ...	7 - HI, 0	8	Rechts	Nominal
14	v151	Numerisch	1	0	GESCHLEC...	{1, MANN}...	Keine	8	Rechts	Nominal
15	v154	Numerisch	3	0	ALTER: BE...	{999, KEIN...	997 - HI	8	Rechts	Skala
16	v173	Numerisch	2	0	ALLGEMEIN...	{1, OHNE ...	97 - HI	8	Rechts	Nominal
17	v186	Numerisch	1	0	BEFRAGTE...	{1, HAUPT...	7 - HI	8	Rechts	Nominal
18	v248	Numerisch	1	0	FAMILIENS...	{1, VERHE...	7 - HI	8	Rechts	Nominal
19	v290	Numerisch	1	0	HABEN SIE ...	{0, TRIFFT...	7 - HI, 0	8	Rechts	Nominal
20	v386	Numerisch	5	0	BEFR.: NET...	{0, KEIN E...	99997 - HI...	8	Rechts	Skala
21	v387	Numerisch	2	0	BEFR.: NET...	{0, KEIN E...	95 - HI, 0	8	Rechts	Skala
22	v390	Numerisch	1	0	MEHRPERS...	{1, MEHR...	7 - HI	8	Rechts	Nominal
23	v560	Numerisch	1	0	NUTZEN SI...	{1, JA}...	Keine	8	Rechts	Nominal
24	v561	Numerisch	1	0	WIE HAEUF...	{0, TRIFFT...	0	8	Rechts	Nominal

Datenansicht | Variablenansicht

Variablenansicht

	v1	v2	v4	v11	v12
1	4600	1	1	7,0	120
2	4600	2	1	7,0	240
3	4600	3	1	6,0	120
4	4600	4	1	7,0	360
5	4600	5	1	2,0	150
6	4600	7	1	7,0	180
7	4600	8	3	7,0	330
8	4600	9	1	7,0	60
9	4600	11	1	7,0	300
10	4600	12	1	5,0	60
11	4600	14	1	7,0	60
12	4600	15	1	7,0	240
13	4600	17	1	7,0	150
14	4600	18	1	7,0	180
15	4600	19	1	4,0	240
16	4600	20	1	7,0	120
17	4600	21	1	7,0	120
18	4600	22	1	5,0	150
19	4600	24	1	7,0	390
20	4600	25	1	7,0	60
21	4600	27	2	4,0	75
22	4600	28	1	4,0	120
23	4600	29	1	5,0	30

Datenansicht | Variablenansicht

Datenansicht

Abb. 11.3 Variablenansicht und Datenansicht des ALLBUS-Datensatzes in SPSS

maschinenlesbaren Papierfragebögen, die über entsprechende Scanner eingelesen werden. Bei einfachen Papierfragebögen muss dagegen eine manuelle Dateneingabe in ein Auswertungsprogramm erfolgen.

Bei studentischen Forschungsprojekten werden solche Daten meist von den Forschenden selbst am Computer eingegeben. Bei größeren wissenschaftlichen Projekten wird die Dateneingabe dagegen an Hilfskräfte delegiert. Zudem ist in Teilbereichen der Forschung (z. B. klinische Forschung) die Datenaufbereitung und Datenanalyse oft aus dem Forschungsteam an Statistiker ausgelagert, was zum einen die Forschenden entlastet und zum anderen jegliche Form der unbewussten Beeinflussung der Datenaufbereitung im Sinne der Wunschhypothesen unterbindet. Im Sinne einer **Dreifachverblindung** wissen dann etwa bei einer klinischen Kontrollgruppenstudie, in der es um die Wirksamkeit eines neuen Medikaments geht, weder die Versuchsleiter, noch die Versuchspersonen, noch die mit der Datenaufbereitung und Datenanalyse befassten Statistiker, wer zu welcher Gruppe gehört und welche Ergebnisse wunschgemäß wären (zu Verblindung im Experiment ▶ Abschn. 7.6.1).

Relativ üblich ist es, die Dateneingabe direkt in das zu verwendende **Datenanalyseprogramm** vorzunehmen, also etwa direkt in die von Programmen wie Excel, Stata oder SPSS angebotene Datenmatrix einzutragen und im Programm auch die Vercodung durch die Vergabe von Labels festzuhalten. Für **Variablennamen** sind bei manchen Programmen nur wenige Zeichen vorgesehen, so dass man entweder eine sinnvoll Abkürzung finden muss (z. B. Variablenname: „lern1“ für Lernerfolg bei der 1. Messung) oder die Fragennummer aus dem Fragebogen übernimmt (z. B. Variablenname: „F16a“ für Frage 16a). Bei beiden Varianten wird der kurze Variablenname durch ein längeres und „sprechendes“ **Variablenlabel** ergänzt (Variablenlabel: „Lernerfolg bei der 1. Messung“). Zudem werden für die einzelnen Werte die zugehörigen Bedeutungen durch **Wertelabels** angegeben. Dieses **Labeling** sowie sonstige Festlegungen der Variablenattribute (z. B. Anzahl der dargestellten Dezimalstellen) erfolgt in der **Variablenansicht**, während die Datenmatrix in der **Datenansicht** zu sehen ist (Abb. 11.3).

Wenn die Dateneingabe von Personen übernommen wird, die mit dem bestreffenden Statistikprogramm nicht vertraut sind, wird statt der direkten Eingabe in die Datenmatrix eine Eingabe über eine vorbereitete **Eingabemaske** präferiert (▶ Tipps zur Dateneingabe bei Paper-Pencil-Fragebögen in Statistikprogrammen). Die Eingabemaske ist übersichtlicher und kann Eingabefehler reduzieren, indem z. B. ungültige Werte (also Werte außerhalb des zulässigen Wertebereichs) gar nicht akzeptiert werden, so dass bestimmte Tippfehler bei der Dateneingabe verhindert werden können. Zur Erstellung von Eingabemasken steht z. B. in SPSS ein Zusatzmo-

Tipps zur Dateneingabe bei Paper-Pencil-Fragebögen in Statistikprogrammen

1. Im Zuge der Eingabe des ersten Fragebogens gemäß des Codeplans wird auch die **elektronische Datei vollständig kommentiert und gelabelt**, d. h. die Variablenlabels und Wertelabels werden vollständig eingegeben. Beachten Sie dabei den Wechsel zwischen Datenansicht (Datenmatrix) und Variablenansicht (Variablenlabels, Wertelabels etc. für die Metainformationen zum Datensatz). Das sorgfältige Labeling jeder einzelnen Variable ist wichtig, damit der elektronische Datensatz auch zu späteren Zeitpunkten und auch von anderen Personen ohne Zusatzinformationen nachvollziehbar ist.
2. Die Original-Fragebögen sollte man beim Eingeben durchnummerieren, in aufsteigender Reihenfolge stapeln und archivieren. Die erste Variable im Datensatz ist die **Fall- bzw. Fragebogennummer**. Auf diese Weise werden die Originalfragebögen mit dem Datensatz verknüpft, was für eine spätere Kontrolle und Bereinigung des Datensatzes notwendig ist.
3. **Daten des Untersuchungsdesigns** (die nicht als Variablen im Fragebogen stehen) sollten ggf. als Variablen in den Datensatz aufgenommen werden (z. B. Erhebungsdatum, Erhebungsort, Erhebungsgruppe etc.)
4. Die Eingabe der Variablen erfolgt **gemäß der Variablenreihenfolge im Fragebogen**, um zeitraubendes Blättern oder fehleranfälliges Umdenken zu vermeiden. Numerische Werte für angekreuzte Antworten bzw. Variablenausprägungen werden ebenfalls immer gemäß Reihenfolge im Fragebogen vergeben (am einfachsten beginnend mit 1), z. B. stimmt gar nicht = 1, stimmt wenig = 2, stimmt ziemlich = 3, stimmt völlig = 4; Raucher: nein = 1, ja = 2.
5. Normalerweise werden in der quantitativen Forschung numerische Variablen eingegeben, Texteingaben sind jedoch ergänzend sinnvoll bei offenen und halboffenen Fragen (**Variablentyp: String**).
6. Nach der Vorbereitung des Datenfiles mit allen Variablen- und Wertelabels beginnt die Dateneingabe aus allen Fragebögen. Gehen Sie **langsam und sorgfältig** vor, vermeiden Sie Tippfehler.
7. Wenn im Fragebogen einzelne Fragen nicht beantwortet wurden und somit **fehlende Werte** („missing values“) auftreten, so wird möglichst bereits bei der Dateneingabe ein numerischer Code für die Ursache des fehlenden Wertes eingegeben. Einzelne fehlende Werte eines Falles können bei der Auswertung als Leerstellen behandelt oder im Zuge der Datentransformation durch Imputationswerte (d. h. statistisch erzeugte Ersatzwerte) ausgetauscht werden (zum Umgang mit fehlenden Werten ► Abschn. 11.4.4). Enthält ein einzelner Fragebogen sehr viele fehlende Werte (z. B. fehlende Angaben bei über 50 % der erhobenen Variablen) und/oder wurde er offensichtlich nicht ernsthaft ausgefüllt (z. B. durchgestrichene Seiten und Items, beleidigende Kommentare, „witzige“ Randbemerkungen), so wird der ganze Fall von der Analyse ausgeschlossen. Dementsprechend entstehen dann **fehlende Fälle** („missing cases“). Die Vorgehensweise ist zu dokumentieren (z. B. Anzahl der ausgeschlossenen Fälle und Gründe). Je nach Umfang und Komplexität des Datensatzes sowie Anspruch der Studie werden also pragmatisch oft bereits während der Dateneingabe **erste Maßnahmen der Datenbereinigung** vorgenommen, indem auf die Eingabe eindeutig fragwürdiger Fälle von vorne herein verzichtet wird. Es können aber im Sinne der Transparenz auch zunächst alle Fragebögen eingegeben und erst später einer systematischen Plausiblitäts- und Verweigereranalyse unterzogen werden, bei der dann einzelne Fälle ausgeschlossen werden.
8. Sollen sich **mehrere Personen die Dateneingabe teilen**, ist es empfehlenswert, wenn zunächst ein leerer Datensatz mit allen Labels erstellt und kopiert wird, damit in die identischen Kopien dann jeweils im Zuge paralleler Datenerfassung die Daten zügig eingegeben werden können. Das Zusammenfügen separat erstellter Datensätze in eine gemeinsame Datei ist möglich, jedoch vorsichtshalber nach der Eingabe erster Datensätze zu testen. Daten, die gemeinsam auszuwerten sind, müssen immer in derselben Datei stehen. Eine SPSS-Analyse von mehreren Dateien gleichzeitig ist nicht möglich.
9. Wird ein Fragebogen mehrmals an derselben Stichprobe erhoben (z. B. Vorher-Nachher-Messung; allgemein: **Messwiederholung**), so sind die Daten der 2. (3., 4. etc.) Messung später in der SPSS-Datei der ersten Messung nachzutragen. Um die erste und die zweite sowie die weiteren Messungen einander zuzuordnen, ist bei allen Messung die Erfassung eines anonymen Identifizierungscodes notwendig (z. B. Geburtstag der Mutter plus 5 Ziffern einer wichtigen Telefonnummer ohne Vorwahl: [24-02-1982-35349]. Die Daten der zweiten Messung werden pro Person an das Ende der jeweiligen Zeile im Datensatz angehängt. Jede Messwiederholung wird in einer neuen Variablen erfasst. Beispiel: Variable Lernerfolg und 3 Messungen pro Person = 3 Variablen: lern1, lern2, lern3.
10. Ein in SPSS eingegebener Datensatz wird im **SPSS-Daten-Format** gespeichert. Eine SPSS-Daten-Datei trägt die Dateinamenserweiterung .SAV. Daten, die z. B. in Excel eingegeben wurden, können von SPSS importiert werden. Auch dies sollte im Vorfeld getestet werden, da die Programmversionen kompatibel sein müssen.

dul („data entry module“) zur Verfügung. Zudem existieren spezielle Dateneingabeprogramme (z. B. EpiData: http://www.epidata.dk). Die Eingabemaske kann auch mit Datenbankprogrammen erstellt werden (z. B. MS Access). Anschließend werden die Daten dann in das Statistikprogramm importiert.

Viele standardisierte Fragebögen enthalten einige **offene Fragen**, in denen alle Befragungspersonen um eine Auskunft in eigenen Worten gebeten werden (z. B. in einem Lehrveranstaltungsevaluationsfragebogen: „Was hat Ihnen an der Vorlesung besonders gut gefallen? …“). Zudem werden nicht selten auch **halboffene Fragen**

verwendet, in denen neben vorgegebenen Antwortalternativen (z. B. Familienstand: alleinstehend/Single, feste Partnerschaft, Ehe) eine zusätzliche Antwortkategorie „anderes, und zwar: . . . " eine Erläuterung in eigenen Worten erbittet (zur Operationalisierung des Familienstandes ▶ Abschn. 8.5.1). Diese kurzen Textantworten der Befragten werden vollständig in den Datensatz übernommen (als Stringvariablen). Während Antworten auf halboffene Fragen, bei denen die Kategorie „anderes, und zwar: . . . " präzisiert werden soll, nur vereinzelt abgegeben werden und den Forschenden als Hintergrundinformation dienen, werden die Antworten auf offene Fragen, die sich an alle Befragten richten, einer systematischen Auswertung unterzogen. Dazu werden sie kodiert, was meist auf eine induktive Gruppierung der Antworten zu einer überschaubaren Anzahl zusammenfassender Kategorien hinausläuft. Die kodierten Antworten können als neue Variable in den Datensatz eingefügt werden, wobei jede Antwortkategorie dann einen numerischen Wert erhält. Wenn die offenen Fragen für die inhaltliche Auswertung sehr wichtig und nicht nur randständig sind, sollte die Kodierung mittels Intercoder-Reliabilität geprüft werden (zu „Quantitative Inhaltsanalyse" ▶ Abschn. 10.6.3).

11.4.2 Anonymisierung quantitativer Datensätze

Bei quantitativen Datensätzen ist durch die standardisierten Antworten die Anonymisierung von vorne herein eher gegeben als bei qualitativen Daten. Problematisch im Sinne der Identifizierbarkeit können bei quantitativen Daten v. a. **sehr seltene Merkmale** oder **Merkmalskombinationen** sein, wenn die Zielgruppe bekannt ist (z. B. Identifizierbarkeit des einzigen Rollstuhlfahrers oder der einzigen Asiatin im Datensatz). Entsprechend identifizierbare Daten sollten möglichst von vorne herein nicht erhoben oder dann im Zuge der Datenaufbereitung eliminiert werden. Bei elektronisch erfassten Datensätzen ist zwecks Anonymisierung darauf zu achten, dass **IP-Adressen der Rechner**, von denen auf den Onlinefragebogen zugegriffen wurde, nicht protokolliert und gespeichert werden.

Über sozialwissenschaftliche **Datenarchive** können die eigenen Primärdaten der Scientific Community für Sekundäranalysen (▶ Abschn. 7.4) zur Verfügung gestellt werden. Dies setzt voraus, dass bei der Datenaufbereitung auch **Datenarchiv-Standards** erfüllt werden, dabei spielen u. a. die Sicherstellung von Anonymität und Datenschutz eine wichtige Rolle. Teilweise bieten die Datenarchive selbst entsprechend Aufbereitungsleistungen an (z. B. GESIS mit ihrem Datenbestandskatalog; http://www.gesis.org). Zudem publizieren die Archive ihre Standards und geben Hinweise zur optimalen Datenaufbereitung über den gesamten Datenlebenszyklus hinweg (z. B. Inter-University Consortium for Political and Social Research, 2009; http://www.icpsr.umich.edu).

11.4.3 Datenbereinigung bei quantitativen Datensätzen

Bei jeder noch so sorgfältigen Dateneingabe schleichen sich Tippfehler und sonstige Irregularitäten ein. Ein kleiner Tippfehler (z. B. „55" statt „5") kann die Ergebnisse der späteren statistischen Datenanalyse stark verzerren. Deswegen ist eine quantitative Datenanalyse erst dann sinnvoll, wenn der Datensatz zuvor gründlich von Fehlern bereinigt wurde. Im Zuge der Datenbereinigung werden für jede Variable zumindest der Wertebereich, die Häufigkeitsverteilung und die vergebenen Wertelabels geprüft. Differenziertere Methoden der Datenbereinigung sind möglich (Schendera, 2007) und insbesondere indiziert bei sehr umfassenden Datensätzen, die aus mehreren Datenquellen zusammengetragen wurden, so dass z. B. Probleme wie Dopplungen oder unterschiedliche nationale Datumskodierungen auftreten können (etwa im Rahmen des „Data Mining"; ▶ Abschn. 12.3.3).

- **Prüfung der Wertebereiche:** Anhand der Häufigkeitsverteilungen der einzelnen Variablen ist ersichtlich, ob alle Daten im zugelassenen Wertebereich bleiben. Wird ein fragwürdiger Wert entdeckt (z. B. Wert „55" bei einer Variable, die nur Ausprägungen von 1–5 haben darf) oder eine unplausible Antwort bei einer offenen Frage (z. B. „111" Kinder), so geht man in der entsprechenden Zeile in der Datenmatrix ganz nach links zur ersten Variable, nämlich der Fallnummer. Hier ist z. B. bei einer Paper-Pencil-Fragebogenerhebung dann die Fragebogennummer zu entnehmen. Der entsprechende Fragebogen wird im sortierten und durchnummerierten Fragebogenstapel gesucht und der Wert in der Datenmatrix gemäß der Antwort auf dem Originalbogen korrigiert. Bei vielen inhaltlichen Items ohne beschränkte Zahl an Antwortalternativen kann es schwierig sein zu entscheiden, ob ein angegebener Wert unplausibel bzw. ein Ausreißer ist, oder ob er noch zu den typischen Werten dazugehört. Beispiel: „Wie viele Personen mit Doktortitel haben Sie in Ihrem Bekanntenkreis?" Ist hier die Antwort „100" ein Ausreißer oder für bestimmte Teilpopulationen vielleicht ein durchaus plausibler Wert? Zur Prüfung des Wertebereiches kann die Häufigkeitsverteilung hinzugezogen werden (z. B. Gibt es mehrere Personen, die „100" oder ähnliche Werte angegeben haben? Damit würde der Wert plausibler). Zudem sind theoretische Überlegungen zu Ausprägung und Verteilung des Merkmals in der Zielpopu-

lation anzustellen, aus denen Plausibilitätsannahmen abgeleitet werden können.

- **Prüfung der Häufigkeitsverteilungen**: Wenn alle Variablen zwecks Überprüfung der Wertebereiche einzeln durchgegangen werden, ist es sinnvoll, gleichzeitig auch die Form der Häufigkeitsverteilungen im Auge zu behalten (d. h. Wie oft kommen die einzelnen Werte innerhalb des Wertebereiches vor?). Variablen mit extremen Verteilungsformen (z. B. ja: 1 %, nein: 99 %) sollte man sich notieren. Variablen mit **extrem verzerrten Verteilungen** können zu verzerrten Ergebnissen bei der statistischen Auswertung führen (z. B. bei der Berechnung von Korrelationen). Generell sollten bei der Datenanalyse jedoch vor der Durchführung jedes einzelnen statistischen Signifikanztests jeweils die Häufigkeitsverteilungen und sonstigen statistischen Voraussetzungen des jeweiligen Tests geprüft werden. Die **statistische Voraussetzungsprüfung** für einzelne Signifikanztests (z. B. Ausreißeranalyse bei Korrelationsanalysen, Varianzhomogenität bei Varianzanalysen, Multikollinearität bei multiplen Regressionen) ist Teil der quantitativen Datenanalyse (► Kap. 12) und nicht der Datenaufbereitung.
- **Prüfung der Wertelabels**: Anhand der Häufigkeitsausgabe lässt sich nicht zuletzt auch noch einmal in Augenschein nehmen, ob alle Wertelabels fehlerfrei und allgemeinverständlich eingegeben wurden. Ungünstige Formulierungen oder Tippfehler sind entsprechend in der Variablenansicht zu korrigieren.

Sehr unvollständig oder offensichtlich nicht ernsthaft ausgefüllte Fragebögen fallen in der Regel beim manuellen Eintippen der Daten sofort auf. Entsprechende Fälle können dann gleich ausgeschlossen werden. Bei elektronischer Datenerhebung muss indessen im Nachgang für **jede einzelne Untersuchungseinheit** geprüft werden, ob **genügend plausible Werte** vorliegen, um Aussagen über diesen Fall treffen zu können, oder ob der gesamte Fall von den weiteren Analysen ausgeschlossen werden soll. Die Anzahl und die Gründe für den Ausschluss von Fällen sind im Methodenteil der Studie zu dokumentieren. Ein geringer Prozentsatz an unvollständig oder nicht ernsthaft bzw. unplausibel ausgefüllten Fragebögen ist typisch. Werden die Daten elektronisch z. B. über Online-Fragebögen erhoben, so kann ein Teil der Datenbereinigung vorweggenommen werden, weil dann z. B. die Eingabe von unzulässigen Werten von vorne herein unterbunden und auch schon automatische **Plausibilitätschecks** durchgeführt werden können (z. B. Analyse auf stereotype Antwortmuster: z. B. Mittelkreuzer, die immer die mittlere Kategorie ankreuzen oder Extremkreuzer, die v. a. die Randkategorien ankreuzen; zudem kann die Zeit für das Ausfüllen des Fragebogens erfasst werden, die ebenfalls ein Indikator für ernsthafte Beantwortung ist).

11.4.4 Datentransformation bei quantitativen Datensätzen

Es ist empfehlenswert, nach Abschluss von Dateneingabe, Kodierung der offenen Fragen und Datenbereinigung den **Masterdatensatz** zu speichern und für die Datentransformation mit einer **Arbeitskopie** weiterzuarbeiten. Die folgenden Schritte umfassen nämlich mehr oder minder weitreichende Transformationen des Rohdatenmaterials, für die teilweise mehrere unterschiedliche Ansätze existieren. Sollen später verschiedene Varianten der Datenaufbereitung verglichen werden oder die Daten für Sekundäranalysen verwendet werden, ist ein Rückgriff auf den Originaldatensatz hilfreich.

Welche weiteren Transformationsschritte an der Arbeitskopie des Datensatzes im Einzelnen vorgenommen werden, sollte möglichst vollständig dokumentiert werden. Dazu besteht in SPSS und anderen Statistikprogrammen (► Abschn. 12.2.3) die Möglichkeit, sich die im Menü ausgewählten Transformationsoptionen auch als **Befehlssyntax** ausgeben zu lassen (jeweils über den Button „Einfügen“ bei den Menüoptionen). Somit können dann die einzelnen Transformationsschritte in Syntaxform im Anhang einer Abschlussarbeit dokumentiert werden. Damit ist der Prozess der Datenaufbereitung transparent und kann von Außenstehenden nachvollzogen und beurteilt werden. Die Dokumentation der Befehlssyntax ist auch bei der späteren Datenanalyse empfehlenswert. Syntaxbefehle nicht nur aus den Menüs heraus zu speichern, sondern selbst zu erstellen, hat für fortgeschrittene Nutzerinnen und Nutzer von Statistikprogrammen wie SPSS den Vorteil, dass sie ihre Analysen noch differenzierter steuern können, weil nicht alle Optionen über Menüs auszuwählen sind (zur Syntaxsteuerung von SPSS Schendera, 2005).

Behandlung fehlender Werte

Fehlende Werte („missing data“) können aus unterschiedlichen Gründen entstehen: Befragungspersonen können ein Fragebogen-Item übersehen, eine heikle Frage (z. B. nach dem Einkommen) bewusst überspringen, absichtlich oder versehentlich eine ungültige Antwort abgeben, bei einem Messzeitpunkt einer Längsschnittstudie nicht erreichbar sein etc. Zudem kann es z. B. im Rahmen von Beobachtungen zu Fehleingaben des Beobachters in den Beobachtungsbogen kommen oder im Rahmen physiologischer Messungen zu Fehlwerten aufgrund mangelnder Kalibrierung des Messgerätes oder äußerer Störeinflüsse. Datenaufbereitung bedeutet zunächst, dass

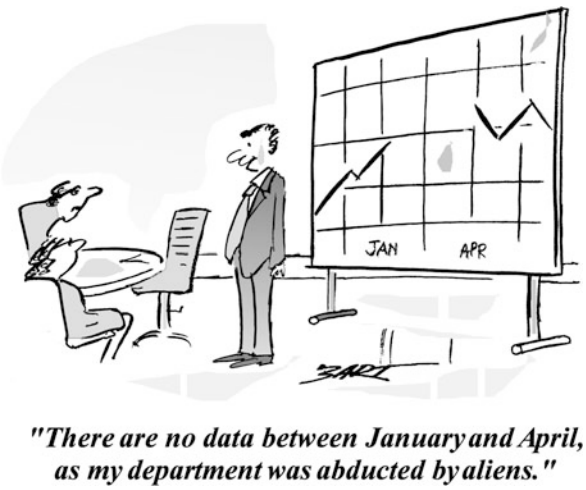


Abb. 11.4 Die Gründe für fehlende Werte sollten bekannt sein und – sofern sie für die Auswertung oder Interpretation der Daten relevant sind – auch vercodet werden. (© BART/Search ID: bron1954, Rights Available from CartoonStock.com)

fehlende Werte als solche **kenntlich gemacht** werden, wobei gleichzeitig auch der **Grund für den fehlenden Wert** vercodet werden kann (Abb. 11.4).

Fehlende Werte sollten in der Datenmatrix möglichst nicht als echte **Leerstellen** auftreten, sondern idealerweise ausschließlich mit **Codes für Missings** (z. B. –99, –77) gekennzeichnet werden. So kann bei Fragebögen z. B. differenziert werden, ob eine Lücke im Datensatz einer Person vorliegt, weil sie a) aufgrund der Filterführung die Frage gar nicht vorgelegt bekommen hat, b) zum betreffenden Sachverhalt ausdrücklich keine Meinung hat, c) zu einer persönlichen Frage die Angabe verweigert etc. Eine entsprechend differenzierte Kodierung fehlender Werte ist jedoch meist nur sinnvoll, wenn diese dann auch bei der Bewertung der Datenqualität und der späteren Auswertung berücksichtigt werden. Dringend abzuraten ist davon, fehlende Werte mit Werten innerhalb eines plausiblen Wertebereiches zu kodieren (z. B. mit Null), da diese Werte, wenn sie versehentlich in die Analyse einbezogen werden, zu Verzerrungen führen werden.

Einzelne fehlende Werte im Datensatz sind meist unproblematisch. Heikel ist es jedoch, wenn **viele Missings** auftreten und wenn diese **systematisch gehäuft** bei bestimmten Personen(gruppen) oder Variablen vorkommen. Per Standardeinstellung gehen Statistikprogramme meist so vor, dass sie fehlende Werte fallweise oder variablenweise von der Analyse ausschließen. **Ausschlussverfahren** führen jedoch nicht selten unbemerkt zu Verzerrungen in den Ergebnissen, insbesondere wenn fehlende Werte bei einer Teilpopulation systematisch auftreten.

In den letzten Jahrzehnten wurden deswegen eine Reihe von statistischen Methoden entwickelt, die darauf abzielen, fehlende Werte nicht einfach auszuschließen, sondern anhand der vorhandenen Daten sowie anhand von Informationen über den Ausfallprozess differenziert zu behandeln. Als „State-of-the-Art" werden heute vor allem zwei Ansätze verfolgt (Enders, 2010):

1. **Imputationsbasierte Verfahren**: Hier werden im einfachsten Fall die fehlenden Werte durch jeweils **einen** Schätzwert **(„single imputation")** ersetzt, etwa den Mittelwert oder einen durch die multiple Regression geschätzten Wert. Aktuelle Ansätze favorisieren die **multiple Imputation**, d. h. für jeden fehlenden Wert werden mehrere plausible Schätzwerte bestimmt und kombiniert zum Auffüllen der Lücken im Datensatz verwendet. Dass für fehlende Werte statistisch erzeugte Ersatzwerte in den Datensatz eingesetzt (imputiert) werden können, mag auf den ersten Blick wie Datenmanipulation wirken. Tatsächlich beeinflusst das gewählte Imputationsverfahren die Ergebnisse. Die Ergebnisse werden indessen auch stark beeinflusst und ggf. verfälscht, wenn fehlende Werte einfach pauschal ausgeschlossen werden. Deswegen ist es notwendig – insbesondere wenn viele Missings auftreten – sich mit der Missing-Data-Problematik und den verschiedenen Imputationsmethoden vertraut zu machen, um mit entsprechender Begründung das geeignete Verfahren auswählen zu können (für entsprechende Hinweise s. Allison, 2001; Little & Rubin, 2002; Schafer, 1997; Schafer & Graham, 2002; Wirtz, 2004). Im Ergebnis liegt nach der Anwendung imputationsbasierter Verfahren ein vollständiger Datensatz vor, der dann mit herkömmlichen Verfahren statistisch ausgewertet werden kann. SPSS bietet im Menü „ Transformieren" die Funktion „Fehlende Werte ersetzen" („single imputation" durch Mittelwert oder Regression) sowie im Menü „Analysieren" die Option „Multiple Imputation".
2. **Modellbasierte Verfahren**. Bei modellbasierten Verfahren, die mit **Strukturgleichungsmodellen** (▶ Kap. 17) realisiert werden, erfolgen die Behandlung der fehlenden Werte und die Datenauswertung bzw. Modellanpassung in einem Schritt. Bei den modellbasierten Methoden werden heute die auf **Maximum-Likelihood-Schätzungen** basierenden Techniken favorisiert (Enders, 2010). Werden Missing-Data-Probleme bereits bei der Untersuchungsplanung antizipiert, empfiehlt es sich, über entsprechende Zusatz-Items die Gründe für Ausfälle genauer zu erheben (z. B. keine Meinung, keine Antwortbereitschaft) und nach Möglichkeit in die Berechnung einzubeziehen.

Tabelle 11.1 Umpolung einer Antwortskala zu dem Item „Wie oft fühlen Sie sich bei der Arbeit gestresst?“

	sehr oft	oft	manch-mal	selten	nie
Kodierung laut Dateneingabe (alte Werte)	1	2	3	4	5
Umpolung für die Datenauswertung (neue Werte)	5	4	3	2	1

Umpolung von Variablen

Wir empfehlen, den Merkmalsausprägungen von ordinal- sowie intervall- und verhältnisskalierten Variablen (zum Skalenniveau ▶ Abschn. 8.4) Messwerte in der Weise zuzuordnen, dass höhere Werte jeweils einer stärkeren Merkmalsausprägung entsprechen. Auf diese Weise lassen sich spätere Analyseergebnisse meist intuitiv verstehen. Stellt sich bei der Datenbereinigung heraus, dass einige Variablenausprägungen gemäß ihrer Reihenfolge im Fragebogen zwecks Vereinfachung der Dateneingabe zunächst kontraintuitiv kodiert wurden, so kann man sie im Nachgang problemlos umpolen.

Wird z. B. die Häufigkeit von „Stresserleben bei der Arbeit“ auf einer 5-stufigen Ratingskala in der Reihenfolge „sehr oft – oft – manchmal – selten – nie“- abgefragt und werden in dieser Reihenfolge bei der manuellen Dateneingabe auch die Codes vergeben (z. B. von sehr oft: 1 bis nie: 5), so ist dies für die Dateneingabe das optimale, weil schnellste und fehlerärmste Vorgehen. Diese Kodierung ist jedoch für die Datenanalyse suboptimal, denn jetzt wird geringere Stressbelastung kontraintuitiv durch höhere Werte ausgedrückt. Würde bei der späteren Auswertung z. B. ein Gruppenvergleich vorgenommen und dieser grafisch dargestellt, so hätte die Gruppe mit dem höheren Stresserleben den niedrigeren Balken im Balkendiagramm. Eine Umpolung würde dies vermeiden, so dass der höhere Balken dann auch dem höheren bzw. häufigeren Stresserleben entspricht. Bei einer Umpolung würden dementsprechend bei allen Untersuchungseinheiten die Werte ausgetauscht (Tab. 11.1), was z. B. im Statistikprogramm SPSS im Menü „Transformieren“ unter dem Befehl „Umkodieren in neue Variable“ möglich ist. Wir empfehlen die Umkodierung in eine **neue Variable** (anstelle des ebenfalls möglichen Überschreibens der alten Variablen). Zu beachten ist, dass für die neue umgepolte Variable auch die **Wertelabels neu zu vergeben** sind.

Sehr wichtig ist die Umpolung, wenn für die Messung eines latenten Konstrukts mehrere Indikatoren verwendet werden, die inhaltlich teils in positive und teils in negative Richtung weisen. Hier muss vor der Berechnung eines Index (▶ Abschn. 8.7) oder einer psychometrischen Skala (▶ Abschn. 8.6) eine Umpolung aller negativen Items erfolgen (Tab. 11.2).

Bildung von neuen Variablen

Wenn die für die Studie relevanten latenten Konstrukte mittels multipler Indikatoren operationalisiert wurden, müssen im Rahmen der Datenaufbereitung die entsprechenden **Skalenwerte** oder **Indexwerte** berechnet werden (▶ Abschn. 8.6 und ▶ Abschn. 8.7). Hierfür ist im Vorfeld die korrekte Polung aller beteiligten Items notwendig. Zudem muss mit fehlenden Werten korrekt umgegangen werden, da sich andernfalls Verzerrungen ergeben können. So müssen Durchschnittswerte in der Weise gebildet werden, dass die Summe der Skalenwerte der Items nicht durch die Anzahl aller Items pro Skala, sondern jeweils durch die Anzahl aller von der jeweiligen Person beantworteten Items dividiert wird (SPSS-Menü „Transformieren“: Option „Variable berechnen“). Wenn z. B. 4 Motivationsitems (mot_A, mot_B, mot_C, mot_D) aus dem Datensatz zu einer Motivationsskala zusammengefasst werden sollen und als Skalenwert der Durchschnittswert dient, so ist zu berechnen: „compute mot_ges = mean (mot_A, mot_B, mot_C, mot_D)“. Demgegenüber würde ein Befehl wie „compute mot_ges = (mot_A + mot_B + mot_C + mot_D)/4“ zu falschen Ergebnissen führen, wenn der Datensatz Missings enthält. Ob es testtheoretisch gerechtfertigt ist, mehrere Items zu einer Skala zusammenzufassen, muss anhand einer Analyse entsprechender Gütekriterien (z. B. Reliabilität, ▶ Abschn. 10.4.1) geprüft werden.

Zur Bildung neuer Variablen kann auch die Transformation einer stetigen oder sehr fein abgestuften Variable (z. B. Reaktionszeiten, Lebensalter, Einkommen) in eine diskrete Variable mit wenigen Kategorien gehören (zu Skalentransformationen ▶ Abschn. 8.4.6). SPSS bietet eine Vielzahl an Funktionen, um neue Variablen zu bilden. Diese müssen im Zuge der Pflege des Datensatzes dann wiederum auch mit Variablen- und Wertelabels versehen werden. Wie bereits erwähnt, empfehlen wir, die SPSS-Syntax, die zur Bildung neuer Variablen führt, entsprechend zu dokumentieren.

Gewichtung von Stichproben

Ebenso wie es möglich ist, fehlende Werte im Datensatz im Nachgang mittels Imputationsverfahren durch Schätzwerte zu ersetzen, kann eine im Hinblick auf die Population verzerrte Stichprobe im Nachhinein durch Gewichtungsverfahren korrigiert werden. In beiden Fällen handelt es sich um einen legitimen Ansatz, der darauf abzielt, mittels Vervielfachung von vorhandenen Daten Angleichungen an die empirische Realität zu modellieren. Denn fehlende Werte und fehlende Fälle kommen in

Tabelle 11.2 Umpolung der negativen Items einer Skala (Auszug aus der Rosenberg-Skala für Selbstwertgefühl). (Aus Rosenberg, 1965)

Self-Esteem Items	Strongly disagree	Disagree	Agree	Strongly agree
1. I feel that I have a number of good qualities. (Positive)	0	1	2	3
2. At times I think I am no good at all. (Negative)	~~0~~ 3	~~1~~ 2	~~2~~ 1	~~3~~ 0
3. I certainly feel useless at times. (Negative)	~~0~~ 3	~~1~~ 2	~~2~~ 1	~~3~~ 0
4. On the whole, I am satisfied with myself. (Positive)	0	1	2	3

allen Studien vor, die mit Untersuchungspersonen (und nicht z. B. mit vorgefundenen Dokumenten) arbeiten, da im Sinne der Forschungsethik (► Abschn. 4.1) die Teilnahme an jeder Studie freiwillig ist. Somit ist mehr oder minder beträchtliche Teilnahme- und Auskunftsverweigerung – je nach Thema und Zielgruppe – in Kauf zu nehmen, was auf verzerrte Stichproben hinausläuft. Statistische Gewichtungsmethoden bieten hier – zumindest in gewissem Rahmen (eine ausreichende Datengrundlage ist notwendig, um diese sinnvoll hochgewichten zu können) – die Möglichkeit zum Ausgleich.

Populationsbeschreibende Studien, die auf zufälligen Stichproben basieren und eine möglichst präzise Abbildung der Populationsverhältnisse anstreben (► Abschn. 9.3.2), sind für den Einsatz von Gewichtungsverfahren prädestiniert. Das Ziel der Gewichtungsverfahren besteht darin, die Verteilung in der Population mithilfe von **personenspezifischen Gewichten** möglichst präzise in der Stichprobe nachzubilden. Arbeitet man mit nicht-probabilistischen Stichproben, deren Repräsentativität für eine Zielpopulation ohnehin offen ist (► Abschn. 9.3.1), erscheinen auch Gewichtungsverfahren nicht sinnvoll. Gewichtungen werden v. a. in der Umfrageforschung eingesetzt.

Das Grundprinzip aller Gewichtungsfaktoren läuft darauf hinaus, dass Fälle, die – im Vergleich zur Population – zu gering in der Stichprobe vertreten sind, eine höheres Gewicht erhalten und gleichzeitig Fälle, die in der Stichprobe überrepräsentiert sind, heruntergewichtet werden. Hierzu sind Daten über die Zusammensetzung der interessierenden Zielpopulation notwendig. Die Stichprobe wird anhand ihrer einzelnen Fälle so gewichtet, dass sie in ihrer Zusammensetzung der Zielpopulation besser entspricht. Je nach Datenlage und intendierter Aussage kann zwischen **unterschiedlichen Gewichtungsverfahren** gewählt werden, die in der Fachliteratur ausführlich diskutiert werden (Moore, Pedlow, Krishnamurty, & Wolter, 2000; Pike, 2007, 2008). Das **SPSS-Modul „Complex Samples“** bietet eine Reihe von Gewichtungsmöglichkeiten, um auf der Basis unterschiedlicher und auch mehrstufiger Zufallsstichproben zu möglichst genauen Parameterschätzungen zu kommen.

Lernquiz

1. Nennen Sie drei Funktionen, die eine Datenaufbereitung zwecks Steigerung der Datenqualität erfüllt!
2. Warum hat die Datenaufbereitung eine zunehmend größere Bedeutung in der sozialwissenschaftlichen Forschung?
3. Wie können Fehler in Interview-Transkripten entdeckt werden?
4. Welche Probleme können bei der Anonymisierung von Transskripten auftreten?
5. Wie sind Messwerte außerhalb des zulässigen Wertebereichs im Zuge der Datenaufbereitung zu verhindern bzw. zu erkennen?
6. Nennen Sie die sechs Kriterien für gute Datenqualität bei quantitativen Daten!
7. Was versteht man unter einem Masterdatensatz?
8. Nennen Sie drei Möglichkeiten, fehlende Werte in einem quantitativen Datensatz zu behandeln!
9. Was soll durch eine Umpolung von Variablen erreicht werden?
10. Erläutern Sie die Funktion von Gewichtungsverfahren und wann man sie einsetzt!

Übungsaufgaben

1. Eine Befragungsperson hat die in Abb. 11.2 dargestellte Kurzskala zur Messung sozialer Erwünschtheit folgendermaßen beantwortet: Item 6: trifft überhaupt nicht zu, Item 2: trifft nicht zu, Item 5: trifft teilweise zu, Item 1 und Item 3: trifft zu, Item 4: keine Angabe. Geben Sie die Daten in SPSS ein, labeln Sie den Datensatz, polen Sie die Variablen um und bestimmen Sie den Skalenwert

der Person. Warum hat die Befragungsperson Item 4 unbeantwortet gelassen, was vermuten Sie?

2. Laden Sie sich den SPSS-Datensatz zum ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) 2012 in der kompakten Version (Studienname: ALLBUScompact-2012_ZA4615) herunter über http://www.gesis.org/allbus/datenzugang/. Von $n = 3\,480$ Befragungspersonen haben laut Variable V248 426 Personen die Frage „Sind Sie im öffentlichen Dienst tätig" mit „Ja" beantwortet. Das sind rechnerisch $426/3\,480 = 12.2\,\%$. Doch diese Prozentangabe ist verzerrt, wenn sich im Datensatz Personen befinden, auf die die Frage gar nicht anwendbar bzw. zutreffend ist (z. B. weil sie nicht erwerbstätig sind) oder die die Frage nicht beantwortet haben. Wie wurden diese beiden Varianten von Missings (Frage trifft nicht zu; Frage trifft zu, wurde aber nicht beantwortet) im Datensatz kodiert? Wie lautet der gültige Prozentsatz der im öffentlichen Dienst Tätigen in der Stichprobe, wenn man die beiden Arten von Missings nicht in die Berechnung einbezieht?
3. Betrachten Sie den in der vorigen Übungsaufgabe genutzten ALLBUS-Datensatz (Studienname: ALLBUScompact-2012_ZA4615) genauer. Er basiert auf einer zweistufigen Zufallsstichprobe der Wohnbevölkerung in Deutschland. Die Daten wurden mit vollstrukturierten mündlichen Interviews erhoben. Nennen Sie drei Variablen im Datensatz, die sich nicht auf die Interview-Fragen beziehen, sondern als Zusatzinformationen dienen. Wurden Missings im Datensatz teilweise durch Imputationswerte ersetzt? Wurde der Datensatz zur Steigerung der Repräsentativität im Nachhinein statistisch gewichtet?

Weiterführende Literatur

- Inter-University Consortium for Political and Social Research (2009). *Guide to social science data preparation and archiving, best practice throughout the data life cycle*. Abgerufen 4. September 2013 von http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf
- Schendera, C. F. G. (2007). *Datenqualität mit SPSS*. München: Oldenbourg.
- Schleicher, A. & Saito, M. (2005). Module 10 – Data preparation and management. In K. N. Ross (Ed.), *Quantitative research methods in educational planning*. (pp. 1–79). Paris: International Institute for Educational Planning/UNESCO. Retrieved September 9, 2013, from http://www.iiep.unesco.org/capacity-development/training/training-materials/quantitative-research.html.
- Taylor, C. & Gibbs, G. R. (2010). *Preparing Data. Online QDA Web Site*. Retrieved September 9, 2013, from http://onlineqda.hud.ac.uk/Intro_QDA/preparing_data.php.

Literatur

Allison, P. D. (2001). *Missing Data*. Thousand Oaks: Sage.

Andrews, R. L. & Currim, I. S. (2005). An experimental investigation of scanner data preparation strategies for costumer choice models. *International Journal of Research in Marketing, 22*(3), 319–331.

Bühl, A. (2009). *PASW 18 (ehemals SPSS): Einführung in die moderne Datenanalyse* (Vol. 12). München: Pearson.

Corti, L., Day, A., & Backhouse, G. (2000). Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Forum Qualitative Social Research, 1*(3), Art. 7. Retrieved October 4, 2013, from http://www.qualitative-research.net/index.php/fqs/article/view/1024/2207

Corti, L., Witzel, A., & Bishop, L. (2005). Potenziale und Probleme der Sekundäranalyse. Eine Einführung in die FQS-Schwerpunktausgabe über die Sekundäranalyse qualitativer Daten. *Forum Qualitative Social Research, 6*(1), Art. 49. Abgerufen 4. September 2013 von http://www.qualitativeresearch.net/index.php/fqs/article/view/498

Dresing, T. & Pehl, T. (2010). *Transkription*. Wiesbaden: VS Verlag.

Eberle, T. S. & Bergman, M. M. (2005). Introduction. *Forum Qualitative Social Research, 6*(2), Art. 30. Retrieved September 4, 2013, from http://www.qualitativeresearch.net/index.php/fqs/article/view/453

Enders, C. K. (2010). *Applied missing data analysis (Methodology in the social sciences)*. New York: Guilford.

Inter-University Consortium for Political and Social Research, I. (2009). Guide to social science data preparation and archiving, best practice throughout the data life cycle. Retrieved September 4, 2013, from http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Janssen, J. & Laatz, W. (2009). *Statistische Datenanalyse mit SPSS für Windows: Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests* (Vol. 7). Berlin: Springer.

LeCompte, M. (2000). Analyzing qualitative data. *Theory Into Practice, 39*(3), 146–154.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: NJ: Wiley.

McLellan, E., MacQueen, K. M., & Neidig, J. L. (2003). Beyond the qualitative interview: Data preparation and transcription. *Field Methods, 15*(1), 63–84.

Moore, W., Pedlow, S., Krishnamurty, P., & Wolter, K. (2000). *National longitudinal survey of youth 1997 (NLSY97)*. Chicago: NORC.

Nideröst, B. (2002). Die technikunterstützte Analyse von qualitativen Daten mit Word. *Forum: Qualitative Social Research, 3*(2), Art. 22. Aufgerufen 4. September, 2013, unter http://www.qualitativeresearch.net/index.php/fqs/article/view/861

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). New York: Academic Press.

Pike, G. R. (2007). Adjusting for nonresponse in surveys. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 411–449). Dordrecht, NL: Springer.

Pike, G. R. (2008). Using weighting adjustments to compensate for survey nonresponse. *Research in Higher Education, 49*(2), 153–171.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton: Princeton University Press.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Schendera, C. F. G. (2005). *Datenmanagement mit SPSS: Kontrollierter und beschleunigter Umgang mit Datensätzen, Texten und Werten*. Heidelberg: Springer.

Schendera, C. F. G. (2006). Analyse einer Hochschulevaluation: Der Studentenspiegel 2004. Die Qualität von Studie, Daten und Ergebnissen. *Zeitschrift für Empirische Pädagogik, 20*(4), 421–437.

Schendera, C. F. G. (2007). *Datenqualität mit SPSS*. München: Oldenbourg.

Schmidt, T. (2005). Datenarchive für die Gesprächsforschung: Perspektiven, Probleme und Lösungsansätze. Gesprächsforschung. *Online-Zeitschrift zur verbalen Interaktion, 6*, 103–126. Aufgerufen 28. Februar, 2012, unter http://www.gespraechsforschungozs.de/heft2005/px--schmidt.pdf

Winkler, N., Kroh, M., & Spiess, M. (2006). Entwicklung einer deutschen Kurzskala zur zweidimensionalen Messung von sozialer Erwünschtheit. *DIW Discussion Papers*, 579, 1–37. Aufgerufen 9. September, 2013, unter http://www.diw.de/documents/publikationen/73/diw_01.c.44281.de/dp579.pdf

Wirtz, M. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und verringert werden kann. *Rehabiliation, 43*(2), 109–115.

Datenanalyse

Nicola Döring, Jürgen Bortz

12.1 Qualitative Datenanalyse – 599

12.2 Quantitative Datenanalyse – 612

12.3 Quantitative Datenanalyse bei explorativen Studien – 621

12.4 Quantitative Datenanalyse bei deskriptiven Studien – 631

12.5 Quantitative Datenanalyse bei explanativen Studien – 657

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_12, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Wissen, was man unter qualitativer Datenanalyse versteht und verschiedene interpretative Auswertungsverfahren kennen.
- Wissen, was man unter quantitativer Datenanalyse versteht und unterschiedliche statistische Auswertungsansätze voneinander abgrenzen können.
- Die Logik des klassischen statistischen Signifikanztests zur Überprüfung von Hypothesen erläutern können.
- Bei quantitativen explorativen (gegenstandserkundenden und theoriebildenden) Studien Methoden der explorativen Datenanalyse beschreiben können.
- Bei quantitativen deskriptiven (populationsbeschreibenden) Studien die Parameterschätzung mittels Punkt- und Intervallschätzung hinsichtlich unterschiedlicher Arten von Parametern und Stichproben erklären können.
- Bei quantitativen explanativen (hypothesenprüfenden) Studien die Hypothesenprüfung mittels klassischem statistischem Signifikanztest hinsichtlich verschiedener Arten von Unterschieds-, Zusammenhangs- und Veränderungs-Hypothesen sowie Einzelfall-Hypothesen erläutern können.

Die **Datenanalyse** („data analysis“) zielt darauf ab, die zum Forschungsproblem formulierten Forschungsfragen systematisch und nachvollziehbar auf der Basis der vorliegenden empirischen Daten zu beantworten bzw. die aufgestellten Forschungshypothesen mithilfe der Daten zu testen oder neue Hypothesen zu bilden. Voraussetzung der Datenanalyse ist stets eine sorgfältige Datenerhebung (► Kap. 10) mit anschließender Datenaufbereitung (► Kap. 11). Bei der Datenanalyse sind in Abhängigkeit davon, ob nicht-numerisches (qualitatives: verbales, visuelles etc.) oder numerisches (quantitatives) Datenmaterial vorliegt, ganz unterschiedliche Vorgehensweisen notwendig. Im Folgenden werden deswegen zunächst die auf nicht-numerisches Material zugeschnittene **interpretative bzw. qualitative Datenanalyse** (► Abschn. 12.1) und dann die auf numerische Daten ausgerichtete **statistische bzw. quantitative Datenanalyse** (► Abschn. 12.2) in ihren Grundzügen vorgestellt. Anschließend werden statistische Auswertungsverfahren vertieft und zwar für quantitative Studien mit unterschiedlichem Erkenntnisinteresse, nämlich für explorative bzw. gegenstandserkundende und theoriebildende Studien (► Abschn. 12.3), für deskriptive bzw. populationsbeschreibende Studien (► Abschn. 12.4) und explanative bzw. hypothesenprüfende Studien (► Abschn. 12.5).

Die **qualitative Datenanalyse ist meist viel zeitaufwändiger** als die quantitative Datenanalyse, denn die interpretative Auswertung von Textmaterial kann (im Unterschied zur statistischen Analyse von Messwerten) durch Software nicht automatisiert, sondern nur unterstützt werden. In qualitativen Studien werden deswegen in der Regel deutlich kleinere Stichproben untersucht als in quantitativen Studien. Bei quantitativen Studien ist dagegen der Zeitaufwand für die Entwicklung und Überprüfung der standardisierten Datenerhebungsinstrumente höher, denn es muss sichergestellt werden, dass die bei der Datenerhebung erzeugten Zahlenwerte, tatsächlich inhaltlich sinnvoll interpretierbare Messwerte der interessierenden Konstrukte darstellen (im Unterschied dazu verlangt eine nicht- oder halbstrukturierte qualitative Datenerhebung weniger intensive Vorbereitung, allerdings müssen die Beobachtenden oder Interviewenden gut geschult sein, damit sie aussagekräftige qualitative Daten generieren).

Trotz aller Unterschiede zwischen qualitativen und quantitativen Auswertungsmethoden sind auch die **Gemeinsamkeiten** nicht aus den Augen zu verlieren (Hardy & Bryman, 2004): Sowohl die qualitative als auch die quantitative Datenanalyse zielen darauf ab, eine systematische Verbindung zwischen Theorien einerseits und der Erfahrungswirklichkeit andererseits herzustellen, wobei die qualitative Forschungsstrategie eher theoriebildend, die quantitative eher theorieprüfend ausgerichtet ist (► Kap. 2). Sowohl qualitative als auch quantitative Datenanalyse müssen Kriterien der Wissenschaftlichkeit folgen (► Abschn. 3.1). Das betrifft das Kriterium der wissenschaftlichen Strenge (d. h. es müssen die zu Forschungsproblem und Daten jeweils passenden Auswertungsmethoden gewählt und sachgerecht umgesetzt werden) sowie das Kriterium der interpersonalen Nachvollziehbarkeit, was hohe Präsentationsqualität verlangt (d. h. die Arbeitsschritte bei der Datenanalyse sowie die Ergebnisse und deren Interpretation müssen genau dokumentiert, begründet und kritisch reflektiert werden). Generell muss die wissenschaftliche Datenanalyse systematisch und ergebnisoffen erfolgen, d. h. absichtliche Verzerrungen der empirischen Befunde in Richtung auf ein „Wunschergebnis“ oder die Überinterpretationen der Daten einer Studie als allgemeingültige Gewissheiten sind gemäß Wissenschaftsethik (► Abschn. 4.2) unzulässig. Inhaltlich konzentrieren sich qualitative und quantitative Analysetechniken ganz zentral auf die Unterschiedlichkeit (Variabilität) in den Daten (z. B. auf die Individualität von Fällen in der qualitativen Forschung, auf Messwertunterschiede in der quantitativen Forschung) sowie auf die Häufigkeit des einzelnen oder gemeinsamen Auftretens bestimmter Kategorien oder Variablenausprägungen.

12.1 Qualitative Datenanalyse

Lernziele

- Wissen, was man unter qualitativer Datenanalyse versteht und verschiedene interpretative Auswertungsverfahren kennen.
- Den Unterschied zwischen spezialisierten und allgemeinen qualitativen Datenanalyseverfahren kennen.
- Zentrale Arbeitsschritte einer qualitativen Datenanalyse kennen.
- Einen Überblick über Software zu qualitativer Datenanalyse gewinnen.

Wenn in der empirischen Sozialforschung von **qualitativem Datenmaterial** die Rede ist, handelt es sich um nicht-numerische (verbale, visuelle etc.) Daten, die mit folgenden Datenerhebungsverfahren generiert wurden:

- Qualitative Beobachtung: Im Zuge einer ethnografischen Feldbeobachtung wurden offene **Beobachtungsprotokolle** bzw. Feldnotizen sowie **Fotos und Videoaufzeichnungen** durch die Forschenden erstellt (▶ Abschn. 10.1.2).
- Qualitatives Interview: Im Zuge von unstrukturierten oder halbstrukturierten Interviews wurden **Interviewtranskripte** produziert, in denen die Äußerungen von Interviewenden und Befragungspersonen wortwörtlich enthalten sind (▶ Abschn. 10.2.2).
- Qualitativer Fragebogen: Im Zuge einer qualitativen Fragebogenerhebung wurden **Freitextantworten und Aufsätze** von den Untersuchungsteilnehmenden geschrieben (▶ Abschn. 10.3.2).
- Qualitative Dokumentenanalyse: Im Zuge einer qualitativen Dokumentenanalyse wurden **vorhandene Text-, Ton-, Bild- und Videodokumente oder sonstige Artefakte** gesammelt: z. B. Zeitungsartikel, TV-Sendungen, Beiträge aus Online-Foren, Hausordnungen, Briefe, Fotoalben oder Kinderzeichnungen (▶ Abschn. 10.6.2).

In der qualitativen Forschung verlaufen Datenerhebung, Datenaufbereitung und Datenanalyse anstelle einer linearen Abfolge idealerweise **zirkulär**, so dass Zwischenergebnisse der qualitativen Datenanalyse die weitere qualitative Datenerhebung und Datenanalyse steuern. So wird sich z. B. im Zuge einer ethnografischen Feldbeobachtung das Verständnis der Forschenden für das Untersuchungsfeld schrittweise vertiefen und somit wird sich auch der Aufmerksamkeitsfokus bei der Beobachtung verschieben. Etwa in dem Sinne, dass zunächst unbeachtete Details in ihrer Bedeutung erkannt und dann stärker in den Beobachtungsprotokollen bzw. Feldnotizen berücksichtigt werden.

Qualitatives Datenmaterial muss nicht unbedingt qualitativ, sondern kann auch quantitativ ausgewertet werden. Dafür ist es notwendig, das qualitative Datenmaterial (Text- und Bilddokumente etc.) zunächst einer **quantitativen Inhaltsanalyse** zu unterziehen, mit deren Hilfe die Ausprägungen einzelner formaler und inhaltlicher Merkmale der Texte mit Hilfe eines zuvor entwickelten standardisierten Kategoriensystems gemessen werden (▶ Abschn. 10.6.3). Die resultierenden Messwerte sind dann statistisch auszuwerten. Auch die quantitative Inhaltsanalyse arbeitet interpretierend Bedeutungen des Text- oder Bildmaterials heraus. Sie tut dies jedoch im Sinne der Forschungslogik des quantitativen Paradigmas empirischer Sozialforschung vor allem anhand vordefinierter theoretischer Konstrukte (▶ Abschn. 8.2). Ein und dasselbe qualitative Datenmaterial kann nicht zuletzt im Rahmen einer Mixed-Methods-Studie ganz oder ausschnittweise sowohl qualitativ als auch quantitativ ausgewertet werden, so dass die qualitativen und quantitativen Ergebnisse aufeinander bezogen und zu einem Gesamtergebnis verknüpft werden können.

Qualitative Datenanalyse – Die **qualitative Datenanalyse** („qualitative data analysis“) wertet qualitatives – in der Regel verbales/textuelles, aber auch visuelles oder sonstiges nicht-numerisches – Datenmaterial im Hinblick auf das Forschungsproblem einer empirischen Studie interpretierend (hermeneutisch) aus. Sie folgt dabei dem explorativen (gegenstandserkundenden, hypothesen- und theoriebildenden) Erkenntnisinteresse qualitativer Forschung und ist stark **induktiv** (datengesteuert) ausgerichtet.
Es liegen eine Reihe von allgemeinen (d. h. auf qualitatives Datenmaterial aller Art anwendbaren) sowie von spezialisierten (d. h. nur für bestimmte Arten von Daten und Fragestellungen einsetzbaren) qualitativen Datenanalysemethoden vor. Bei der sehr verbreiteten **kategorienbildenden Analyse** wird das Material in sinnvolle Analyseeinheiten segmentiert (z. B. Sätze, Absätze), den Analyseeinheiten werden Codes zugeordnet (Prozess der Kodierung) und die Codes werden zu übergeordneten Kategorien zusammengefasst. Neben der kategorienbildenden existiert noch die **sequenzielle Analyse**, bei der Ablaufstrukturen aus den Daten herausgearbeitet werden.
Die qualitative Datenanalyse erfolgt in der Regel im ersten Schritt **fallbezogen** und im zweiten Schritt **fallübergreifend**, indem sie meist Themen, Typen oder Theorien generiert.
Schließlich kann eine qualitative Datenanalyse eines einzelnen Datensatzes noch erweitert werden, indem sie – im Rahmen der Bearbeitung desselben Forschungsproblems – mit anderen qualitativen Auswertungen **(Methoden-Triangulation)** sowie mit quantitativen Analysen **(Mixed-Methods-Ansatz)** verknüpft wird.

Im Folgenden werden kursorisch verschiedene Ansätze der qualitativen Datenanalyse vorgestellt (▶ Abschn. 12.1.1), zentrale Arbeitsschritte einer qualitativen Datenanalyse beschrieben (▶ Abschn. 12.1.2) sowie abschließend Computerprogramme für qualitative Datenauswertungen (▶ Abschn. 12.1.3) vorgestellt.

12.1.1 Verschiedene Ansätze der qualitativen Datenanalyse

Eine systematische, nach unterschiedlichen Kriterien geordnete **Klassifikation qualitativer Datenanalyseverfahren** hat sich bislang in der Methodenliteratur nicht etabliert. Allerdings zeigt eine Durchsicht der Lehr- und Handbücher zur qualitativen Datenanalyse, dass dort ein bestimmter Grundkanon an Ansätzen immer wieder auftaucht. Einen groben Anhaltspunkt für die Bedeutung der verschiedenen qualitativen Datenanalysemethoden speziell in der psychologischen Fachliteratur liefern auch die in den folgenden Abschnitten berichteten Trefferzahlen aus der wissenschaftlichen Literaturdatenbank PsycINFO.

Manche qualitativen Auswertungsmethoden entstammen der **Psychologie** (z. B. Interpretative Phänomenologische Analyse, qualitative Inhaltsanalyse), die meisten aber der **Soziologie** (z. B. Konversationsanalyse, Dokumentarische Methode, Grounded-Theory-Methodologie). Teilweise werden auch Auswertungsansätze aus der **Sprach-, Kunst- und Filmwissenschaft** einbezogen. Eine Grobeinteilung der Verfahren lässt sich danach vornehmen, ob die jeweilige Auswertungsmethode auf bestimmte Arten von Datenmaterial und/oder bestimmte inhaltliche Fragestellungen spezialisiert ist, oder aber relativ breit anwendbar ist (◘ Tab. 12.1).

Im Folgenden werden – ohne Anspruch auf Vollständigkeit – einige besonders verbreitete Ansätze der qualitativen Datenanalyse kurz präsentiert, wobei aus didaktischen Gründen die Unterschiede akzentuiert werden. Drei der Methoden – die qualitative Inhaltsanalyse, die Interpretative Phänomenologische Analyse und die Kodierung gemäß Grounded-Theory-Methodologie – werden zudem in ► Abschn. 10.6.2 zur qualitativen Dokumentenanalyse anhand von Studienbeispielen ausführlicher behandelt.

Zu den Details der Verfahren sei auf die umfangreiche deutsch- und englischsprachige **Fachliteratur** verwiesen (Bernard & Ryan, 2009; Gibbs, 2008; Grbich, 2007), manche Bücher sind dabei auf qualitative Datenanalyse in der Psychologie fokussiert (Lyons & Coyle, 2007; Mey & Mruck, 2010). Neben Monografien und Sammelbänden seien auch **Übersichtsartikel** empfohlen, die verschiedene Verfahren darstellen und Empfehlungen zur Auswahl geben (Leech & Onwuegbuzie, 2007, 2008). Nicht zuletzt sind **Methodenvergleichsstudien** aufschlussreich, seien sie theoretisch oder empirisch angelegt, die anhand derselben Fragestellung oder desselben Datenmaterials die Prozesse und Ergebnisse verschiedener qualitativer Auswertungsverfahren kontrastieren (Frost et al., 2011: Vergleich von Interpretativer Phänomenologischer Analyse, Kodierung gemäß Grounded-Theory-Methodologie, narrativer Analyse und Kritischer Diskursanalyse anhand von Interviewtranskripten; Wertz et al., 2011, Kap. 10). Eine generelle Überlegenheit einer Methode ist dabei nicht zu konstatieren, vielmehr sind die Passung der Methode hinsichtlich Forschungsproblem und Datenmaterial entscheidend sowie die Methodenkompetenz der Anwenderinnen und Anwender, die eine fehlerfreie und reflektierte Umsetzung der Methode sicherstellen müssen.

In der qualitativen Forschung ist es nicht unüblich, eine eingeführte Datenanalysetechnik in ihrer konkreten Umsetzung den Bedingungen der eigenen Untersuchung behutsam anzupassen. Besondere Herausforderungen stellen sich dabei, wenn in einer Studie mehrere qualitative Datenerhebungsmethoden zum Einsatz kommen, dementsprechend mehrere qualitative Datensätze bereitstehen, die jeweils separat sowie integriert bzw. mit Bezug zueinander auszuwerten sind. Man spricht in der qualitativen Sozialforschung von **Triangulation**, wenn bei der Untersuchung eines Forschungsproblems verschiedene theoretische Perspektiven (Theorientriangulation), verschiedene Forschende (Forschertriangulation), verschiedene Datensätze (Datenriangulation) und/oder verschiedene Methoden der Datenerhebung und Datenanalyse (Methodentriangulation) miteinander kombiniert werden, um den Erkenntnisgewinn der Studie zu steigern. Während sich die Triangulation u. a. der Kombination verschiedener qualitativer Methoden widmet (Flick, 2008, 2010), konzentriert sich der **Mixed-Methods-Ansatz** auf die Verknüpfung von qualitativen und quantitativen Forschungsstrategien (Bryman, 2006; Tashakkori & Teddlie, 2003). Was bei der qualitativen Datenanalyse zu beachten ist, wenn die Auswertung eines einzelnen Datensatzes gemäß Methodentriangulation oder Mixed-Methods-Ansatz mit anderen Auswertungsverfahren und/oder weiteren Datensätzen verknüpft werden soll, ist oft nicht allgemein zu beantworten, sondern mit Blick auf die jeweilige Studie und deren Forschungsproblem abzustimmen.

Spezialisierte qualitative Datenanalyseverfahren

Manche qualitative Auswertungsmethoden sind auf spezielle Arten von qualitativem Datenmaterial zugeschnitten, etwa auf Zeichnungen und speziell Kinderzeichnungen (zur **qualitativen Analyse von Kinderzeichnungen** s. Billmann-Mahecha, 2010) oder auf Videomaterial (zur **qualitativen Analyse von Videomaterial** s. Knoblauch, Schnettler, Raab, & Soeffner, 2006).

Andere Methoden fokussieren einzelne Merkmale des qualitativen Datenmaterials, etwa die Verwendung von Metaphern im Rahmen der **Metaphernanalyse** (für ein Studienbeispiel s. Erjavec & Volčič, 2010; Schmitt, 2010).

Die meisten spezialisierten Verfahren operieren jeweils nur mit bestimmten Arten von qualitativem Da-

Tabelle 12.1 Einordnung qualitativer Datenanalyseverfahren (Auswahl)

Klassifikation	Qualitative Datenanalyseverfahren (Auswahl)
Spezialisierte Verfahren, die auf bestimmte Arten von qualitativem Datenmaterial und/oder bestimmte inhaltliche Fragestellungen zugeschnitten sind	1. Qualitative Analyse von Kinderzeichnungen
	2. Qualitative Analyse von Videomaterial
	3. Metaphernanalyse
	4. Narrative Analyse
	5. Interpretative Phänomenologische Analyse[a]
	6. Konversationsanalyse
	7. Kritische Diskursanalyse
	8. Tiefenhermeneutik
Allgemeine Verfahren, die relativ breit für unterschiedliches qualitatives Datenmaterial und für unterschiedliche inhaltliche Fragestellungen einsetzbar sind	9. Objektive Hermeneutik
	10. Qualitative Inhaltsanalyse[a]
	11. Dokumentarische Methode
	12. Grounded-Theory-Methodologie[a]

[a] siehe zu diesen Verfahren die jeweils ausführliche Darstellung in ▶ Abschn. 10.6.2 im Kontext der qualitativen Dokumentenanalyse

tenmaterial und konzentrierten sich gleichzeitig auf bestimmte Themenbereiche:

- Die **narrative Analyse** („narrative analysis“) widmet sich in den Sozial- und Humanwissenschaften primär dem autobiografischen Erzählen und wertet erzählte Episoden inhaltlich und formal aus, etwa im Hinblick auf die Identitätskonstruktionen der Erzählerinnen und Erzähler (Lucius-Hoene, 2010; Lucius-Hoene & Deppermann, 2004; Riessman, 2008). Als Spezialform narrativer Analyse dient die **Methode der biografischen Fallrekonstruktion** primär der Auswertung narrativer Interviews (▶ Abschn. 10.2.2) im Rahmen der Biografieforschung (Schulze, 2010). In der Literaturwissenschaft werden narrative Analysen dagegen zur Auswertung fiktionaler Texte eingesetzt. Auch Medienprodukte wie z. B. Computerspiele lassen sich hinsichtlich ihrer Erzählstrukturen mittels narrativer Analyse untersuchen. „Narrative Inquiry“ ist eine führende Fachzeitschrift. Die APA-Literaturdatenbank PsycINFO verzeichnet 1 133 Beiträge mit dem Stichwort „narrative analysis“ im Abstract (Stand: August 2013).
- Die **Interpretative Phänomenologische Analyse** („interpretative phenomenological analysis“, IPA) dient ebenfalls der Auswertung autobiografischer Erzählungen, wobei im Mittelpunkt die Frage steht, wie Menschen kritische biografische Ereignisse interpretieren und verarbeiten, sei es das Erleben einer Schizophrenie-Erkrankung, der Tod eines Kindes oder das altersbedingte Ausscheiden aus dem Erwerbsleben. Die gelebte Erfahrung („lived experience“) soll aus Sicht der Betroffenen möglichst anschaulich herausgearbeitet werden (▶ Abschn. 10.6.2). Die APA-Literaturdatenbank PsycINFO verzeichnet 744 Beiträge mit dem Stichwort „interpretative phenomenological analysis“ im Abstract (Stand: August 2013).
- Die **Konversationsanalyse** („conversation analysis“, auch: Gesprächsanalyse, diskursive Analyse) konzentriert sich weniger auf die Biografie und Identität von Individuen, sondern stärker auf interpersonale Kommunikation und Interaktion (Deppermann, 1999, 2010; Hutchby & Wooffitt, 2008). Sie ist geeignet, um Alltagsgespräche, Gruppendiskussionen (zu Fokusgruppen ▶ Abschn. 10.2.2) oder Diskussionen in Online-Foren auszuwerten, wobei dann besonderes Augenmerk darauf liegt, wie die verschiedenen Personen in ihrem nonverbalen und paraverbalen Verhalten sowie in ihren Wortbeiträgen wechselseitig aufeinander reagieren (z. B. Sprecherwechsel, Unterbrechungen). Die Auswertung erfolgt nicht durch Bildung einzelner Codes und Kategorien, sondern sequenziell, indem Ablaufmuster herausgearbeitet werden. „Gesprächsforschung“ ist eine einschlägige deutschsprachige Fachzeitschrift (Informationsportal: http://www.gespraechsforschung.de). Die APA-Literaturdatenbank PsycINFO verzeichnet 1 040 Beiträge mit dem Stichwort „conversation analysis“ im Abstract (Stand: August 2013).

- Die **Kritische Diskursanalyse** („critical discourse analysis"; nicht zu verwechseln mit der zuvor erwähnten diskursiven Analyse) zielt in der Tradition des französischen Philosophen und Psychologen Michel Foucault (1982) v. a. auf das Herausarbeiten mehr oder minder verdeckter gesellschaftlicher und sozialer Machtverhältnisse ab, die durch Diskurse entstehen und sich in ihnen widerspiegeln (Allolio-Näcke, 2010; Burman & Parker, 1993). Besonders gut geeignet ist die Methode u. a. für die Auswertung von medialen Darstellungen (z. B. TV-Sendungen, Zeitungsartikeln), sofern diese Dokumente als partielle Manifestationen eines Diskurses aufgefasst werden können. Die mediale Darstellung politischer Konflikte, Epidemien, Umweltkatastrophen oder gesellschaftlicher Minoritäten gehören zu den Forschungsthemen, die mit der Kritischen bzw. Foucaultschen Diskursanalyse gut zu bearbeiten sind, weil hier Machtfragen besonders virulent sind. Die führende internationale Fachzeitschrift heißt „Discourse and Society" (http://das.sagepub.com/). Die APA-Literaturdatenbank PsycINFO verzeichnet 647 Beiträge mit dem Stichwort „critical discourse analysis" im Abstract (Stand: August 2013).
- Die **Tiefenhermeneutik** („deep structure hermeneutics", „deep hermeneutic interpretation", „deep hermeneutic approach") basiert auf psychoanalytischen Theorien und widmet sich der Auswertung kultureller Artefakte wie z. B. literarischer Texte oder Filme (Lorenzer, 1986, zur Bedeutung der Psychoanalyse in der Sozialforschung s. auch Leithäuser & Volmerg, 1988). Im Zuge sog. szenischen Verstehens sollen die latenten Bedeutungsgehalte erschlossen werden (für tiefenhermeneutische Analysen der Filme „Terminator II" und „Basic Instinct" s. König, 1994a, 1994b, 1994c, 1995). Da der Ansatz im deutschsprachigen Raum beheimatet ist, verzeichnet die APA-Literaturdatenbank PsycINFO nur 5 Beiträge mit dem Stichwort „deep hermeneutic" im Abstract (Stand: August 2013).

Allgemeine qualitative Datenanalyseverfahren

Schließlich existieren neben den spezialisierten auch **allgemeine qualitative Auswertungsmethoden**, die sich für die Analyse aller Arten von qualitativem Datenmaterial und die Bearbeitung verschiedenster Forschungsprobleme eignen, allerdings auf jeweils unterschiedlichen erkenntnistheoretischen Positionen basieren:

- Die **Objektive Hermeneutik** („objective hermeneutics") geht davon aus, dass sich psychische, soziale und kulturelle Erscheinungen (z. B. Gedanken, Gefühle) nur über ihre Ausdrucksgestalten (z. B. Erzählung, Tagebucheintrag, aber auch visuelles Material) untersuchen lassen. Die latenten Sinnstrukturen und objektiven Bedeutungsstrukturen dieser Ausdrucksgestalten sind Gegenstand der Objektiven Hermeneutik. Bei der qualitativen Datenanalyse gemäß Objektiver Hermeneutik geht es ausdrücklich nicht darum, das subjektive Erleben der Textproduzenten verstehend nachzuvollziehen und zu Kategorien zusammenzufassen, sondern die Texte bzw. Protokolle hinsichtlich ihrer algorithmischen Erzeugungsregeln zu analysieren. Dies geschieht nicht kategorienbildend, sondern sequenzanalytisch. Sequenz meint dabei nicht einfach zeitliches Nacheinander, sondern Strukturbildung in dem Sinne, dass jede Einzelhandlung bestimmte Optionen ausschließt und neue Optionen eröffnet. Die wissenschaftstheoretischen Grundlagen sowie das konkrete **sequenzanalytische Vorgehen** der Objektiven Hermeneutik unterscheiden sich deutlich von den anderen, kategorienbildenden Analyseverfahren (Oevermann, 2001, 2002). Da der Ansatz im deutschsprachigen Raum beheimatet ist, verzeichnet die APA-Literaturdatenbank PsycINFO bislang nur 10 Beiträge mit dem Stichwort „objective hermeneutics" im Abstract (Stand: August 2013). Die objektive Hermeneutik ist u. a. in der qualitativen Pädagogik bzw. Erziehungswissenschaft stark verbreitet.
- Die **qualitative Inhaltsanalyse** („qualitative content analysis") zielt darauf ab, aus qualitativem Text- oder Bildmaterial systematisch v. a. die manifesten Inhalte durch Kategorienbildung herauszuarbeiten und diese bei Bedarf auch zu quantifizieren (▶ Abschn. 10.6.2). Sie nimmt somit zwischen qualitativer und quantitativer Forschung eine Zwischenposition ein und wird in der Forschungspraxis auch oft mit der quantitativen Inhaltsanalyse (▶ Abschn. 10.6.3) kombiniert. Die APA-Literaturdatenbank PsycINFO verzeichnet 1 080 Beiträge mit dem Stichwort „qualitative content analysis" im Abstract (Stand: August 2013).
- Die **Dokumentarische Methode** („documentary method") basiert auf der erkenntnistheoretischen Position der **Praxeologie**, dergemäß soziale Sachverhalte weder als objektive noch als subjektive Phänomene verstanden werden, sondern als Ergebnisse kollektiver Handlungen und Sinnkonstruktionen. Sie geht ursprünglich auf den Soziologen und Begründer der Ethnomethodologie Harold Garfinkel zurück und wurde im deutschsprachigen Raum vor allem durch den Soziologen Ralf Bohnsack weiterentwickelt – zunächst für die Auswertung von Gruppendiskussionen. Heute wird die Methode breit angewendet, etwa auch für Datenmaterial aus teilnehmenden Beobachtungen, narrativen und Leitfadeninterviews sowie auf vorgefundene Dokumente einschließlich Fotos und Videos (Bohnsack, 2009; Bohnsack, Nentwig-Gesemann & Nohl, 2007). Die Dokumentarische Methode unter-

zieht das qualitative Material zunächst einer sog. formulierenden Interpretation, dann einer reflektierenden Interpretation und mündet schließlich in eine fallübergreifende komparative Analyse und Typenbildung. Sie ist in der Soziologie und Erziehungswissenschaft sehr viel etablierter als in der Psychologie und wird auch in der Evaluationsforschung eingesetzt (Bohnsack & Nentwig-Gesemann, 2010). Eine Verbreitung über den deutschsprachigen Raum hinaus hat erst jüngst begonnen (Bohnsack, Pfaff, & Weller, 2010). Die APA-Literaturdatenbank PsycINFO verzeichnet 10 Beiträge mit dem Stichwort „documentary method" im Abstract (Stand: August 2013).

- Die **Grounded-Theory-Methodologie** („grounded theory methodology" GTM) stellt eine eigene Forschungsstrategie dar, die auf die Entwicklung gegenstandsverankerter Theorien hinausläuft und damit einen höheren Anspruch verfolgt als eine reine Beschreibung des Untersuchungsgegenstandes. Die GTM geht über eine qualitative Datenauswertungsmethode hinaus und macht Vorgaben zur Gestaltung des gesamten qualitativen Forschungsprozesses. Der in den 1960er Jahren in der Soziologie entwickelte Grounded-Theory-Ansatz ist im Bereich der qualitativen Datenanalyse mit großem Abstand der populärste und auch in der psychologischen Forschung etabliert: Die APA-Literaturdatenbank PsycINFO verzeichnet 8 275 Beiträge mit dem Stichwort „grounded theory" im Abstract (Stand: August 2013). Die zugrunde liegende **Kodiermethode** ist die **Methode des permanenten Vergleichs** („constant comparison coding/analysis") mit den Techniken des offenen Kodierens, des axialen Kodierens (Verknüpfung der Kategorien des offenen Kodierens zu Hauptkategorien) und des selektiven Kodierens (Entwicklung von Kernkategorien aus den Hauptkategorien des axialen Kodierens; Strauss & Corbin, 1996), jeweils begleitet durch umfangreiches Schreiben von Memos („memoing", „memowriting"; für ein Auswertungsbeispiel ► Abschn. 10.6.2).

Qualitative Datenanalyse wird entweder **manuell** oder **computergestützt** („computer aided/assisted qualitative data analysis": CAQDAS) durchgeführt (zu entsprechenden Computerprogrammen ► Abschn. 12.1.3).

12.1.2 Zentrale Arbeitsschritte der qualitativen Datenanalyse

Generell lassen sich bei den meisten Varianten der qualitativen Datenanalyse **zwei Auswertungsebenen** differenzieren (für entsprechende Beispiele aus empirischen Studien s. Ayres, Kavanaugh, & Knafl, 2003): Die fallbezogene Auswertung sowie die fallübergreifende Auswertung, die in der Praxis teilweise auch ineinander greifen.

Fallbezogene Auswertung

In der Regel wird das qualitative Datenmaterial zunächst fallweise ausgewertet („within-case analysis"). Dabei stellt z. B. ein einzelnes Beobachtungsprotokoll, ein Interviewtranskript oder ein Online-Tagebuch aus dem gesamten Materialkorpus einen Fall dar. Die Fallbearbeitung erfolgt **sequenziell**, das heißt z. B., dass ein Interviewtranskript von vorne bis hinten durchgearbeitet wird. Zudem wird im Sinne des sog. „hermeneutischen Zirkels" **iterativ** gearbeitet, d. h. ein Text wird mehrfach von vorne bis hinten durchgearbeitet, so dass ein erstes Grobverständnis des Textes im Zuge der Wiederholungsdurchgänge verfeinert werden kann und sich die Bedeutungen einzelner Textstellen im Kontext des Gesamttextes immer besser erschließen.

Nachdem der auszuwertende Text gründlich gelesen wurde, beginnt die qualitative Datenanalyse mit dem Prozess der Kodierung. Beim Kodieren wird das gesamte Material (oder nur die als besonders wichtig erachteten Materialstellen) zunächst in sinnvolle Einheiten **segmentiert**. Dabei können sehr kleine Textsegmente bzw. Analyseeinheiten gebildet werden (z. B. einzelne Wörter oder Wortgruppen) oder auch größere Einheiten (z. B. ganze Sätze oder Absätze). Während man in der quantitativen Inhaltsanalyse (► Abschn. 10.6.3) meist den Gesamttext als Analyseeinheit und die einzelnen Textsegmente als Kodiereinheiten bezeichnet, hat es sich in der qualitativen Forschung eingebürgert, die Textsegmente als Analyseeinheiten zu bezeichnen. Jede **Analyseeinheit** (d. h. sinnvoll abgrenzbare Textstelle) wird daraufhin betrachtet, welche inhaltlichen und/oder formalen Merkmale sie aufweist, auf welche dahinterliegenden Prozesse oder Phänomene sie hinweist, für welches übergeordnete Konzept sie ein Beispiel darstellt etc. Eine solche Interpretation der Textstelle erfolgt vor dem Hintergrund des Forschungsproblems der Studie.

Im Ergebnis wird der Textstelle (also einem Teil des qualitativen Datenmaterials) ein **Code** zugewiesen, der ihr eine zusammenfassende oder erklärende Eigenschaft zuschreibt. Die Zuordnung von Codes zu Textstellen nennt man **Kodierung**. Codes sollten möglichst sprechende Namen tragen. Jeder Code wird in einer **Codeliste** vermerkt und kann zudem durch **Codenotizen** bzw. ein ganzes **Codebuch** (als Zusammenfassung der Codenotizen) näher erläutert werden (z. B. Abgrenzung von ähnlichen Codes, Regeln zur Anwendung der Codes). Codes können **induktiv** aus den Daten gewonnen, aber auch **deduktiv** auf der Basis von Theorien an das Material herangetra-

Tabelle 12.2 Beispiele für Textstellen, diesen jeweils zugeordnete Codes und eine aus den Codes gebildete übergeordnete Kategorie. (▶ Abschn. 10.6.2; Daten aus Williams & Merten, 2009)

Textstellen (aus Nachrichten von Hinterbliebenen auf Online-Profilen Verstorbener)	Codes	Übergeordnete Kategorie
„I just wanted to say that I am a mess right now … Everything is hitting me so hard and no one can see it or care to notice that I'm falling through the cracks and living on the edge."	„depression"	„Indicators of emotional or cognitive coping strategies"
„hate the fact that u did this to yourself. I hate it that u didn't tell anyone. But there's nothing I can do. I can be mad all I want."	„anger"	
„So essentially I'm feeling very guilty right now that I have some amount of happiness. How can I be happy when you're dead? I feel like just the idea of being happy is a betrayal to you."	„guilt"	

Studienbeispiel

Fallbezogene Auswertung

Eine Evaluationsstudie ging der Forschungsfrage nach, ob und wie berufsvorbereitende sog. „Schülerfirmen" benachteiligte Jugendliche darin unterstützen, den zukünftigen Anforderungen des Arbeitsmarktes besser gewachsen zu sein (Ahrbeck, Lehmann, Fickler-Stang, Kretschmer, & Maué, 2009). Bei der Evaluationsstudie wurden quantitative und qualitative Methoden kombiniert. Im qualitativen Studienteil wurden 69 Schülerinnen und Schüler mittels Leitfadeninterviews befragt. Die Transkripte wurden mithilfe der qualitativen Inhaltsanalyse ausgewertet. Im Zuge der fallbezogenen Auswertungen wurden für alle Befragten Fallbeschreibungen wie die folgende erstellt:

„C., eine junge Frau im Alter von 19 Jahren, besuchte wegen Lernschwierigkeiten sieben Jahre lang die Schule mit dem Förderschwerpunkt Lernen. Sie hat die Schule mit dem berufsorientierenden Abschluss verlassen und möchte die Berufsvorbereitung nutzen, um ihren Hauptschulabschluss nachzuholen. In der Schülerfirma war sie zwei Jahre tätig und hat dort ihre Computerkenntnisse trainiert und erweitert. Ihre Praktika hat sie überwiegend im Einzelhandel absolviert, das habe aber mit ihrem Berufswunsch nichts zu tun, weil sie lieber im Garten- und Landschaftsbau arbeiten möchte. Dieses Berufsfeld wurde ihr auch in der Berufsberatung empfohlen. Sie lernt und arbeitet mit mittlerem Aufwand. Besondere Probleme habe sie in Mathematik und könne deshalb noch nicht abschätzen, ob sie den Anforderungen des Hauptschulabschlusses gerecht werden könne. Außerdem habe sie große Angst vor Prüfungen, da sie gelernte Inhalte nicht lange behalten könne. Sie möchte noch so lange bei ihren Eltern wohnen, bis sie sich eine eigenständige Lebensführung leisten kann, weiter könne sie ja noch nicht planen, meint sie. Die junge Frau möchte unbedingt arbeiten und auf keinen Fall Sozialleistungen in Anspruch nehmen müssen. Aber bisher kann sie sich persönlich keine Chancen ausrechnen. Außerdem weiß sie noch nicht, ob sie den Hauptschulabschluss schaffen wird. Deshalb wird sie abwarten, welche weiteren Maßnahmen ihr von der Arbeitsagentur angeboten werden." (Ahrbeck et al., 2009, S. 151).

gen werden. Neben Induktion und Deduktion als zentralen Varianten des Schlussfolgerns wird in der qualitativen Forschung auch die **Abduktion** (▶ Abschn. 2.1.1) als Strategie zur Ableitung von Erklärungen diskutiert. Codes stellen eine erste Abstraktionsstufe vom Datenmaterial dar, sind aber noch sehr datennah. Bei sog. in-vivo-Codes wird sogar ein Zitat aus den Daten als Codename verwendet. Meist wird im Zuge der qualitativen Datenanalyse eine stärkere Abstraktion vom Datenmaterial angestrebt, indem ähnliche Codes zu **übergeordneten Kategorien** gebündelt werden (Tab. 12.2).

Unterschiedliche qualitative Datenanalysemethoden unterscheiden sich dahingehend, nach welchen Regeln und Prinzipien sowie mit welchen Zielen die Kodierung im Einzelnen vorgenommen wird. Der **Arbeitsaufwand** kann dabei sehr unterschiedlich sein: Wenn a) wenig Material vorliegt, b) dieses in große Analyseeinheiten segmentiert wird und c) bereits bei den ursprünglichen Codes eine recht hohe Abstraktionsstufe gewählt wird, ist die Analyse sehr viel schneller abgeschlossen, als wenn umfangreiches Datenmaterial zugrunde gelegt wird, dieses in sehr kleine Einheiten segmentiert wird, zunächst sehr datennahe Codes gebildet und diese dann schrittweise unter hohem theoretischem Reflexionsaufwand zu ausdifferenzierten übergeordneten Kategorien verdichtet werden. Wie akribisch vorgegangen werden soll und kann, hängt von den vorhandenen personellen und zeitlichen Ressourcen sowie von der Zielsetzung der Studie ab. Bei explorativen qualitativen Studien im Rahmen eines Mixed-Method-Designs oder bei der qualitativen Datenanalyse einzelner offener Fragenbogenitems im Zuge einer quantitativen Studie ist eine gröbere Kodierung eher gerechtfertigt, während gegenstandsnahe Theoriebildung einen sehr viel gründlicheren Kodierprozess erfordert.

Ergänzend zur detaillierten Kodierung des Materials und darauf aufbauender Kategorienbildung wird zu jedem Fall eine **summarische Fallbeschreibung** („case summary") erstellt. Bei einem einzelnen Interview kann die Fallbeschreibung z. B. eine Kennzeichnung der Befragungsperson anhand soziodemografischer Merkmale, eine Zusammenfassung der wichtigsten Interviewäuße-

Studienbeispiel

Thematische Analyse

Wie gelingt es Psychotherapeutinnen und Psychotherapeuten, die mit schwer traumatisierten Patienten arbeiten, selbst nicht in ihrem Wohlbefinden beeinträchtigt zu werden und keine stellvertretende Traumatisierung („vicarious traumatization") zu erleiden? Um diese Forschungsfrage zu beantworten, wurde eine bewusste Stichprobe von 6 berufserfahrenen Therapeutinnen und Therapeuten mit Leitfadeninterviews befragt (Harrison & Westwood, 2009). Im Zuge der Kodierung und fallübergreifenden Auswertung der Transkripte ließen sich insgesamt **neun zentrale Themen** bzw. Bewältigungsstrategien – z. B. aktiver Optimismus, ganzheitliche Selbstfürsorge, klare Abgrenzung – herausarbeiten. Jede Strategie wird dabei zusammenfassend beschrieben und durch einzelne Zitate illustriert.

So werden z. B. für die **Themenkategorie „ganzheitliche Selbstfürsorge"** diverse **Subkategorien** angeführt und erläutert. Die befragten Therapeutinnen und Therapeuten berichteten etwa über körperliche (z. B. genügend Schlaf, gesunde Ernährung), mentale (z. B. regelmäßige Weiterbildung), emotionale (z. B. vertrauensvolle Beziehungen), spirituelle (z. B. Meditation) und ästhetische (z. B. Umfeld schön gestalten) Aspekte der Selbstfürsorge, um die beruflichen Belastungen bei der Arbeit mit traumatisierten Patienten zu bewältigen. In einem die Kategorie illustrierenden Interview-Zitat berichtet eine Befragungsperson, wie sie jeden Morgen mit einem Freund einen langen Spaziergang unternimmt, dabei die Natur erlebt, das Zusammensein genießt, Gespräche führt und dadurch Kraft für den Tag tankt.

Die Ergebnisse der Studie können als Handlungsempfehlungen zur Prävention stellvertretender Traumatisierung im Therapieberuf dienen und z. B. in die Therapieausbildung einbezogen werden.

rungen zum Forschungsthema sowie Bemerkungen zur Interviewsituation beinhalten.

Am Ende der fallbezogenen Auswertung liegen für alle untersuchten Fälle jeweils ein durchkodierter qualitativer Datensatz (z. B. Interviewtranskript) sowie eine summarische Fallbeschreibung vor (für ein ▶ Studienbeispiel „Fallbezogene Auswertung"). Die fallbezogene Auswertung beinhaltet darüber hinaus auch fallübergreifende Auswertungsschritte und liefert im Ergebnis eine fallübergreifende Codeliste („master code list") samt Codenotizen bzw. Codebuch sowie erste aus den Codes entwickelte übergeordnete Kategorien und Memos (schriftliche Notizen zu Gedanken und Ideen der Forschenden bzgl. der Kategorien, die während der Arbeit am Material erstellt werden).

Fallübergreifende Auswertung

Die fallübergreifende Auswertung („cross-case analysis") setzt voraus, dass das qualitative Datenmaterial aller Fälle bereits einzeln durchgearbeitet und kodiert wurde. Die Arbeit an den übergeordneten Kategorien auf der Basis der Codeliste findet teilweise bereits während der fallbezogenen Auswertung statt, weil die definierten Codes aus der Codeliste auf unterschiedliche Fälle angewendet und dementsprechend überarbeitet werden. Zudem entstehen im Zuge der Auswertung der einzelnen Fälle auch sukzessive Ideen zur Verknüpfung von Codes zu **übergeordneten Kategorien**. Diese übergeordneten Kategorien werden bei der fallübergreifenden Analyse weiter elaboriert, in Subkategorien ausdifferenziert und mit ihren jeweiligen Dimensionen (d. h. qualitativen oder quantitativen Merkmalsausprägungen) beschrieben.

▪ Thematische Analyse

Eine Variante der fallübergreifenden Auswertung ist die thematische Analyse, bei der über alle Fälle hinweg zusammengefasst wird, welche zentralen Themen oder Aspekte des untersuchten Phänomens im Zusammenhang mit dem Forschungsproblem in der untersuchten Stichprobe zum Ausdruck kommen. Dabei spielen auch Quantifizierungen in dem Sinne eine Rolle, dass die Themenkategorien daraufhin geordnet werden können, bei wie vielen Fällen sie eine Rolle spielen (für ein ▶ Studienbeispiel „Thematische Analyse").

▪ Typenbildung

Anstelle einer Themensammlung kann die fallübergreifende Analyse auch auf einen kontrastierenden Vergleich der einzelnen Fälle hinauslaufen. Dabei werden dann im Sinne einer Typenbildung einander ähnliche Fälle zu Typen zusammengefasst und anhand ihrer Charakteristika beschrieben (zur qualitativen Typenbildung s. Kluge, 1999, 2000). In qualitativen Studien werden meist etwa drei bis fünf Typen gebildet (für ein ▶ Studienbeispiel „Typenbildung").

▪ Theoriebildung

Die anspruchsvollste Form der fallübergreifenden qualitativen Auswertung stellt die Theoriebildung dar, wie sie gemäß Grounded-Theory-Methodologie verfolgt wird (▶ Abschn. 6.3.5). Hier haben die auf der Basis der Codes gebildeten übergreifenden Kategorien von vorne herein nicht nur beschreibenden Charakter, sondern den Anspruch theoretischer Konzepte. Im Zuge der **Kodiermethode des permanenten Vergleichs** (▶ Abschn. 10.6.2) werden gemäß der im deutschsprachigen Raum verbreiteten GTM-Variante von Strauss & Corbin (1996) mittels axialen Kodierens sowie selektiven Kodierens Kategorien herausgearbeitet und gemäß

Tabelle 12.3 Ausgewählte Merkmale der drei Typen von Jugendlichen: Gewinner, Gewinner-Verlierer und Verlierer der deutschen Einheit. (Ausschnitt; mod. nach Steinke, I. (2000). Gütekriterien qualitativer Forschung. In: U. Flick, E. von Kardorff u. I. Steinke (Hrsg.), „Qualitative Forschung: Ein Handbuch". Copyright © 2000 Rowohlt Taschenbuchverlag GmbH, Reinbek bei Hamburg, S. 148ff.)

	Gewinner der Transformation	teils Gewinner/teils Verlierer der Transformation	Verlierer der Transformation
Bewertung der Transformation	+	+/−	−
Strategien, um Anerkennung zu bekommen, Werte, Ziele	– kreative Aktivitäten – gute Schulnoten – Demonstrationen – Kleidungsstil – wechselnde Aktivitäten	– konformes Verhalten – Berufe mit legitimer Machtausübung (z. B. Polizei) – Prestigeobjekte – einfaches, geordnetes Leben	– körperliche Überlegenheit – körperliche Arbeit – Kampfsport, Gewichtheben – einfache Berufe – Prestigeobjekte (oft durch Diebstahl angeeignet) – Provokation – Kleidungsstil
Familiärer Hintergrund	– Gebildete Eltern – kritische Sichtweise elterlicher Werte	– Eltern sind Arbeiter, Angestellte – starke Einbindung in Familie wird gefordert – außerfamiliäre Kontakte müssen gerechtfertigt werden	– Eltern mit geringer Bildung – problematische Familiensituation – oft aus der Familie herausgedrängt wegen Platz-, Geldmangel, Bevorzugung von Geschwistern
Soziokulturelle Ressourcen durch die Familie	+ + +	++	+

12

Studienbeispiel

Typenbildung

Wie haben Jugendliche aus der ehemaligen DDR die deutsche Wiedervereinigung im Jahr 1989 und die damit verbundene gesellschaftliche Transformation erlebt? Eine Theoretische Stichprobe von 64 Ostberliner Jugendlichen im Alter zwischen 14 und 19 Jahren aus verschiedenen Jugendszenen wurden ethnografisch beobachtet und mithilfe von Leitfadeninterviews befragt (Steinke, 2000).

Im Zuge der qualitativen Datenanalyse wurden bei der fallübergreifenden Auswertung schließlich **drei Typen** gebildet: Die Gewinner, die Gewinner-Verlierer und die Verlierer (Tab. 12.3). Jeder Typus wird ausführlich anhand von familiärem Hintergrund, sozialem Netzwerk, Zukunftsvorstellungen, Werten und Zielen sowie Einstellungen gegenüber der Wiedervereinigung charakterisiert, zudem werden die drei Typen in einer tabellarischen Übersicht voneinander abgegrenzt, wobei manche Kategorien mit ihren qualitativen Dimensionen dargestellt werden (z. B. familiärer Hintergrund), andere dagegen in ihrer Ausprägung auf einer Skala quantifiziert werden (z. B. Bewertung der Transformation, familiäre Ressourcen).

Kodierparadigma in ein theoretisches Modell integriert, das Ursachen und Konsequenzen, Kontextbedingungen sowie Strategien im Hinblick auf das untersuchte Phänomen verdeutlicht (für ein ▶ Studienbeispiel „Theoriebildung").

Studienbeispiel

Theoriebildung

In einer pflegewissenschaftlichen Studie wurde untersucht, wie Angehörige damit umgehen, wenn ein erwachsenes Familienmitglied im Krankenhaus intensivmedizinisch betreut werden muss (Vandall-Walker & Clark, 2011).

Insgesamt 35 betroffene Familienmitglieder wurden nach Prinzipien der Theoretischen Stichprobenbildung aus dem Umfeld von Intensivstationen (Warteräume, Hinweise durch das Pflegepersonal etc.) rekrutiert und mündlich befragt. Die Inteviewtranskripte wurden gemäß der Methode des Grounded-Theory-Ansatzes (Strauss & Corbin, 1996) kodiert (offenes, axiales und selektives Kodieren).

Im Ergebnis stellte sich heraus, dass die Angehörigen vier Arten von Arbeit verrichteten, um die Krankheitsphase ihrer Verwandten durchzustehen: 1) Sie bemühten sich um Zugang zu dem auf der Intensivstation liegenden Patienten („work of gaining access") und 2) betreuten ihn teilweise direkt („patient related work"). Zudem 3) arbeiteten sie mit dem Pflegepersonal zusammen („nurse/physician related work") und widmeten sich nicht zuletzt 4) verstärkt der Selbstfürsorge in dieser schwierigen Zeit („self-related work"). Jede dieser vier Schlüsselkategorien wird in Subkategorien ausdifferenziert. Zudem werden in dem so erarbeiteten theoretischen Modell vier Kontextfaktoren angeführt (Verantwortung, Ressourcen, Krankheitsverlauf, Unterstützung durch medizinisches Personal), die das zentrale Bedürfnis der Angehörigen, für den Kranken da zu sein („need to be there"), mitbeeinflussen. Das theoretische Modell wird im Fließtext erläutert sowie in einer Grafik veranschaulicht (Abb. 12.1). Die Ergebnisse der Studie liefern dem Pflegepersonal wichtige Hinweise zur Situation der Angehörigen von Patienten und können z. B. für Maßnahmen genutzt werden, um die Kooperation mit diesen zu verbessern.

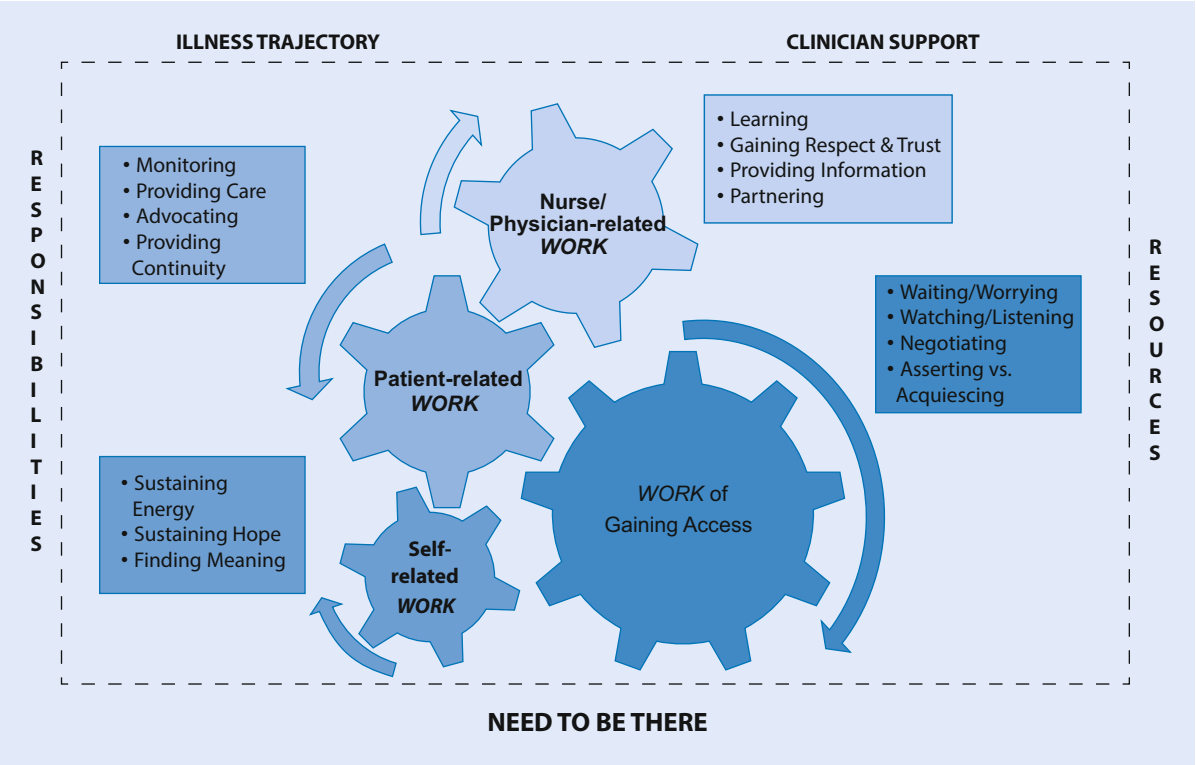


Abb. 12.1 Eine gegenstandsverankerte Theorie („grounded theory") zum Umgang von Familienmitgliedern mit kritischen Krankheiten bei ihren Angehörigen. Die Kategorien und Subkategorien der Arbeiten, die von Familienmitgliedern im Zuge des Umgangs mit kritischen Krankheiten von erwachsenen Verwandten geleistet werden, sind als Zahnräder *innerhalb des gestrichelten Rahmens* dargestellt. Kontextbezogene, externe Einflüsse sind *außerhalb des gestrichelten Rahmens* dargestellt. (Aus Vandall-Walker & Clark, 2011, S. 159. Mit freundlicher Genehmigung von SAGE Publications)

12.1.3 Computerprogramme für die qualitative Datenanalyse

Hinsichtlich technischer Unterstützung lassen sich drei Vorgehensweisen bei der qualitativen Datenanalyse unterscheiden: Die manuelle qualitative Datenanalyse, die qualitative Datenanalyse mit Bürosoftware sowie die qualitative Datenanalyse mit spezialisierten Computerprogrammen.

Manuelle qualitative Datenanalyse

Bei der sog. manuellen Datenanalyse („manual qualitative data analysis") werden die auf Papier vorliegenden Textdokumente durchgelesen und die Kodierungen durch **Textmarker**, **Unterstreichungen** und **handschriftliche Randbemerkungen** vorgenommen. Für eine fallübergreifende Analyse werden teilweise auch die den Kategorien zugeordneten Codes bzw. Textstellen aus den Dokumenten der einzelnen Fälle ausgeschnitten und nebeneinander gelegt. Memos (d. h. Notizen mit Erläuterungen oder weiterführenden Ideen zu einzelnen Codes oder Textstellen) werden auf **Karteikarten** notiert und den Texten beigefügt. Ähnlich wird mit Bildern verfahren. Für die manuelle Auswertung von Videos müssen Kodierbögen erstellt und die kodierten Videosequenzen durch Zeitmarken ausgewiesen werden.

Insbesondere bei größeren Fallzahlen ist eine manuelle Vorgehensweise recht mühsam. Sie erschwert auch ein Arbeiten in räumlich verteilten Teams. Deswegen hat sich mittlerweile eine computergestützte bzw. elektronische qualitative Datenanalyse durchgesetzt (für forschungspraktische Vergleiche zwischen manueller und elektronischer Vorgehensweise s. Davis & Meyer, 2009; Tehmina, 2003).

Wenn heute noch manuell ausgewertet wird, dann zum einen aus **forschungsökonomischen Gründen**, weil man bei der konkreten Studie davon ausgeht, die Datenaufbereitung für ein Computerprogramm sowie die Einarbeitung in das Programm würden zu großen Zusatzaufwand in Relation zum Nutzen bedeuten oder auch weil die Kosten für das Programm zu hoch sind. Zum anderen bestehen vereinzelt auch grundsätzliche **erkenntnistheoretische Einwände** gegen eine computergestützte Auswertung, etwa weil man befürchtet, dass diese durch ihre vordefinierten Funktionen die Auswertung in ei-

ne bestimmte Richtung lenkt, die Vielfalt qualitativer Analyseverfahren homogenisiert oder durch ihr regelgeleitetes Vorgehen die den qualitativen Forschungsprozess besonders kennzeichnende Kreativität und Offenheit beschneidet (Coffey, Holbrook, & Atkinson, 1996; Roberts & Wilson, 2002). Hier ist jedoch entgegen zu halten, dass auch ein manuelles Vorgehen mit Unterstreichungen und Ausschneiden letztlich regelgeleitet erfolgt und die Kreativität noch stärker einschränken kann, weil es sehr viel unflexibler ist: Computergestützt können Fehlkodierungen leicht korrigiert und Codes sowie übergeordnete Kategorien allein oder im Team immer wieder per Mausklick überarbeitet werden. Bei fehlkodierten Papierdokumenten ist dies nicht so leicht möglich, was dem iterativen Prozess qualitativer Datenanalyse im Wege steht.

Qualitative Datenanalyse mit Bürosoftware

Zuweilen wird vorgeschlagen und an Beispielen detailliert demonstriert, wie allgemeine Bürosoftware – namentlich Microsoft **Word** (La Pelle, 2004; Nideröst, 2002) oder Microsoft **Excel** (Amozurrutia & Servós, 2011; Meyer & Avery, 2009) – für die qualitative Datenanalyse genutzt werden können („qualitative data analysis with general-purpose software"). Der Hauptvorteil besteht darin, dass **Zusatzkosten** für spezialisierte qualitative Datenanalyseprogramme und zudem die **Einarbeitung** in die Spezialtools entfallen, weil Bürosoftware meist bereits zur Verfügung steht und vertraut ist. Die Nutzung von Bürosoftware ist v. a. dann sinnvoll, wenn sehr einfache Analysen anzufertigen sind (z. B. qualitative Auswertung *einer* offenen Frage innerhalb eines standardisierten Fragebogens; La Pelle, 2004).

Dafür muss jedoch auf eine Reihe von Funktionen verzichtet werden (z. B. Visualisierungen von Codes, Analyse von Fotos oder Videos). Zudem ist nicht unbeträchtlicher Aufwand notwendig, um qualitatives Datenmaterial in Word oder Excel für eine systematische qualitative Datenanalyse aufzubereiten (zur Datenaufbereitung ▶ Kap. 11) und die gewünschten Analyseschritte dann automatisiert umzusetzen (z. B. Verwendung von Makros). Im Rahmen der akademischen Ausbildung wird man vermutlich die Einarbeitung in professionelle qualitative Datenanalyseprogramme im Sinne des Erwerbs von Forschungskompetenz eher befürworten. Zudem stehen für Hochschulen bzw. Studierende meist kostengünstige Lizenzen für professionelle qualitative Analysesoftware zur Verfügung.

Qualitative Datenanalyse mit QDA-Software

Der Regelfall ist heute in der Forschungspraxis die Nutzung spezialisierter **QDA-Software** (qualitative Datenanalysesoftware; „qualitative data analysis with special-purpose software"; computer aided/assisted qualitative data analysis software", **CAQDAS**). Im Unterschied zu Statistikprogrammen für die quantitative Datenanalyse, die tatsächlich die statistischen Auswertungen vornehmen und Ergebnisse präsentieren, können Computerprogramme bei der qualitativen Datenanalyse die sinnrekonstruierenden Interpretationsschritte nicht automatisch vollziehen und dementsprechend auch keine Analyseergebnisse im engeren Sinne erzeugen. Stattdessen erbringen sie für die weiterhin von den Forschenden selbst durchgeführte inhaltliche Interpretationsarbeit in erster Linie eine **Unterstützungs- und Systematisierungsleistung** (Kuckartz & Rädiker, 2010, S. 735; für eine wissenschaftstheoretische Reflexion des computergestützten Vorgehens s. Konopásek, 2008). Bei der Auswahl und Nutzung von QDA-Software sind neben Fragen der Datenauswertung im engeren Sinne auch Fragen des qualitativen Untersuchungsdesigns zu beachten (Gregorio & Davidson, 2008).

Die führende QDA-Software (Tab. 12.4), die teilweise aus Deutschland stammt (Atlas.ti, MaxQDA), weist heute eine hohe Nutzerfreundlichkeit bzw. Gebrauchstauglichkeit auf, d. h. die Einarbeitungszeit ist nicht allzu hoch und viele Optionen sind selbsterklärend. Zudem erweitert sich der Funktionsumfang fortlaufend, aktuell etwa im Sinne einer Integration quantifizierender Funktionen für Mixed-Methods-Studien oder verbesserter Möglichkeiten der Analyse von Multimedia- und Geo-Daten. Dass durch QDA-Software die Datenanalyse effizienter erfolgt, größere und heterogenere Datenmengen verarbeitet werden können, Teamarbeit erleichtert ist und der gesamte Auswertungsprozess durch Dokumentation transparenter wird, sind die unbestrittenen Vorteile des computergestützten Vorgehens. Neben den auf den Programm-Websites zu findenden Informationen finden sich in der Fachliteratur Anleitungen und Rezensionen einzelner Programme (Hwang, 2008; Leech & Onwuegbuzie, 2011) sowie systematische Softwarevergleiche (Evers, Silver, Mruck, & Peeters, 2011; Lewis, 2004). Eine wichtige Informationsquelle hierzu ist das **CAQDAS Network Project** an der Universität Surrey, das Fachleute und Ressourcen zum Thema bündelt (http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/).

QDA-Software stellt v. a. **folgende Funktionen** zur Verfügung (Kuckartz, 2009):

1. Datenmanagement,
2. Code- und Kategorienmanagement,
3. Memos,
4. Suchfunktionen,
5. quantitative Variablen,
6. interne und externe Links bzw. Querverweise,
7. Funktionen für die Datenanalyse im Team (z. B. farbliche Markierungsmöglichkeit für unterschiedliche Kodierer),

Tabelle 12.4 Auswahl führender Computerprogramme für die qualitative Datenanalyse. (Stand: August 2013; Treffer PsycINFO: Suche nach Programmnamen im Abstract)

Software (alphabetisch)	Version	Website	Betriebssystem	Demoversion	Treffer PsycInfo
Atlas.ti	7	http://www.atlasti.de	Win	zeitlich unbegrenzt, limitierte Zahl an Fällen und Codes	317
Ethnograph	6.0	http://www.qualisresearch.com	Win	zeitlich unbegrenzt, limitierte Zahl an Projekten und Dateien	56
HyperResearch	3.5	http://www.researchware.com	Win/Mac	zeitlich unbegrenzt, limitierte Zahl an Fällen und Codes	35
MaxQDA	11	http://www.maxqda.de	Win	30 Tage, voller Funktionsumfang	21
NVivo (früher: Nud*ist)	10	http://www.qsrinternational.com	Win	30 Tage, voller Funktionsumfang	557

8. wortbasierte Häufigkeitsauswertungen,
9. neuere Funktionen (Kuckartz & Rädiker, 2010): Datendisplay und Visualisierungsmöglichkeiten, Multimedia-Integration, „Mixed Methods“, Geo-Referencing.

QDA-Software will durch Strukturierungsleistungen die interpretative Arbeit unterstützen und ist geeignet für **unterschiedliche Ansätze der qualitativen Datenanalyse**. So können z. B. MaxQDA oder Atlas.ti genutzt werden, um eine narrative Analyse, eine Auswertung gemäß Grounded-Theory-Methodologie, eine qualitative Inhaltsanalyse oder eine Interpretativ-Phänomenologische Analyse durchzuführen. Dementsprechend ist die computergestützte Vorgehensweise im Detail der jeweiligen Analysemethode anzupassen, wobei die generellen Arbeitsschritte der fallbezogenen Auswertung (Kodierung, Codenotizen, Kategorienbildung, Memos) sowie der fallübergreifenden Analyse (z. B. thematische Analyse, Typenbildung, Theoriebildung) sich meist wiederfinden (Kuckartz & Rädiker, 2010; Lewins & Silver, 2007).

Lernquiz

1. Nennen Sie vier Beispiele für Datenmaterial, das einer qualitativen Datenanalyse zu unterziehen ist!
2. Worin unterscheiden sich spezialisierte und allgemeine Verfahren der qualitativen Datenanalyse?
3. Grenzen Sie die diskursive Analyse und die Kritische Diskursanalyse voneinander ab und nennen Sie eine Gemeinsamkeit!
4. Grenzen Sie Tiefenhermeneutik und Objektive Hermeneutik als qualitative Analysemethoden voneinander ab und nennen Sie eine Gemeinsamkeit!
5. Schildern Sie die Arbeitsschritte bei der kategorienbildenden qualitativen Datenanalyse und benennen Sie die nicht-kategorienbildende Auswertungsstrategie!
6. Welche Auswertungsebenen werden bei der qualitativen Datenanalyse unterschieden?
7. Nennen Sie drei Varianten der fallübergreifenden qualitativen Datenauswertung!
8. Mit welchen Hilfsmitteln kann eine qualitative Datenanalyse erfolgen?
9. Nennen Sie zwei QDA-Programme, die in Deutschland verbreitet sind!
10. Lässt sich qualitatives Datenmaterial auch statistisch auswerten? (Begründung)

Übungsaufgaben

1. Installieren Sie die Testversion von Atlas.ti und arbeiten Sie Online-Tutorials (z. B. auf Videoplattformen) durch!
2. Installieren Sie die Testversion von MaxQDA und arbeiten Sie Online-Tutorials durch!
3. Installieren Sie die Testversion von NVivo und arbeiten Sie Online-Tutorials durch!

Weiterführende Literatur

- Kuckartz, U. (2009). *Einführung in die computergestützte Analyse qualitativer Daten*. Wiesbaden: VS.
- Kuckartz, U., Grunenberg, H. & Dresing, T. (Hrsg.) (2007). *Qualitative Datenanalyse: computergestützt. Methodische Hintergründe und Beispiele aus der Forschungspraxis* (2. Aufl.). Wiesbaden: VS.
- Lamnek, S. (2010). *Qualitative Sozialforschung* (5. Aufl.). Weinheim: Beltz.

- Mey, G. & Mruck, K. (Hrsg.) (2010). *Handbuch Qualitative Forschung in der Psychologie.* Wiesbaden: VS.
- Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed). Thousand Oaks: Sage.

Literatur

Ahrbeck, B., Lehmann, R., Fickler-Stang, U., Kretschmer, A., & Maué, E. (2009). *ENEBS. Evaluationsstudie Berliner Schülerfirmen*. Berlin: Humboldt-Universität zu Berlin.

Allolio-Näcke, L. (2010). Diskursanalyse. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 662–675). Wiesbaden: VS.

Amozurrutia, J. & Servós, C. (2011). Excel spreadsheet as a tool for social narrative analysis. *Quality and Quantity, 45*(4), 953–967.

Ayres, L., Kavanaugh, K., & Knafl, K. A. (2003). Within-case and across-case approaches to qualitative data analysis. *Qualitative Health Research, 13*(6), 871–883.

Bernard, H. R. & Ryan, G. W. (2009). *Analyzing qualitative data: Systematic approaches*. Thousand Oaks: Sage.

Billmann-Mahecha, E. (2010). Auswertung von Zeichnungen. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 707–722). Wiesbaden: VS.

Bohnsack, R. (2009). *Qualitative Bild- und Videointerpretation. Die dokumentarische Methode*. Stuttgart: UTB.

Bohnsack, R. & Nentwig-Gesemann, I. (Hrsg.). (2010). *Dokumentarische Evaluationsforschung: Theoretische Grundlagen und Beispiele aus der Praxis*. Opladen: Budrich.

Bohnsack, R., Nentwig-Gesemann, I., & Nohl, A.-M. (2007). Einleitung: Die dokumentarische Methode und ihre Forschungspraxis. In R. Bohnsack (Hrsg.), *Die dokumentarische Methode und ihre Forschungspraxis: Grundlagen qualitativer Sozialforschung* (S. 9–27). Wiesbaden: VS.

Bohnsack, R., Pfaff, N., & Weller, W. (Eds.). (2010). *Qualitative research and documentary method in educational science – results from Brazilian-German cooperations*. Opladen: Budrich.

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97–113.

Burman, E. & Parker, I. (Eds.). (1993). *Discourse analytic research: Repertoires and readings of texts in action*. London: Routledge.

Coffey, A., Holbrook, B., & Atkinson, P. (1996). Qualitative data analysis: Technologies and representations. *Sociological Research Online, 1*(1), 44. Retrieved August 22, 2013, from http://www.socresonline.org.uk/1/1/4.html

Davis, N. W. & Meyer, B. B. (2009). Qualitative data analysis: A procedural comparison. *Journal of Applied Sport Psychology, 21*(1), 116–124.

Deppermann, A. (1999). *Gespräche analysieren: Eine Einführung in konversationsanalytische Methoden*. Wiesbaden: Leske & Budrich.

Deppermann, A. (2010). Konversationsanalyse und diskursive Psychologie. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 643–661). Wiesbaden: VS.

Erjavec, K. & Volčič, Z. (2010). 'Target', 'cancer' and 'warrior': Exploring painful metaphors of self-presentation used by girls born of war rape. *Discourse and Society, 21*(5), 524–543.

Evers, J. C., Silver, C., Mruck, K., & Peeters, B. (2011). Introduction to the KWALON experiment: Discussions on qualitative data analysis software by developers and users. *Forum Qualitative Social Research, 12*(1). Retrieved August 22, 2013 , from http://www.qualitativeresearch.net/index.php/fqs/issue/view/36

Flick, U. (2008). *Triangulation. Eine Einführung.* (Vol. 2.). Wiesbaden: VS.

Flick, U. (2010). Triangulation. I. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 279–289). Wiesbaden: VS.

Foucault, M. (1982). Subjekt und Macht. In D. Defert & F. Ewald (Hrsg.), *Michel Foucault Schriften in vier Bänden Dits et Ecrits.* (Vol. 2., S. 269–294). Frankfurt/Main: Suhrkamp.

Frost, N. A., Holt, A., Shinebourne, P., Esin, C., Nolas, S.-M., Mehdizadeh, L., et al. (2011). Collective findings, individual interpretations: An illustration of a pluralistic approach to qualitative data analysis. *Qualitative Research in Psychology, 8*(1), 93–113.

Gibbs, G. R. (2008). *Analyzing qualitative data*. Thousand Oaks: Sage.

Grbich, C. (2007). *Qualitative data analysis: an introduction*. Thousand Oaks: Sage Publications.

Gregorio, S. D. & Davidson, J. (2008). *Qualitative research design for software users*. New York: McGraw-Hill.

Hardy, M. & Bryman, A. (Eds.). (2004). *The handbook of data analysis*. Thousand Oaks: Sage.

Harrison, R. L. & Westwood, M. J. (2009). Preventing vicarious traumatization of mental health therapists: Identifying protective practices. *Psychotherapy: Theory, Research, Practice, Training, 46*(2), 203–219.

Hutchby, I. & Wooffitt, R. (2008). *Conversation analysis*. Cambridge: Polity Press.

Hwang, S. (2008). Utilizing qualitative data analysis software. *Social Science Computer Review, 26*(4), 519–527.

Kluge, S. (1999). *Empirisch begründete Typenbildung. Zur Konstruktion von Typen und Typologien in der qualitativen Sozialforschung*. Opladen: Leske & Budrich.

Kluge, S. (2000). Empirisch begründete Typenbildung in der qualitativen Sozialforschung. *Qualitative Social Research, 1*(1). Abgerufen 22. August, 2013, unter http://www.qualitative-research.net/index.php/fqs/article/view/1124/2498

Knoblauch, H., Schnettler, B., Raab, J., & Soeffner, H.-G. (2006). *Video-analysis. Methodology and methods – qualitative audiovisual analysis in sociology*. Frankfurt/Main: Lang.

König, H.-D. (1994a). Mutter und Sohn und ein Mann aus Stahl. Tiefenhermeneutische Rekonstruktion von ‚Terminator II'. Teil 1. *medien praktisch, 1*, 12–18.

König, H.-D. (1994b). Mutter und Sohn und ein Mann aus Stahl. Tiefenhermeneutische Rekonstruktion von ‚Terminator II'. Teil 2. *medien praktisch, 2*, 45–49.

König, H.-D. (1994c). Mutter und Sohn und ein Mann aus Stahl. Tiefenhermeneutische Rekonstruktion von ‚Terminator II'. Teil 3. *medien praktisch, 3*, 52–60.

König, H.-D. (1995). Sexualität zwischen Lust und Tod. Tiefenhermeneutische Rekonstruktion des in dem Film Basic Instinct inszenierten Geschlechterkampfes. In S. Müller-Doohm & K. Neumann-Braun (Hrsg.), *Kulturinszenierungen* (S. 141–165). Frankfurt/Main: Suhrkamp.

Konopásek, Z. (2008). Making thinking visible with Atlas.ti: Computer assisted qualitative analysis as textual practices. *Forum Qualitative Social Research, 9*(2), Art. 12. Retrieved August 22, 2013, from http://www.qualitativeresearch.net/index.php/fqs/article/view/420

Kuckartz, U. (2009). *Einführung in die computergestützte Analyse qualitativer Daten*. Wiesbaden: VS.

Kuckartz, U. & Rädiker, S. (2010). Computergestützte Analyse (CAQDAS). In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 734–750). Wiesbaden: VS.

La Pelle, N. (2004). Simplifying qualitative data analysis using general purpose software tools. *Field Methods, 16*(1), 85–108.

Leech, N. L. & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School Psychology Quarterly, 22*(4), 557–584.

Leech, N. L. & Onwuegbuzie, A. J. (2008). Qualitative data analysis: A compendium of techniques and a framework for selection for school psychology research and beyond. *School Psychology Quarterly, 23*(4), 587–604.

Leech, N. L. & Onwuegbuzie, A. J. (2011). Beyond constant comparison qualitative data analysis: Using NVivo. *School Psychology Quarterly, 26*(1), 70–84.

Leithäuser, T. & Volmerg, B. (1988). *Psychoanalyse in der Sozialforschung*. Opladen: Westdeutscher Verlag.

Lewins, A. & Silver, C. (2007). *Using software in qualitative research: A step-by-step guide*. Thousand Oaks: Sage.

Lewis, R. B. (2004). NVivo 2.0 and ATLAS.ti 5.0: A comparative review of two popular qualitative data-analysis programs. *Field Methods, 16*(4), 439–464.

Lorenzer, A. (1986). Tiefenhermeneutische Kulturanalyse. In A. Lorenzer (Hrsg.), *Kultur-Analysen* (S. 11–98). Frankfurt/Main: Fischer.

Lucius-Hoene, G. (2010). Narrative Analysen. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 584–600). Wiesbaden: VS.

Lucius-Hoene, G. & Deppermann, A. (2004). *Rekonstruktion narrativer Identität: Ein Arbeitsbuch zur Analyse narrativer Interviews* (2nd Aufl.). Wiesbaden: VS.

Lyons, E. & Coyle, A. (2007). *Analysing qualitative data in psychology*. Thousand Oaks: Sage.

Mey, G. & Mruck, K. (Eds.). (2010). *Grounded theory reader*. Wiesbaden: VS.

Meyer, D. Z. & Avery, L. M. (2009). Excel as a qualitative data analysis tool. *Field Methods, 21*(1), 91–112.

Nideröst, B. (2002). Die technikunterstütze Analyse von qualitativen Daten mit Word [27 Absätze]. *Forum Qualitative Social Research, 3*(2), Art. 22. Abgerufen 22. August, 2013, unter http://www.qualitative-research.net/index.php/fqs/article/view/861

Oevermann, U. (2001). Strukturprobleme supervisorischer Praxis. Eine objektiv hermeneutische Sequenzanalyse zur Überprüfung der Professionalisierungstheorie. In U. Oevermann, R. Burkholz, C. Gärtner & F. Zehentreiter (Hrsg.), Forschungsbeiträge aus der objektiven Hermeneutik (Vol. 2.). Abgerufen 18. November, 2011, unter https://ssl.humanities-online.de/download/supervision.html.

Oevermann, U. (2002). Klinische Soziologie auf der Basis der Methodologie der objektiven Hermeneutik – Manifest der objektiv hermeneutischen Sozialforschung. Abgerufen 18. November, 2011, unter http://www.ihsk.de/publikationen/Ulrich_Oevermann-Manifest_der_objektiv_hermeneutischen_Sozialforschung.pdf

Reichertz, J. (2003). *Die Abduktion in der qualitativen Sozialforschung*. Opladen: Leske & Budrich.

Reichertz, J. (2004). Abduktion, Deduktion und Induktion in der qualitativen Forschung. In U. Flick, E. von Kardorff & I. Steinke (Hrsg.), *Qualitative Forschung: Ein Handbuch* (S. 276–286). Reinbek: Rowohlt.

Riessman, C. (2008). *Narrative methods for social sciences*. Thousand Oaks: Sage.

Roberts, K. A. & Wilson, R. W. (2002). ICT and the research process: Issues around the compatibility of technology with qualitative data analysis. *Forum Qualitative Social Research, 3*(2). Retrieved August 22, 2013, from http://www.qualitative-research.net/index.php/fqs/article/view/862

Schmitt, R. (2010). Metaphernanalyse. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 676–691). Wiesbaden: VS.

Schulze, H. (2010). Biografische Fallrekonstruktion. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 569–583). Wiesbaden: VS.

Steinke, I. (2000). Gütekriterien qualitativer Forschung. In U. Flick, E. von Kardorff & I. Steinke (Hrsg.), *Qualitative Forschung: Ein Handbuch* (S. 319–331). Hamburg: Rowohlt.

Strauss, A. L. & Corbin, J. (1996). *Grundlagen Qualitativer Sozialforschung*. Weinheim: Beltz.

Tashakkori, A. & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social &. behavioral research*. Thousand Oaks: Sage.

Tehmina, B. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational Research, 45*(2), 143–154.

Vandall-Walker, V. & Clark, A. M. (2011). It starts with access! A grounded theory of family members working to get through critical illness. *Journal of Family Nursing, 17*(2), 148–181.

Wertz, F. J., Charmaz, K., McMullen, L. M., Josselson, R., Anderson, R., & McSpadden, E. (2011). *Five ways of doing qualitative analysis: Phenomenological psychology, grounded theory, discourse analysis, narrative research, and intuitive inquiry*. New York: Guilford.

Williams, A. L. & Merten, M. J. (2009). Adolescents' online social networking following the death of a peer. *Journal of Adolescent Research, 24*(1), 67–90.

12.2 Quantitative Datenanalyse

Lernziele

- Wissen, was man unter quantitativer Datenanalyse versteht.
- Unterschiedliche statistische Auswertungsansätze voneinander abgrenzen können.
- Die zentralen Arbeitsschritte der quantitativen Datenanalyse verstehen und umsetzen können.
- Software zur quantitativen Datenanalyse kennen.

Wenn von **quantitativem Datenmaterial** die Rede ist, handelt es sich um numerisches Datenmaterial, das auf folgenden Wegen gewonnen wurde und statistisch auszuwerten ist:

- strukturierte Beobachtung mittels standardisiertem Beobachtungsplan oder automatische strukturierte Beobachtung mittels numerischen Protokolldateien (Log-Files, ▶ Abschn. 10.1.3),
- strukturierte mündliche Befragung (Interview) mittels standardisierten Fragebögen (▶ Abschn. 10.2.3),
- strukturierte schriftliche Befragung mittels standardisierten selbstadministrierten Fragebögen (▶ Abschn. 10.3.3),
- psychometrische Tests (▶ Abschn. 10.4.3),
- physiologische Messungen (▶ Abschn. 10.5) oder
- quantitative Inhaltsanalyse von Dokumenten (▶ Abschn. 10.6.3).

Quantitative Datenanalyse – Im Rahmen der quantitativen Datenanalyse („quantitative data analysis“) wird numerisches Datenmaterial im Hinblick auf das Forschungsproblem einer empirischen Studie statistisch ausgewertet. Die dabei eingesetzten statistischen Verfahren gliedern sich in zwei große Felder:
Die **Deskriptivstatistik** („descriptive statistics“, beschreibende Statistik) fasst die Stichprobendaten anhand von Stichprobenkennwerten (z. B. Mittelwerte, Prozentwerte etc.) zusammen und stellt diese bei Bedarf in Tabellen und Grafiken anschaulich dar.
Die **Inferenzstatistik** („inferential statistics“, schließende Statistik) schließt anhand von Stichprobendaten auf Populationsverhältnisse. Dabei wird zum einen die Ausprägung einzelner Variablen in der Population geschätzt (Methoden der statistischen Parameterschätzung) und zum anderen werden Hypothesen zu Relationen zwischen Variablen in der Population geprüft (Methoden der statistischen Hypothesenprüfung).

Die Wahl der passenden statistischen Auswertungsstrategie hängt entscheidend vom erkenntnistheoretischen Interesse der Studie ab, wobei wir zwischen explorativen, deskriptiven und explanativen Studien unterscheiden.

Die statistische Analyse eines einzelnen numerischen Datensatzes kann im **Mixed-Methods-Ansatz** noch erweitert werden durch Verknüpfungen mit der Analyse zusätzlicher quantitativer und qualitativer Datensätze, die im Rahmen derselben Studie erhoben wurden. Im Unterschied zu **Primäranalysen**, die selbst erhobene Datensätze auswerten, können auch vorliegende Datensätze (etwa aus entsprechenden wissenschaftlichen Datenbanken) einer neuen Analyse unterzogen werden **(Sekundäranalyse)**. Zudem lassen sich die Befunde inhaltlich homogener früherer Studien im Rahmen einer **Metaanalyse** zu einem Gesamtbefund statistisch aggregieren (▶ Kap. 16).

Im Folgenden werden zunächst kursorisch verschiedene Ansätze der quantitativen Datenanalyse vorgestellt, zentrale Arbeitsschritte beschrieben sowie die Auswahl und Nutzung von Statistiksoftware behandelt.

12.2.1 Verschiedene Ansätze der quantitativen Datenanalyse

Bei der Wahl der statistischen Auswertungsstrategie (deskriptiv- und/oder inferenzstatistische Verfahren) ist an erster Stelle das Erkenntnisinteresse der jeweiligen quantitativen Studie ausschlaggebend (explorative, deskriptive oder explanative Studien). Somit ergibt sich ein 2 × 3-Felder-Schema für statistische Auswertungsstrategien unterschiedlicher quantitativer Studien (◘ Tab. 12.5).

Explorative Studien zur Bildung von Hypothesen

Explorative quantitative Studien („exploratory studies“) dienen der Erkundung eines Gegenstandes, zu dem wenige theoretische Vorannahmen sowie wenige frühere Studien vorliegen. Es geht also um die Beantwortung offener Forschungsfragen und nicht um die Prüfung von theoretisch abgeleiteten Hypothesen zum untersuchten Phänomen. Ziel einer explorativen Studie ist es, den Forschungsgegenstand besser beschreiben zu können und die Hypothesen- bzw. Theoriebildung voranzutreiben. Explorative Studien werden oft als qualitative Studien, zuweilen aber auch als quantitative Studien oder als Mixed-Methods-Studien durchgeführt (z. B. qualitative Interviewstudie plus quantitative Fragebogenstudie). Typischerweise werden im Rahmen einer quantitativen Explorationsstudie **sehr viele verschiedene Variablen** zum Forschungsgegenstand an einer leicht zugänglichen (oft nicht-repräsentativen Stichprobe) erhoben, in der Hoffnung, auf diese Weise im Rahmen der quantitativen Datenanalyse interessante neue Effekte zu entdecken.

Zur Auswertung quantitativer explorativer Studien werden v. a. deskriptivstatistische Verfahren eingesetzt, die die Stichprobendaten übersichtlich anhand von Kennwerten, Tabellen und Grafiken darstellen. Die **explorative Datenanalyse** liefert darüber hinaus zusätzliche Visualisierungstechniken sowie multivariate Auswertungsverfahren, um eine große Zahl von Variablen induktiv zu strukturieren. Ergänzend werden die gefundenen Stichprobeneffekte zuweilen inferenzstatistisch mit dem klassischen Signifikanztest gegen den Zufall abgesichert. Dies

Tabelle 12.5 Übersicht über quantitative Studienarten gemäß Erkenntnisinteresse und die zugehörigen statistischen Auswertungsstrategien

	Deskriptivstatistik (beschreibende Statistik)	Inferenzstatistik (schließende Statistik)
	Die Deskriptivstatistik stellt numerische Stichprobendaten zusammenfassend in Form von **Stichprobenkennwerten** dar (v. a. Maße der zentralen Tendenz: Modalwert, Medianwert, arithmetischer Mittelwert; Maße der Dispersion: Variationsbreite, Varianz, Standardabweichung; Zusammenhangsmaße: bivariater Korrelationskoeffizient; Häufigkeiten: absolute Häufigkeiten, relative Häufigkeiten, kumulierte absolute oder relative Häufigkeiten). Die Kennwerte werden im **Fließtext** berichtet oder auch in **Tabellen** und **Grafiken** (z. B. Balkendiagramm, Liniendiagramm, Kreisdiagramm, Histogramm) präsentiert.	Die Inferenzstatistik schließt auf der Basis von Stichprobendaten auf **Populationseffekte**. Es werden Populationsparameter geschätzt (**Parameterschätzung**: z. B. Schätzung des Populationsmittelwertes auf der Basis des Stichprobenmittelwertes) oder theoriebasiert postulierte Relationen zwischen Populationsparametern geprüft (**Hypothesenprüfung**: z. B. Prüfung einer Unterschiedshypothese, dergemäß der Mittelwert in der einen Teilpopulation größer ist als in der anderen Teilpopulation).
Explorative Studie (dient der Bildung von Hypothesen)	Stichprobenbeschreibung Spezielle Techniken der Deskriptivstatistik: **Explorative Datenanalyse** und **Data-Mining** zur Identifikation von Effekten in Datensätzen mit sehr vielen Variablen	Beantwortung der Forschungsfragen mit Visualisierungstechniken, explorativen multivariaten Verfahren und teilweise auch mit statistischen Signifikanztests (diese haben dann aber nicht den Status der Hypothesenprüfung)
Deskriptive Studie (dient der Beschreibung von Populationen)	Stichprobenbeschreibung	**Parameterschätzung** für die laut Forschungsproblem interessierenden Variablen mittels Verfahren a) der Punktschätzung und/oder b) der Intervallschätzung (Konfidenzintervall)
Explanative Studie (dient der Prüfung von Hypothesen)	Stichprobenbeschreibung Ergänzende deskriptive Darstellung der Stichprobenkennwerte zu den in den Hypothesen geprüften Effekten	**Hypothesenprüfung** gemäß der zum Forschungsproblem formulierten Unterschieds-, Zusammenhangs-, Veränderungs- und Einzelfallhypothesen, wobei unterschiedliche inferenzstatistische Ansätze zur Verfügung stehen: – Klassische Signifikanztests – Minimum-Effektgrößen-Tests (► Kap. 15) – Strukturgleichungsmodelle (► Kap. 17) – Resampling-Verfahren – Verfahren der Bayes-Statistik

ist jedoch forschungslogisch **keine** Hypothesenprüfung, sondern soll lediglich Hinweise auf die Stabilität der Effekte liefern und eine Hypothesen- und Theoriebildung vorbereiten. Liegen sehr umfangreiche Datensätze vor (z. B. Protokolldateien), in denen Effekte entdeckt werden sollen, so werden auch spezielle Techniken wie das Data-Mining eingesetzt (für eine ausführliche Darstellung der explorativen Datenanalyse ► Abschn. 12.3).

Deskriptive Studien zur Beschreibung von Populationen

Bei populationsbeschreibenden quantitativen Studien („descriptive studies"), die auf repräsentativen (zufälligen bzw. probabilistischen) Stichproben basieren, kommt es darauf an, die interessierenden einzelnen **Populationsparameter** wie z. B. Mittelwerte oder Prozentwerte (z. B. durchschnittliche tägliche TV-Nutzungszeit oder Prozentsatz der Depressionskranken in der Bevölkerung eines Landes) möglichst genau auf der Basis der entsprechenden **Stichprobenkennwerte** (z. B. durchschnittliche tägliche TV-Nutzungszeit oder Prozentsatz der Depressionskranken in der untersuchten bevölkerungsrepräsentativen Zufallsstichprobe) zu schätzen. Dieser Rückschluss von den empirischen Stichprobenkennwerten auf die unbekannten Populationsparameter erfordert das Instrumentarium der Inferenzstatistik („inferential statistics") und zwar in Form der Parameterschätzung.

Für die **Parameterschätzung** werden inferenzstatistische Verfahren der **Punktschätzung** sowie der **Intervallschätzung** (Konfidenzintervalle) eingesetzt, die den Rückschluss von Stichprobenkennwerten auf Populationsparameter absichern. Diese Methoden der Parameterschätzung können im Falle probabilistischer Stichproben noch präzisiert werden, wenn die Art der probabilistischen Stichprobe (z. B. geschichtete Stichprobe oder Klumpenstichprobe, ► Abschn. 9.3.2 zu Stichprobenar-

ten) bei der Schätzung der Populationsparameter berücksichtigt wird (für eine vertiefte Darstellung der statistischen Parameterschätzung ◘ Tab. 12.4).

Explanative Studien zur Prüfung von Hypothesen

Bei explanativen quantitativen Studien („explanative studies") kommt es darauf an, einen Satz von theoretisch abgeleiteten Hypothesen anhand der Stichprobendaten zu testen. Dabei ist v. a. sicherzustellen, dass die in der Stichprobe möglicherweise gefundenen hypothesenkonformen Effekte keine Zufallsprodukte sind, die auf die besondere Zusammensetzung der betrachteten Stichprobe zurückgehen, sondern tatsächlich auf einen entsprechenden systematischen Effekt in der Population hindeuten (überzufälliges bzw. statistisch signifikantes Ergebnis). Für diese inferenzstatistische Auswertung, in deren Verlauf nicht nur einzelne Populationsparameter geschätzt, sondern v. a. **Relationen zwischen verschiedenen Populationsparametern** untersucht werden (z. B. Differenz zwischen 2 Populationsmittelwerten), existieren mehrere Ansätze: 1.) der klassische Signifikanztest, 2.) der Minimum-Effektgrößen-Test, 3.) Strukturgleichungsmodelle, 4.) Resampling-Verfahren sowie 5.) Verfahren der Bayes-Statistik (◘ Tab. 12.5).

1. Hypothesenprüfung mit klassischen Signifikanztests

Der klassische Signifikanztest basiert auf einem Hypothesenpaar. Die Forschungs- bzw. Alternativhypothese (H_1) postuliert den theoriebasiert erwarteten Effekt (z. B. Unterschied zwischen 2 Mittelwerten, Zusammenhang zwischen 2 Variablen). Die komplementäre Nullhypothese (H_0) negiert diesen Effekt und behauptet, dass entweder gar kein oder ein gegengerichteter Effekt auftritt. In der Logik des klassischen statistischen Signifikanztests werden die Daten gegen ein sog. Nullhypothesen-Modell bzw. **H_0-Modell** getestet, demgemäß kein Effekt in der Population besteht (sog. Nil-Nullhypothese). Passen die Daten gut zum H_0-Modell, so ist das Ergebnis **statistisch nicht signifikant** und die Forschungshypothese bzw. **Alternativhypothese gilt als nicht bestätigt**. Sich bei einem nicht-signifikanten Ergebnis des Signifikanztests für die H_0 zu entscheiden ist nur zulässig, sofern (z. B. durch einen relativ großen Stichprobenumfang) ausreichende **Teststärke** ($1 - \beta > 80\,\%$) vorhanden ist, andernfalls ist das Ergebnis nicht interpretierbar.

Passen die Daten indessen nicht gut zum Nullhypothesenmodell bzw. ist die Wahrscheinlichkeit, unter Annahme der Gültigkeit der H_0 die empirischen Daten oder noch extremere Ergebnisse zu finden (Irrtumswahrscheinlichkeit p), kleiner oder gleich 5 % (Signifikanzniveau α), so gilt das Ergebnis als **statistisch signifikant**. Bei einem statistisch signifikanten Ergebnis wird die H_0 verworfen und man entscheidet sich für die H_1, d. h. die **Forschungshypothese gilt als vorläufig bestätigt**. Ob der statistisch überzufällige Effekt auch praktisch bedeutsam ist, muss anhand seiner **Effektgröße** diskutiert werden. Man unterscheidet grob zwischen kleinen, mittleren und großen Effekten (▶ Abschn. 14.2). Als Ergänzung oder Alternative zum klassischen Signifikanztest werden in den letzten Jahren verstärkt die Konfidenzintervalle der interessierenden Effekte bzw. standardisierten Effektgrößen berechnet und berichtet (Kline, 2005).

Je nach Art der Forschungshypothese (Unterschieds-, Zusammenhangs-, Veränderungs-, Einzelfallhypothese) und Datenlage (Anzahl der Variablen, Skalenniveau der Variablen, Verteilungsform der Variablen, Stichprobengröße) muss der jeweils passende statistische Signifikanztest ausgewählt und – nach einer Überprüfung der statistischen Voraussetzungen der Daten – sachgerecht durchgeführt werden. Der klassische Signifikanztest basiert auf **theoretischen Prüfverteilungen**, wobei typischerweise – je nach Test – die Normalverteilung, die t-Verteilung, die F-Verteilung oder die Chi-Quadrat-Verteilung zum Einsatz kommen.

2. Hypothesenprüfung mit Minimum-Effektgrößen-Tests

Der traditionelle Signifikanztest hat sich zwar stark etabliert, weist jedoch eine Reihe von Problemen auf. Diese betreffen einerseits Anwendungs- und Interpretationsfehler (insbesondere mangelndes Verständnis des Konzepts der „statistischen Signifikanz"), andererseits aber auch die gesamte Logik des Verfahrens. Der gravierendste Einwand betrifft die Nullhypothese, die im klassischen Signifikanztest postuliert, dass in der Population überhaupt kein Effekt vorliegt **(Nil-Nullhypothese)**. Somit kann dann ein noch so kleiner und praktisch unbedeutender Effekt statistisch signifikant werden, wenn nur der Stichprobenumfang (und damit die Teststärke) entsprechend groß ist. Überzeugender ist deswegen in vielen Fällen ein Signifikanztest, dessen Nullhypothese nicht nur exakt den Null-Effekt, sondern auch alle sehr kleinen, praktisch unbedeutenden Effekte umfasst **(Minimum-Effekt-Nullhypothese)**. Ein derartiger Minimum-Effektgrößen-Test würde nur dann ein signifikantes Ergebnis liefern, wenn die Daten auf einen Effekt hinweisen, der nicht nur größer als Null ist, sondern auch größer als eine vorher festgelegte Mindestgröße ist. Der Minimum-Effektgrößen-Test wird in ▶ Kap. 15 ausführlicher dargestellt. Der Minimum-Effektgrößen-Test folgt derselben Logik wie der traditionelle Signifikanztest (Testen der Stichprobendaten gegen eine Nullhypothese anhand theoretischer Prüfverteilungen), allerdings sind modifizierte Prüfverteilungen, sog. **nichtzentrale Prüfverteilungen**

einzusetzen. Diese sind in den etablierten Statistikprogrammen nicht implementiert, so dass auf entsprechend spezialisierte Software-Tools sowie austabellierte Prüfverteilungen zurückzugreifen ist.

3. Hypothesenprüfung mit Strukturgleichungsmodellen

Während sowohl beim klassischen Signifikanztest als auch beim Minimum-Effektgrößen-Test jeweils gegen eine einzelne Nullhypothese getestet wird, werden bei der quantitativen Datenanalyse mit Strukturgleichungsmodellen **komplexe Strukturmodelle**, die aus **vielen Einzelhypothesen** bestehen, zunächst spezifiziert und dann mit den empirischen Daten verglichen. Ein Modell gilt als bestätigt, wenn sich die empirischen Daten gut mit dem Modell vereinbaren lassen bzw. wenn sie nicht überzufällig vom Modell abweichen. Beim klassischen Signifikanztest ist das Verwerfen des H_0-Modells das im Sinne der Forschungshypothese gewünschte Ergebnis, bei Strukturgleichungsmodellen strebt man umgekehrt eine möglichst gute Passung der Daten an das postulierte Modell an **(Modell-Fit)**, da das Modell die Forschungshypothesen darstellt. Ein Annehmen der Nullhypothese, dergemäß sich Daten und Modell nicht unterscheiden, ist somit wünschenswert (zur Nullhypothese als Wunschhypothese ► Abschn. 15.3).

Durch Strukturgleichungsmodelle lassen sich nicht nur mehrere Hypothesen zusammen in einem Gesamtmodell prüfen, sondern gleichzeitig auch die den Daten zugrunde liegenden **Messmodelle** testen (z. B. wenn mehrere Items eines psychometrischen Fragebogens oder Tests gemeinsam zur Messung eines Konstrukts verwendet werden). Die Überprüfung des Messmodells bezieht sich auf die Feststellung von Reliabilität und Validität (z. B. faktorielle Validität, ► Abschn. 10.4.1).

Strukturgleichungsmodelle erlauben also komplexe quantitative Datenanalysen, die mit einfachen Signifikanztests oder Minimum-Effektgrößen-Tests gar nicht möglich sind. Die höhere Komplexität des Ansatzes hat jedoch aus Anwendungssicht auch Nachteile: Es ist deutlich schwieriger, sich in diese Datenanalysemethode einzuarbeiten, Ergebnisse von Strukturgleichungsmodellierungen zu verstehen oder gar selbst entsprechende Analysen sachgerecht durchzuführen (für eine ausführliche Darstellung ► Kap. 17). Auch die statistischen Voraussetzungen, die an Daten gestellt werden, mit denen Strukturgleichungsmodelle gerechnet werden sollen, sind höher als bei klassischen Signifikanztests und müssen bereits bei der Untersuchungsplanung und Datenerhebung mitbedacht werden. Nicht zuletzt werden Strukturgleichungsmodelle aus methodologischer Sicht auch kritisch diskutiert, etwa im Hinblick auf die verwendeten Schätzverfahren oder Modellanpassungstests.

4. Hypothesenprüfung mit Resampling-Verfahren

Bei Resampling-Signifikanztests ersetzt eine **empirische Prüfverteilung** die theoretische Prüfverteilung einer Zufallsvariablen (z. B. Normalverteilung, t-Verteilung, F-Verteilung), wie sie beim traditionellen Signifikanztest zum Einsatz kommt. Dazu werden aus der empirischen Stichprobe (Ausgangsstichprobe) systematisch Unterstichproben gezogen. Aus den einzelnen Ergebnissen der Unterstichproben wird dann die Testverteilung gebildet. Das empirische Stichprobenergebnis der Ausgangsstichprobe lässt sich somit im Licht der möglichen anderen Stichprobenergebnisse der Unterstichproben bewerten. Handelt es sich um ein sehr unwahrscheinliches Ergebnis, so spricht dies gegen die Nullhypothese und für die Existenz eines systematischen Effekts im Sinne der Forschungshypothese (statistisch signifikantes Ergebnis).

Bei kleinen Stichprobenumfängen werden Resampling-Signifikanztests als sog. **Randomisierungs-Tests** (auch: Permutations-Tests, exakte Tests) konstruiert: Die Unterstichproben werden **ohne Zurücklegen** systematisch so gebildet, dass jede mögliche Ergebniskonstellation genau einmal vorkommt. Randomisierungs-Tests spielen im Kontext der sog. nonparametrischen bzw. verteilungsfreien Statistik eine wichtige Rolle. Sie sind auf sehr kleine Stichproben anwendbar sowie auf Datensätze, in denen übliche Verteilungsannahmen (z. B. Normalverteilung) nicht erfüllt sind (Bortz, Lienert, & Boehnke, 2008). Randomisierungs-Tests sind einschlägig zur Prüfung von Einzelfallhypothesen (► Abschn. 12.5.5).

Bei großen Stichproben wird aus der Gruppe der Resampling-Verfahren das **Bootstrapping** angewendet. Man spricht auch von **Monte-Carlo-Studien**. Hierbei wird aus der Ausgangsstichprobe **mit Zurücklegen** zufällig eine vordefinierte Anzahl von Unterstichproben gebildet, um das empirische Ergebnis wiederum im Kontext der Verteilung möglicher anderer Ergebnisse zufallskritisch zu bewerten (Efron & Tibshirani, 1993). Randomisierungs-Tests und Bootstrapping-Verfahren lassen sich mit konventionellen Statistik-Programmen wie z. B. SPSS durchführen.

5. Hypothesenprüfung mit Verfahren der Bayes-Statistik

Die Besonderheit der Bayes-Statistik besteht darin, dass **Vorwissen** über die Gültigkeit der zu prüfenden Hypothese in die statistische Signifikanzberechnung eingeht. Im Unterschied zur Wahrscheinlichkeit als **relative Häufigkeit** (frequentistisches Wahrscheinlichkeitskonzept) verwendet der Bayes-Ansatz **subjektive Wahrscheinlichkeiten**, die den Grad der Überzeugung der Forschenden von der Richtigkeit einer Aussage kennzeichnen. Diese A-priori-Wahrscheinlichkeit für die Gültigkeit der Hypothese leitet sich z. B. aus dem Forschungsstand ab. Bei

Hypothesenprüfungen nach dem Bayes-Ansatz wird die A-priori-Wahrscheinlichkeit der Gültigkeit der Hypothese um empirische Stichprobeninformationen ergänzt und zu einer A-posteriori-Wahrscheinlichkeit der Gültigkeit der Hypothese integriert. Dabei kommt das **Bayes'sche Wahrscheinlichkeitstheorem** zum Einsatz (Gl. 12.1):

$$p(\mathrm{A} \mid \mathrm{B}) = \frac{p(\mathrm{A} \cap \mathrm{B})}{p(\mathrm{B})} = \frac{p(\mathrm{A}) \cdot p(\mathrm{B} \mid \mathrm{A})}{p(\mathrm{B})} \qquad (12.1)$$

Dabei ist $p(\mathrm{A} \mid \mathrm{B})$ die bedingte Wahrscheinlichkeit, dass A auftritt, nachdem B eingetreten ist. Analog ist $p(\mathrm{B} \mid \mathrm{A})$ die bedingte Wahrscheinlichkeit, dass B auftritt, nachdem A eingetreten ist. Die A-priori-Wahrscheinlichkeiten von A und B lauten $p(\mathrm{A})$ und $p(\mathrm{B})$, wobei $p(\mathrm{B})$ nicht das unmögliche Ereignis sein darf, d. h. $p(\mathrm{B}) > 0$. Übertragen wir die allgemeine Formel des Bayes-Theorems nun auf die Frage, wie wahrscheinlich es ist, dass die **Nullhypothese H_0** gilt, wenn ein bestimmtes empirisches **Stichprobenergebnis D** vorliegt: Diese Wahrscheinlichkeit ist eine A-Posteriori-Wahrscheinlichkeit $p(H_0 \mid D)$, die nach dem Bayes-Theorem zu bestimmen ist (Gl. 12.2 Kline, 2005):

$$p(H_0 \mid D) = \frac{p(H_0) \cdot p(D \mid H_0)}{p(D)} \qquad (12.2)$$

Ist $p(H_0 \mid D)$ klein, so spricht dies gegen die Gültigkeit der Nullhypothese und für die Gültigkeit der Alternativhypothese. Die bedingte Wahrscheinlichkeit $p(H_0 \mid D)$ ist eine Funktion zweier A-priori-Wahrscheinlichkeiten, $p(H_0)$ und $p(D)$ sowie der bedingten Wahrscheinlichkeit $p(D \mid H_0)$. Letztere ist die Wahrscheinlichkeit für das Vorkommen der gegebenen Datenlage unter der Voraussetzung der Gültigkeit der Nullhypothese und entspricht dem p-Wert (bzw. der Irrtumswahrscheinlichkeit) im traditionellen Signifikanztest. Bei $p(H_0)$ handelt es sich um die angenommene Wahrscheinlichkeit der Nullhypothese, bevor die Daten erhoben wurden, $p(D)$ ist die Wahrscheinlichkeit für das Auftreten der gegebenen Datenlage unabhängig davon, ob die Nullhypothese falsch oder richtig ist. Die Hypothesenprüfung nach dem Bayes-Verfahren läuft also darauf hinaus, zunächst anhand von Vorinformationen aus dem Forschungsstand die Wahrscheinlichkeit der Gültigkeit der Hypothese zu bestimmen $p(H_0)$, diese dann mit empirischen Daten aus der empirischen Studie zu kombinieren und dadurch zu einer genaueren Einschätzung der Gültigkeit der Hypothese zu kommen $p(H_0 \mid D)$.

Für die Forschungspraxis bedeutet dies, dass man, um den Bayes-Ansatzes der Hypothesenprüfung nutzen zu können, nicht nur Vorstellungen über denjenigen Parameter haben muss, der am „wahrscheinlichsten" erscheint und dies in den A-priori-Hypothesen angeben muss, sondern dass man zusätzlich Angaben darüber machen muss, für wie „wahrscheinlich" oder plausibel man alle übrigen denkbaren Ausprägungen des Parameters hält. Kurz formuliert: Man muss bei diskreten Zufallsvariablen Informationen über die Wahrscheinlichkeitsfunktion des Parameters und bei stetigen Zufallsvariablen Informationen über dessen Dichtefunktion (▶ Abschn. 12.4.1) haben (für ein Tutorial zur Durchführung von Bayes-Hypothesentests s. z. B. Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; für Einführungen in die statistische Hypothesenprüfung nach dem Bayes-Ansatz s. Koch, 2012; Lee, 2012).

Während beim traditionellen Signifikanztest die Irrtumswahrscheinlichkeit $p(D \mid H_0)$ berechnet wird und von dieser Wahrscheinlichkeitsaussage über Daten ein Rückschluss auf die Gültigkeit von Hypothesen erfolgt, wird im Bayes-Ansatz direkt die Wahrscheinlichkeit der Hypothese $p(H_0 \mid D)$ bestimmt – somit ist die Logik des Verfahrens stringenter.

Die Vor- und Nachteile sowie die Indikation des traditionellen Signifikanztests und der Bayes-Hypothesenprüfung werden kontrovers diskutiert. Kritiker wenden z. B. ein, dass die Anwendbarkeit von Bayes-Verfahren dadurch begrenzt sei, dass die A-priori-Wahrscheinlichkeiten für Daten und Hypothesen in vielen Fällen nicht zuverlässig angegeben werden können (Trafimow, 2003, 2005). Befürworter betonen indessen, dass Bayes-Verfahren keineswegs auf unsicheren subjektiven Wahrscheinlichkeiten beruhen, sondern auf rationalen Plausibilitätsangaben, dass die Priorverteilungen allgemein bestimmbar seien und dass in Fällen, in denen die Voraussetzungen klassischer Signifikanztests nicht erfüllt sind (z. B. sehr kleiner Stichprobenumfang, nichtnormalverteilte Variablen), Bayes-Verfahren eingesetzt werden können (Lee & Wagenmakers, 2005).

12.2.2 Zentrale Arbeitsschritte der quantitativen Datenanalyse

Die quantitative Datenanalyse verläuft definitionsgemäß sehr viel standardisierter als die qualitative Datenanalyse. Folgende Arbeitsschritte sind bei der Auswertung jeder quantitativen Studie zu durchlaufen:

1. **Datenbereinigung**: Schon im Zuge der Datenaufbereitung (▶ Kap. 11) sollte dafür Sorge getragen werden, dass der quantitative Datensatz um Fehler (z. B. Messwerte außerhalb des gültigen Wertebereiches oder unplausible Angaben) bereinigt worden ist.
2. **Stichprobenbeschreibung**: Jede quantitative empirische Studie bedient sich zunächst deskriptivstatistischer Methoden, um die untersuchte Stichprobe an-

hand zentraler soziodemografischer Merkmale zu beschreiben (z. B. Angaben zu Häufigkeitsverteilungen, Mittelwerten und Streuungen hinsichtlich Geschlecht, Alter, Beruf, Bildung etc.). Dazu genügt in der Regel ein einziger Absatz Fließtext, um die Kennwerte darzustellen. Umfangreiche Tabellen und Grafiken sind hier meist nicht sinnvoll, da die Stichprobenbeschreibung nur eine Hintergrundinformation zur Studie darstellt.

3. **Dateninspektion und deskriptivstatistische Analysen**: Bei explorativen Studien stehen deskriptivstatistische Analysen im Mittelpunkt. Auch bei explanativen Studien, die auf inferenzstatistische Hypothesentests hinauslaufen, sollte man sich vor der Durchführung komplexerer Analysen zunächst die Daten im Detail ansehen (Dateninspektion) und für die interessierenden Variablen und Variablenrelationen die Stichprobenverhältnisse zur Veranschaulichung tabellarisch und/oder grafisch darstellen. Entsprechende Informationen sind auch in der späteren Ergebnisdarstellung wichtig. So ist es meist unbefriedigend, wenn man z. B. der Darstellung einer explanativen Studie nur entnehmen kann, **dass** 2 Variablen hypothesenkonform überzufällig korrelieren oder **dass** sich 3 Gruppen überzufällig unterscheiden, ohne dass man erfährt, **wie** die Variablen in der Stichprobe eigentlich ausgeprägt bzw. verteilt waren.
4. **Inferenzstatistische Analysen**: Bei deskriptiven Studien stehen inferenzstatistische Verfahren der **Parameterschätzung** im Mittelpunkt, seien es Punktschätzungen, bei denen der unbekannte Populationsparameter (z. B. Populationsmittelwert) mit dem zugehörigen Stichprobenkennwert (z. B. Stichprobenmittelwert) direkt geschätzt wird, oder Intervallschätzungen, bei denen ein Konfidenz- bzw. Vertrauensintervall für den Populationsparameter berechnet wird.
   Bei explanativen Studien hat sich der **klassische statistische Signifikanztest** als inferenzstatistisches Auswertungsverfahren stark etabliert. Je nachdem, welche Art von Forschungshypothese geprüft werden soll (Zusammenhangs-, Unterschieds-, Veränderungshypothese oder auch Einzelfallhypothese) und wie die Datenlage beschaffen ist (Anzahl der Variablen, Skalenniveau der Variablen, Verteilungsform der Variablen etc.), muss das passende Verfahren ausgewählt werden (Indikation). Dabei sind auch die **statistischen Voraussetzungen** (z. B. Varianzhomogenität der verglichenen Gruppen; Normalverteilung der berücksichtigten Variablen) zu prüfen, die erfüllt sein müssen, damit die Analyse aussagekräftige Resultate bringt. Für die Prüfung der Voraussetzungen werden teilweise eigene Signifikanztests durchgeführt (z. B. Voraussetzungstests wie der Varianzhomogenitätstest oder der Test auf Normalverteilung).
   Unabhängig davon, ob bei explanativen Studien mit dem klassischen Signifikanztest oder alternativ mit Minimum-Effektgrößen-Tests oder Strukturgleichungsmodellen gearbeitet wird, entscheidend für eine erfolgreiche computergestützte quantitative Datenanalyse ist stets a) die korrekte Wahl des jeweiligen Verfahrens mit Blick auf das Forschungsproblem und die vorliegenden Daten, b) die korrekte Durchführung der Analyse im Sinne von Voraussetzungsprüfungen und sachgerechter Einstellung von Optionen im Statistikprogramm sowie c) die Extraktion der relevanten Ergebnisse aus dem meist umfangreichen und unübersichtlichen Programm-Output.
5. **Inhaltliche Interpretation der statistischen Ergebnisse**: Die relevanten deskriptiv- und inferenzstatistischen Ergebnisse, die aus dem Output des verwendeten Statistik-Programms entnommen wurden, sind mit Bezug zum Forschungsproblem bzw. den einzelnen Forschungsfragen und Forschungshypothesen zu interpretieren. Dabei ist zunächst für **jede einzelne Forschungsfrage bzw. Hypothese** herauszuarbeiten, wie das Ergebnis lautet. Bei Hypothesentests mittels klassischem Signifikanztest ist z. B. zu entscheiden, ob die Nullhypothese anlässlich eines statistisch nichtsignifikanten Ergebnisses beibehalten werden muss (dies setzt ausreichende **Teststärke** voraus, da andernfalls das Ergebnis uneindeutig ist), oder ob die Nullhypothese anlässlich eines statistisch signifikanten Ergebnisses verworfen und die Forschungshypothese vorläufig angenommen werden kann (dabei ist dann neben der statistischen Signifikanz auch die **Effektstärke** im Sinne der praktischen Bedeutung des Effekts zu diskutieren).
   Nach einer Betrachtung der Einzelergebnisse ist ein Gesamtfazit notwendig. Dabei ist das **Ergebnismuster** zu betrachten: Wird im Rahmen einer explanativen Studie ein ganzer Satz von Hypothesen getestet, so fallen in der Regel einige statistische Ergebnisse theoriekonform, andere dagegen theorie- bzw. hypothesenkonträr aus. Aus den teilweise widersprüchlichen Ergebnissen muss also ein Gesamtbefund herausgearbeitet werden. Dabei sind Überinterpretationen zu vermeiden. Dies gilt sowohl für **unzulässige Generalisierungen** begrenzter Befunde auf nicht untersuchte Situationen, Personengruppen oder Phänomene als auch für **spekulative post-hoc-Erklärungen**, in denen plausibel klingende, aber theoretisch nicht fundierte, mehr oder minder beliebige Begründungen für unerwartete Ergebnisse nachgeliefert werden. Eine seriöse inhaltliche Interpretation berücksichtigt immer

ausdrücklich die **Limitationen der Studie** und die Grenzen ihrer Aussagekraft, weist auf unsichere Befunde oder spekulative Erklärungen als solche hin. Ein selbstkritischer Umgang mit den Befunden der eigenen Studie gehört gemäß Regeln guter wissenschaftlicher Praxis zu den ethischen Anforderungen, die an Forschende gestellt werden (zur Wissenschaftsethik ▶ Abschn. 4.2). Wurden die statistischen Einzelergebnisse sowie der Gesamtbefund hinsichtlich ihrer Bedeutung für das Forschungsproblem erläutert (wissenschaftliche **Diskussion** der Befunde), so gehören zur Interpretation meist auch Empfehlungen für aussagekräftige Anschlussstudien sowie Schlussfolgerungen für die Praxis (**Ausblick** auf zukünftige Forschung sowie praktische Anwendungen).

12.2.3 Computerprogramme für die quantitative Datenanalyse

Eine **manuelle Analyse** ist in der qualitative Datenanalyse zwar zuweilen anzutreffen (▶ Abschn. 12.1.3), bei statistischen Auswertungen dagegen gänzlich unüblich. Der Einsatz **allgemeiner Bürosoftware** (v. a. Tabellenkalkulationsprogramme wie Microsoft [MS] Excel) ist in der Praxis der akademischen Forschung sowie der privatwirtschaftlichen Markt- und Meinungsforschung teilweise verbreitet. Für einfache quantitative Analysen (Deskriptivstatistik) und grafische Aufbereitungen von numerischem Datenmaterial ist teure und komplexe Statistiksoftware tatsächlich oft überflüssig. Sich im Studium in die statistische Datenanalyse mit MS Excel einzuarbeiten ist insofern sinnvoll, als in vielen Unternehmen damit gearbeitet wird. Für anspruchsvollere quantitative Studien und insbesondere inferenzstatistische Analysen sind spezialisierte **Statistik-Programmpakete** unumgänglich.

Im Unterschied zur QDA-Software, die nur strukturierend bei der interpretativen Datenauswertung hilft (▶ Abschn. 12.1.3), führt Statistik-Software die statistischen Berechnungen eigenständig durch. Die Leistung der Forschenden ist trotzdem nicht zu unterschätzen: Sie müssen zum einen anhand der Datenlage sowie der jeweiligen Fragestellung und Hypothese das korrekte Auswertungsverfahren auswählen, die Details der Analyse durch sachgerechte Wahl von Optionen steuern, aus dem oft umfangreichen und komplexen Programm-Output die relevanten Ergebnisse herausfiltern und diese angemessen interpretieren. Die Statistiksoftware selbst ist heute durch eine grafische Oberfläche und Menüführung zwar recht benutzungsfreundlich, solide **Statistik-Vorkenntnisse** sind für die Bedienung jedoch notwendig.

Es steht eine große und wachsende Zahl an Statistikprogrammen zur Verfügung, die sich für die quantitative sozial- und humanwissenschaftliche Forschung eignen (◘ Tab. 12.6). Wir unterscheiden **Statistik-Programmpakete** mit einem breiten Funktionsumfang von **spezialisierten Statistik-Programmen**, die nur einzelne Verfahren anbieten.

Für die heutigen Anforderungen in der sozial- und humanwissenschaftlichen Ausbildung, Forschung und Praxis hat sich **SPSS**, dessen erste Version 1968 an der Stanford-Universität (USA) unter dem Namen „Statistical Package for the Social Sciences“ entwickelt wurde, als ein vergleichsweise leicht bedienbares und mächtiges Statistikpaket etabliert. Es ist per Menü und Maus sowie über eine Kommandosprache steuerbar. Mit SPSS lassen sich – neben deskriptivstatistischen Auswertungen – im Bereich der Inferenzstatistik Parameterschätzungen, klassische Signifikanztests sowie mit dem Zusatzmodul **SPSS AMOS** auch Strukturgleichungsmodelle rechnen. Sinnvoll ergänzt wird das kommerzielle SPSS-Paket durch das kostenlose Spezial-Tool **G*Power**, mit dem Teststärken und optimale Stichprobenumfänge bestimmt werden können.

Neben SPSS existieren noch weitere kommerzielle Statistikprogrammpakete wie SAS oder Stata, die alle relativ kostspielig sind (◘ Tab. 12.6). Das Statistik-Programmpaket **R**, das als **Open-Source-Projekt** an der Universität Auckland (Neuseeland) von Robert Gentleman und Ross Ihaka („R & R“) ins Leben gerufen wurde und seit 1997 von einem 20-köpfigen Kernprogrammierteam weiterentwickelt wird, kann hier eine echte nichtkommerzielle Alternative bieten. R hat teilweise sogar mehr Funktionen als kommerzielle Produkte und wird fortlaufend erweitert. Die Vorteile von R werden jedoch durch einen etwas geringeren Bedienungskomfort bezahlt. Allerdings liegen inzwischen deutschsprachige Einführungsbücher vor (Hatzinger, Hornik & Nagel, 2011; Luhmann, 2010; Manderscheid, 2011). Wie man Strukturgleichungsmodelle mit R rechnen kann, wird in ▶ Kap. 17 exemplarisch illustriert.

Generell stehen zur Einarbeitung in Statistik-Programme unterschiedliche Hilfsmittel zur Verfügung: Online-Hilfen auf den Websites der Programme, Handbücher, Skripte, Workshops etc. Für welches Programm man sich entscheidet, wird nicht zuletzt auch von Traditionen der jeweiligen Fachdisziplin bzw. des Forschungsfeldes bestimmt.

Tabelle 12.6 Auswahl führender Statistik-Software in der Reihenfolge der internationalen Verbreitung (Stand: Dezember 2013). (Aus PsycINFO)[a]

Software	Version	Website	Betriebs-system	Demoversion	Treffer PsycINFO[a]
Allgemeine Statistik-Programmpakete für Deskriptiv- und Inferenzstatistik in den Sozialwissenschaften					
SPSS	21	http://www.spss.com	Win, Mac, Linux	14 Tage, voller Funktionsumfang (Basis-version und alle Module)	1 470
SAS	9.4	http://www.sas.com	Win, Linux	nicht verfügbar	13[b]
Stata	13	http://www.stata.com	Win, Mac, Linux	30 Tage, voller Funktionsumfang und technische Unterstützung.	118
R	2.15.3	http://www.r-project.org	Win, Mac, Linux	Open Source Projekt	49[c]
STATISTICA	10	http://www.statsoft.de	Win	30 Tage, voller Funktionsumfang	24
Strukturgleichungsmodelle					
LISREL	9.1	http://www.ssicentral.com	Win	15 Tage	1 323
AMOS (von SPSS)	22	http://www.spss.com	Win	14 Tage, voller Funktionsumfang	313
EQS	6.1	http://www.mvsoft.com	Win, Mac, Unix	zeitlich unbeschränkt, limitierter Funk-tionsumfang (kein Kopieren, Speichern, Drucken)	145
MPlus	7.11	http://www.statmodel.com	Win, Mac, Unix	zeitlich unbeschränkt, limitierte Variablen-anzahl	184
PLS-Graph	3.0	http://www.plsgraph.com	Win	kostenlose Beta-Version	7
SmartPLS	2.0 (beta)	http://www.smartpls.de	Win	Free Beta-Version	7
Bestimmung von Teststärke und optimalem Stichprobenumfang					
G*Power	3	http://www.gpower.hhu.de	Win, Mac	Freeware	2

[a] Namen der Statistik-Programme als Suchworte im Abstract von Peer-reviewed-Artikeln. Die Nennung im Abstract ist dabei nur ein ungenauer Indikator für die Verbreitung in der Fachliteratur. Da die Software-Namen teilweise auch in anderer Bedeutung auftauchen, wurden einige Suchbegriffe modifiziert:
[b] „SAS program“ or „SAS software“ or „SAS macro“ or „SAS version“ or „SAS v“ or „SAS 9“;
[c] „program R“ or „R program“

Lernquiz

1. Wovon hängt die statistische Auswertungsstrategie bei quantitativen Studien ab?
2. Mit welchen Datenerhebungsmethoden wird quantitatives Datenmaterial in den Human- und Sozialwissenschaften üblicherweise gewonnen?
3. Welche Ansätze der inferenzstatistischen Hypothesenprüfung sind zu unterscheiden?
4. Die Stichprobenbeschreibung einer Studie sollte mit Tabellen und Grafiken veranschaulicht werden. Diskutieren Sie diese Aussage!
5. Mit der inferenzstatistischen Datenanalyse ist die statistische Prüfung der Forschungshypothesen gemeint. Richtig oder falsch?
6. Was stimmt?
   a) Der statistische Signifikanztest stellt die einzige Möglichkeit der Prüfung quantitativer Hypothesen dar.
   b) Zur Deskriptivstatistik zählen die Bestimmung von Stichprobenkennwerten sowie von Populationsparametern.

c) Wenn ein Ergebnis statistisch nicht signifikant ist, dann gilt automatisch die Nullhypothese als bestätigt.
d) Wenn ein Ergebnis statistisch signifikant ist, dann handelt es sich um einen praktisch bedeutsamen Effekt.
e) Ein Ergebnis ist signifikant, wenn das Alpha-Fehlerniveau unter 5 % liegt.

7. Grenzen Sie den klassischen Signifikanztest von Resampling-Verfahren und Verfahren der Bayes-Statistik ab!
8. Nennen Sie mindestens drei allgemeine Statistik-Programmpakete, die in den Sozialwissenschaften verbreitet sind!
9. Deskriptivstatistische Methoden werden im Rahmen jeder quantitativen Studie eingesetzt. Wozu?
10. Das Ergebnis einer empirisch-quantitativen Studie zeigt sich im Output des verwendeten Statistik-Programms. Diskutieren Sie diese Aussage!

Übungsaufgaben

Beschäftigen Sie sich mit verbreiteten Statistik-Programmen, indem Sie die kostenlosen Testversionen herunterladen und installieren und a) die mitgelieferten Anleitungen, b) im Internet frei verfügbare Online-Tutorials (z. B. Linksammlung der Universität Princeton: http://dss.princeton.edu/training/) und/oder c) einschlägige Lehrbücher bearbeiten. Führen Sie dieselben Berechnungen am selben Datensatz vergleichend mit unterschiedlichen Statistik-Programmen durch!

1. Installieren Sie die Testversion von SPSS!
2. Installieren Sie die Testversion von Stata!
3. Installieren Sie R!

Weiterführende Literatur

- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Heidelberg: Springer.
- Bortz, J., Lienert, G. A. & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Heidelberg: Springer.
- Bühl, A. (2011). *SPSS 20: Einführung in die moderne Datenanalyse* (13. Aufl.). München: Pearson Studium.
- Hatzinger, R., Hornik, K. & Nagel, H. (2011). *R – Einführung durch angewandte Statistik*. München: Pearson Studium.
- Kline, R. B. (2005). *Beyond significance testing. Reforming data analysis methods in behavioral research* (2nd ed.). Washington, DC: American Psychological Association.

Literatur

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Heidelberg: Springer.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton: CRC Press.

Hatzinger, R., Hornik, K., & Nagel, H. (2011). *R – Einführung durch angewandte Statistik*. München: Pearson.

Kline, R. B. (2005). *Beyond significance testing. Reforming data analysis methods in behavioral research* (2nd ed.). Washington: American Psychological Association.

Koch, R. (2012). *Einführung in die Bayes-Statistik*. Heidelberg: Springer.

Lee, P. M. (2012). *Bayesian statistics. An introduction* (4th ed.). New York: Wiley.

Lee, M. D. & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112*(3), 662–668.

Luhmann, M. (2010). *R für Einsteiger: Einführung in die Statistiksoftware für die Sozialwissenschaften*. Weinheim: Beltz.

Manderscheid, K. (2011). *Sozialwissenschaftliche Datenanalyse mit R: Eine Einführung*. Wiesbaden: VS.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110*(3), 526–535.

Trafimow, D. (2005). The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers (2005). *Psychological Review, 112*(2), 669–674.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*(3), 158–189.

12.3 Quantitative Datenanalyse bei explorativen Studien

Lernziele

- Zielsetzung und Vorgehensweise der statistischen Datenanalyse bei explorativen Studien kennen.
- Visualisierungstechniken für quantitative Daten kennen.
- Einen Überblick über explorative multivariate Analysetechniken gewinnen.
- Verfahren des Data-Mining einordnen können.
- Die Bedeutung von Signifikanztests in explorativen Studien richtig einschätzen können.

Explorative Studien zielen darauf ab, einen neuen oder bislang wenig erforschten Sachverhalt zu erkunden. Folglich kann bei explorativen Studien kaum auf etablierte Theorien oder einen umfangreichen Forschungsstand zurückgegriffen werden. Anstelle der Überprüfung theoretisch abgeleiteter Hypothesen tritt daher die detaillierte **Beschreibung des interessierenden Sachverhaltes** und das **Erarbeiten neuer Hypothesen und Theorien** in den Vordergrund.

Im Rahmen von explorativen Studien kommen oft qualitative Datenerhebungsmethoden zum Einsatz, da diese eine offene Herangehensweise an den Untersuchungsgegenstand ermöglichen (z. B. ethnografische Feldbeobachtungen, Leitfaden-Interviews). Jedoch sind auch quantitative Erkundungsstudien möglich, deren Daten mit strukturierten Beobachtungen und Befragungen, physiologischen Messungen oder Dokumentenanalysen gesammelt werden. Quantitative Erkundungsstudien basieren im Unterschied zu populationsbeschreibenden Studien meist auf **leicht zugänglichen nicht-probabilistischen Stichproben** und dienen im Unterschied zu explanativen Studien nicht der Überprüfung von Hypothesen, sondern der **Beantwortung von offenen Forschungsfragen**. Oft werden im Rahmen explorativer quantitativer Studien sehr viele verschiedene Variablen erhoben. Aus den genannten Zielsetzungen explorativer Studien ergeben sich die drei typischen statistischen Datenanalysestrategien für diesen Studientyp:

1. **Explorative Datenanalyse** und deskriptivstatistische Methoden werden verwendet, um das quantitative Datenmaterial hinsichtlich der Vielzahl der Variablen zu strukturieren sowie besondere und vielleicht auch unerwartete Effekte zu finden (► Abschn. 12.3.1).
2. **Inferenzstatistische Verfahren** werden zur Beantwortung von Forschungsfragen und zufallskritischen Betrachtung von Effekten eingesetzt, dürfen aber nicht als Hypothesentests verstanden werden (► Abschn. 12.3.2).
3. Das **Data-Mining** dient der Entdeckung von Mustern in sehr großen Datensätzen (► Abschn. 12.3.3).

12.3.1 Deskriptivstatistische und explorative Datenanalyse

Die explorative Datenanalyse (**„exploratory data analysis“**, EDA) ist v. a. mit dem Namen des amerikanischen Statistikers John Tukey (1977) verbunden. EDA dient dazu, Strukturen, Trends und Muster in einem Satz quantitativer Daten zu entdecken, die ohne technische Hilfsmittel leicht übersehen werden (Behrens & Yu, 2003). Während man sich in hypothesenprüfenden Untersuchungen auf die Präsentation und Analyse des laut Hypothese interessierenden Effektes beschränkt (**„confirmatory data analysis“**, CDA), dienen EDA-Techniken dazu, ein möglichst vollständiges und übersichtliches Bild des gesamten Datensatzes zu geben, indem die einzelnen Variablen, ihre Verteilungen und Zusammenhänge detailliert betrachtet werden. Eine wichtige Gruppe von EDA-Verfahren sind **Visualisierungstechniken**, mit denen die Verteilungen einzelner oder mehrerer Variablen grafisch darstellbar sind. Zudem existieren **explorative multivariate Auswertungstechniken**, die induktiv das Datenmaterial strukturieren.

Visualisierungstechniken

Die **Deskriptivstatistik** dient dazu, die Merkmalsausprägungen und Merkmalsverteilungen in einem quantitativen Datensatz durch Kennwerte im Fließtext sowie durch Tabellen und Grafiken übersichtlich darzustellen. Dabei kann man die vorliegenden Daten zunächst variablenweise hinsichtlich ihrer Verteilungsformen betrachten, zudem Gruppenunterschiede hinzuziehen sowie Zusammenhänge zwischen zwei und mehr Variablen aufzudecken versuchen. Anhand der Anzahl der jeweils betrachteten Variablen lassen sich uni-, bi- und multivariate Verfahren unterscheiden. Im Hinblick auf Datenvisualisierung sind nach ihrer Funktion Präsentationsgrafiken (stellen identifizierte statistische Befunde übersichtlich dar) von **Analysegrafiken** zu unterscheiden, die dem Auffinden von Effekten dienen und somit für die explorative Datenanalyse einschlägig sind (Chen, Härdle, & Unwin, 2008; Degen, 2010; Schnell, 1994; Unwin, Theus, & Hofmann, 2006).

Im ersten Schritt wird man bei einer quantitativen explorativen Studie die **interessierenden Variablen** einzeln genauer betrachten, indem man sich deren **(univariate) Häufigkeitsverteilungen** ausgeben, ihre zentralen Tendenzen (z. B. Modalwert bei nominalskalierten Variablen, Mittelwert bei kardinalskalierten Variablen) sowie die zugehörigen Dispersionsmaße berechnen lässt. Eine

absolute Häufigkeit	Stamm und Blätter
3	0. 589
6	1. 000139
11	2. 00000556667
15	3. 000000445566778
4	4. 0008
1	5. 0

Abb. 12.2 Stem-and-Leaf-Plot der Ergebnisse eines Mathematiktests (0–50 Punkte) von $n = 40$ Personen

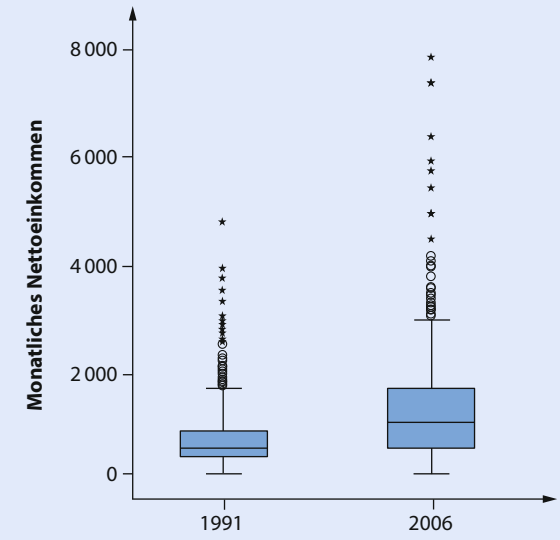


Abb. 12.3 Gruppierte Box-Plots zum monatlichen Nettoeinkommen in Euro aus dem ALLBUS 1991 und 2006. (Aus Degen, 2010, S. 98)

grafische Veranschaulichung – z. B. die Darstellung der Häufigkeitsverteilung einer Variable als **Histogramm** – ist dabei besonders hilfreich, um Auffälligkeiten im Datensatz zu entdecken (Bortz & Schuster, 2010, S. 41). Wird bei einem betrachteten kardinalskalierten Merkmal eine Normalverteilung erwartet, so kann bei der grafischen Datenanalyse eine Normalverteilungskurve über das Histogramm gelegt werden. **Stem-and-Leaf-Plots** (Stängel-Blatt-Diagramme) sind Histogramme, bei denen die Häufigkeit der einzelnen Merkmalsausprägungen nicht einfach durch die Höhe von „blanken" Balken veranschaulicht wird, sondern durch Balken, die mit entsprechenden Messwerten „gefüllt" sind. Ein Stem-and-Leaf-Plot enthält somit alle Messwerte in alphanumerischer Form und ordnet diese grafisch übersichtlich an. Dem Stem-and-Leaf-Plot ist zu entnehmen, welche Werte besonders häufig oder selten vertreten sind, wo das Zentrum der Verteilung liegt, ob sich die Werte in Subgruppen aufteilen, ob die Verteilung symmetrisch oder schief ist und wie stark die Werte streuen.

Messwerte jeden Skalenniveaus können als Stem-and-Leaf-Plot dargestellt werden. Der „Stamm" des Stem-and-Leaf-Plot ist die x-Achse und bildet die Merkmalskategorien ab, die „Blätter" sind die einzelnen, vom Stamm „abzweigenden" Messwerte innerhalb der Kategorien. Um die Lesbarkeit zu erleichtern, werden Stem-and-Leaf-Plots wie liegende Histogramme (d. h. mit senkrechter x-Achse) dargestellt (zum Stem-and-Leaf-Plot s. Bortz & Schuster, 2010, S. 43; Emerson & Hoaglin, 1983; Tukey, 1977, Kap. 1; Velleman & Hoaglin, 1981). Den Stem-and-Leaf-Plot der Ergebnisse eines Mathematiktests von $n = 40$ Personen, in dem 0–50 Punkte erreichbar waren, zeigt Abb. 12.2. Der Stängel des Stem-and-Leaf-Plots wird durch die Zehnerziffern, die durch einen Punkt abgetrennten Blätter werden durch die Einerziffern gebildet: Die Angabe „4. 0008" bedeutet somit, dass drei Fälle mit dem Wert 40 und ein Fall mit dem Wert 48 vorliegen. Zur Darstellung muss eine Schrift gewählt werden, bei der alle Ziffern dieselbe Druckbreite haben (also keine Proportionalschrift; Degen, 2010, S. 95). Anhand der Betrachtung des Stängel-Blatt-Diagramms kann z. B. das gehäufte oder fehlende Auftreten einzelner Werte auffallen wie im Beispiel die starke Präsenz „glatter" Ergebnisse (10, 20, 30, 40 oder 50 Punkte), die bei der Darstellung als Histogramm verborgen bliebe.

Die variablenweise Inspektion des Datensatzes wird meist aussagekräftiger, wenn jeweils nicht die komplette Stichprobe zugrunde gelegt wird, sondern diese in für den Untersuchungsgegenstand **relevante Gruppen aufgeteilt** wird. Die Verteilung eines Merkmals lässt sich dann gruppenweise betrachten und vergleichen. Gruppenunterschiede in der zentralen Tendenz eines Merkmals sind wahrscheinlich die am häufigsten betrachteten Effekte. Zu ihrer grafischen Veranschaulichung werden oft **Balkendiagramme** eingesetzt, bei denen die Gruppenmittelwerte durch die Höhe der Balken symbolisiert sind. Derartige Aggregatvergleiche sollten nur dann visualisiert werden, wenn die empirischen Messwertverteilungen eine weitgehend homogene Gruppenstruktur nahelegen. Eine einfache, optische Verteilungsprüfung ermöglichen sog. Box-Plots. **Box-Plots** (Box-and-Whisker-Plots) gehen auf Tukey (1977, S. 39ff.) zurück und stellen – vereinfacht gesagt – den Median, die mittleren 50 % der Werte (Interquartil-Bereich), die Streuung und die Ausreißer einer Verteilung dar (Abb. 12.3). Sie geben damit sowohl über die zentrale Tendenz als auch über die Verteilungsform in komprimierter Weise Auskunft (Bortz & Schuster,

2010, S. 44; Schnell, 1994, S. 18ff.). Abb. 12.3 zeigt die Boxplots der Einkommensverteilungen für 1991 und 2006 anhand der ALLBUS-Umfragedaten. Es wird deutlich, dass binnen 15 Jahren der Median des Einkommens gestiegen ist, die Schiefe im Interquartilbereich zurückgegangen ist (2006 liegen innerhalb der mittleren 50 % der Verteilung ähnlich viele Messwerte ober- und unterhalb des Medians) und die Streuung der Einkommenswerte sich stark vergrößert hat.

Häufig interessieren auch Zusammenhänge zwischen unterschiedlichen Variablen, wobei im einfachsten Fall zwei Variablen beteiligt sind **(bivariater Zusammenhang)**. Bei mehr als zwei Variablen spricht man von multivariaten Zusammenhängen. Ein **Scatter-Plot** (Streudiagramm) visualisiert den Zusammenhang zwischen zwei kardinalskalierten Variablen, indem auf der einen Achse des Koordinatensystems die Werte der einen Variablen, auf der anderen Achse die der anderen Variable abgetragen werden. Die Messwertpaare jedes Untersuchungsobjektes sind als Punkte im Scatter-Plot sichtbar. Dabei generiert die gesamte betrachtete Gruppe eine Punktewolke, deren Form über die Art des Zusammenhangs Auskunft gibt. Der bivariate Zusammenhang kann statistisch durch einen bivariaten Korrelationskoeffizienten sowie eine bivariate Regressionsgleichung (grafisch darstellbar als Regressionsgerade) dargestellt werden. Betrachtet man Scatter-Plots, so lässt sich jeweils die zugehörige Regressionsgerade anzeigen (Bortz & Schuster, 2010, S. 185ff.). Besonders wichtig bei der Betrachtung von Scatter-Plots ist auch die Frage, ob überhaupt ein linearer Zusammenhang vorliegt oder die Form der Punktwolke nicht eher auf einen nicht-linearen bzw. kurvilinearen Zusammenhang hinweist, etwa einen umgekehrt U-förmigen Zusammenhang, so dass entsprechend eine nicht-lineare Regression zu berechnen wäre.

Im **multivariaten Anwendungsfall** kann man mit einer **Scatter-Plot-Matrix** (Streudiagrammmatrix) arbeiten (Abb. 12.4). Hierbei werden für alle beteiligten Variablen bivariate Scatter-Plots gebildet. In Ergänzung zu einer Korrelationsmatrix, in der die bivariaten Korrelationskoeffizienten eingetragen sind, gibt die Scatter-Plot-Matrix auf einen Blick Auskunft über die Form der Zusammenhänge.

Die genannten Beispiele visueller explorativer Datenanalyse mittels Stem-and-Leaf-Plot, Box-Plot sowie der Scatter-Plot-Matrix mögen genügen, um zu verdeutlichen, dass die **optische Inspektion uni-, bi- oder multivariater Merkmalsverteilungen** erheblich mehr „Denkanstöße“ zur Hypothesenbildung vermitteln kann als rein numerische Deskriptionen.

Die Vielfalt bislang entwickelter Typen und **Varianten von Grafiken (Plots)**, die für EDA-Zwecke verwendet werden, ist beeindruckend; ihre Namen sind jedoch häufig

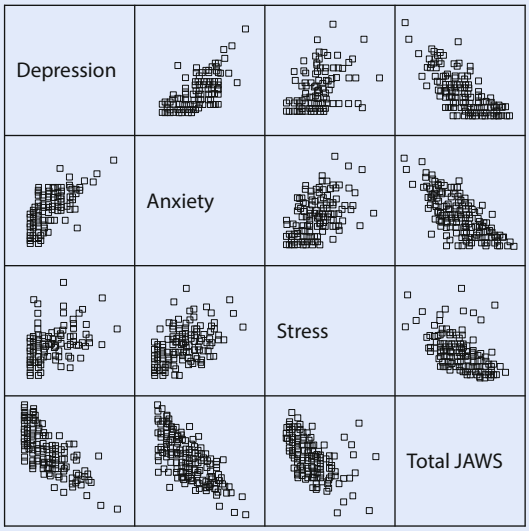


Abb. 12.4 Scatter-Plot-Matrix für Depression, Angst und Stress sowie emotionale Zufriedenheit im Beruf (JAWS: „job affective well-being scale“) bei $n = 274$ türkischen Ärztinnen und Ärzten. (Aus Uncu, Bayram, & Bilgel, 2007. Mit freundlicher Genehmigung von Oxford Journals)

eher verwirrend: Es gibt z. B. Q-Q-Plots und Jittered-Dot-Plots, Box-Dot-Plots, Kernel-Smoothed-Quantile-Plots, Coplots und Andrew-Plots, Poissonness-Plots und Voronoi-Plots. Genau wie bei inferenzstatistischen Verfahren taucht hier das Problem der **Indikation** auf: Welcher Plot ist im konkreten Fall besser, welcher schlechter geeignet, den relevanten Informationsgehalt der Daten zur Geltung zu bringen? Grundkenntnisse und eigene Erfahrungen in grafischer Datenanalyse sind hier unabdingbar. Für die vertiefte Einarbeitung in die Materie sei auf die Arbeiten von Tukey verwiesen (Hoaglin, Mosteller, & Tukey, 1983, 1985; Tukey, 1977) sowie auf zahlreiche Hand- und Lehrbücher und Übersichtsartikel (Behrens, 1997; Behrens & Yu, 2003; Chen et al., 2008; Cleveland, 1993; Degen, 2010; Oldenbürger, 1996; Polasek, 1994; Schnell, 1994; Unwin et al., 2006).

Die visuelle Datenanalyse dient nicht nur der Exploration, sondern wird auch im Zuge explanativer Studien eingesetzt, etwa zur Prüfung von Verteilungsvoraussetzungen inferenzstatistischer Verfahren (z. B. Analyse von Residualwerten).

Explorative multivariate Auswertungstechniken

Neben der grafischen Darstellung univariater Verteilungen sowie bivariater und multivariater Zusammenhänge werden im Rahmen der explorativen Datenanaly-

se auch einige multivariate statistische Auswertungsverfahren eingesetzt, die **datenreduzierend** bzw. **datenstrukturierend** wirken. Diese explorativen multivariaten Analysetechniken – explorative Faktorenanalyse, Clusteranalyse, multidimensionale Skalierung und multiple Regressionsanalyse – liefern meist mehrere Lösungsmöglichkeiten, von denen die Forschenden die am besten interpretierbare auswählen.

Explorative Faktorenanalyse Die explorative Faktorenanalyse („exploratory factor analysis", EFA) bündelt üblicherweise kardinalskalierte **Variablen** gemäß ihrer **Interkorrelationen** zu neuen übergeordneten bzw. abstrakteren Variablen, sog. **Faktoren**. Die exploratorische Faktorenanalyse verläuft dabei meist in zwei Schritten, für die jeweils unterschiedliche Berechnungsvarianten existieren: 1. Faktorextraktion (z. B. Hauptkomponentenanalyse) und 2. Faktorrotation (z. B. Varimax-Rotation; Bortz & Schuster, 2010, S. 405ff.; Wolff & Bacher, 2010).

Fasst die Faktorenanalyse z. B. einen Satz von 40 Variablen in 3 Faktoren zusammen, prüft man, welche Variablen zu einem Faktor gehören (d. h. hohe **Faktorladungen** auf dem Faktor aufweisen) und versucht zu ergründen, in welcher Hinsicht sich eben diese Variablen ähneln. Erfahrungsgemäß wirken faktorenanalytische Ergebnisse ausgesprochen inspirierend, d. h., es fällt meistens nicht schwer, diverse Hypothesen darüber zu generieren, was ein Faktor inhaltlich bedeutet bzw. was „hinter" den Variablen eines Faktors steht (zur Interpretationsproblematik faktorenanalytischer Ergebnisse s. Holz-Ebeling, 1995). Die Subjektivität der Interpretationen ist darin begründet, dass viele explorative faktorenanalytische Lösungen aus mathematischer Sicht gleichwertig sind, so dass entsprechend viele Interpretationen gleichberechtigt nebeneinander stehen, ohne dass objektive Entscheidungskriterien bestimmte Lösungen favorisieren. „Sinnvolle Interpretierbarkeit" ist ein wichtiges Entscheidungskriterium bei der Wahl des Faktorenmodells; aber was sich „sinnvoll interpretieren" lässt, hängt eben stark von der Perspektive der Deutenden ab (zur Überprüfung faktorieller Strukturhypothesen mittels konfirmatorischer Faktorenanalysen ▶ Abschn. 17.2.1).

Neben der Auswahl der Variablen hat auch die Auswahl der Untersuchungsteilnehmenden bzw. Untersuchungsobjekte entscheidenden Einfluss auf das Ergebnis. Nicht selten zeichnen sich **differenzielle Faktorenstrukturen** in unterschiedlichen Populationen bzw. Teilgruppen ab. Aus diesem Grund werden Clusteranalyse und Faktorenanalyse zuweilen auch nacheinander durchgeführt: Mittels Clusteranalyse identifiziert man zunächst homogene Teilgruppen, für die dann jeweils separat Faktorenanalysen zu rechnen sind. Dies ist natürlich nur möglich, wenn man eine entsprechend große Fallzahl bearbeitet, so dass sich Aufteilungen überhaupt lohnen. In jeder Teilgruppe muss die Anzahl der Untersuchungsobjekte größer sein als die der Variablen.

Studienbeispiel

Explorative Studie mit explorativer Faktorenanalyse

Im Rahmen einer bevölkerungsrepräsentativen Umfrage (ALLBUS) wurden die Bundesbürgerinnen und -bürger u. a. nach ihrem Vertrauen in elf verschiedene Institutionen (z. B. Gesundheitswesen, Zeitungswesen, Polizei) befragt. Der Grad des Vertrauens wurde jeweils auf einer 7-stufigen Ratingskala (1: überhaupt kein Vertrauen, 7: sehr großes Vertrauen) angegeben. Es sollte mittels explorativer Faktorenanalyse erkundet werden, ob und wie sich die einzelnen Vertrauensbekundungen bündeln lassen (dieses Beispiel stammt aus einem Methodenbeitrag von Wolff & Bacher, 2010).

Im Zuge der explorativen Untersuchung werden zunächst die deskriptivstatistischen Ergebnisse für jede der 11 Variablen betrachtet, nämlich Mittelwert und Streuung sowie die Korrelationen mit den anderen Variablen (◘ Abb. 12.5). Es zeigt sich, dass das Vertrauen in die katholische Kirche am vergleichsweise geringsten ($M = 2.99$) und das Vartrauen in die Hochschulen am höchsten ($M = 4.94$) ausfiel. Die Interkorrelationen waren durchgängig positiv und wiesen geringe ($r = +.12$: Zusammenhang zwischen Vertrauen in die evanglische Kirche und in das Gesundheitswesen) bis hohe ($r = +.70$: Zusammenhang zwischen Vertrauen in die evangelisch und in die katholische Kirche) Effektgrößen auf.

Die Faktorisierung der Korrelationsmatrix – Faktorextraktion mit anschließender Varimax-Rotation – ergab eine Drei-Faktoren-Lösung (◘ Abb. 12.6).

Betrachtet man, welche Variablen auf den Faktoren hoch laden, so lässt sich Faktor 1 als „Vertrauen in staatliche Institutionen", Faktor 2 als „Vertrauen in Kirchen" und Faktor 3 als „Vertrauen in Medien" interpretieren. Dass Faktor 3 als „Vertrauen in Medien" interpretiert und etikettiert wird, leitet sich daraus ab, dass auf Faktor 3 die Variablen Vertrauen in Fernsehen ($a = .78$) und Vertrauen in Zeitungswesen ($a = .77$) sehr hoch, die anderen Variablen dagegen deutlich geringer laden.

Die Faktorenanalyse arbeitet jeweils mit einer Korrelationsmatrix. SPSS berechnet diese anhand der Produkt-Moment-Korrelationskoeffizienten, so dass entweder kardinalskalierte oder dichotom-nominalskalierte Variablen einbezogen werden können (für ein ▶ Studienbeispiel „Explorative Studie mit explorativer Faktorenanalyse"). Polytom nominalskalierte oder ordinalskalierte Variablen sind jedoch auch faktorenanalytisch zu bündeln, indem man zunächst die für das jeweilige Skalenniveau passenden Korrelationskoeffizienten berechnet (▶ Abschn. 12.5.2) und die resultierende Korrelationsmatrix dann faktorisiert (Bühner, 2011, S. 343).

Clusteranalyse Die Clusteranalyse teilt die **Untersuchungsobjekte** anhand ihrer Merkmalsausprägungen in homogene Gruppen **(Cluster)** ein, die sich untereinander

	M	SD	1	2	3	4	5	6	7	8	9	10	11
1) Gesundheitswesen	4.34	1.44	1										
2) Bundesverfassungsgericht	4.82	1.47	.26	1									
3) Bundestag	3.73	1.41	.31	.52	1								
4) Kath. Kirche	2.99	1.79	.14	.18	.33	1							
5) Ev. Kirche	3.44	1.78	.12	.18	.32	.70	1						
6) Justiz	4.19	1.48	.29	.46	.46	.22	.24	1					
7) Fernsehen	3.34	1.46	.26	.10	.22	.15	.17	.24	1				
8) Zeitungswesen	3.78	1.37	.18	.15	.22	.15	.17	.25	.64	1			
9) Hochschulen	4.94	1.16	.27	.28	.27	.11	.12	.32	.26	.33	1		
10) Bundesregierung	3.61	1.43	.32	.38	.62	.23	.24	.44	.26	.26	.34	1	
11) Polizei	4.78	1.31	.29	.33	.35	.23	.25	.45	.23	.23	.34	.37	1

Anmerkungen: $N = 2\,400$, Angaben auf siebenstufiger Likert-Skala mit 1 = überhaupt kein Vertrauen, 7 = sehr großes Vertrauen. Alle Korrelationen sind signifikant mit $p < .01$

Abb. 12.5 Interkorrelationsmatrix von 11 Items zum Vertrauen in Institutionen. (Aus Wolff & Bacher, 2010, S. 364)

	1	2	3
1) Gesundheitswesen	**.42**	.05	.19
2) Bundesverfassungsgericht	**.64**	.08	–.01
3) Bundestag	**.71**	.24	.08
4) Kath. Kirche	.18	**.82**	.07
5) Ev. Kirche	.18	**.81**	.09
6) Justiz	**.64**	.13	.15
7) Fernsehen	.17	.08	**.78**
8) Zeitungswesen	.19	.08	**.77**
9) Hochschulen	**.43**	.02	.30
10) Bundesregierung	**.66**	.14	.17
11) Polizei	**.53**	.15	.18

Abb. 12.6 Vertrauen in Institutionen, varimaxrotierte Faktorladungsmatrix. (Aus Wolff & Bacher, 2010, S. 359; Hervorhebungen hinzugefügt)

Studienbeispiel

Explorative Studie mit Clusteranalyse

In einer gesundheitswissenschaftlichen Studie zu Rückenbeschwerden wurden 15 Variablen zu beruflichen und häuslichen Tätigkeiten und psychosozialen Lebensbedingungen erhoben. Auf der Basis dieser Variablen wurden die $n = 1\,341$ schwedischen Untersuchungspersonen explorativ in 11 Cluster eingeteilt, deren differenzielle Betroffenheit von Rückenbeschwerden betrachtet wurde (Leijon, Härenstam, Waldenström, Alderling, & Vingård, 2006). Im Ergebnis zeigte sich, dass vor allem drei Typen von Personen verstärkt von Rückenschmerzen betroffen waren: Personen, die im Servicebereich beschäftigt sind (vor allem Frauen), die selbstständig tätig sind (vor allem Männer) und die eine hohe familiäre Arbeitsbelastung haben (vor allem Frauen). Die Typenbildung kann dabei helfen, Präventionskampagnen gegen Rückenschmerzen zielgruppenspezifisch auszurichten.

Studienbeispiel

Explorative Studie mit multidimensionaler Skalierung

In einer Studie in den USA wurden 10 verschiedene Hamburger-Schnellrestaurants von $n = 66$ Studierenden hinsichtlich ihrer Ähnlichkeit auf einer Skala von 0 (sehr ähnlich) bis 10 (sehr unähnlich) eingeschätzt (Mano & Davis, 1990). Eine Auswertung mittels multidimensionaler Skalierung zeigte, dass sich der Wahrnehmungsraum auf zwei Dimensionen abbilden lässt, von denen die erste als Prototypizität der Produkte und die zweite als wahrgenommene Qualität der Fastfood-Restaurants interpretiert wurde (◘ Abb. 12.7).

Studienbeispiel

Explorative Studie mit multipler Regressionsanalyse

In einer explorativen Tourismusstudie in der französischen Region Champagne wurden Touristen zu ihren Reiseerfahrungen, soziodemografischen Variablen sowie dazu befragt, wie viel Geld sie vor Ort auf Weingütern beim Kauf von Champagner ausgeben würden. Mittels multipler Regressionsanalyse (Einschlussmethode und induktiv-schrittweise Rückwärtstechnik) wurde ermittelt, durch welche von 13 Merkmalen der Touristen (z. B. Einkommen, Nationalität, Geschlecht, Bewertung der lokalen Restaurants, Anzahl der Wiederholungskäufe) der akzeptierte Kaufpreis für eine Flasche Champagner bestimmt wird (Charters & Menival, 2011).

maximal unterscheiden (für ein ▶ Studienbeispiel „Explorative Studie mit Clusteranalyse“). In Abhängigkeit vom Skalenniveau der Variablen sowie den Prinzipien der Clusterzusammenfassung liegen sehr viele unterschiedliche Varianten der Clusteranalyse vor (Bortz & Schuster, 2010, S. 453f.). Eine Clusterbildung ist nur möglich, wenn alle Variablen auf demselben Skalenniveau operationalisiert sind, so dass vor der Berechnung möglicherweise eine Skalentransformation notwendig ist (▶ Abschn. 8.4.6). Beachte: Empirische Klassifikationen sind stets abhängig von der Art der untersuchten Objekte und Variablen. Es ist also nicht auszuschließen, dass sich Objekte anders gruppieren, wenn ein anderer Satz von Beschreibungsmerkmalen verwendet wird. In die Auswahl der Variablen fließen mehr oder minder bewusst reflektierte theoretische Vorannahmen ein, so dass man auch bei explorativer Vorgehensweise niemals ganz „bei Null“ anfängt.

Multidimensionale Skalierung Die Multidimensionale Skalierung (MDS) ist anwendbar, wenn mehrere **Urteilsobjekte** (z. B. Berufe, Automarken, Prominente etc.) von einer Stichprobe von Personen hinsichtlich ihrer Ähnlichkeit oder Unähnlichkeit auf einer Ratingskala eingestuft wurden (für ein ▶ Studienbeispiel „Explorative Studie mit multidimensionaler Skalierung“). Die Multidimensionale Skalierung stellt die Urteilsobjekte dann in einem **2- oder 3-dimensionalen Wahrnehmungsraum** dar, wobei die Ähnlichkeitsurteile als Distanzen der Objekte abgebildet werden (Borg & Groenen, 2005; Mathar, 1997; Torgerson, 1958). Die zwei oder drei Dimensionen der Konfiguration lassen sich inhaltlich interpretieren und liefern dadurch Hinweise auf die Wahrnehmung der Objekte durch die untersuchte Stichprobe. Das Verfahren wird u. a. in der Marktforschung eingesetzt. Werden die Ähnlichkeitsurteile nicht kardinalskaliert erhoben, sondern nur ordinalskaliert, so ist die nonmetrische multidimensionale Skalierung (NMDS) indiziert (▶ Webangebot zu diesem Buch).

Multiple Korrelations- und Regressionsanalyse Die multiple Korrelations- und Regressionsanalyse bestimmt, wie gut die Ausprägungen einer Zielvariable **(Kriteriumsvariable)** durch einen Satz erklärender Variablen **(Prädiktorvariablen)** vorhersagbar sind (für ein ▶ Studienbeispiel „Explorative Studie mit multipler Regressionsanalyse“). Für alle Prädiktorvariablen wird berechnet, wie eng und in welcher Richtung sie – im Kontext der anderen betrachteten Prädiktorvariablen – mit dem Kriterium zusammenhängen. Der multiple Gesamtzusammenhang (multipler Korrelationskoeffizient R; multipler Determinationskoeffizient R^2) sowie die Einflüsse der einzelnen Prädiktoren, die durch β-Gewichte ausgedrückt werden, lassen sich auf Signifikanz prüfen. Während bei einer theorieprüfenden Vorgehensweise vorgegeben wird, welche Prädiktoren wie stark und in welcher Richtung wirken (theorieprüfende hierarchische Regressionsanalyse), betrachtet man beim explorativen Vorgehen induktiv-empirisch, welche Prädiktoren für die Vorhersage besonders wichtig oder unwichtig sind (z. B.

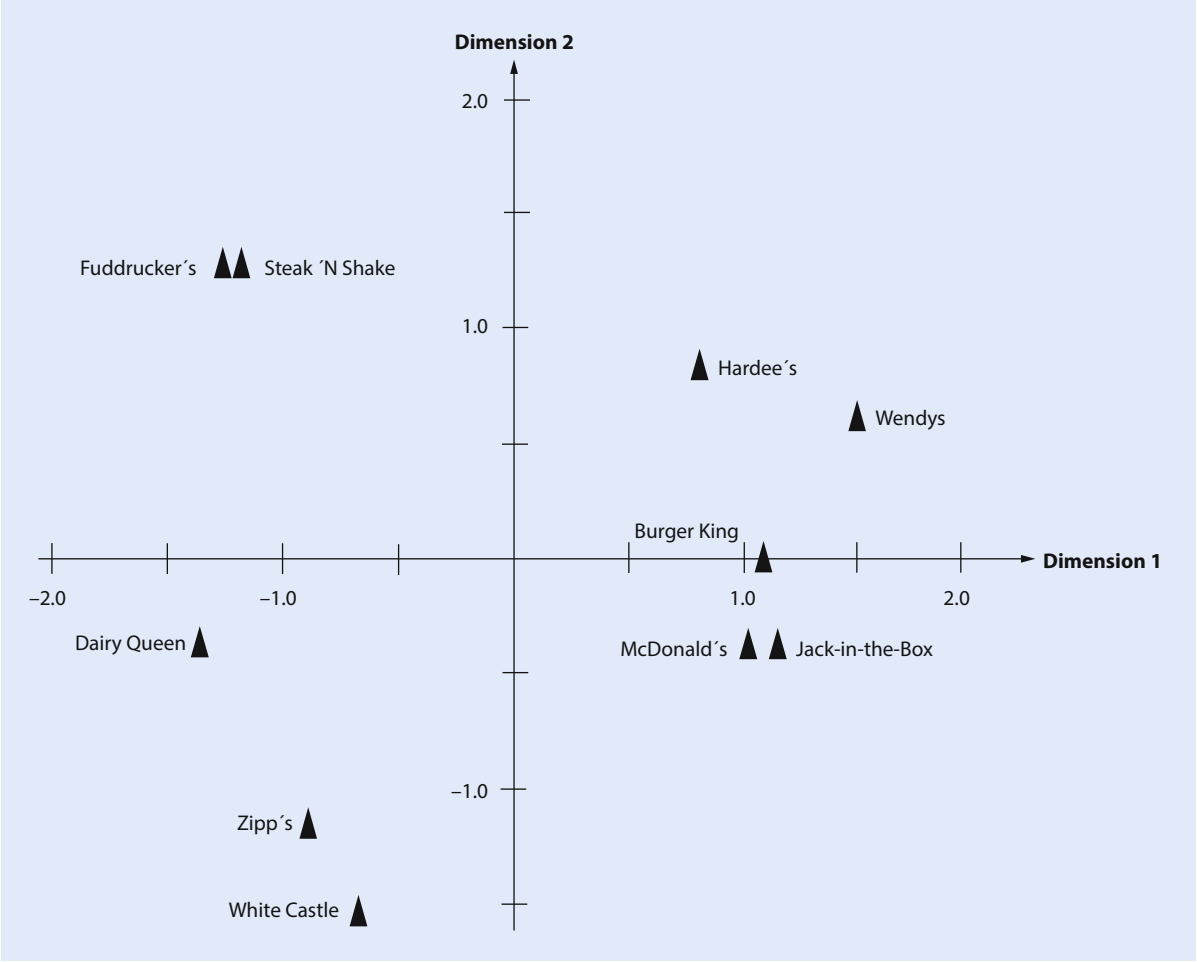


Abb. 12.7 Beispiel einer multidimensionalen Skalierung von Fastfood-Restaurants. Dimension 1: Prototypizität der Produkte, Dimension 2: Wahrgenommene Qualität der Fastfood-Restaurants. (Aus Mano & Davis, 1990. ADVANCES IN CONSUMER RESEARCH by ASSOCIATION FOR CONSUMER RESARCH (U.S.) Reproduced with permission of ASSOCIATION FOR CONSUMER RESEARCH, in the format Republish in a book via Copyright Clearance Center)

mit der Einschlussmethode oder der induktiven schrittweisen Regressionsanalyse). Zur multiplen Korrelations- und Regressionsanalyse und ihrem Einsatz zur Hypothesenprüfung ► Abschn. 12.5.2.

12.3.2 Inferenzstatistische Überprüfung von Forschungsfragen

Explorative Studien testen keine theoretisch abgeleiteten Hypothesen, sondern zielen auf die Bildung neuer Hypothesen ab. Dazu werden offene Forschungsfragen formuliert und anhand des Datensatzes beantwortet. Solche offenen Forschungsfragen beziehen sich z. B. darauf, ob zwischen zwei oder mehr Gruppen Unterschiede bestehen, ob es Zusammenhänge in einem Satz von Variablen gibt oder ob Veränderungen in den Variablenausprägungen über die Zeit feststellbar sind. Hier werden ergänzend zur deskriptivstatistischen Darstellung der Ergebnisse oft auch statistische Signifikanztests berechnet, um die gefundene Effekte gegen den Zufall abzusichern. Dass ein Signifikanztest gerechnet wird, besagt dabei keineswegs automatisch, dass es sich auch um einen Hypothesentest handelt, denn dieser liegt nur dann vor, wenn die getesteten Hypothesen **vor** der Datenerhebung formuliert wurden (sog. **A-priori-Hypothesen**) und somit ein bestimmtes Ergebnis vorhersagen. Ein **„Signifikanztest auf Probe“** für einen Effekt, der nicht theoretisch vorhergesagt, sondern im Datensatz gefunden wurde, darf nicht als Bestätigung der Hypothese gelten, sondern er dient – sofern er ein signifikantes Ergebnis liefert – lediglich als Grundlage für die Formulierung von A-priori-Hypothesen für zukünftige Studien (vgl. hierzu auch Oldenbürger, 1996, S. 72).

Von dieser Form des „Signifikanztests auf Probe“, der ausdrücklich dazu verwendet wird, neue A-priori-Hypothesen für zukünftige Studien zu bilden, sind Pseudo-Hypothesentests zu unterscheiden (► Studienbeispiel „‚Signifikanztest auf Probe‘ und Ex-post-Erklärung bei einer explorativen Studie“). Es handelt sich um einen **Pseudo-Hypothesentest**, wenn man in einem Datensatz nach signifikanten Effekten sucht, zu diesen nachträglich Hypothesen und theoretische Erklärungen konstruiert und somit eine eigentlich explorative Studie dann als hypothesenprüfende Studie darstellt. Diese Vorgehensweise widerspricht wissenschaftstheoretischen Regeln des Erkenntnisgewinns und ist zudem wissenschaftsethisch als Manipulation zu verurteilen (► Abschn. 4.2.2). Pseudo-Hypothesentests werden auch als **HARKing** („hypothesizing after the effects are known“) (Kerr, 1998) sowie als **Signifikanzen-Fischen** („fishing for significance“, „*p*-value fishing“) bezeichnet. Sucht man lange genug in einem Datensatz, so finden sich immer signifikante Effekte und diese lassen sich im Nachgang auch immer irgendwie plausibilisieren, als Beleg für die Gültigkeit von Hypothesen und Theorien kann dieses Vorgehen aber nicht dienen.

Studienbeispiel

„Signifikanztest auf Probe“ und Ex-post-Erklärung bei einer explorativen Studie

Angenommen, im Zuge einer explorativen Untersuchung von $n = 20$ Einzelkindern und $n = 20$ Geschwisterkindern findet man im Datensatz einen Effekt in dem Sinne, dass die untersuchten Einzelkinder eine geringere soziale Kompetenz aufweisen als die Geschwisterkinder. Dieser – nicht durch eine Hypothese vorhergesagte, sondern induktiv gefundene – Effekt möge sich als statistisch signifikant herausstellen. Man mag nun im Nachgang spekulieren, dass Kinder, die ohne Geschwister aufwachsen, möglicherweise weniger soziale Fähigkeiten erwerben als Kinder, die mit Geschwistern großwerden. Als **„Signifikanztest auf Probe“** und als ausdrücklich spekulative **Ex-post-Erklärung** dürfen diese Ergebnisse einer explorativen Studie so berichtet werden.

Nicht vertretbar wäre es jedoch, würde man im Ergebnisbericht nachträglich eine Hypothese einfügen, dergemäß Einzelkinder weniger sozial kompetent sind als Geschwisterkinder und dann – wissenschaftstheoretisch falsch und wissenschaftsethisch unzulässig – angeben, diese Hypothese sei durch die Studie bestätigt worden. Wenn man nämlich das empirische Ergebnis nicht im Vorfeld durch eine gerichtete Hypothese korrekt vorhersagen konnte, sind nachträgliche Erklärungen sehr beliebig. Hätte die explorative Studie umgekehrt gezeigt, dass die Einzelkinder über mehr Sozialkompetenz verfügen als die Geschwisterkinder, so hätte man auch dies ex-post plausibel erklären können, etwa damit, dass Einzelkinder, die darauf angewiesen sind, sich Spielkameraden außerhalb des eigenen Zuhause zu suchen, deswegen bessere soziale Fähigketen entwickeln. Im Nachhinein gelingt es mühelos, so gut wie jedes Ergebnis plausibel zu erklären. Nur wenn sich wohlbegründete A-priori-Hypothesen empirisch bestätigen lassen, gilt dies im wissenschaftstheoretischen Verständnis des Kritischen Rationalismus als Hypothesentest (► Abschn. 2.2.9).

12

12.3.3 Data-Mining

Das Auffinden von Mustern und Zusammenhängen in sehr großen, typischerweise in elektronischen Datenbanken verwalteten Datensätzen, nennt man **Data-Mining** oder auch **„knowledge discovery in databases“** KDD (Berry & Linoff, 2000). Hier geht es darum, entweder im Forschungsprozess generierte Daten (z. B. umfangreiche physiologische Messwertreihen: Dolce, Riganello, Quintieri, Candelieri, & Conforti, 2008) sowie v. a. auch vorgefundenes bzw. in der computerisierten Gesellschaft im Alltag in wachsender Menge anfallendes quantitatives Datenmaterial (Kundendaten von Unternehmen) zur Hypothesengewinnung zu nutzen. Werden etwa in einer Kaufhauskette alle Einkäufe über das elektronische Kassensystem registriert und die entsprechenden Verkaufsdaten in eine Datenbank eingespielt, so kann man mittels Data-Mining typische Kaufmuster identifizieren. Durch Mobilfunkdaten lassen sich geografische Bewegungsmuster der Kunden herausarbeiten. Ein großes Anwendungsfeld des Data-Mining sind dementsprechend die Datenbanken von Unternehmen, die sich durch eine explorative Analyse neue Erkenntnisse über ihre Kundschaft versprechen. Somit stellen sich im Zuge des Data-Minings mit personalisierten Kundendaten auch forschungsethische Fragen der Anonymisierung und Vertraulichkeit (Payne & Trumbach, 2009; zur Forschungsethik ► Abschn. 4.1).

Die rapide wachsenden Datenmengen einerseits und die gesteigerten Verarbeitungskapazitäten durch leistungsfähige Hard- und Software andererseits führten in den letzten Jahren zu einem Boom des Data-Minings, bei dem statistische Analysen, komplexe Datenbankabfragen sowie Methoden der künstlichen Intelligenz (z. B. neuronale Netzwerke) zum Einsatz kommen. Beim sog. **prädiktiven Data-Mining** wird in drei Schritten vorgegangen:

1. zunächst erfolgt eine explorative Datenanalyse des Materials, um relevante Variablen herauszufiltern,
2. anschließend wird ein Modell entwickelt und validiert, mit dem die Untersuchungsobjekte (z. B. Kunden) hinsichtlich ihrer Merkmalsausprägungen auf verschiedenen Variablen in Gruppen eingeteilt werden,
3. abschließend wird das Modell auf einen neuen Datensatz angewendet (prädiktives Data-Mining).

Sozialwissenschaftlich relevantes Datenmaterial entsteht in großem Stil etwa auch im Zuge der Nutzung von Computernetzwerken: Nach welchen Stichworten Perso-

nen wann und wie oft das Web durchforsten, welchen Links sie folgen oder nicht folgen, wie lange sie auf einer Seite verweilen etc., derartige Informationen fallen im Sinne nonreaktiver bzw. automatischer Beobachtung in Form von Logfiles (► Abschn. 10.1.3) an und lassen sich via Data-Mining auswerten (**„web usage mining“**). Die Entdeckung von Regelläufigkeiten im protokollierten Webnutzungsverhalten könnte etwa Hinweise auf das Informationsbedürfnis oder die Medienkompetenz liefern. Auch Web-Inhalte (**„web content mining“**) oder Verlinkungen zwischen Websites (**„web structure mining“**) lassen sich mittels Data-Mining – bzw. konkreter: Web-Mining – erkunden.

Man beachte, dass die Metapher des Data-Minings problematisch ist, wenn sie nahelegt, allein durch maschinelle Intelligenz sei es nun möglich, in den weltweit vorliegenden quantitativen Datenbeständen die Erkenntnisse einfach wie Edelmetalle abzubauen. Tatsächlich werden interessante Muster im Datensatz, sofern sie überhaupt gefunden werden, nur dann zum wissenschaftlich oder praktisch verwertbaren „Wissen“, wenn sie zu Hypothesen verdichtet und diese dann zu Theorien weiterentwickelt werden (MacKay, 1993, für eine Gegenüberstellung von theoriebasiertem und empiriebasiertem Erkenntniszugewinn). Bislang ist das Feld des Data-Minings stark technikgetrieben im Sinne der Entwicklung neuer Algorithmen. Dementsprechend wird eine stärkere Orientierung an relevanten Forschungsfragen gefordert (Pechenizkiy, Puuronen, Tsymbala, & Cunningham, 2008). Einschlägige Informationen zum Data-Mining – von Begriffsdefinitionen bis zu Auswertungstools – findet man z. B. bei http://www.kdnuggets.com. Werden vorhandene quantitative Datenbanken aus dem Bereich der Sozialarbeit ausgewertet, so spricht man von **„clinical data mining“** (Epstein, 2009), einem Ansatz, der v. a. auch der Evaluation der Sozialarbeit dient.

Unter dem Oberbegriff **Big Data** wird heute die Sammlung, Archivierung und Analyse sehr großer Datenmengen verstanden, die durch automatische Datenaufzeichnung entstehen. Neben einer Analyse zu wirtschaftlichen Zwecken, wird der aktuellen und zukünftigen Nutzung von Big Data für den (sozial)wissenschaftlichen Erkenntnisgewinn sehr viel Bedeutung beigemessen. Erste wissenschaftliche Fachzeitschriften zu Big Data wurden im Jahr 2013 gegründet, etwa *Big Data* (Mary Ann Liebert Publishers) oder das *Journal of Big Data* (Springer). Die Auswertung von Big Data kann der Bildung von Hypothesen, aber auch der Populationsbeschreibung oder Hypothesenprüfung dienen. Viele Fragen nach sinnvollen sozialwissenschaftlichen Forschungsstrategien, neuen Aspekten der Forschungsethik sowie notwendigen Software-Tools für die Datenanalyse sind im Zusammenhang mit Big Data aktuell noch offen.

Lernquiz

1. Was ist der Unterschied zwischen der explorativen Datenanalyse (EDA) und der konfirmativen Datenanalyse (CDA)?
2. Welche Techniken werden im Rahmen der explorativen Datenanalyse eingesetzt?
3. Welche Werte können Boxplots entnommen werden?
4. Grenzen Sie die Zielsetzungen einer explorativen Faktorenanalye und einer Clusteranalye voneinander ab.
5. Was unterscheidet einen „Signifikanztest auf Probe“ von einem „Pseudo-Signifikanztest“?
6. Welches Hauptmerkmal kennzeichnet eine A-priori-Hypothese?
7. Handelt es sich bei Auswertungen mit dem Tool Google „Trends“ (http://www.google.com/trends/) um „web usage mining“? (Begründung)
8. Was versteht man unter Big Data?
9. Wozu nutzt man die Multidimensionale Skalierung?
10. Welcher Statistiker hat die explorative Datenanalyse namhaft propagiert?

Übungsaufgaben

Analysieren Sie den quantitativen Datensatz ALLBUS 2010 (zum Download bereitgestellt unter http://www.gesis.org/allbus/allgemeine-informationen/aktuelles-allbus-2010/) mittels Visualisierungstechniken:

1. Stellen Sie Variable „Allgemeine Lebenszufriedenheit“ (V749) gegliedert nach dem Schulabschluss der Befragten (V327) mittels Boxplots dar!
2. Lassen Sie sich das Histogramm für die Variable „Wirtschaftslage in der BRD heute“ (V79) anzeigen!
3. Analysieren Sie den Zusammenhang zwischen den Variablen „Allgemeine Lebenszufriedenheit (V749) und „Ausländer: Politische Betätigung untersagen“ (V320) mittels Scatter-Plot.

Weiterführende Literatur

- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Heidelberg: Springer.
- Hoaglin, D. C., Tukey, J. & Mosteller, F. (2000). *Understanding robust and exploratory data analysis*. New York: Wiley.

- Myatt, G. J. (2006). *Making sense of data: A practical guide to exploratory Data analysis and data mining*. New York: Wiley.
- Polasek, W. (1994). *EDA Explorative Datenanalyse: Einführung in die deskriptive Statistik*. Heidelberg: Springer.
- Schnell, R. (1994). *Graphisch gestützte Datenanalyse*. München, Wien: Oldenbourg.

Literatur

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160.

Behrens, J. T. & Yu, C.-H. (2003). Exploratory Data Analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology – Volume 2: Research methods in psychology* (Vol. 2, pp. 33–64). New Jersey: Wiley.

Berry, M. J. A. & Linoff, G. (2000). *Mastering data mining, The art and science of customer relationship management*. New York: Wiley.

Borg, I. & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Ed.). Berlin, Heidelberg: Springer

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (Vol. 3). München: Pearson.

Charters, S. & Menival, D. (2011). Wine tourism in Champagne. *Journal of Hospitality and Tourism, 35*(1), 102–118.

Chen, C., Härdle, W., & Unwin, A. (Eds.). (2008). *Handbook of data visualization*. Heidelberg: Springer.

Cleveland, W. S. (1993). *Visualizing data*. Summit: Hobart.

Degen, H. (2010). Graphische Datenexploration. In C. Wolf & H. Best (Hrsg.), *Handbuch sozialwissenschaftliche Datenanaylse Teil III* (S. 91–116). Wiesbaden: VS.

Dolce, G., Riganello, F., Quintieri, M., Candelieri, A., & Conforti, D. (2008). Personal interaction in the vegetative state: A data-mining study. *Journal of Psychophysiology, 22*(3), 150–156.

Emerson, J. D. & Hoaglin, D. (1983). Steam-and-Leaf-Displays. In D. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 1–32). New York: Wiley.

Epstein, I. (2009). Promoting harmony where there is commonly conflict: Evidence-informed practice as an integrative strategy. *Social Work in Health Care, 48*(3), 216–231.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1985). *Exploring data,tables, trends, and shapes.* New York: Wiley.

Holz-Ebeling, F. (1995). Faktorenanalyse und was dann? Zur Frage der Validität von Dimensionsinterpretationen. *Psychologische Rundschau, 46*, 18–35.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychological Review, 2*, 196–217.

Leijon, O., Härenstam, A., Waldenström, K., Alderling, M., & Vingård, E. (2006). Target groups for prevention of neck/shoulder and low back disorders: An exploratory cluster analysis of working and living conditions. *Work: Journal of Prevention, Assessment and Rehabilitation, 27*(2), 189–204.

MacKay, D. G. (1993). The theoretical epistemology: A new perspective on some long-standing methodological issues in psychology. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological Issues* (pp. 229–255). Hillsdale: Erlbaum.

Mano, H. & Davis, S. M. (1990). The effects of familiarity on cognitive maps. *Advances in Consumer Research, 17*, 275–282.

Mathar, R. (1997). *Multidimensionale Skalierung: Mathematische Grundlagen und algorithmische Aspekte*. Stuttgart: Teubner.

Oldenbürger, H. A. (1996). Exploratorische, graphische und robuste Datenanalyse. In E. E., R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 71–86). Weinheim: Beltz.

Payne, D. & Trumbach, C. C. (2009). Data mining: proprietary rights, people and proposals. *Business Ethics: A European Review, 18*(3), 241–252.

Pechenizkiy, M., Puuronen, S., Tsymbala, A., & Cunningham, P. (2008). Dynamic integration of classifiers for handling concept drift. *Information Fusion, 9*(1), 56–68.

Polasek, W. (1994). *EDA Explorative Datenanalyse: Einführung in die deskriptive Statistik*. Heidelberg: Springer.

Schnell, R. (1994). *Graphisch gestützte Datenanalyse*. München, Wien: Oldenbourg.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading/MA: Addison-Wesley.

Uncu, Y., Bayram, N., & Bilgel, N. (2007). Job related affective well-being among primary health care physicians. *European Journal of Public Health, 17* (5), 514–519.

Unwin, A., Theus, M. & Hofmann, H. (2006). *Graphics of large datasets. Visualizing a Million.* Heidelberg: Springer.

Velleman, P. F. & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston, Massachusetts: Duxbury.

Wolff, H-G. & Bacher, J. (2010). Hauptkomponentenanalyse und explorative Faktorenanalyse. In C. Wolf & H. Best (Hrsg.), *Handbuch sozialwissenschaftliche Datenanalyse, Teil III* (S. 333–365). Wiesbaden: VS.

12.4 Quantitative Datenanalyse bei deskriptiven Studien

Lernziele

- Wissen, welche Zielsetzung deskriptive quantitative Studien haben und welche Datenanalyseverfahren hierfür indiziert sind.
- Das Prinzip von Punktschätzungen verstehen und erläutern können.
- Die Logik von Intervallschätzungen kennen.
- Konfidenzintervalle für Mittelwerte und Anteilswerte berechnen können.

Deskriptive quantitative Studien zielen darauf ab, Aussagen über die Verteilung oder Ausprägung bestimmter Merkmale in einer interessierenden Zielpopulation zu treffen. Dafür ist es notwendig, eine für die Zielpopulation möglichst **repräsentative Stichprobe** zu ziehen und die interessierenden Merkmale mit möglichst reliablen und validen Messinstrumenten (meist psychometrische Fragebögen oder psychologische Tests) zu erfassen. Anhand der Stichprobenkennwerte (z. B. Mittelwerte oder Anteils-/Prozentwerte der interessierenden Merkmale in der Stichprobe) werden dann die zugehörigen Populationsparameter geschätzt. Als Schätzverfahren stehen dafür **Punktschätzungen** (ein einzelner Schätzwert wird bestimmt) oder **Intervallschätzungen** (ein Wertebereich – das sog. Konfidenzintervall – für den gesuchten Populationsparameter wird berechnet) zur Verfügung (Tab. 12.7; ausführliche Erklärungen sind den folgenden Abschnitten zu entnehmen).

Liegt z. B. von der interessierenden Variable ein Stichprobenmittelwert M bzw. $\bar{x}$ (sprich: x quer) aus einer möglichst repräsentativen Stichprobe vor, so kann dieser als Punktschätzer für den unbekannten Populationsparameter μ (my) genutzt werden: $\hat{\mu}$ ($\hat{\mu}$ sprich: my Dach) $= \bar{x}$. Ebenso kann aus dem Stichprobenmittelwert ein Konfidenzintervall Δ_{krit} (sprich: Delta krit) berechnet werden, das einen Wertebereich für den gesuchten Populationsparameter angibt.

Um aussagekräftige Schätzwerte berechnen zu können, sollte schon bei der Untersuchungsplanung der optimale Stichprobenumfang (▶ Abschn. 14.3) für das Schätzproblem bestimmt und umgesetzt werden. Eine Präzisierung von Parameterschätzungen ist zudem möglich, wenn spezifische Merkmale der Stichprobe berücksichtigt werden (z. B. geschichtete Stichprobe oder Klumpenstichprobe).

12.4.1 Punktschätzung von Populationsparametern

In empirischen Studien untersuchen wir in der Regel diejenige Population, über die wir Aussagen treffen wollen, nicht vollständig (Vollerhebung), sondern arbeiten mit Ausschnitten (Stichproben). Dabei werden **probabilistische Stichproben**, die nach echten Zufallsprinzipien aus der Population gezogen werden und diese – bei entsprechender Mindestgröße des Samples – relativ gut abbilden (hohe Repräsentativität) von nicht-probabilistischen Stichproben, die nicht-zufällig gezogen werden und somit mehr oder minder starke Verzerrungen im Vergleich zur Population aufweisen (geringe Repräsentativität) unterschieden (zu Stichproben-Typen ▶ Kap. 9). Wenn wir uns im Rahmen **deskriptiver** bzw. **populationsbeschreibender Studien** dafür interessieren, wie einzelne Merkmale in der Population (z. B. in der Gesamtbevölkerung eines Landes, in einer Berufsgruppe oder in einer Altersgruppe) ausgeprägt sind, stellt sich die **inferenzstatistische Frage**, wie von Daten bzw. Kennwerten aus einer möglichst repräsentativen (d. h. probabilistischen) Stichprobe (z. B. Stichprobenmittelwert) zuverlässig auf die unbekannten Populationsparameter (z. B. Populationsmittelwert) zu schließen ist.

Eine Methode dieses Rückschlusses stellen die **inferferenzstatistischen Verfahren der Punktschätzung** dar, die angeben, wie die Stichprobendaten zu einem einzelnen Schätzwert des gesuchten unbekannten Populationsparameters zu verrechnen sind. Dafür stehen eine Reihe **unterschiedlicher Schätzmethoden** zur Verfügung. Wichtige klassische Schätzmethoden sind a) die Momenten-Methode, b) die Kleinstquadrat-Methode und c) die Maximum-Likelihood-Methode. Hinzu kommen Verfahren der Punktschätzung nach Resampling- bzw. Bootstrap-Verfahren sowie nach dem Bayes-Ansatz. Die Verfahren führen zu Punktschätzungen, die sich – je nach Datenlage – teilweise in ihrer Aussagekraft unterscheiden. Die **Qualität eines Punktschätzers** wird über die Kriterien Erwartungstreue, Konsistenz, Effizienz, Suffizienz und Robustheit definiert.

Punktschätzung – Bei einer Punktschätzung („point estimation") wird ein unbekannter Populationsparameter mittels eines einzelnen Stichprobenkennwertes (Punktschätzer, „point estimator") geschätzt. Die Qualität einer Punktschätzung wird über die Kriterien Erwartungstreue, Konsistenz, Effizienz, Suffizienz und Robustheit ermittelt. Wichtige klassische Schätzmethoden sind a) die Momenten-Methode, b) die Kleinstquadrat-Methode und c) die Maximum-Likelihood-Methode. Hinzu kommen Punktschätzungen mit Resampling- sowie Bayes-Verfahren.

Tabelle 12.7 Punkt- und Intervallschätzung für Mittelwert und Anteilswert

	Populations-parameter	Stichproben-kennwert	Punktschätzung des Populations-parameters durch den Stichproben-kennwert	Intervallschätzung des Populationsparameters durch Konfidenzintervalle
Arithmetischer Mittelwert	μ (my)	$M = \bar{x}$	$\hat{\mu} = \bar{x}$	Bei bekannter Populationsvarianz: $\Delta_{krit(1-\alpha)} = \bar{x} \pm z_{(\frac{\alpha}{2})} \cdot \sigma_{\bar{x}}$ mit $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$ Bei unbekannter Populationsvarianz und $n \geq 30$: $\Delta_{krit(1-\alpha)} = \bar{x} \pm z_{(\frac{\alpha}{2})} \cdot \hat{\sigma}_{\bar{x}}$ mit $\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}^2}{n}}$ Bei unbekannter Populationsvarianz und $n < 30$ (Voraussetzung: normalverteilte Population): $\Delta_{krit(1-\alpha)} = \bar{x} \pm t_{(\frac{\alpha}{2}, df)} \cdot \hat{\sigma}_{\bar{x}}$ mit $df = n - 1$ und $\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}^2}{n}}$
Anteilswert	π (pi)	p	$\hat{\pi} = p$	$\Delta_{krit(1-\alpha)} = p \pm z_{(\frac{\alpha}{2})} \cdot \hat{\sigma}_p$ mit $\hat{\sigma}_p = \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}} = \sqrt{\frac{p \cdot (p-1)}{n}}$ Eine präzisere Schätzung ermöglicht diese aufwändigere Formel: $\Delta_{krit(1-\alpha)} = \frac{n}{n + z^2_{(\frac{\alpha}{2})}} \cdot \left[p + \frac{z^2_{(\frac{\alpha}{2})}}{2 \cdot n} \pm z_{(\frac{\alpha}{2})} \cdot \sqrt{\frac{p \cdot (1-p)}{n} + \frac{z^2_{(\frac{\alpha}{2})}}{4 \cdot n^2}} \right]$

12

Im Folgenden werden wir das **Prinzip der Punktschätzung** an einer einfachen Zufallsstichprobe erläutern, die **Gütekriterien für Punktschätzer** vorstellen und auf verschiedene **Methoden der Punktschätzung** eingehen.

Das Prinzip der Punktschätzung

Im 100-Meter-Lauf möge eine Zufallsstichprobe von einhundert 16-jährigen Schülerinnen eine Durchschnittszeit von 15 Sekunden erzielt haben. Als Modalwert (dies ist die am häufigsten gestoppte Zeit) werden Mo = 14 Sekunden und als Medianwert (dies ist die Laufzeit, die 50 % aller Schülerinnen mindestens erreichen) Md = 14.50 Sekunden ermittelt. Was sagen diese Zahlen über die durchschnittliche Laufzeit der Population aller 16-jährigen Schülerinnen aus? Beträgt sie – wie in der Stichprobe – ebenfalls 15 Sekunden oder ist sie vielleicht eher besser, weil die am häufigsten registrierte Zeit 14 Sekunden betrug, oder kommen vielleicht völlig andere Zahlen in Betracht, weil sich in der Stichprobe zufällig besonders langsame Läuferinnen befanden?

Zufallsexperimente und Zufallsvariablen

Die Erläuterung der Prinzipien von Punkt- und Intervallschätzungen wird durch die Einführung einiger in der Statistik gebräuchlicher Bezeichnungen erleichtert. Wird z. B. die Zeit einer beliebigen Schülerin gemessen, so bezeichnen wir dies als ein Zufallsexperiment. Allgemein versteht man im Kontext der Wahrscheinlichkeitstheorie unter einem **Zufallsexperiment** einen Vorgang, dessen Ergebnis in der Weise vom Zufall abhängt, dass man vor dem Experiment nicht weiß, zu welchen der möglichen Ergebnisse das Experiment führen wird. Das Zufallsexperiment hat immer mindestens zwei mögliche Ausgänge läuft unter definierten, gleichbleibenden Bedingungen ab, die eine beliebige Wiederholung gleichartiger Experimente gestatten (Helten, 1974, S. 15). Das hier definierte wahrscheinlichkeitstheoretische Zufallsexperiment ist nicht zu verwechseln mit dem wissenschaftlichen Experiment als einem bestimmten Studiendesign (randomisiertes Kontrollgruppen-Design), bei dem Probanden per Zufallsprinzip verschiedenen Bedingungen zugeordnet und unterschiedlich behandelt werden, um Ursache-Wirkungs-Relationen zu prüfen (▶ Abschn. 7.6).

Ein Zufallsexperiment wird durchgeführt, um ein bestimmtes Merkmal (im Beispiel die Laufzeit) beobachten zu können. Die verschiedenen, im Zufallsexperiment potenziell beobachtbaren Merkmalsausprägungen heißen **Elementarereignisse**, und die Menge aller Elementarereignisse bildet den **Merkmalsraum** (im Beispiel wären dies alle von 16-jährigen Schülerinnen überhaupt erreichbaren Zeiten im 100-Meter-Lauf).

Zufallsexperiment – Ein Zufallsexperiment („experiment") meint im Kontext der Wahrscheinlichkeitstheorie einen Versuch, dessen Ablauf bekannt ist und der mehr als ein Ergebnis hervorbringen kann (die möglichen Ergebnisse müssen bekannt sein), wobei aber bei jedem Versuchsdurchgang das konkrete Ergebnis vorab unbekannt ist. Beispiele für Zufallsexperimente sind das Würfeln (vorab ist nicht bekannt, welche der sechs Augenzahlen erwürfelt wird) oder die Befragung (vorab ist bei einer zufällig ausgewählten Person, die man nach ihrer Depressivität fragt, nicht bekannt, welche Depressionsausprägung sie hat).

Das Ergebnis eines Zufallsexperimentes bzw. das Elementarereignis kann kardinalskaliert (z. B. Zeit im 100-Meter-Lauf; Grad der Depressivität) oder nominalskaliert sein (z. B. Haarfarbe; Parteipräferenz). Jedem **Elementarereignis *e*** wird nach einer eindeutigen Regel eine reelle Zahl x(e) zugeordnet. Die Zuordnungsvorschrift bzw. die Funktion, die jedes Elementarereignis mit einer bestimmten Zahl verbindet, bezeichnen wir als **Zufallsvariable** X. (Hier und im Folgenden verwenden wir für **Zufallsvariablen Großbuchstaben** und für eine konkrete Ausprägung der Zufallsvariablen bzw. eine **Realisation der Zufallsvariablen Kleinbuchstaben**, es sei denn, der Kontext macht diese Unterscheidung nicht erforderlich.) Sind die Elementarereignisse selbst numerisch, können die erhobenen Zahlen direkt eine Zufallsvariable darstellen. Wenn eine Schülerin z. B. 13.80 Sekunden ($e = 13.80$) läuft, wäre eine mögliche Zuordnungsvorschrift z. B. $X(e = 13.80) = 13.80$. Eine andere Zuordnungsvorschrift, die nur ganzzahlig gerundete Werte verwendet, lautet $X(e = 13.80) = 14$.

In gleicher Weise legt die Zufallsvariable auch bei nicht-numerischen bzw. nominalskalierten Elementarereignissen fest, welche Zahlen den Elementarereignissen zuzuordnen sind. Für Haarfarben könnte diese Zuordnungsvorschrift z. B. lauten: $X(e_1 = \text{schwarz}) = 0$, $X(e_2 = \text{blond}) = 1$, $X(e_3 = \text{braun}) = 2$, $X(e_4 = \text{rot}) = 3$.

Zufallsvariable – Eine Zufallsvariable („random variable") ordnet allen Ergebnissen eines Zufallsexperiments eine Wahrscheinlichkeit (bei abzählbaren Ergebnissen) oder eine Wahrscheinlichkeits-Dichte (bei nicht abzählbaren Ergebnissen) zu. Sie ist eine Abbildung der Menge aller Elementarereignisse (d. h. aller möglichen Ergebnisse eines Zufallsexperiments) in die reellen Zahlen. Eine Zufallsvariable wird oft mit Großbuchstaben, die konkrete Ausprägung der entsprechenden Zufallsvariablen mit Kleinbuchstaben bezeichnet.

■ Verteilung von Zufallsvariablen

Eine Zufallsvariable ist **diskret**, wenn sie nur endlich (oder abzählbar) viele Werte aufweist (Beispiel: Anzahl der Geschwister). **Stetige** (oder kontinuierliche) Zufallsvariablen können jeden Wert annehmen, der zwischen zwei beliebigen Werten der Zufallsvariablen liegt. Ihre Werte sind deshalb nicht abzählbar (Beispiele: Zeit-, Längen- oder Gewichtsmessungen; Ausprägungsgrade latenter Merkmale wie Extraversion oder Neurotizismus). Die Anzahl möglicher Werte ist hierbei theoretisch unbegrenzt, sie hängt praktisch jedoch von der Genauigkeit des Messinstrumentes ab.

Der Ausdruck $P(X = x)$ symbolisiert die Wahrscheinlichkeit, dass die Zufallsvariable X den Wert x annimmt. Diese Wahrscheinlichkeit entspricht der Wahrscheinlichkeit des Elementarereignisses e, dem der Wert x der Zufallsvariablen X zugeordnet ist. Beim Münzwurf-Experiment seien die Elementarereignisse z. B. mit $e_1 =$ Zahl und $e_2 =$ Kopf definiert. Eine Zufallsvariable X(e) ordnet e_1 die Zahl 0 und e_2 die Zahl 1 zu: $X(e_1) = 0$ und $X(e_2) = 1$. Die Wahrscheinlichkeit, dass die Zufallsvariable X den Wert 0 annimmt, lautet dann $P(X = 0) = 1/2$.

Die Liste aller möglichen Werte einer diskreten Zufallsvariablen zusammen mit den ihnen zugeordneten Wahrscheinlichkeiten bezeichnet man als **Wahrscheinlichkeitsfunktion**. Beispiel: Die Wahrscheinlichkeitsfunktion eines Würfelexperimentes lautet $P(X = 1) = 1/6$; $P(X = 2) = 1/6 \ldots P(X = 6) = 1/6$. Für alle übrigen Werte $1 > x > 6$ ist $P(1 > x > 6) = 0$. Denn es ist unmöglich (Wahrscheinlichkeit 0 %), dass mit einem herkömmlichen Würfel eine Augenzahl unter 1 oder über 6 gewürfelt wird.

Bei stetigen Zufallsvariablen beziehen sich die Wahrscheinlichkeitsangaben auf Intervalle (z. B. die Wahrscheinlichkeit, dass eine 16-jährige Schülerin eine Laufzeit zwischen 14 und 15 Sekunden im 100-Meter-Lauf erreicht). Mit kleiner werdenden Intervallen sinkt die Wahrscheinlichkeit. Sie nimmt für einen einzelnen Punkt der Zufallsvariablen den Wert Null an. (Die Wahrscheinlichkeit, dass eine zufällig herausgegriffene Schülerin 100 Meter in exakt 14.3238712… Sekunden läuft, ist nahezu Null.) Anstatt von Wahrscheinlichkeitsfunktion spricht man bei stetigen Zufallsvariablen von der **Dichtefunktion** der Zufallsvariablen. Der zu einem einzelnen Wert der Zufallsvariablen gehörende Ordinatenwert heißt **(Wahrscheinlichkeits-)Dichte** dieses Wertes. Wahrscheinlichkeitsfunktion und Dichtefunktion einer Zufallsvariablen werden – wenn der Kontext eindeutig ist – auch kurz **„Verteilung einer Zufallsvariablen"** genannt.

Summiert (kumuliert) man bei einer diskreten Zufallsvariablen die durch die Wahrscheinlichkeitsfunktion definierten Einzelwahrscheinlichkeiten, resultiert eine kumulierte Wahrscheinlichkeitsfunktion, die üblicherweise **Verteilungsfunktion** genannt wird. (Würfelbeispiel: $P(X = 1) = 1/6$; $P(X \leq 2) = 2/6$; $P(X \leq 3) = 3/6$ etc.) Die Summe der Einzelwahrscheinlichkeiten ergibt den Wert 1, d. h. die Wahrscheinlichkeit, dass beim Würfeln eine Augenzahl zwischen 1 und 6 gewürfelt wird, beträgt 100 %.

Bei stetigen Zufallsvariablen ist die Verteilungsfunktion als das Integral (in Analogie zur Summe bei diskreten

Wahrscheinlichkeitsfunktion – Dichtefunktion – Verteilungsfunktion

Wahrscheinlichkeitsfunktion. Gegeben sei eine diskrete Zufallsvariable X mit abzählbar vielen Werten a_i $(i = 1, 2 \ldots k)$, für die gilt:

$$P(x = a_i) > 0$$

und

$$\sum_{i=1}^{k} P(x = a_i) = 1.$$

Die Wahrscheinlichkeitsfunktion der Zufallsvariablen X lautet dann

$$P(x = a_i) = \begin{cases} p_i & \text{für } x = a_i \\ & (i = 1, 2, \ldots, k) \\ 0 & \text{für alle übrigen } x. \end{cases}$$

Beispiel: Wahrscheinlichkeitsfunktion der Zufallsvariablen X beim Würfeln mit einem Würfel:

$$P(x = i) = \begin{cases} \frac{1}{6} & \text{für } i = 1, 2 \ldots 6) \\ 0 & \text{für alle übrigen } x. \end{cases}$$

In Abb. 12.8 wird die Wahrscheinlichkeitsfunktion der Zufallsvariablen X beim Würfeln mit zwei Würfeln veranschaulicht.

Dichtefunktion. Gegeben sei eine stetige Zufallsvariable X. Bei stetigen Zufallsvariablen spricht man nicht von der Wahrscheinlichkeit eines bestimmten Wertes (diese ist bei einer stetigen Zufallsvariablen gleich Null), sondern von der Wahrscheinlichkeit eines bestimmten Intervalls der Zufallsvariablen:

$$P(a \le x \le b) = \int_a^b f(x)\mathrm{d}x$$

(lies: Integral von $f(x)$ zwischen den Grenzen von a und b).

$f(x)$ ist hierbei die Dichtefunktion der Zufallsvariablen, die für jeden X-Wert einen Ordinatenwert (= Dichte) definiert. Mit dem Integral in den Grenzen a und b wird eine Fläche bestimmt, die wegen der Normierungsvorschrift

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

der Wahrscheinlichkeit entspricht, dass ein Wert der Zufallsvariablen X in den Bereich a bis b fällt. (Das Integral entspricht dem Summenzeichen bei diskreten Wahrscheinlichkeitsfunktionen.)

Beispiel: In Abb. 12.9 dargestellt ist die (geschätzte) Dichtefunktion für die Zufallsvariable „100-m-Zeiten“ von 16-jährigen Mädchen. Die Wahrscheinlichkeit, dass eine 16-jährige Schülerin eine 100-m-Zeit von 14–15 s erreicht, entspricht der *schraffierten Fläche*. Sie beträgt (geschätzt): $P(14 \le x \le 15\text{ s}) = .38$. Für $x = 13$ s ergibt sich eine Dichte von $f(x = 13\text{ s}) = 0.12$.

Für die weiteren Überlegungen sind in erster Linie die Flächenanteile unter der Dichtefunktion von Interesse. Die Dichten der einzelnen x-Werte sind vorerst von nachgeordneter Bedeutung. (Der „Dichte“-Begriff geht auf eine Analogie zwischen Wahrscheinlichkeits- und Massenverteilung zurück. Näheres hierzu z. B. bei Kreyszig, 1973, Kap. 27.).

Verteilungsfunktion. Unter einer diskreten Verteilungsfunktion versteht man die kumulierte Wahrscheinlichkeitsfunktion und unter einer stetigen Verteilungsfunktion das Integral der Dichtefunktion. Die Verteilungsfunktion lautet im diskreten Falle

$$F(a) = P(x \le a) = \sum_{a_i \le a} P(x = a_i)$$

und im stetigen Falle

$$F(a) = P(x \le a) = \int_{-\infty}^{a} f(x)\mathrm{d}x.$$

$F(a)$ gibt damit die Wahrscheinlichkeit an, dass die Zufallsvariable X einen Wert annimmt, der höchstens so groß ist wie a. Aus den Definitionen für Wahrscheinlichkeitsfunktion und Dichtefunktion folgt, dass $0 \le F(a) \le 1$. Verteilungsfunktionen sind monoton steigend (oder zumindest nicht monoton abnehmend).

Beispiel: In Abb. 12.10 wird die Verteilungsfunktion der Zufallsvariablen X „Würfeln mit 2 Würfeln“ gezeigt, deren Wahrscheinlichkeitsverteilung in Abb. 12.8 zu sehen ist.

Der Abb. 12.10 ist z. B. zu entnehmen, dass die Wahrscheinlichkeit, mit 2 Würfeln höchstens 8 Punkte zu würfeln, ca. 72 % beträgt: $P(x \le 8) = 26/36$.

Die Verteilungsfunktion der Zufallsvariablen X für „100-m-Zeiten“ gibt Abb. 12.11 wieder.

Für 15 s ergibt sich aufgrund der Verteilungsfunktion ein Ordinatenwert von 0.77, d. h., eine 16-jährige Schülerin benötigt mit einer Wahrscheinlichkeit von 77 % höchstens 15 s für die 100-m-Strecke.

Zufallsvariablen) der Dichtefunktion definiert, wobei der Gesamtfläche unter der Dichtefunktion der Wert 1 zugewiesen wird. Die Ordinate des Wertes $X = a$ einer stetigen Verteilungsfunktion gibt an, mit welcher Wahrscheinlichkeit Werte $X \le a$ auftreten. (Würde im o. g. Beispiel der Medianwert der Stichprobe den Populationsmedianwert richtig schätzen, hätte der Wert $X = 14.5$ in der Verteilungsfunktion der Zufallsvariablen „Laufzeit“ einen Funktionswert von $P = 0.5$.)

Eine zusammenfassende Darstellung der Begriffe Wahrscheinlichkeitsfunktion, Dichtefunktion und Verteilungsfunktion findet man in ▶ Wahrscheinlichkeitsfunktion – Dichtefunktion – Verteilungsfunktion.

Gütekriterien für Punktschätzer

Ausgerüstet mit diesem begrifflichen Vorwissen wenden wir uns erneut der Frage zu, wie tauglich ein statistischer Stichprobenkennwert für die Beschreibung einer Population ist. Können wir davon ausgehen, dass z. B. der Mittelwert $\bar{X}$ bzw. M einer Zufallsstichprobe dem Mittelwert μ der Population entspricht? Sicherlich nicht, denn es ist leicht einzusehen, dass statt der erhobenen Stich-

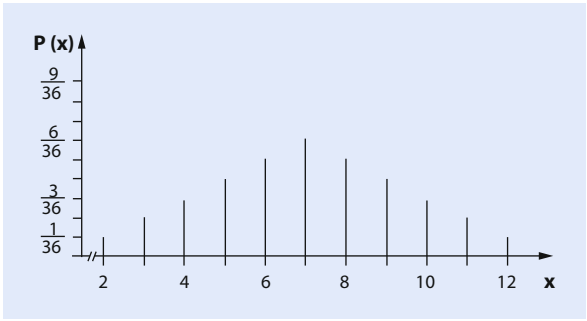


Abb. 12.8 Wahrscheinlichkeitsfunktion

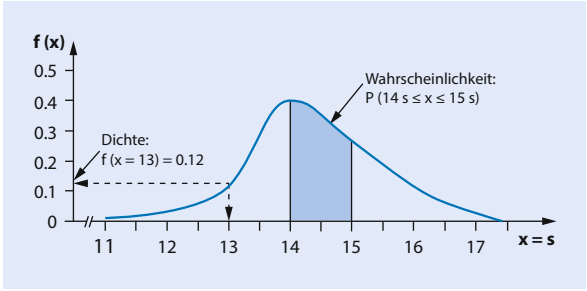


Abb. 12.9 Dichtefunktion

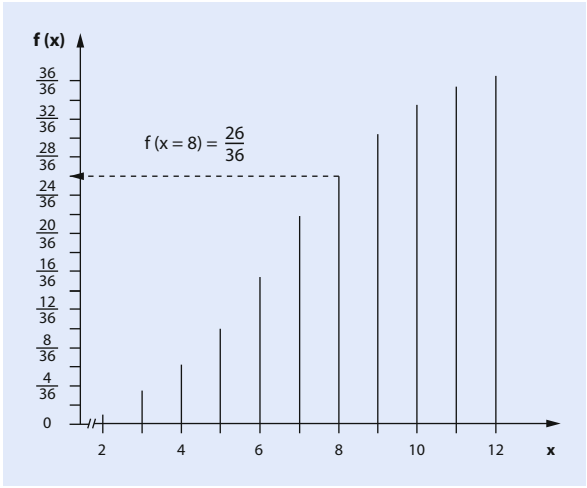


Abb. 12.10 Diskrete Verteilungsfunktion

probe auch eine andere Stichprobe hätte gezogen werden können, deren Mittelwert keineswegs mit dem Mittelwert der ersten Stichprobe identisch sein muss. Für den Populationsparameter μ lägen damit zwei verschiedene Schätzungen vor. Würde man die Untersuchung beliebig häufig mit immer wieder anderen Zufallsstichproben wiederholen, erhielte man eine Häufigkeitsverteilung verschiedener Stichprobenmittelwerte. Genauso wie im Beispiel jede Zeitmessung eine Realisation der Zufalls-

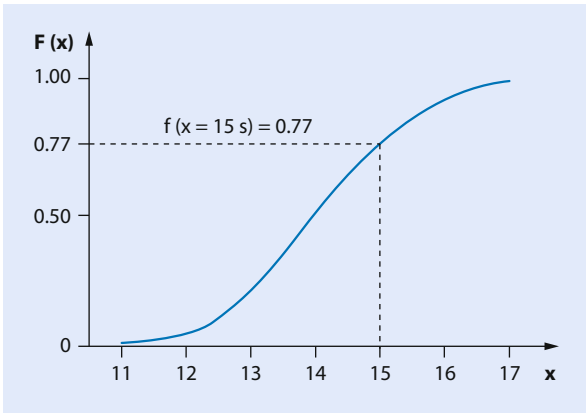


Abb. 12.11 Stetige Verteilungsfunktion

variablen „100-Meter-Zeit“ darstellt, ist jeder einzelne Mittelwert eine Realisation der stetigen Zufallsvariablen „durchschnittliche 100-Meter-Zeit bei Stichproben mit $n = 100$“.

Es interessiert also zunächst die Frage, wie tauglich ein einzelner Stichprobenkennwert, d. h. also ein „Punkt“ der Zufallsvariablen „Stichprobenmittelwerte“ zur Schätzung eines Populationsparameters μ ist. Für die Beurteilung der Tauglichkeit oder Güte eines Punktschätzers hat der britische Statistiker Ronald Fisher (1925) vier Schätzkriterien vorgeschlagen: Erwartungstreue, Konsistenz, Effizienz und Suffizienz (Bortz & Schuster, 2010, S. 88ff.; Wirtz & Nachtigall, 2009, S. 111ff.), die noch durch das Kriterium der Robustheit zu ergänzen sind.

Kriterien für die Güte eines Punktschätzers – Die Güte eines Punktschätzers (d. h. der Schätzmethode) wird v. a. über fünf Kriterien definiert:

1. **Erwartungstreue** (Unverzerrtheit, Unverfälschtheit, „mean-unbiasedness“): Ein erwartungstreuer Schätzer entspricht im Mittel dem Populationsparameter.
2. **Konsistenz** („consistency“, „asymptotic consistency“): Ein konsistenter Schätzer wird mit wachsendem Stichprobenumfang genauer, d. h. er ist erwartungstreu und seine Varianz sinkt mit wachsendem Stichprobenumfang.
3. **Effizienz** („efficiency“): Ein effizienter Schätzer ist genauer (d. h. weist eine geringere Varianz auf) als alternative erwartungstreue Schätzer.
4. **Suffizienz** („sufficiency“): Ein suffizienter/erschöpfender Schätzer berücksichtigt alle relevanten Informationen in der Stichprobe.
5. **Robustheit** („robustness“): Ein robuster Schätzer ist gegenüber Ausreißern bzw. Extremwerten in der Stichprobe unempfindlich.

▪ Erwartungstreue („mean-unbiasedness“)

Ein Stichprobenkennwert k schätzt den Parameter κ (kappa) einer Population erwartungstreu, wenn der Mittelwert der k-Werte für zufällig aus der Population gezogene Stichproben mit dem Populationsparameter κ identisch ist. Das arithmetische Mittel $\bar{X}$ ist eine erwartungstreue Schätzung von μ, und die relative Häufigkeit P ist eine

erwartungstreue Schätzung von π. Die Stichprobenvarianz S^2 hingegen schätzt die Populationsvarianz σ^2 (sigma quadrat) nicht erwartungstreu. Ihr Erwartungswert unterschätzt die Populationsvarianz um den Faktor $(n-1)/n$. Eine erwartungstreue Schätzung für σ^2 erhalten wir, indem wir diesen „Bias" korrigieren (Wirtz & Nachtigall, 2009, S. 85):

$$\hat{\sigma}^2 = \frac{n}{n-1} \cdot S^2. \tag{12.3}$$

Eine erwartungstreue Schätzung von σ^2 wird durch $\hat{\sigma}^2$ symbolisiert (lies: Sigma Dach Quadrat).

■ Konsistenz („consistency", „asymptotic consistency")

Ein Stichprobenkennwert k schätzt den Parameter κ einer Population konsistent, wenn k mit wachsendem Umfang der Stichprobe ($n \rightarrow \infty$) gegen κ konvergiert. Die Stichprobenkennwerte $\bar{X}$, P und S^2 sind konsistente Schätzungen der entsprechenden Populationsparameter μ, π und σ^2.

Formal lässt sich das Konsistenzkriterium folgendermaßen darstellen:

$$P(|k - K| < \varepsilon) \rightarrow 1 \quad \text{für } n \rightarrow \infty. \tag{12.4}$$

Die Wahrscheinlichkeit, dass der absolute Differenzbetrag zwischen dem Schätzwert k und dem Parameter K kleiner ist als eine beliebige Größe ε (epsilon), geht gegen 1, wenn n gegen unendlich geht.

■ Effizienz („efficiency")

Ein Stichprobenkennwert schätzt einen Populationsparameter effizient, wenn die Varianz der Verteilung dieses Kennwertes kleiner ist als die Varianzen der Verteilungen anderer, zur Schätzung dieses Parameters ebenfalls geeigneter Kennwerte. Die Effizienz charakterisiert damit die Genauigkeit einer Parameterschätzung.

Stehen für die Schätzung eines Populationsparameters verschiedenartige Kennwerte zur Verfügung (z. B. Mittelwert und Medianwert als Schätzer für μ in symmetrischen Verteilungen), gibt die relative Effizienz an, welcher der beiden Kennwerte zur Schätzung des Populationsparameters vorzuziehen ist. Die relative Effizienz ist als Quotient der Varianzen der zu vergleichenden Kennwerte definiert. Sie lautet z. B. für einen Kennwert k im Vergleich zu einem Kennwert g (in Prozent ausgedrückt):

$$\text{rel. Effizienz von } k = \frac{\sigma_g^2}{\sigma_k^2} \cdot 100\,\%. \tag{12.5}$$

Bei normalverteilten Zufallsvariablen beträgt die relative Effizienz des Medianwertes in Bezug auf das arithmetische Mittel 64 % (Bortz & Schuster, 2010, S. 90). Das arithmetische Mittel schätzt damit μ erheblich präziser als der Medianwert. Die relative Effizienz von 64 % kann so interpretiert werden, dass der Medianwert einer Stichprobe des Umfanges $n = 100$ aus einer normalverteilten Population den Parameter μ genauso präzise schätzt wie das arithmetische Mittel einer Stichprobe des Umfanges $n = 64$.

■ Suffizienz („sufficiency")

Ein Schätzwert ist suffizient (erschöpfend), wenn er alle in den Daten einer Stichprobe enthaltenen Informationen berücksichtigt, so dass man durch Berechnung eines weiteren statistischen Kennwertes keine zusätzlichen Informationen über den zu schätzenden Parameter erhält. Der Kennwert P ist ein suffizienter Schätzer des Parameters π und die Kennwerte $\bar{X}$ und $\hat{\sigma}^2$ sind zusammen genommen bei normalverteilten Zufallsvariablen (aber nicht jeder für sich allein) suffiziente Schätzer der Populationsparameter μ und σ^2. Auf eine genauere Darstellung suffizienter bzw. erschöpfender Statistiken soll hier verzichtet werden (Fischer, 1974, S. 22ff.; Kendall & Stuart, 1973, S. 184ff.).

■ Robustheit („robustness")

Ein Schätzwert ist robust, wenn er durch Abweichungen von Verteilungsannahmen oder durch Ausreißer – d. h. einzelne Extremwerte innerhalb der Stichprobe – wenig beeinflusst wird. Eine Reihe von Punktschätzern weisen nur geringe Robustheit auf und reagieren sehr sensibel auf Ausreißerwerte. Dies trifft z. B. auch auf den arithmetischen Mittelwert zu, der als Schätzer der zentralen Tendenz eines kardinalskalierten Merkmals in der Population wesentlich weniger robust ist als der Medianwert. Ein Maß für die Robustheit von Schätzern ist der **Bruchpunkt** („breakdown point"). Er gibt an, welcher Anteil der Stichprobendaten maximal aus Ausreißern bestehen kann, bevor der Schätzer zu stark verzerrt und somit unbrauchbar wird bzw. „zusammenbricht" (Huber & Ronchetti, 2009, S. 8; Marronna, Martin, & Yohai, 2006, S. 58).

Der **arithmetische Mittelwert** hat einen Bruchpunkt von 0 %, d. h. schon ein einziger Extremwert kann den gesamten Mittelwert so stark verzerren, dass er als Punktschätzer unbrauchbar wird. Der **Medianwert** dagegen hat einen Bruchpunkt von 50 %, d. h. er liefert auch dann noch eine relativ unverzerrte Punktschätzung der zentralen Tendenz der Population, wenn die Hälfte der Daten aus Ausreißern besteht. Das Beispiel weist gleichzeitig darauf hin, dass die verschiedenen Qualitätskriterien von Punktschätzern in einem Spannungsverhältnis zueinander stehen: Der Medianwert ist zwar ein deutlich robusterer, aber gleichzeitig auch ein viel weniger effizienter Schätzer der zentralen Tendenz einer Verteilung in der Population als der Stichprobenmittelwert. Der Medianwert sollte also nur dann als robuster Schätzer herangezogen werden, wenn die Stichprobendaten stark von Ausreißerwerten belastet sind und diese im Rahmen der Datenbe-

reinigung (► Abschn. 11.4) nicht sinnvoll auszuschließen sind. Zudem existieren als robuste Alternativen zum arithmetischen Mittelwert sog. **gestutzte** oder **winsorisierte Mittelwerte**, bei deren Berechnung die Extremwerte am oberen und unteren Rand der Verteilung entweder ausgeschlossen oder durch weniger extreme Werte ersetzt werden.

Zusammenfassend erweisen sich die Stichprobenkennwerte $\bar{X}$, $\hat{\sigma}^2$ und P für die meisten Schätzprobleme als optimale Punktschätzer der entsprechenden Populationsparameter μ, σ^2 und π. (Die Ausnahmen sind von so geringer praktischer Bedeutung, dass sie hier unerwähnt bleiben können.)

Methoden der Punktschätzung

Zur Berechnung von Punktschätzungen liegen unterschiedliche Verfahren vor:

1. **Momenten-Methode** („method of moments"): In der Statistik werden Kenngrößen von Zufallsvariablen als Momente bezeichnet. Bei der Momenten-Methode der Punktschätzung werden die Momente der Verteilung der Population durch die entsprechenden Momente der Stichprobe geschätzt. Die Punktschätzung des Populationsmittelwertes durch den Stichprobenmittelwert ist ein Beispiel dieses Verfahrens: Hier wird das erste Moment der Populationsverteilung (der Populationsmittelwert) durch das erste Moment der Stichprobenverteilung (den Stichprobenmittelwert) der betrachteten Variable geschätzt (zu Details der Momenten-Methode s. Assenmacher, 2000, S. 217; Hays & Winkler, 1970). Momentschätzer sind nicht immer erwartungstreu (z. B. ist die Stichprobenvarianz kein erwartungstreuer Momentschätzer der Populationsvarianz).
2. **Kleinstquadrat-Methode** („method of least squares"): Bei der Methode der kleinsten Quadrate werden die Schätzungen so vorgenommen, dass die quadrierten Abweichungen der geschätzten Werte von dem Modell – d. h. die quadrierten Residuen – minimal sind. Die Kleinstquadrat-Methode wird v. a. im Zuge der **Regressionsanalyse** eingesetzt, um die Regressionsgerade so durch die Punktewolke der Messwertpaare zu legen, dass die quadrierten Abweichungen der Messwerte von der Regressionsgerade minimiert werden. Geschätzt werden hierbei auf der Basis der Stichprobendaten die Regressions- sowie Korrelationskoeffizienten in der Population (Bortz & Schuster, 2010, S. 90; oder ausführlicher Daniel & Wood, 1971). Kleinstquadratschätzer zeichnen sich durch Konsistenz und Erwartungstreue aus.
3. **Maximum-Likelihood-Methode** (ML-Methode; „maximum likelihood method"): Mit der Maximum-Likelihood-Methode finden wir heraus, welcher der möglichen Populationsparameter angesichts eines Stichprobenergebnisses die höchste Likelihood („Plausibilität") aufweist. Maximum-Likelihood-Schätzungen sind nicht unbedingt erwartungstreu, allerdings sind sie effizient, konsistent und suffizient. Der **Prozentwert** P der Stichprobe ist z. B. ein ML-Schätzer des Populationsprozentwertes π (► Die Maximum-Likelihood-Methode der Punktschätzung). Parameterschätzungen nach der Maximum Likelihood-Methode spielen u. a. im Rahmen der probabilistischen Testtheorie (► Abschn. 10.4.4) sowie bei Sturkturgleichungsmodellen (► Kap. 17) eine wichtige Rolle.

Neben diesen „klassischen" Verfahren der Parameterschätzung, bei denen die Stichprobendaten für die Punktschätzung zu einem Schätzwert verrechnet werden, existieren noch zwei weitere Ansätze. Diese sind nicht nur für Punkt-, sondern auch für Intervallschätzungen geeignet:

1. **Resampling-Verfahren**: Bei Punktschätzungen nach dem Resampling-Verfahren wird die vorliegende Stichprobe als Ausgangsstichprobe verwendet, aus der wiederum (mit oder ohne Zurücklegen) **Unterstichproben** gezogen werden. Will man etwa anhand eines Stichprobenmittelwertes den Populationsmittelwert schätzen, so wird aus der empirisch erhobenen Ausgangsstichprobe (z. B. einer Stichprobe mit dem Stichprobenumfang $N = 1\,000$) eine große Zahl von Unterstichproben gezogen: Beispielsweise werden zufällig 5 000 Unterstichproben mit Umfang $N = 300$ aus der untersuchten Ausgangsstichprobe von $N = 1\,000$ gezogen, für die jeweils der Unterstichprobenmittelwert bestimmt wird. Die Zusammenfassung der 5 000 Unterstichprobenmittelwerte wird dann als Schätzer des Populationsmittelwertes verwendet. Beim Resampling-Ansatz handelt es sich also nicht um einen analytischen, sondern um einen empirischen Lösungsansatz: die Lösung wird aus einer vorliegenden Stichprobe heraus generiert, indem durch die wiederholte Ziehung von Unterstichproben eine empirische Stichprobenkennwerteverteilung entsteht. Man spricht auch von der **Bootstrap-Methode**, da sich der Lügenbaron Münchhausen im Märchen am eigenen Schopf bzw. an der Stiefelschlaufe (Bootstrap) aus dem Sumpf gezogen hat (Efron & Tibshirani, 1993).
2. **Bayes-Ansatz**: Bei Punktschätzungen nach dem Bayes-Ansatz – benannt nach dem englischen Mathematiker Thomas Bayes – wird neben den Stichprobendaten auch das **Vorwissen der Forschenden** – im Sinne einer subjektiven bzw. aus dem Forschungsstand abgeleiteten Wahrscheinlichkeitsbe-

wertung unterschiedlicher Parameterausprägungen – in die Schätzung integriert. Das Vorwissen wird in einer sog. **Prior-Verteilung** dargestellt, die durch Berücksichtigung der Stichprobendaten in eine **Posterior-Verteilung** transformiert wird. Eine genauere Darstellung der Methode ist dem ▶ Webangebot zu diesem Buch zu entnehmen.

Die Maximum-Likelihood-Methode der Punktschätzung

Der Grundgedanke der Maximum-Likelihood-Methode sei im Folgenden an einem Beispiel erläutert: Ein Student hat Schwierigkeiten mit seinem Studium und fragt sich, wie vielen Kommilitonen es wohl ähnlich ergeht. Er entschließt sich zu einer kleinen Umfrage, die ergibt, dass von 100 zufällig aus dem Immatrikulationsverzeichnis ausgewählten Studierenden 40 bekunden, ebenfalls mit dem Studium nicht zurechtzukommen. Dieses Ergebnis macht es ihm leichter, mit seinen eigenen Schwierigkeiten fertig zu werden, denn es haben – so behauptet er – immerhin „ca. 40 % aller Studierenden" ähnliche Schwierigkeiten wie er.

Wie kommt der Student zu dieser Behauptung? Offensichtlich hat er intuitiv erfasst, dass der **Merkmalsanteil p in einer Stichprobe** der beste Schätzwert für den **unbekannten Parameter π** ist. Seine Einschränkung, dass nicht „exakt" 40 %, sondern „ca." 40 % der Studierenden Studienschwierigkeiten haben, begründet er damit, dass er schließlich nur die Aussagen einiger und nicht die aller Studierender kenne.

Welche Alternativen hätte ein Student mit seinem Anliegen, den Populationsparameter π richtig zu schätzen? Er könnte z. B. behaupten, dass alle Studierenden, also 100 %, Studienschwierigkeiten eingestehen. Diese Behauptung wäre jedoch unsinnig, weil dann – wie auch bei $\pi = 0\,\%$ – niemals ein Stichprobenergebnis mit $P = 40\,\%$ resultieren könnte. Andere Parameter, wie z. B. 90 % kommen demgegenüber jedoch zumindest theoretisch infrage. Es ist aber wenig **plausibel** („likely"), dass sich in einer Zufallsstichprobe von 100 Studenten aus einer Population, in der 90 % Studienschwierigkeiten haben, nur 40 % mit Studienschwierigkeiten befinden. Die **höchste Plausibilität** („Maximum Likelihood") hat die Annahme, dass der Populationsparameter π dem Stichprobenkennwert p entspricht.

Wahrscheinlichkeit und Likelihood An dieser Stelle sind Erläuterungen angebracht, warum man nicht von der Wahrscheinlichkeit („probability") eines Parameters spricht, sondern von seiner Likelihood. (Dieser Ausdruck bleibt üblicherweise in deutschsprachigen Texten unübersetzt.) Fisher (1922; zit. nach Yamane, 1976, S. 177), auf den die Bezeichnung „Maximum Likelihood" zurückgeht, schreibt hierzu (wobei er den Anteilsparameter mit P und nicht mit p bezeichnet):

> » Wir müssen zu dem Faktum zurückkehren, daß ein Wert von P aus der Verteilung, über die wir nichts wissen, ein beobachtetes Ergebnis dreimal so häufig hervorbringt wie ein anderer Wert von P. Falls wir ein Wort benötigen, um diese relative Eigenschaft verschiedener Werte von P zu charakterisieren, würde ich vorschlagen, daß wir, um Verwirrung zu vermeiden, von der Likelihood eines Wertes P sprechen, dreimal die Likelihood eines anderen Wertes auszumachen, wobei wir stets berücksichtigen müssen, daß Likelihood hier nicht vage als synonym für Wahrscheinlichkeit (probability) verwendet wird, sondern einfach die relativen Häufigkeiten ausdrücken soll, mit der solche Werte der hypothetischen Quantität tatsächlich die beobachteten Stichproben erzeugen würden.

Die hier getroffene Unterscheidung zwischen Wahrscheinlichkeit und Likelihood findet im deduktiven bzw. induktiven Denkansatz ihre Entsprechung. Wird eine Population durch π gekennzeichnet, lässt sich hieraus deduktiv ableiten, mit welcher Wahrscheinlichkeit bestimmte, einander ausschließende Stichprobenergebnisse auftreten können. Die Summe dieser Wahrscheinlichkeiten (bzw. das Integral der Dichteverteilung bei stetig verteilten Stichprobenergebnissen) ergibt 1. (Die Wahrscheinlichkeit, dass bei $\pi = 50\,\%$ eine Stichprobe mit beliebigem p gezogen wird, ist 1.)

Umgekehrt sprechen wir von der Likelihood (L), wenn ausgehend von einem Stichprobenergebnis induktiv die **Plausibilität** verschiedener Populationsparameter gemeint ist. Dass es sich hierbei nicht um Wahrscheinlichkeiten handeln kann, geht aus der einfachen Tatsache hervor, dass die Summe aller möglichen, einander ausschließenden Likelihoods **nicht** – wie für Wahrscheinlichkeiten gefordert – eins ergibt. Die Summe der Likelihoods für alle Populationsparameter, die angesichts eines Stichprobenergebnisses möglich sind, ist größer als eins. Die Weiterführung des letzten Beispiels zeigt diese Besonderheit von Likelihoods.

Binomialverteilung als Beispiel Bisher gingen wir nur von der Plausibilität (bzw. der geschätzten Likelihood) verschiedener Populationsparameter bei gegebenem Stichprobenergebnis aus. Hierbei erschien uns der Populationsparameter $\pi = .40$ bei einem Stichprobenkennwert von $P = .40$ am plausibelsten. Dass dieser Parameter

tatsächlich die höchste Likelihood besitzt, lässt sich auch rechnerisch zeigen.

Tritt in einem Zufallsexperiment eine Ereignisalternative mit einer Wahrscheinlichkeit von π auf (z. B. $\pi = .50$ für das Ereignis „Zahl“ beim Münzwurf), kann die Wahrscheinlichkeit, dass die Häufigkeit X für das Auftreten dieses Ereignisses den Wert k annimmt, nach folgender Beziehung bestimmt werden:

$$p(X = k \mid \pi; n) = \binom{n}{k} \cdot \pi^{k} \cdot (1 - \pi)^{n-k}. \qquad (12.6)$$

(Die linke Seite der Gleichung wird gelesen als: Die Wahrscheinlichkeit für $X = k$ unter der Bedingung von π und n; zur Berechnung von $\binom{n}{k}$ Gl. 9.1.)

Die Wahrscheinlichkeiten für die einzelnen k-Werte bei gegebenem n und π konstituieren eine Wahrscheinlichkeitsfunktion, die unter dem Namen Binomialverteilung bekannt ist (zur Herleitung der Binomialverteilung Bortz, 2005, Abschn. 2.4.1).

Für $N = 5$, $X = 2$ und $P = .50$ (also 2-mal Zahl bei 5 Münzwürfen) ergibt sich folgende Wahrscheinlichkeit:

$$\begin{aligned} &p(x = 2 \mid \pi = .50; n = 5) \\ &\quad = \binom{5}{2} \cdot .50^{2} \cdot .50^{3} = .31. \end{aligned} \qquad (12.7)$$

Diese Beziehung können wir auch verwenden, wenn die Likelihood verschiedener Populationsparameter für ein bestimmtes Stichprobenergebnis zu berechnen ist. Für die Parameter $\pi_1 = .40$, $\pi_2 = .41$, $\pi_3 = .10$ und $\pi_4 = .90$ z. B. ergeben sich die folgenden Likelihoods, wenn – wie im Beispiel „Studienschwierigkeiten“ – $N = 100$ und $X = 40$ sind:

$$\begin{aligned} L_1(X = 40 \mid \pi_1 = .40; n = 100) &= \binom{100}{40} \cdot .40^{40} \cdot .60^{60} = .0812 \\ L_2(X = 40 \mid \pi_2 = .41; n = 100) &= \binom{100}{40} \cdot .41^{40} \cdot .59^{60} = .0796 \\ L_3(X = 40 \mid \pi_3 = .10; n = 100) &= \binom{100}{40} \cdot .10^{40} \cdot .90^{60} = 2.4703 \cdot 10^{-15} \\ L_4(X = 40 \mid \pi_4 = .90; n = 100) &= \binom{100}{40} \cdot .90^{40} \cdot .10^{60} = 2.0319 \cdot 10^{-34}. \end{aligned} \qquad (12.8)$$

Die Beispiele bestätigen, dass die Likelihood für den Parameter $\pi_1 = .40$ tatsächlich am höchsten ist. Sie verdeutlichen aber auch, dass die Summe der Likelihoods nicht 1 ergeben kann. Für $\pi_1 = .40$ resultiert $L_1 = .0812$ und für $\pi_2 = .41$ errechnen wir $L_2 = .0796$. Zwischen diesen beiden Parametern befinden sich jedoch beliebig viele andere Parameter (z. B. .405 oder .409117), deren Likelihoods jeweils zwischen .0812 und .0796 liegen. Allein für die Menge dieser Parameter ergibt sich eine Likelihood-Summe, die gegen unendlich tendiert.

Maximierung der Likelihood-Funktion. Wie aber kann man sicher sein, dass tatsächlich kein anderer Wert für den Parameter π existiert, der eine größere Likelihood aufweist als der Parameter $\pi = .40$? Um dieses Problem zu lösen, muss die Funktion, die die Likelihoods für variable π-Werte bestimmt, die sog. **Likelihood-Funktion**, bekannt sein. Sie lautet in unserem Beispiel:

$$L(X = k \mid \pi; n) = \binom{n}{k} \cdot \pi^{k} \cdot (1 - \pi)^{n-k}. \qquad (12.9)$$

Wir suchen denjenigen π-Wert, dessen Likelihood bei einem bestimmten k-Wert (im Beispiel $k = 40$) für ein bestimmtes n (im Beispiel $N = 100$) maximal ist – den π-Wert mit maximaler Likelihood.

Den Maximalwert einer Funktion bestimmt man mit der Differenzialrechnung. Aus rechnerischen Gründen differenzieren wir jedoch im Beispiel der Binomialverteilung nicht die Likelihood-Funktion, sondern die zur Basis e logarithmierte Likelihood-Funktion. (Diese Vorgehensweise ist zulässig, denn der Logarithmus eines positiven Arguments ist eine monotone Funktion des Arguments. Das Maximum der ursprünglichen Funktion entspricht damit dem Maximum der logarithmierten Funktion.)

$$\ln L = \ln \binom{n}{k} + k \cdot \ln \pi + (n - k) \cdot \ln(1 - \pi). \qquad (12.10)$$

Die nach π differenzierte Funktion heißt

$$\frac{d \ln L}{d \pi} = \frac{k}{\pi} - \frac{n - k}{1 - \pi}. \qquad (12.11)$$

Wir setzen die erste Ableitung null und ermitteln für π:

$$\begin{aligned} \frac{k}{\pi} - \frac{n - k}{1 - \pi} &= 0 \\ \frac{k \cdot (1 - \pi) - \pi \cdot (n - k)}{\pi \cdot (1 - \pi)} &= 0 \\ k \cdot (1 - \pi) - \pi \cdot (n - k) &= 0 \\ k - k \cdot \pi - \pi \cdot n + \pi \cdot k &= 0 \\ \pi &= \frac{k}{n}. \end{aligned} \qquad (12.12)$$

Die zweite Ableitung ist negativ, d. h., der durch k/n geschätzte Parameter π hat die größte Likelihood; k/n ist die Maximum-Likelihood-Schätzung des Parameters π.

In ähnlicher Weise lässt sich zeigen, dass $\bar{X}$ bei normalverteilten Zufallsvariablen eine Maximum-Likelihood-Schätzung des Parameters μ darstellt. Für die Populationsvarianz σ^2 resultiert als Maximum-Likelihood-Schätzung die Stichprobenvarianz S^2. Dieses Beispiel zeigt, dass Maximum-Likelihood-Schätzungen nicht immer erwartungstreue Schätzungen sind: S^2 ist, wie oben bereits erwähnt wurde,keine erwartungstreue Schätzung von σ^2.

Der Wahrscheinlichkeit (*Probability*) ist zu entnehmen, wie häufig verschiedene, einander ausschließende Stichprobenergebnisse zustande kommen, wenn ein bestimmter Populationsparameter vorliegt (*deduktiver Ansatz*). Umgekehrt spricht man von der *Likelihood*, wenn es darum geht, anhand eines Stichprobenergebnisses zu schätzen, wie plausibel („wahrscheinlich") verschiedene Populationsparameter als die Erzeuger dieses Wertes anzusehen sind (*induktiver Ansatz*).

12.4.2 Intervallschätzung von Populationsparametern

Bisher wurde der Parameter einer Population (z. B. μ) durch einen einzigen Wert ($\bar{x}$) geschätzt. Wir haben diese Schätzung Punktschätzung genannt. Obwohl $\bar{x}$ die bestmögliche Schätzung für μ darstellt, dürfte es ohne weiteres einsichtig sein, dass $\bar{x}$ in der Regel nicht mit μ identisch ist, denn schließlich wird $\bar{x}$ aus einer zufällig gezogenen Stichprobe errechnet, deren Werte von Stichprobe zu Stichprobe unterschiedlich ausfallen. Angesichts dieser Tatsache wäre es nun für die Beschreibung von Populationen durch Stichproben hilfreich, wenn man wüsste, wie genau $\bar{x}$ den Parameter μ schätzt.

Hierfür wurde von dem Statistiker Jerzy Neyman (1937) ein Verfahren vorgeschlagen, mit dessen Hilfe ein Parameter durch ein Intervall vom Typ

$$a < \mu < b \qquad (12.13)$$

geschätzt wird, d. h. ein Verfahren, das Grenzen ermittelt, innerhalb derer sich der wahre Populationsparameter mit hoher Plausibilität befindet (**Konfidenzintervall**). Derartige Schätzungen nennt man Intervallschätzungen. Wie sicher eine Intervallschätzung erfolgen soll, wird mit dem **Konfidenzkoeffizienten** (z. B. 99 % oder 95 %) festgelegt.

Intervallschätzung – Bei einer Intervallschätzung („interval estimation") wird ein unbekannter Populationsparameter durch einen auf der Basis der Stichprobenergebnisse konstruierten Wertebereich (**Konfidenzintervall**, Vertrauenintervall, „confidence interval", CI) geschätzt. Dieser Wertebereich wird so konstruiert, dass sich in ihm $1 - \alpha = 99\,\%$ oder $1 - \alpha = 95\,\%$ (Konfidenzkoeffizient, Vertrauenswahrscheinlichkeit) aller möglichen Populationsparameter befinden, die den empirisch ermittelten Stichprobenkennwert erzeugt haben können. Der **Konfidenzkoeffizient** $1 - \alpha$ ist das Komplement zum **Signifikanzniveau** α, das als zulässige Obergrenze der Irrtumswahrscheinlichkeit konventionell auf 1 % bzw. 5 % festgelegt ist.

Konfidenzintervall und Signifikanztest – Zwischen dem Konfidenzintervall und dem klassischen Signifikanztest besteht insofern eine Parallele, als ein Konfidenzintervall, das nicht die Null umschließt, gleichzeitig auch impliziert, dass der bestreffende Parameter statistisch signifikant ist (Ablehnung des H_0-Modells; ▶ Abschn. 12.5.1).

Glaubwürdigkeitsintervall – Bei der Intervallschätzung nach dem Bayes-Ansatz spricht man statt vom Konfidenzintervall vom Glaubwürdigkeitsintervall („credible intervall"), in dessen Berechnung neben den Stichprobendaten auch Vorinformationen der Forschenden (auf der Basis des Forschungsstandes) einfließen.

Im Folgenden werden die Berechnung des **Konfidenzintervalls des arithmetischen Mittels** (bei bekannter sowie bei unbekannter Varianz) und die Berechnung des **Konfidenzintervalls eines Populationsanteils** dargestellt. Zudem werden Hinweise dazu gegeben, welcher **optimale Stichprobenumfang** für die Berechnung eines aussagekräftigen Konfidenzintervalls bereits bei der Untersuchungsplanung einzukalkulieren ist.

Konfidenzintervall des arithmetischen Mittels bei bekannter Varianz

Zunächst nehmen wir an, die Verteilung eines Merkmals X (z. B. **Intelligenzquotient**) in einer Population (z. B. Abiturienten) sei bekannt. Ihr Mittelwert betrage $\mu = 110$ und ihre Varianz $\sigma^2 = 144$. Wie die Verteilung dieses Merkmals aussehen könnte, zeigt ◘ Abb. 12.12.

▪ Zentrales Grenzwerttheorem

Die in ◘ Abb. 12.12a gezeigte Verteilung weist unregelmäßige Schwankungen auf und ist linkssteil. Aus dieser Population werden wiederholt Stichproben des Umfanges $N = 2$ gezogen. Die Verteilung der Mittelwerte dieser „Miniaturstichproben" bezeichnen wir als „**Stichprobenkennwerteverteilung**" („sampling distribution"). Sie ist in ◘ Abb. 12.12b wiedergegeben. Im Vergleich zur ursprünglichen Merkmalsverteilung hat diese Verteilung eine geringere Streuung und weist zudem weniger Irregularitäten auf. Entnehmen wir Stichproben des Umfanges $N = 5$ bzw. $N = 20$, zeigt ◘ Abb. 12.12b, dass die Verteilung der Zufallsvariablen $\bar{X}$ mit wachsendem Stichprobenumfang weniger streut und zunehmend deutlicher in eine Normalverteilung übergeht. Dies ist ein für die schließende Statistik grundlegender Befund. (Eine zusätzliche Voraussetzung besagt, dass Mittelwert und Varianz der Grundgesamtheit endlich sind.) Dieser

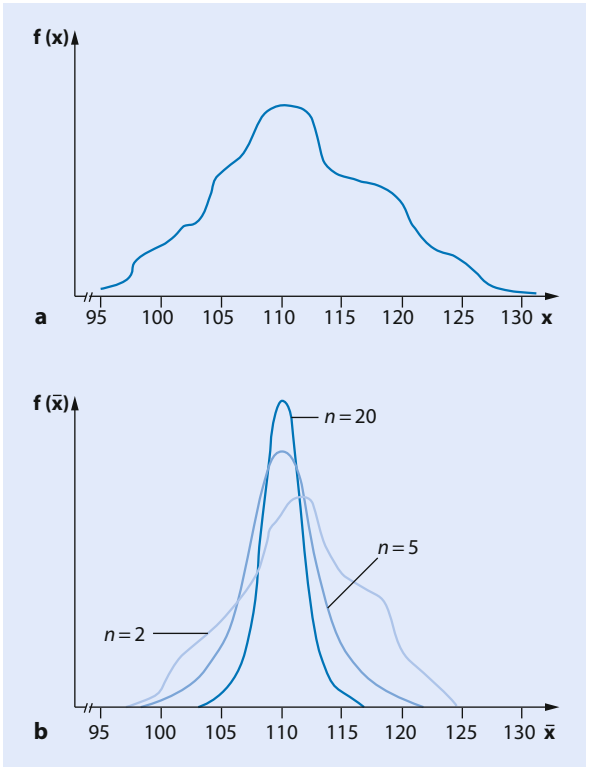


Abb. 12.12 Verteilung der Intelligenzquotienten von Abiturienten in der Population **(a)** sowie Verteilung von Mittelwerten von aus der Population gezogenen Stichproben unterschiedlichen Umfangs **(b)**

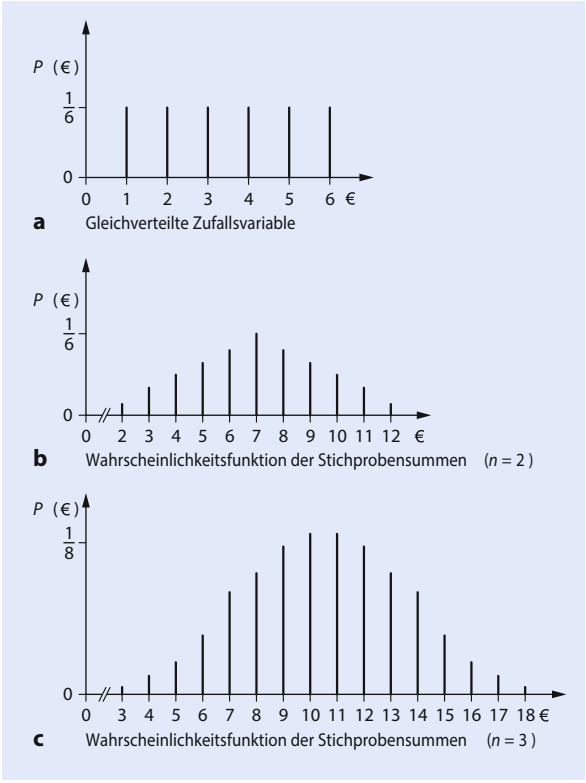


Abb. 12.13 Gleich verteilte Zufallsvariable **(a)** und deren Wahrscheinlichkeitsfunktionen für Stichprobensummen **(b, c)**

Sachverhalt wird als das „zentrale Grenzwerttheorem" bezeichnet. Für praktische Zwecke können wir davon ausgehen, dass die Mittelwerteverteilung auch für extrem von der Normalität abweichende Grundgesamtheiten hinreichend normal ist, wenn $n \geq 30$ ist.

Zentrales Grenzwerttheorem – Das zentrale Grenzwerttheorem (zentraler Grenzwertsatz; „central limit theorem) besagt, dass die Verteilung von Mittelwerten aus Stichproben des Umfanges n, die einer beliebig verteilten Grundgesamtheit entnommen werden, einer Normalverteilung entspricht – vorausgesetzt, n ist genügend groß (mindestens $n \geq 30$).

Die mathematische Herleitung des zentralen Grenzwertsatzes ist zu kompliziert, um sie in diesem Zusammenhang behandeln zu können (vgl. hierzu Kendall & Stuart, 1973). Wir wollen uns damit begnügen, die „Arbeitsweise" dieses wichtigen Satzes anhand eines kleinen Beispiels zu veranschaulichen.

Ein Obststand verkauft **sechs verschiedene Obstsorten**, die wir einfachheitshalber mit A, B, C, D, E und F bezeichnen. Als Obstpreise verlangt der Händler für je ein Pfund € 1,–, € 2,–, € 3,–, € 4,–, € 5,– und € 6,– (in gleicher Reihenfolge). Ferner gehen wir davon aus, dass jede Obstsorte gleich häufig bzw. mit gleicher Wahrscheinlichkeit ($P = 1/6$) gekauft wird. Die Zufallsvariable „Obstpreise" ist damit gleichverteilt: $p(\text{Obstpreis} = €\ 1{,}–) = 1/6$; $p(\text{Obstpreis} = €\ 2{,}–) = 1/6$ etc. (Abb. 12.13a).

Wie verteilen sich nun die Kaufsummen, wenn wir davon ausgehen, dass jeder Käufer zufällig 2 Obstsorten wählt? Insgesamt sind $6 \cdot 6 = 36$ verschiedene „Stichproben" möglich. (Hierbei werden die Stichproben AB und BA, AC und CA etc. getrennt gezählt.) Die billigste Stichprobe kostet € 2,– (2× Sorte A) und die teuerste € 12,– (2× Sorte F). Insgesamt erhalten wir für die Zufallsvariable „Kaufsumme" die in Tab. 12.8 dargestellte Wahrscheinlichkeitsfunktion.

Abb. 12.13b veranschaulicht diese Wahrscheinlichkeitsfunktion grafisch. Sie zeigt, dass sich die Stichprobensummen (und damit natürlich auch die Mittelwerte) anders verteilen als die einzelnen Kaufpreise. Die Kaufpreise sind gleichverteilt, und die Summen verteilen sich eingipflig und symmetrisch. Lassen wir die Stichprobengröße wachsen, nähert sich die Summenverteilung einer Normalverteilung. Dies verdeutlicht Abb. 12.13c, die

Tabelle 12.8 Verteilung der Zufallsvariablen „Kaufsumme“

Kaufsumme (Euro)	Stichprobe (Anzahl)	P
2,–	AA (1)	$\frac{1}{36}$
3,–	AB, BA (2)	$\frac{2}{36}$
4,–	AC, BB, CA (3)	$\frac{3}{36}$
5,–	AD, BC, CB, DA (4)	$\frac{4}{36}$
6,–	AE, BD, CC, DB, EA (5)	$\frac{5}{36}$
7,–	AF, BE, CD, DC, EB, FA (6)	$\frac{6}{36}$
8,–	BF, CE, DD, EC, FB (5)	$\frac{5}{36}$
9,–	CF, DE, ED, FC (4)	$\frac{4}{36}$
10,–	DF, EE, FD (3)	$\frac{3}{36}$
11,–	EF, FE (2)	$\frac{2}{36}$
12,–	FF (1)	$\frac{1}{36}$

die Wahrscheinlichkeitsfunktion der Summen aus Stichproben mit $N = 3$ zeigt.

Unter den genannten Umständen können wir also davon ausgehen, dass die Verteilung der Zufallsvariablen „Stichprobenmittelwerte“ normal ist. Der Erwartungswert dieser Verteilung ist $E(\bar{X}) = \mu$. Die Streuung der Stichprobenkennwerteverteilung des Mittelwertes heißt **Standardfehler des Mittelwertes** ($\sigma_{\bar{x}}$). Man berechnet $\sigma_{\bar{x}}$ wie folgt (Bortz & Schuster, 2010, S. 85):

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}. \qquad (12.14)$$

Stichprobenkennwerteverteilung des Mittelwertes – Die Stichprobenkennwerteverteilung des Mittelwertes folgt gemäß zentralem Grenzwerttheorem ab einer Mindeststichprobengröße von $N = 30$ einer Normalverteilung. Der Mittelwert dieser Stichprobenkennwerteverteilung ist der **Erwartungswert der Stichprobenmittelwerte** μ und die Streuung dieser Stichprobenkennwerteverteilung ist der **Standardfehler des Mittelwertes** $\sigma_{\bar{x}}$.

Da die Parameter μ und $\sigma_{\bar{x}}$ eine Normalverteilung eindeutig bestimmen, ist die Dichtefunktion der Mittelwerte bekannt, d. h., wir können errechnen, mit welcher Wahrscheinlichkeit Stichprobenmittelwerte bestimmter Größe bei gegebenem μ und $\sigma_{\bar{x}}$ auftreten (zur Dichtefunktion einer Normalverteilung s. Bortz & Schuster, 2010, S. 71). Die Wahrscheinlichkeit für Mittelwerte der Größe $\bar{X} > a$ z. B. entspricht dem Integral der Dichtefunktion zwischen a und ∞ (Abb. 12.14).

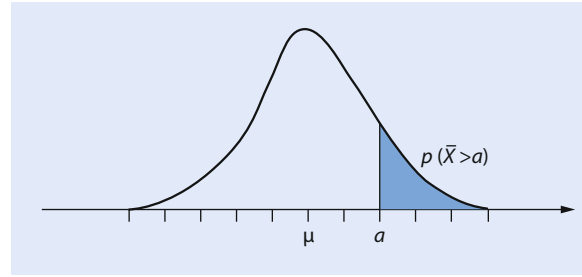


Abb. 12.14 Wahrscheinlichkeit für Mittelwerte $\bar{X} > a$

Standardnormalverteilung

Der folgende Gedankengang erleichtert die Bestimmung von Flächenanteilen einer Normalverteilung. Jede beliebige Zufallsvariable X mit dem Mittelwert μ und der Streuung σ lässt sich durch folgende Transformation (Standardisierung) in eine Zufallsvariable z mit $\mu = 0$ und der Streuung $\sigma = 1$ überführen (**z-Transformation**).

$$z = \frac{X - \mu}{\sigma}. \qquad (12.15)$$

Man beachte, dass sich durch die z-Transformation nur Mittelwert und Streuung der Verteilung ändern, nicht jedoch die Verteilungsform.

Wenden wir die z-Transformation auf die normalverteilte Zufallsvariable $\bar{X}$ an, resultiert mit

$$z_{\bar{x}} = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} \qquad (12.16)$$

eine normalverteilte Zufallsvariable mit einem Mittelwert von Null und einer Streuung von Eins. Diese Normalverteilung heißt Standardnormalverteilung. Die Flächenanteile der Standardnormalverteilung liegen in tabellierter Form vor (Tab. 1, ▶ Webangebot zu diesem Buch).

***z*-Transformation** – Die z-Transformation („z transformation“) überführt jede beliebige Verteilung in eine entsprechende Verteilung mit dem Mittelwert $\mu = 0$ und der Streuung $\sigma = 1$.
Die Anwendung der z-Transformation auf eine Normalverteilung führt zu einer Normalverteilung mit dem Mittelwert von $\mu = 0$ und der Streuung von $\sigma = 1$. Diese z-transformierte Normalverteilung wird als **Standardnormalverteilung** bezeichnet.

Damit lässt sich die Wahrscheinlichkeit, mit der Mittelwerte $\bar{X} > a$ auftreten, leicht bestimmen. Interessiert in dem oben genannten Beispiel der Verteilung von Intelligenzquotienten von Abiturienten (Abb. 12.12) die Wahrscheinlichkeit von Stichprobenmittelwerten $\bar{X} > 115$, ergeben sich für Stichproben mit $N = 36$ die folgenden Werte:

Durchschnittlicher IQ: $\mu = 110$
Varianz der IQ-Werte: $\sigma^2 = 144$
Standardfehler der Mittelwertverteilung:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{144}{36}} = \sqrt{4} = 2.00$$
$$z_{\bar{x}} = \frac{115 - 110}{2} = 2.50. \tag{12.17}$$

Folglich entspricht dem Wert $\bar{x} = 115$ der Wert $z_{\bar{x}} = 2.50$ in der Standardnormalverteilung. Wir fragen nun nach der Wahrscheinlichkeit für $z_{\bar{x}} = 2.50$, also dem Flächenanteil der Standardnormalverteilung zwischen 2.50 und ∞. Dieser lautet gem. Tab. 1 im ▶ Webangebot zu diesem Buch:

$$p(z_{\bar{x}} > 2.50) = .0062. \tag{12.18}$$

Die Wahrscheinlichkeit, in einer Stichprobe des Umfanges $N = 36$ einen Mittelwert $\bar{X} = 115$ zu erhalten, beträgt 0.62 %, wenn $\mu = 110$ und $\sigma^2 = 144$ sind.

Die Wahrscheinlichkeit, dass ein Stichprobenmittelwert mindestens 5 IQ-Punkte von μ abweicht, ermitteln wir auf ähnliche Weise:

$$\frac{105 - 110}{2} = -2.50,$$
$$p(z_{\bar{x}} > 2.50) = .0062, \tag{12.19}$$
$$p(z_{\bar{x}} < -2.50) = .0062.$$

Und

$$p(-2.50 > z_{\bar{x}} > 2.50) = .0062 + .0062 = .0124. \tag{12.20}$$

Die Wahrscheinlichkeit beträgt 1.24 %.

▪ $\bar{X}$-Wertebereiche

Hiervon ausgehend, können wir nun auch dasjenige Intervall bestimmen, in dem sich ein bestimmter Anteil p aller Stichprobenmittelwerte befindet. Setzen wir $p = .95$, benötigen wir diejenigen $Z_{\bar{x}}$-Werte, die von der Standardnormalverteilungsfläche an beiden Seiten 2.50 % abschneiden, so dass eine Restfläche von 95 % (bzw. $p = .95$) verbleibt. Die Standardnormalverteilungstabelle zeigt, dass die Werte $z_{\bar{x}} = -1.96$ und $z_{\bar{x}} = +1.96$ diese Bedingung erfüllen.

$$p(-1.96 < z_{\bar{x}} < 1.96) = .95. \tag{12.21}$$

Über die z-Transformation resultieren für $z_{\bar{x}} = -1.96$ bzw. $z_{\bar{x}} = +1.96$ die folgenden Mittelwerte:

$$-1.96 = \frac{\bar{x}_u - 110}{2};$$
$$\bar{x}_u = 2 \cdot (-1.96) + 110 = 106.08 \tag{12.22}$$

bzw.

$$1.96 = \frac{\bar{x}_0 - 110}{2}$$
$$\bar{x}_0 = 2 \cdot 1.96 + 110 = 113.92. \tag{12.23}$$

Das Intervall hat als untere Grenze ($\bar{x}_u$) den Wert 106.08 und als obere Grenze ($\bar{x}_0$) den Wert 113.92. Für eine Population mit $\mu = 110$ und $\sigma^2 = 144$ treten Mittelwerte aus Stichproben des Umfanges $N = 36$ mit 95 %iger Wahrscheinlichkeit im Bereich 106.08–113.92 auf. Mit $a = 2 \cdot 196$ ergibt sich dieser Bereich zu $\mu \pm a = 110 \pm 3.92$. Wir bezeichnen diesen Bereich zukünftig einfachheitshalber als den $\bar{X}$-Wertebereich von μ.

Nun sind jedoch auch andere Bereiche denkbar, in denen sich 95 % aller Stichprobenmittelwerte befinden. Der Standardnormalverteilungstabelle entnehmen wir z. B., dass sich zwischen den Werten $z_u = 1.75$ und $z_0 = 2.33$ (oder z. B. $z_u = -2.06$ und $z_0 = 1.88$) ebenfalls 95 % der Gesamtfläche befinden, d. h., auch innerhalb dieser Grenzen erwarten wir $z_{\bar{x}}$-Werte mit einer Wahrscheinlichkeit von $p = .95$. Unter den theoretisch unendlich vielen Bereichen der Form $a < \mu < b$ ist jedoch – wie man sich leicht überzeugen kann – das Intervall $\mu \pm 1.96$ das kürzeste: Für $a = -1.75$ und $b = 2.33$ erhalten wir eine Intervallbreite von $2.33 + 1.75 = 4.08$ (bzw. für $a = -2.06$ und $b = 1.88$ von 3.94). Setzen wir $a = -1.96$ und $b = +1.96$, resultiert die **minimale Intervallbreite** von $1.96 + 1.96 = 3.92$. Werden diese Werte wie oben mit $a = 1.96 \cdot \sigma_{\bar{x}}$ multipliziert, erhält man die entsprechenden $\bar{X}$-Wertebereiche. Den kürzesten $\bar{X}$-Wertebereich bevorzugen wir, weil dieser – wie wir noch sehen werden – zu der genauesten Schätzung des Parameters μ führt.

▪ Bestimmung des Konfidenzintervalls

In der Regel ist nicht der Parameter μ, sondern nur **ein** Stichprobenmittelwert x bekannt. Es werden nun für diejenigen $\bar{X}$-Wertebereiche, in denen sich der bekannte x-Wert befindet, die entsprechenden Parameter gesucht. Wir fragen also, bei welchen Parametern der gefundene x-Wert im 95 %igen $\bar{X}$-Wertebereich liegt.

Hierfür kommen offensichtlich alle Parameter im Bereich $x \pm a$ infrage. Nehmen wir für μ den Wert $\bar{x} + a$ an, begrenzt der gefundene $\bar{x}$-Wert den $\bar{X}$-Wertebereich dieses Parameters linksseitig, und nehmen wir für μ den Wert $\bar{x} - a$ an, wird der $\bar{X}$-Wertebereich dieses Parameters rechtsseitig durch $\bar{x}$ begrenzt. Die Parameter im Bereich $\bar{x} \pm a$ weisen damit $\bar{X}$-Wertebereiche auf, in denen sich mit Sicherheit auch der gefundene $\bar{x}$-Wert befindet.

Nun stellt jedoch – wie bereits erwähnt – $\bar{X}$ eine Zufallsvariable dar, d. h., auch $\bar{X} \pm a$ ist eine Zufallsvariable. Wir erhalten bei wiederholter Stichprobenentnahme verschiedene $\bar{x}$-Werte bzw. verschiedene Bereiche $\bar{x} \pm a$.

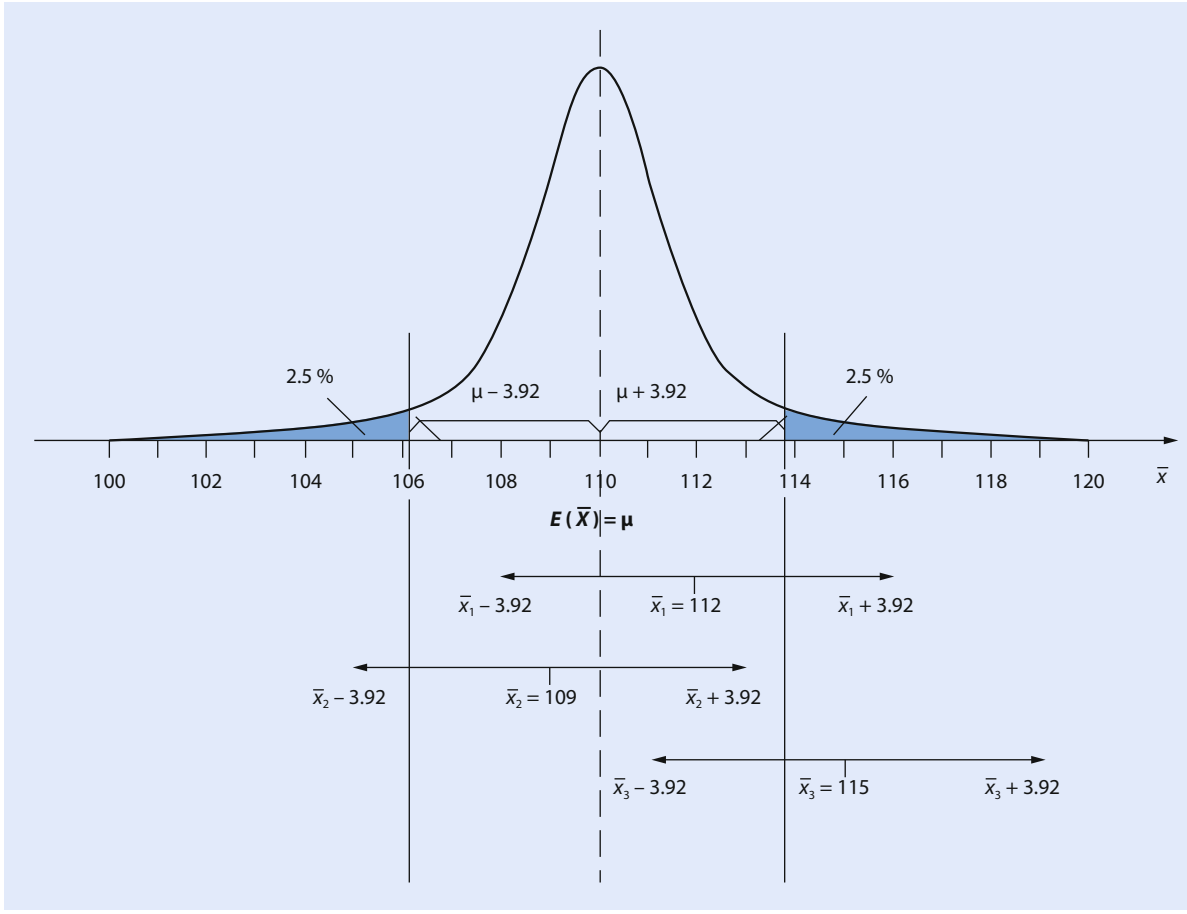


Abb. 12.15 Vergleich verschiedener Realisierung der Zufallsvariablen $\bar{X} + a$

Die Wahrscheinlichkeit des Auftretens eines bestimmten $\bar{x}$-Wertes hängt davon ab, wo sich der wahre Parameter μ befindet. $\bar{x}$-Werte, die stark von μ abweichen, sind unwahrscheinlicher als $\bar{x}$-Werte in der Nähe von μ. Je nachdem wie stark ein $\bar{x}$-Wert von μ abweicht, resultieren Parameterbereiche $\bar{X} \pm a$, in denen sich μ befindet oder in denen sich μ nicht befindet. Dies verdeutlicht Abb. 12.15.

Erhalten wir $\bar{x}_1 = 112$, kommen – wieder bezogen auf das oben genannte Beispiel – Parameter im Bereich 112 ± 3.92 infrage. In diesem Bereich befindet sich auch der wahre Parameter $\mu = 110$. Ähnliches gilt für das Stichprobenergebnis $\bar{x}_2 = 109$. Zu den Parametern, die dieses $\bar{x}$ ermöglichen, zählt auch $\mu = 110$. Ziehen wir hingegen eine Stichprobe mit $\bar{x}_3 = 115$, zählt $\mu = 110$ nicht zu den Parametern, die $\bar{x}_3 = 115$ mit 95 %iger Wahrscheinlichkeit erzeugt haben. Aufgrund des Stichprobenmittelwertes $\bar{x}_3 = 115$ würden wir also ein Intervall möglicher μ-Werte angeben, in dem sich der wahre μ-Wert tatsächlich nicht befindet.

Ziehen wir viele Stichproben des Umfanges n, erhalten wir viele mehr oder weniger verschiedene Parameterbereiche vom Typ $\bar{X} \pm a$. 95 % dieser Parameterbereiche sind richtig, denn sie umschließen den wahren Parameter, und 5 % der Parameterbereiche sind falsch, weil sich der wahre Parameter μ außerhalb dieser Bereiche befindet. Kennen wir – wie üblich – nur einen Stichprobenmittelwert $\bar{x}$, zählt der entsprechende Parameterbereich $\bar{x} \pm a$ entweder zu den richtigen oder den falschen Intervallen. Da wir aber durch die Berechnung dieses Intervalls dafür gesorgt haben, dass 95 % aller vergleichbaren Intervalle den wahren Parameter umschließen, ist es sehr plausibel oder wahrscheinlich, dass das gefundene Intervall zu den richtigen zählt.

Konfidenzintervall – Das Konfidenzintervall („confidence intervall", CI) kennzeichnet denjenigen Bereich von Merkmalsausprägungen, in dem sich 95 % (99 %) aller möglichen Populationsparameter befinden, die den empirisch ermittelten Stichprobenkennwert erzeugt haben können.

Bei der **Interpretation von Konfidenzintervallen** existieren zwei Lehrmeinungen, je nach zugrunde gelegtem Wahrscheinlichkeitskonzept:

- Gemäß **klassischem (frequentistischem) Wahrscheinlichkeitsbegriff** kann man für ein einzelnes auf der Basis empirischer Daten berechnetes Konfidenzintervall nur aussagen, dass es den wahren Populationsparameter entweder beinhaltet oder nicht, womit die Wahrscheinlichkeit dieses Einzelereignisses nur 1 oder 0 betragen kann.
- Gemäß **subjektivem Wahrscheinlichkeitsbegriff** kann man jedoch durchaus sagen, dass der gesuchte Parameter mit einer Wahrscheinlichkeit von 95 % in dem berechneten Konfidenzintervall liegt. Wenn sich in 95 % aller Ziehungen der wahre Populationsparameter innerhalb der jeweiligen Konfidenzintervalle befindet, ist die Wahrscheinlichkeit p, dass sich der wahre Populationsparameter in dem 95 % Konfidenzintervall der eben gezogenen Stichprobe befindet gleich .95; lediglich mit einer Wahrscheinlichkeit $p = .05$ befindet sich der wahre Populationsparameter außerhalb des Konfidenzintervalls: Die Wahrscheinlichkeit für ein Einzelereignis kann gemäß subjektivem Wahrscheinlichkeitskonzept alle Werte zwischen 0 und 1 annehmen (Brandstätter, 1999, S. 7; zum Verhältnis zwischen frequentistischem und subjektivem Wahrscheinlichkeitsbegriff in der Wissenschaft s. Stegmüller, 1985).

Neyman (1937) hat Intervalle des Typus $\bar{X} \pm a$ Konfidenzintervalle genannt. Die Wahrscheinlichkeit, dass ein beliebiges Intervall zu denjenigen zählt, die auch den wahren Populationsparameter μ enthalten, bezeichnete er als **Konfidenzkoeffizienten**. Für den Konfidenzkoeffizienten werden üblicherweise die Werte $p = .95$ oder $p = .99$ angenommen. Für $p = .95$ ermitteln wir $a = 1.96 \cdot \sigma_{\bar{x}}$. Für den Konfidenzkoeffizienten $p = .99$ entnehmen wir der Standardnormalverteilungstabelle die z-Werte ± 2.58, die jeweils 0.5 % (also zusammen 1 %) von den Extremen der Normalverteilungsfläche abschneiden. Das 99 %ige Konfidenzintervall lautet damit $\bar{x} \pm 2.58 \cdot \sigma_{\bar{x}}$.

Allgemein erhalten wir für das Konfidenzintervall (Δ_{krit}) des arithmetischen Mittelwerts bei bekannter Populationsvarianz oder geschätzter (bei Stichprobenumfang $n \geq 30$) Populationsvarianz:

$$\Delta_{krit(1-\alpha)} = \bar{X} \pm Z_{(\alpha/2)} \cdot \sigma_{\bar{x}}. \qquad (12.24)$$

Würde man im oben erwähnten Beispiel μ durch eine Stichprobe des Umfanges $N = 36$ mit $\bar{x} = 112$ schätzen, hätte das 99 %ige Konfidenzintervall die Grenzen

$$112 - 5.16 \leq \mu \leq 112 + 5.16 \qquad (12.25)$$

bzw.

$$106.84 \leq \mu \leq 117.16. \qquad (12.26)$$

Die hier beschriebene Vorgehensweise zur Ermittlung eines Konfidenzintervalls geht davon aus, dass der Umfang der Gesamtpopulation N im Verhältnis zum Stichprobenumfang n sehr groß ist. Für praktische Zwecke sind die hier aufgeführten Bestimmungsgleichungen (wie auch die folgenden) hinreichend genau, wenn der **Auswahlsatz** $n/N < 0.05$ ist (Schwarz, 1975).

Das *Konfidenzintervall* (Δ_{krit}) des *Populationsmittelwertes* μ lässt sich bei *bekannter Populationsvarianz* über die *Standardnormalverteilung* bestimmen, d. h. der kritische Wert für den Konfidenzkoeffizienten $1 - \alpha$ wird der Standardnormalverteilung entnommen ($z_{\alpha/2}$):

$$\Delta_{krit(1-\alpha)} = \bar{x} \pm z_{(\frac{\alpha}{2})} \cdot \sigma_{\bar{x}} \quad \text{mit} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}.$$

Die kritischen z-Werte sind – unabhängig vom Stichprobenumfang – für das 95 %ige Konfidenzintervall $z_{(2.5\,\%)} = 1.96$ und für das 99 %ige Konfidenzintervall $z_{(0.5\,\%)} = 2.58$ (► Wie umfangreich sind Diplomarbeiten? Parameterschätzung auf der Basis einer einfachen Zufallsstichprobe).

Konfidenzintervall des arithmetischen Mittels bei unbekannter Varianz

Der oben ausführlich dargestellte Fall der Bestimmung des Konfidenzintervalls des Populationsmittelwertes bei bekannter Populationsvarianz ist in der Forschungspraxis eher die Ausnahme. Häufig ist die **Populationsvarianz unbekannt** und muss über die Stichprobendaten geschätzt werden. Dementsprechend muss auf der Basis der geschätzten Populationsvarianz dann auch der Standardfehler geschätzt werden.

Im Falle unbekannter Populationsvarianz ist bei der Bestimmung des Konfidenzintervalls zu unterscheiden, ob eine kleine Stichprobe ($n < 30$) oder eine große Stichprobe ($n \geq 30$) vorliegt. Bei größerem Stichprobenumfang kann auf die Standardnormalverteilung zurückgegriffen werden, bei einer kleinen Stichprobe muss die t-Verteilung genutzt werden. Die Formeln für die Konfidenzintervalle werden im Folgenden erläutert.

Das *Konfidenzintervall* (Δ_{krit}) des *Populationsmittelwertes* μ wird bei *unbekannter Populationsvarianz* und *großen Stichproben* ($n \geq 30$) über die *Standardnormalverteilung* berechnet, d. h. der kritische Wert für den Konfidenzkoeffizienten $1 - \alpha$

Wie umfangreich sind Diplomarbeiten? Parameterschätzung auf der Basis einer einfachen Zufallsstichprobe

Eine studentische Arbeitsgruppe möchte wissen, wie umfangreich Diplom- bzw. Masterarbeiten im Fach Psychologie sind. Dazu werden in den Bibliothekskatalogen deutscher Hochschulen, die den Studiengang Psychologie anbieten, die an den psychologischen Instituten abgeschlossenen Diplom- und Masterarbeiten recherchiert. Anhand dieser Aufstellungen definiert man als Population aller Diplom- und Masterarbeiten der vergangenen 10 Jahre. Die Seitenzahlen von 100 zufällig ausgewählten Abschlussarbeiten führen zu folgenden statistischen Angaben:

$$n = 100$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 92.00$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = 1\,849.$$

Als Punktschätzung für μ resultiert also $\bar{x} = 92$. Zusätzlich möchte man wissen, bei welchen Parametern dieses Stichprobenergebnis mit 99 %iger Wahrscheinlichkeit zustande kommen kann, d. h., man interessiert sich für das 99 %ige Konfidenzintervall.

Für dessen Bestimmung ist die t-Verteilung mit $df = 99$ Freiheitsgraden heranzuziehen, denn die unbekannte Populationsvarianz σ^2 muss aus den Stichprobendaten geschätzt werden. Da jedoch $n \geq 30$ ist, entspricht diese Verteilung praktisch der Standardnormalverteilung, d. h., man ermittelt das 99 %ige Konfidenzintervall einfachheitshalber über den z-Wert, der von der Standardnormalverteilung $1\,\%/2 = 0.5\,\%$ abschneidet:

$$\bar{x} \pm z_{(0.5\,\%)} \cdot \hat{\sigma}_{\bar{x}} = 92.00 \pm 2.58 \cdot \sqrt{\frac{1\,849}{100}} \approx 92.00 \pm 11.$$

Als Grenzen des Konfidenzintervalls resultieren damit 81 Seiten und 103 Seiten. Die richtige durchschnittliche Seitenzahl liegt entweder innerhalb dieser Grenzen oder außerhalb. Aufgrund der Art der Konfidenzintervallbestimmung stehen die Chancen jedoch 99 zu 1, dass das ermittelte Konfidenzintervall den Parameter tatsächlich umschließt.

Wie sich die Parameterschätzung ändert, wenn andere probabilistische Stichprobenarten als die einfache Zufallsstichprobe genutzt werden (z. B. geschichtete Stichprobe, Klumpenstichprobe), wird im ▶ Webangebot zu diesem Buch anhand des in dieser Box präsentierten Beispiels erläutert.

wird der Standardnormalverteilung entnommen ($z_{\alpha/2}$):

$$\Delta_{krit(1-\alpha)} = \bar{x} \pm z_{(\frac{\alpha}{2})} \cdot \hat{\sigma}_{\bar{x}} \quad \text{mit} \quad \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

Die kritischen z-Werte sind – unabhängig vom Stichprobenumfang – für das 95 %ige Konfidenzintervall $z_{(2.5\,\%)} = 1.96$ und für das 99 %ige Konfidenzintervall $z_{(0.5\,\%)} = 2.58$.

Das *Konfidenzintervall* (Δ_{krit}) des *Populationsmittelwertes* μ wird bei *unbekannter Populationsvarianz* und *kleinen Stichproben* ($n < 30$) über die *t-Verteilung* berechnet (Voraussetzung: normalverteilte Population), d. h. der kritische Wert für den Konfidenzkoeffizienten $1 - \alpha$ wird der t-Verteilung mit den entsprechenden Freiheitsgraden $df = n - 1$ entnommen ($t_{\alpha/2;df}$):

$$\Delta_{krit(1-\alpha)} = \bar{x} \pm t_{(\frac{\alpha}{2},df)} \cdot \hat{\sigma}_{\bar{x}} \quad \text{mit} \quad \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}^2}{n}}.$$

Die kritischen t-Werte sind jeweils vom Stichprobenumfang bzw. von der Anzahl der daraus abgeleiteten Freiheitsgrade $df = n - 1$ abhängig.

t-Verteilung

Die z-transformierte Zufallsvariable $\bar{X}$ ist nach dem zentralen Grenzwerttheorem normalverteilt mit $\mu = 0$ und $\sigma = 1$:

$$z_{\bar{x}} = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}. \tag{12.27}$$

Diese Gleichung setzt eine bekannte Populationsstreuung σ voraus – eine Annahme, die für die Praxis unrealistisch ist. Üblicherweise sind wir darauf angewiesen, den unbekannten Parameter σ^2 durch Stichprobendaten zu schätzen. Mit $\hat{\sigma}^2$ als erwartungstreue Schätzung für σ^2 (▶ Abschn. 12.4.1) resultiert statt der normalverteilten Zufallsvariablen $z_{\bar{x}}$ die folgende Zufallsvariable t (t-Verteilung):

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}. \tag{12.28}$$

Sowohl $\bar{X}$ als auch $\hat{\sigma}$ sind stichprobenabhängig, d. h., dieser Ausdruck enthält nicht nur im Zähler, sondern auch im Nenner eine Zufallsvariable (im Unterschied zur Definition von $z_{\bar{x}}$, bei der im Nenner die Konstante $\frac{\sigma}{\sqrt{n}}$ steht). Die Eigenschaften der Zufallsvariablen t sind mathematisch kompliziert, es sei denn, $\bar{X}$ und $\hat{\sigma}$ sind voneinander unabhängig. Dies ist der Fall, wenn sich die Zufallsvariable X in der Population normalverteilt.

Der britische Statistiker William Sealy Gossett (Student, 1908) konnte zeigen, dass die Dichtefunktion der Zufallsvariablen t unter der Voraussetzung einer normalverteilten Zufallsvariablen X Eigenschaften aufweist, die denen der Standardnormalverteilung sehr ähneln. (Gossett publizierte unter dem Pseudonym „Student“; die t-Verteilung wird deshalb auch Student-Verteilung genannt.)

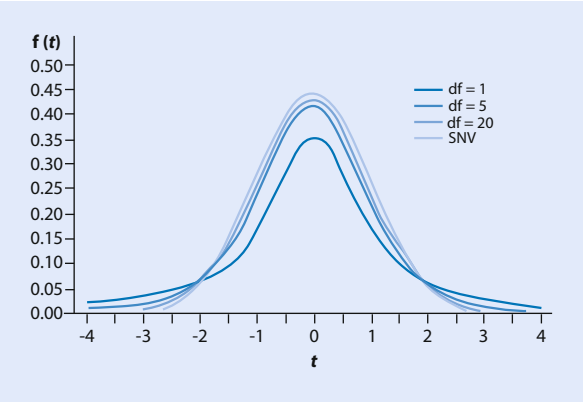


Abb. 12.16 Ausschnitt aus der Familie der t-Verteilungen im Vergleich zur Standardnormalverteilung (SNV)

Wie die Standardnormalverteilung ist auch die t-Verteilung symmetrisch und eingipflig mit einem Erwartungswert (Mittelwert) von $\mu = 0$. Ihre Standardabweichung ist durch $\sigma = \sqrt{(n-1)/(n-3)}$ (für $n > 3$) definiert, d. h., sie ist abhängig vom Umfang der Stichprobe bzw. – genauer – von der Anzahl der Abweichungen $(x_i - \bar{x})$, die bei der Ermittlung der Varianzschätzung $\hat{\sigma}^2$ frei variieren können (Abb. 12.16).

Freiheitsgrade

Die Anzahl frei variierender Abweichungen bezeichnet man als **Freiheitsgrade der Varianz**. Wie man sich leicht überzeugen kann, sind bei einer Stichprobe des Umfanges n nur $n-1$ Abweichungen frei variierbar, d. h., die Varianz hat $n-1$ Freiheitsgrade (**df** $= n-1$; df steht für „**degrees of freedom**“).

Auch hierzu ein kleines Beispiel: Von 4 Messungen weichen 3 in folgender Weise vom Mittelwert ab: $x_1 - \bar{x} = 2$, $x_2 - \bar{x} = -3$ und $x_3 - \bar{x} = -5$. Da die Summe aller 4 Differenzen null ergeben muss, resultiert für $x_4 - \bar{x}$ zwangsläufig der Wert 6, denn es gilt $2 + (-3) + (-5) + (x_4 - \bar{x}) = 0$ oder $x_4 - \bar{x} = (-2) + 3 + 5 = 6$. Von den 4 (allgemein n) Abweichungen sind also nur 3 (allgemein $n-1$) frei variierbar.

Wir erhalten damit eine **„Familie“ verschiedener t-Verteilungen**, deren Streuungen von der Anzahl der Freiheitsgrade der Varianzschätzung abhängen. Die Standardnormalverteilung (SNV) im Vergleich zu t-Verteilungen mit df $= 1$, df $= 5$ und df $= 20$ ist in Abb. 12.16 dargestellt.

Die Abbildung verdeutlicht, dass die t-Verteilungen mit wachsender Anzahl von Freiheitsgraden in die Standardnormalverteilung übergehen. Bei df > 30 ist die Ähnlichkeit beider Verteilungen bereits so groß, dass ohne besondere Genauigkeitseinbuße statt der t-Verteilung die Standardnormalverteilung verwendet werden kann. Bei großen Stichproben ist es zudem praktisch unerheblich, wie das Merkmal in der Population verteilt ist (Abb. 12.12).

Ausgewählte Flächenanteile der Verteilungsfunktionen für t-Verteilungen mit unterschiedlichen Freiheitsgraden enthält das ▶ Webangebot zu diesem Buch.

Bestimmung des Konfidenzintervalls

Die Bestimmung von Konfidenzintervallen $(\Delta_{krit(1-\alpha)})$ des Mittelwertes auf der Basis von t-Verteilungen (also bei unbekanntem bzw. durch $\hat{\sigma}^2$ geschätztem σ^2 und $n < 30$) erfolgt völlig analog der bereits behandelten Konfidenzintervallbestimmung. Wird σ^2 durch $\hat{\sigma}^2$ geschätzt, resultiert als Schätzung des Standardfehlers

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}. \tag{12.29}$$

Der z-Wert der Standardnormalverteilung (1.96 bzw. 2.58) wird durch denjenigen t-Wert ersetzt, der von der t-Verteilung mit $n-1$ Freiheitsgraden an beiden Seiten 2.5 % (für das 95 %ige Konfidenzintervall) bzw. 0.5 % (für das 99 %ige Konfidenzintervall) abschneidet. Wir erhalten dann:

$$\Delta_{krit(95\,\%)} = \bar{x} \pm t_{(\frac{5\,\%}{2},df)} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \tag{12.30}$$

(95 %iges Konfidenzintervall)

bzw.

$$\Delta_{krit(99\,\%)} = \bar{x} \pm t_{(\frac{1\,\%}{2},df)} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \tag{12.31}$$

(99 %iges Konfidenzintervall).

Ein Beispiel: Für eine Stichprobe des Umfanges $N = 9$ aus einer normalverteilten Population wurden $\bar{x} = 45$ und $\hat{\sigma}^2 = 49$ ermittelt, d. h., wir errechnen $\hat{\sigma}_x = \sqrt{\frac{49}{9}} = \frac{7}{3} = 2.33$. Der Tabelle zu t-Verteilungen im ▶ Webangebot zu diesem Buch entnehmen wir, dass der Wert $t_{(2.5\,\%;8)} = 2.306$, 2.5 % der Fläche der t-Verteilung für 8 Freiheitsgrade abschneidet. Das Konfidenzintervall heißt damit: $45 \pm 2.306 \cdot 2.33$ bzw. 45 ± 5.37. Für das 99 %ige Konfidenzintervall lesen wir in der t-Tabelle den Wert $t_{(0.5\,\%;8)} = 3.355$ ab, d. h., das Konfidenzintervall lautet $45 \pm 3.355 \cdot 2.33$ bzw. 45 ± 7.82.

Man beachte allerdings, dass Parameterschätzungen auf der Basis sehr kleiner Stichproben ($n < 30$) für praktische Zwecke der Populationsbeschreibung in deskriptiven Studien in der Regel zu ungenau sind.

Konfidenzintervall eines Populationsanteils

Interessiert als Populationsparameter nicht der Mittelwert einer kardinalskalierten Variable, sondern der Anteils-

bzw. Prozentwert π (pi) der einzelnen Ausprägungen einer nominalskalierten Variable, so kann dieser als Punktschätzung über den Stichprobenprozentwert (p) geschätzt werden. Ähnlich wie $\bar{X}$ ist jedoch auch P stichprobenabhängig, d. h., die Punktschätzung P wird π in der Regel fehlerhaft schätzen. Erneut ist es deshalb von Vorteil, wenn ein Intervall angegeben werden kann, in dem sich alle möglichen π-Werte befinden, für die der gefundene p-Wert mit einer Wahrscheinlichkeit von 95 % (oder 99 %) auftreten kann: das **Konfidenzintervall des Populationsanteils**.

Eine Repräsentativbefragung möge ergeben haben, dass sich 35 % von 200 befragten Studierenden für mündliche Gruppenprüfungen als die angenehmste Prüfungsart aussprechen. (Welche bzw. wie viele Prüfungsformen die restlichen 65 % bevorzugen, ist in diesem Zusammenhang unerheblich.) Welche Informationen lassen sich aus diesen Zahlen bzgl. des unbekannten Parameters π (Anteil der Befürworter mündlicher Gruppenprüfungen in der gesamten Studentenschaft) ableiten?

Für die Beantwortung dieser Frage müssen wir – wie auch beim Mittelwert $\bar{X}$ – die Verteilung der Zufallsvariablen P (für ein gegebenes π) bzw. die Stichprobenkennwerteverteilung von P kennen. Diese Verteilung ist unter der Bezeichnung **Binomialverteilung** in vielen Statistikbüchern tabelliert (z. B. Wirtz & Nachtigall, 1998, S. 222). Die Tabelle enthält die Wahrscheinlichkeiten, mit denen die geprüfte Merkmalsalternative bei gegebenem π und n 0-mal, 1-mal, 2-mal … oder allgemein k-mal auftritt. Der relative Merkmalsanteil p entspricht dann dem Quotienten k/n.

Die exakten Binomialverteilungstabellen beziehen sich allerdings nur auf kleinere Stichprobenumfänge, mit denen sich der unbekannte Parameter π nur sehr ungenau schätzen lässt. Bei größeren Stichproben kann man von der Tatsache Gebrauch machen, dass die Binomialverteilung für $n \times p \cdot (1-p) > 9$ hinreichend gut durch eine Normalverteilung approximiert werden kann (Sachs, 2002, S. 228), was die Bestimmung von Konfidenzintervallen erheblich erleichtert.

Die Binomialverteilung hat – bezogen auf Anteilswerte – einen Mittelwert von π und eine Streuung (Standardfehler) von

$$\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}. \tag{12.32}$$

Schätzen wir π durch p, folgt für das Konfidenzintervall:

$$p - z \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \leq \pi \leq p + z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}. \tag{12.33}$$

Erneut ist z derjenige Wert, der von den Extremen der Standardnormalverteilung 2.5 % (für das 95 %ige Konfidenzintervall) bzw. 0.5 % (für das 99 %ige Konfidenzintervall) abschneidet ($z = 1.96$ bzw. $z = 2.58$).

Im Beispiel ermitteln wir für das 95 %ige Konfidenzintervall:

$$.35 - 1.96 \cdot \sqrt{\tfrac{.35 \cdot .65}{200}} \leq \pi \leq .35 + 1.96 \cdot \sqrt{\tfrac{.35 \cdot .65}{200}}$$
$$\text{bzw. } .35 \pm .066. \tag{12.34}$$

Das *Konfidenzintervall* (Δ_{krit}) für *Populationsanteile* (π) bestimmt man nach folgender Formel, wobei die *Standardnormalverteilung* als Approximation der Binomialverteilung genutzt wird unter der Voraussetzung $n \cdot p \cdot (1-p) > 9$:

$$\Delta_{krit(1-\alpha)} = p \pm z_{(\alpha/2)} \cdot \hat{\sigma}_p$$
$$\text{mit } \hat{\sigma}_p = \sqrt{\frac{\hat{\pi} \cdot (1-\hat{\pi})}{n}} = \sqrt{\frac{p \cdot (1-p)}{n}}. \tag{12.35}$$

In Tab. 12.9 sind die Konfidenzintervalle (95 % und 99 %) für ausgewählte Stichprobenumfänge und p-Werte wiedergegeben. Die Werte in den hervorgehobenen Bereichen stellen nur grobe Schätzungen dar, weil hier die Beziehung $n \cdot p \cdot (1-p) > 9$ nicht erfüllt ist. Der Tabelle ist z. B. zu entnehmen, dass das 99 %ige Konfidenzintervall bei einem Stichprobenanteil von $P = .60$ und $N = 100$ von .47–.72 reicht. Konfidenzintervalle für hier nicht aufgeführte Werte sind relativ einfach durch Interpolation zu ermitteln. Es wird deutlich, dass sich die Konfidenzintervalle ab einem Stichprobenumfang von $N = 1\,000$ nur noch unwesentlich verkleinern. Für ein Studienbeispiel zu Konfidenzintervallen von Populationsanteilen ▶ Studienbeispiel „Berechnung von Konfidenzintervallen von Populationsanteilen: Bevölkerungsrepräsentative Daten zur Nutzung elektronischer Medien im Jugendalter in Deutschland“.

Optimale Stichprobenumfänge für Konfidenzintervalle

Zur Planung populationsbeschreibender Untersuchungen gehören auch Überlegungen, wie groß die zu erhebende Stichprobe sein soll. Eindeutige Angaben über den **„optimalen“ Stichprobenumfang** (▶ Abschn. 14.3) sind jedoch ohne weitere Zusatzinformationen nicht möglich. Die optimale Größe der Stichprobe hängt von der gewünschten Schätzgenauigkeit und natürlich auch von den finanziellen und zeitlichen Rahmenbedingungen der Untersuchung ab.

Die Ausführungen über Konfidenzintervalle machten deutlich, dass die Genauigkeit der Schätzungen von Populationsparametern mit wachsendem Stichprobenumfang

Tabelle 12.9 Konfidenzintervalle (KI) für Populationsanteile bei variablen n und p. (1. Zeile: KI 95 %, 2. Zeile: KI 99 %; fettgedruckte KI: ungenaue Schätzungen, da $n \cdot p \cdot (1 - p) > 9$ nicht erfüllt)

$n\,p$	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
50	**.02–.15**	**.04–.21**	**.11–.33**	.19–.44	.28–.54	.37–.63	.46–.72	.56–.81	**.67–.89**	**.79–.96**	**.85–.98**
	.01–.19	**.03–.26**	**.09–.38**	.16–.48	.24–.58	.33–.67	.42–.76	.52–.84	**.62–.91**	**.74–.97**	**.81–.99**
60	**.02–.14**	**.05–.20**	.12–.32	.20–.43	.29–.53	.38–.62	.47–.71	.57–.80	.68–.88	**.80–.95**	**.86–.98**
	.01–.18	**.04–.24**	.10–.36	.17–.47	.25–.57	.34–.66	.43–.75	.53–.83	.64–.90	**.76–.96**	**.82–.99**
70	**.02–.13**	**.05–.19**	.12–.31	.20–.42	.29–.52	.39–.61	.48–.71	.58–.79	.69–.88	**.81–.95**	**.87–.98**
	.01–.16	**.04–.23**	.11–.35	.18–.45	.26–.55	.35–.65	.45–.74	.55–.82	.65–.89	**.77–.96**	**.84–.99**
80	**.02–.12**	**.05–.19**	.13–.30	.21–.41	.30–.51	.39–.61	.49–.70	.59–.79	.70–.87	**.81–.95**	**.88–.98**
	.02–.15	**.04–.22**	.11–.34	.19–.44	.27–.54	.36–.64	.46–.73	.56–.81	.66–.89	**.78–.96**	**.85–.99**
90	**.02–.12**	**.05–.18**	.13–.29	.22–.40	.30–.50	.40–.60	.50–.70	.60–.78	.71–.87	**.82–.95**	**.88–.98**
	.02–.15	**.04–.21**	.11–.33	.19–.43	.28–.54	.37–.63	.46–.72	.57–.81	.67–.89	**.79–.96**	**.85–.98**
100	**.02–.11**	.06–.17	.13–.29	.22–.40	.31–.50	.40–.60	.50–.69	.60–.78	.71–.87	.83–.94	**.89–.98**
	.02–.14	.05–.20	.12–.32	.20–.43	.28–.53	.38–.62	.47–.72	.57–.80	.68–.88	.80–.95	**.86–.98**
150	**.03–.10**	.06–.16	.14–.27	.23–.38	.33–.48	.42–.58	.52–.68	.62–.77	.73–.86	.84–.94	**.90–.98**
	.02–.12	.05–.18	.13–.30	.21–.40	.30–.51	.40–.60	.49–.70	.60–.79	.70–.87	.82–.95	**.88–.98**
200	.03–.09	.07–.15	.15–.26	.24–.37	.33–.47	.43–.57	.53–.67	.63–.76	.74–.85	.85–.93	.91–.97
	.02–.11	.06–.17	.14–.28	.22–.39	.31–.49	.41–.59	.51–.68	.61–.78	.72–.86	.83–.94	.89–.98
300	.03–.08	.07–.14	.16–.25	.25–.35	.35–.46	.44–.56	.54–.65	.65–.75	.75–.84	.86–.93	.92–.97
	.03–.09	.06–.15	.15–.27	.24–.37	.33–.47	.43–.57	.53–.67	.63–.76	.73–.85	.85–.94	.91–.97
400	.03–.08	.07–.13	.16–.24	.26–.35	.35–.45	.45–.55	.55–.65	.65–.74	.76–.84	.87–.93	.92–.97
	.03–.09	.07–.15	.15–.26	.24–.36	.34–.46	.44–.56	.54–.66	.64–.76	.74–.85	.85–.93	.91–.97
500	.03–.07	.08–.13	.17–.24	.26–.34	.36–.44	.46–.54	.56–.64	.66–.74	.76–.83	.87–.92	.93–.97
	.03–.08	.07–.14	.16–.25	.25–.36	.35–.46	.44–.56	.54–.65	.64–.75	.75–.84	.86–.93	.92–.97
1 000	.04–.07	.08–.12	.18–.23	.27–.33	.37–.43	.47–.53	.57–.63	.67–.73	.77–.82	.88–.92	.93–.96
	.03–.07	.08–.13	.17–.23	.26–.34	.36–.44	.46–.54	.56–.64	.66–.74	.77–.83	.88–.92	.93–.97
2 000	.04–.06	.09–.11	.18–.22	.28–.32	.38–.42	.48–.52	.58–.62	.68–.72	.78–.82	.89–.91	.94–.96
	.04–.06	.08–.12	.18–.22	.27–.33	.37–.43	.47–.53	.57–.63	.67–.73	.78–.82	.88–.92	.94–.96
5 000	.04–.06	.09–.11	.19–.21	.29–.31	.39–.41	.49–.51	.59–.61	.69–.71	.79–.81	.89–.91	.94–.96
	.04–.06	.09–.11	.19–.22	.28–.32	.38–.42	.48–.52	.58–.62	.68–.72	.79–.81	.89–.91	.94–.96

zunimmt (die Konfidenzintervalle werden kleiner), woraus zu folgern wäre, dass der Stichprobenumfang immer möglichst groß sein sollte. Auf der anderen Seite wurde bereits festgestellt, dass die Genauigkeit nicht proportional zum Stichprobenumfang zunimmt: Der Zugewinn an Genauigkeit ist bei Vergrößerung einer Stichprobe von 1 000 auf 1 100 unverhältnismäßig kleiner als bei Vergrößerung einer Stichprobe von 100 auf 200 (Tab. 12.9).

Studienbeispiel

Berechnung von Konfidenzintervallen von Populationsanteilen: Bevölkerungsrepräsentative Daten zur Nutzung elektronischer Medien im Jugendalter in Deutschland

Ziel des Kinder- und Jugendgesundheitssurveys (KiGGS), das von 2003–2006 durch das Robert Koch-Institut (RKI) durchgeführt wurde, war es, erstmals umfassende und bundesweit repräsentative Daten zum Gesundheitszustand von Kindern und Jugendlichen im Alter von 0–17 Jahren zu erheben. An der Studie haben insgesamt 17 641 Kinder und Jugendliche (8 985 Jungen und 8 656 Mädchen) aus 167 für die Bundesrepublik repräsentativen Städten und Gemeinden teilgenommen. Erfasst wurden einerseits diverse Aspekte der Gesundheit sowie andererseits des Lebensstils.

Für die Analysen zur Mediennutzung wurde auf Daten aus einer schriftlichen Befragung von $N = 6\,813$ 11- bis 17-jährigen Jungen und Mädchen zurückgegriffen. Erhoben wurden Mediennutzungszeiten, die im Rahmen weiterer Analysen z. B. mit sportlicher Fitness oder Übergewicht in Verbindung gebracht wurden. Die Anteilswerte der Gruppe der Vielnutzer (mehr als 3 Stunden Nutzungszeit des jeweiligen Mediums pro Tag) werden in Tab. 12.10 differenziert nach Geschlecht und Sozialstatus dargestellt. Der Sozialstatus wurde anhand eines mehrdimensionalen Index bestimmt, der auf der Basis von Angaben der Eltern zu deren schulischer und beruflicher Ausbildung, beruflichen Stellung sowie zum Haushaltsnettoeinkommen ermittelt wurde und eine Einteilung in eine niedrige, mittlere und hohe Statusgruppe ermöglicht.

Zur Interpretation der linken oberen Zelle: Von den untersuchten 11- bis 17-jährigen Mädchen mit niedrigem Sozialstatus gaben 34 % im Fragebogen an, mehr als drei Stunden pro Tag fernzusehen. Die Punktschätzung von 34 % für den Populationsanteilswert wird ergänzt durch die Intervallschätzung. Das 95 %ige Konfidenzintervall zeigt, dass der wahre Populationsanteil zwischen 30.5 % und 37.5 % anzusetzen ist.

Dieses Konfidenzintervall überlappt stark mit dem Konfidenzintervall der statusniedrigen Jungen von 28.2 % bis 35 %. Der wahre Anteilswert beider Gruppen könnte also bei 31 %, 32 %, 33 %, 34 % oder 35 % liegen. Anders ist es beim Anteil der Intensivnutzer von Computer und Mobiltelefon: Hier überlappen sich die Konfidenzintervalle der statusniedrigen Jungen und Mädchen nicht, so dass davon auszugehen ist, dass in der Population ein überzufälliger – nicht auf Stichprobeneffekte zurückführbarer – Geschlechtsunterschied vorliegt (zur Nutzung von Konfidenzintervallen zur Überprüfung von Unterschiedshypothesen ▶ Abschn. 12.5.3).

Demgegenüber dürften sich die Kosten einer Untersuchung mehr oder weniger proportional zum Stichprobenumfang ändern.

Genauigkeit und Kosten einer Untersuchung hängen damit wechselseitig, wenn auch nicht proportional voneinander ab. Steht zur Finanzierung einer Untersuchung ein bestimmter Betrag fest, lässt sich der maximal untersuchbare Stichprobenumfang ermitteln, der seinerseits die Genauigkeit der Untersuchung bestimmt. Ist umgekehrt die Genauigkeit, mit der die Population beschrieben werden soll, vorgegeben, sind hieraus der erforderliche Stichprobenumfang und damit auch die notwendigen Untersuchungskosten abschätzbar.

Mit wachsendem Stichprobenumfang steigt die Genauigkeit von Parameterschätzungen; gleichzeitig vergrößern sich aber auch Kosten und Aufwand der Untersuchung erheblich. Dies bedeutet, dass Vorstellungen über die Präzision der Parameterschätzung und über den Untersuchungsaufwand in der Planungsphase aufeinander abgestimmt werden sollten. Bei *zu kleinen Stichproben* ist das Ergebnis des Konfidenzintervalls nicht aussagekräftig (das Konfidenzintervall ist zu breit), bei *zu großen Stichproben* wird unnötig hoher Aufwand betrieben. Ideal ist es deswegen, bereits bei der Untersuchungsplanung den *optimalen Stichprobenumfang* festzulegen.

Optimaler Stichprobenumfang für das Konfidenzintervall eines Populationsanteils

Ein unbekannter Populationsanteil lässt sich auf der Basis eines empirischen Stichprobenanteils umso genauer schätzen (d. h. mit engerem Konfidenzintervall), je größer die untersuchte Zufallsstichprobe ist (Tab. 12.9). Wenn man vor der Durchführung einer populationsbeschreibenden Studie festlegt, wie genau man die Populationsparameter schätzen will (d. h. Vorgabe des maximal tolerierbaren Konfidenzintervalls), so kann man den dafür notwendigen Mindeststichprobenumfang kalkulieren. Dies setzt allerdings voraus, dass man bereits vor der Untersuchung eine Vorstellung über die Größe des Populationsanteils hat. Je nach Fragestellung greift man hierfür auf Untersuchungen mit ähnlicher Thematik oder Erfahrungswerte zurück. Ist dies nicht möglich, sind kleinere Voruntersuchungen angebracht, die zumindest über die Größenordnung des π-Wertes informieren. Das ▶ Studienbeispiel „Ermittlung des optimalen Stichprobenumfangs für das Konfidenzintervall des Populationsanteils“ soll zeigen, wie Abb. 12.17 für die Bestimmung des erforderlichen Stichprobenumfangs eingesetzt werden kann.

Optimaler Stichprobenumfang für das Konfidenzintervall des arithmetischen Mittels

Als Nächstes fragen wir, welcher Stichprobenumfang erforderlich ist, um einen Mittelwertparameter μ mit vorgegebener Genauigkeit schätzen zu können. Hierzu lösen wir Gl. (12.16) nach n auf:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{e}{\frac{\sigma}{\sqrt{n}}}$$
$$n = \frac{z^2 \cdot \sigma^2}{e^2}, \tag{12.36}$$

wobei e den **Schätzfehler** $\bar{X} - \mu$ symbolisiert.

Tabelle 12.10 Starke Nutzung elektronischer Medien (3 Stunden und mehr pro Tag) bei 11- bis 17-jährigen Mädchen und Jungen nach Sozialstatus: Punktschätzungen und Intervallschätzungen (95 %ige Konfidenzintervalle) der Anteilswerte (in %) in der deutschen Bevölkerung. (Daten aus Lampert, Sygusch & Schlack, 2007, S. 647f.; man beachte, dass jede Zelle ein unterschiedliches *n* aufweist, das hier nicht berichtet wird)

	Fernsehen/Video		Computer/Internet		Mobiltelefon	
Sozialstatus	**Mädchen**	**Jungen**	**Mädchen**	**Jungen**	**Mädchen**	**Jungen**
Niedrig	34.0 30.5–37.5	31.5 28.2–35.0	8.0 5.9–10.7	16.2 13.7–19.0	16.7 14.0–19.7	8.9 6.9–11.3
Mittel	22.6 20.3–25.0	21.0 18.7–23.5	6.3 5.0–8.0	18.4 16.2–20.8	13.5 11.7–15.4	4.9 3.9–6.2
Hoch	13.3 10.8–16.3	12.4 10.1–15.2	4.8 3.2–7.0	13.7 11.3–16.5	6.9 5.2–9.0	1.9 1.2–3.0

	Größe des Schätzfehlers (in σ-Einheiten)																			
	0.01		0.02		0.03		0.04		0.05		0.06		0.07		0.08		0.09		0.10	
95 %iges Konfidenzintervall	38 416		9 604		4 268		2 401		1 537		1 067		784		600		474		384	
99 %iges Konfidenzintervall	66 564		16 641		7 396		4 160		2 663		1 849		1 358		1 040		821		665	
	0.12	0.14	0.16	0.18	0.20	0.22	0.24	0.26	0.28	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.70	0.80	0.90	1.0
95 %iges Konfidenzintervall	266	196	150	119	96	79	67	57	49	43	31	24	19	15	13	11	9	6	5	4
99 %iges Konfidenzintervall	426	339	260	205	166	138	116	98	85	74	54	42	33	27	22	18	14	10	8	7

Abb. 12.17 Stichprobenumfänge für Konfidenzintervalle von μ mit unterschiedlichen Schätzfehlern

Aus Gl. (12.36) ergibt sich,

- dass der optimale Stichprobenumfang mit **abnehmendem Schätzfehler** quadratisch wächst: Je genauer geschätzt werden soll bzw. je enger das Konfidenzintervall ausfallen soll, umso größer muss die optimale Stichprobe sein.
- dass der optimale Stichprobenumfang mit **größerem Konfidenzkoeffizienten** bzw. z-Wert quadratisch wächst: Für ein 99 %iges Konfidenzintervall ($z = 2.58$) muss also eine größere optimale Stichprobe gewählt werden als für ein 95 %iges Konfidenzintervall ($z = 1.96$).
- dass der optimale Stichprobenumfang mit **größerer Populationsstreuung** quadratisch wächst. Wenn ein Merkmal eine größere Streuung aufweist, ist ein größerer Stichprobenumfang notwendig, um ein enges Konfidenzintervall zu bestimmen. Ein Schätzfehler von $e = 5$ bei einer Streuung von $\sigma = 10$ entspricht dabei hinsichtlich des optimalen Stichprobenumfangs z. B. einem Schätzfehler von $e = 50$ bei einer Streuung von $\sigma = 100$. Dies verdeutlichen z. B. Längenmessungen in Metern und in Zentimetern. Einer Streuung von $\sigma = 1$ m entspricht eine Streuung von $\sigma = 100$ cm. Damit ist ein Schätzfehler von 0.1 m auf der Meterskala einem Schätzfehler von 10 cm auf der Zentimeterskala gleichwertig. Er beträgt in beiden Fällen 10 % der Streuung.

Soll z. B. der Schätzfehler e nicht größer als 10 % der Merkmalsstreuung sein, ist für das 95 %ige Konfidenzintervall folgender Stichprobenumfang erforderlich:

$$\sqrt{n} = \frac{1.96 \cdot \sigma}{0.10 \cdot \sigma} \tag{12.37}$$

oder

$$n = \frac{1.96^2 \cdot \sigma^2}{0.01 \cdot \sigma^2} = \frac{1.96^2}{0.01} \approx 384. \tag{12.38}$$

Auf der Basis dieser Bestimmungsgleichung fasst Abb. 12.17 diejenigen Stichprobenumfänge zusammen, die benötigt werden, um einen Parameter μ mit unterschiedlicher Genauigkeit zu schätzen.

Die Benutzung dieser Tabelle sei an einem Beispiel demonstriert (▶ Studienbeispiel „Ermittlung des optimalen Stichprobenumfangs für das Konfidenzintervall des arithmetischen Mittelwerts“).

Das Beispiel macht deutlich, dass die Streuung des Merkmals in der Population ungefähr bekannt sein muss, um den zulässigen Schätzfehler vorab festlegen und damit auch den optimalen Stichprobenumfang bestimmen

Studienbeispiel

Ermittlung des optimalen Stichprobenumfangs für das Konfidenzintervall des Populationsanteils

Der Vorstand einer Gewerkschaft plant, den Gewerkschaftsmitgliedern Fortbildungskurse zum Datenschutz anzubieten. Um die Anzahl der hierfür erforderlichen Lehrkräfte, das benötigte Unterrichtsmaterial, Räume, Kosten etc. abschätzen zu können, beschließt man, das Interesse der Gewerkschaftsmitglieder an Fortbildungskursen zum Datenschutz durch eine Umfrage zu erkunden. Da eine eventuelle Fehlplanung erhebliche organisatorische Schwierigkeiten und finanzielle Zusatzbelastungen nach sich ziehen würde, wird ein Stichprobenergebnis gefordert, das den Anteil derjenigen Mitglieder, die später tatsächlich an dem Fortbildungskurs zum Datenschutz teilnehmen, möglichst genau schätzt. Man hält eine Fehlertoleranz von $\pm 5\,\%$ gerade noch für zumutbar. Für das Intervall $p \pm .05$ wird ein Konfidenzkoeffizient von 99 % vorgegeben. Es stellt sich nun die Frage, welcher Stichprobenumfang diese Schätzgenauigkeit gewährleistet.

Aus der Vergangenheit sei bekannt, dass ähnliche Fortbildungsmaßnahmen von ca. 40 % aller Mitglieder wahrgenommen werden. Die Kurse fallen jedoch in die Sommermonate, und man schätzt deshalb den Anteil der Interessierten eher niedriger ein. Tatsächlich zeigt eine vor der eigentlichen Untersuchung durchgeführte kleine Befragung von 50 Mitgliedern, dass nur 10 Personen, also 20 %, bereit wären, an den Kursen teilzunehmen. Man kann also davon ausgehen, dass der Populationsanteil π zwischen 20 % und 40 % liegt.

Aus Tab. 12.9 ist zu entnehmen, dass für $P = .40$ ein Stichprobenumfang von $N = 500$ ausreichen würde, um den Parameter mit der angestrebten Fehlertoleranz schätzen zu können. (Für $P = .40$ und $N = 500$ hat das 99 %ige Konfidenzintervall die Grenzen .35 und .46). Sollte der Populationsanteil π jedoch den kleinsten, gerade noch für möglich gehaltenen Wert von $\pi = .20$ annehmen, würden 400 Personen genügen, um den Parameter mit der gewünschten Genauigkeit zu schätzen. (Das entsprechende Konfidenzintervall lautet .15–.26.) Man entschließt sich, eine Zufallsstichprobe von $N = 500$ Gewerkschaftsmitgliedern (die per Los aus dem Mitgliederverzeichnis gezogen werden) zu befragen, weil diese auch im ungünstigsten Fall (für $P = .40$) eine akzeptable Schätzgenauigkeit gewährleistet. Angesichts der Problematik teilweise geringer Rücklaufquoten ist zu beachten, dass für eine Netto-Stichprobe von $N = 500$ eine größere Zahl von Mitgliedern zur Befragung einzuladen ist.

Studienbeispiel

Ermittlung des optimalen Stichprobenumfangs für das Konfidenzintervall des arithmetischen Mittelwerts

Eine Lehrerin interessiert sich für die Frage, wie viel Zeit 11-jährige Schulkinder täglich für ihre Hausaufgaben aufwenden. Für ihre Untersuchung nimmt sie in Kauf, dass die wahre Durchschnittszeit um maximal 5 Minuten verschätzt wird. Das zu ermittelnde Konfidenzintervall soll mit einem Konfidenzkoeffizienten von 95 % abgesichert werden.

Um erste Anhaltspunkte über die Streuung des Merkmals „Zeit für Hausaufgaben“ zu erhalten, befragt sie zunächst 20 Kinder ihrer Schule. Die Angaben schwanken zwischen 10 Minuten und 2 Stunden. Anhand dieser Werte schätzt die Lehrerin eine Streuung von $\sigma = 35$ min (als Orientierungshilfe für die Streuungsschätzung; Abb. 12.18a–e). Der Fehlergröße von 5 min entspricht damit ein Streuungsanteil von $1/7\sigma$ bzw. ca. 0.14σ. Abb. 12.17 zeigt, dass bei dieser angestrebten Schätzgenauigkeit eine Zufallsstichprobe mit $N = 196$ zu befragen wäre.

zu können. Hierin liegt die Schwierigkeit bei der Kalkulation von optimalen Stichprobenumfängen für Intervallschätzungen des arithmetischen Mittels. Folgende drei Vorgehensweisen sind möglich:

1. **Direkte Schätzung der Populationsstreuung aufgrund von Vorinformationen:** Entweder der Literatur sind Angaben zur Streuung des interessierenden Merkmals in der Zielpopulation zu entnehmen oder eine eigene Vorstudie wird durchgeführt und die Populationsstreuung aus der korrigierten Stichprobenstreuung geschätzt.
2. **Indirekte Schätzung der Populationsstreuung über den Range**: Liegen zumindest Vorinformationen über den Range und die Verteilungsform des interessierenden Merkmals in der Zielpopulation vor, so lässt sich die Populationsstreuung indirekt aus dem Range bzw. der Variationsbreite ableiten (▶ Punktschätzung von Populationsstreuungen auf der Basis der Variationsbreite).
3. **Schätzung der Populationsstreuung während der Hauptuntersuchung**: Liegen überhaupt keine Angaben über die mutmaßliche Größe von σ oder R vor, bleibt letztlich nur die Möglichkeit, den endgültigen Stichprobenumfang erst während der Datenerhebung festzulegen. Man errechnet z. B. aus den ersten 20 Messwerten eine vorläufige Streuungsschätzung, die für eine erste Schätzung des erforderlichen Stichprobenumfangs herangezogen wird. Liegen weitere Messwerte (z. B. insgesamt 40 Messwerte) vor, wird die Streuung erneut berechnet und die Stichprobengröße ggf. korrigiert. Dieser Korrekturvorgang setzt sich so lange fort, bis sich die Streuungsschätzung stabilisiert oder der zuletzt errechnete Stichprobenumfang erreicht ist.

12.4.3 Präzisierung von Parameterschätzungen

Bisher erfolgte die Beschreibung von Populationen bzw. die Schätzung von Populationsparametern aufgrund **einfacher Zufallsstichproben** („random samples“), was die Methode der Wahl ist, wenn man tatsächlich über eine vollständige Liste aller Objekte der Grundgesamtheit verfügt, so dass alle Objekte mit gleicher Wahrscheinlichkeit Mitglied der Stichprobe werden können (zur Zufallsstich-

Punktschätzung von Populationsstreuungen auf der Basis der Variationsbreite

Liegen keine direkten Anhaltspunkte zur Populationsstreuung vor, so lässt sich diese aus einer Transformationen der **Variations-/Streubreite** bzw. des **Range** R ableiten.
Der Range ist bei stetigen Merkmalen als die Differenz des größten erwarteten Wertes (X_{max}) und des kleinsten erwarteten Wertes (X_{min}) definiert ($R = X_{max} - X_{min}$). Bei diskreten Merkmalen entspricht X_{min} der unteren Kategoriengrenze des ersten Intervalls und X_{max} der oberen Kategoriengrenze des letzten Intervalls. Wenn zusätzlich auch die **Verteilungsform des Merkmals** ungefähr bekannt ist, lässt sich aus dem Range relativ einfach die Streuung abschätzen. Hierfür gelten die folgenden Regeln (Schwarz, 1975; Sachs, 2002; oder ausführlicher Schwarz, 1960, 1966):
2-Punkte-Verteilung. Die größte Streuung resultiert, wenn jeweils die Hälfte aller Merkmalsträger die Werte X_{max} und X_{min} annehmen (Abb. 12.18a). Sie hat dann den Wert

$$\hat{\sigma} = \frac{X_{max} - X_{min}}{2} = \frac{R}{2} = 0.5 \cdot R.$$

Sind die beiden extremen Merkmalsausprägungen nicht gleich häufig besetzt, reduziert sich die Streuung. Sie lässt sich bestimmen, wenn zusätzlich die Größenordnung des Mittelwertes $\bar{X}$ der Verteilung bekannt ist:

$$\hat{\sigma} = \sqrt{(X_{max} - \bar{X}) \cdot (\bar{X} - X_{min})}.$$

Gleichverteilung. Merkmale, die zwischen X_{min} und X_{max} in etwa gleichverteilt sind (Abb. 12.18b) haben eine Streuung von

$$\hat{\sigma} = \frac{1}{\sqrt{3}} \cdot \frac{R}{2} = \frac{R}{\sqrt{12}} = 0.289 \cdot R.$$

Die Streuung von Merkmalen, die in zwei Bereichen unterschiedlich gleichverteilt sind (konstante Dichte in einem Bereich und konstante, aber andere Dichte in einem anderen Bereich), lässt sich ermitteln, wenn sich für $\bar{X}$ ein plausibler Wert angeben lässt:

$$\hat{\sigma} = \frac{1}{\sqrt{3}} \cdot \sqrt{(X_{max} - \bar{X}) \cdot (\bar{X} - X_{min})}.$$

Dreiecksverteilung. Merkmale, deren Dichte von einem Merkmalsextrem zum anderen kontinuierlich sinkt (oder steigt), heißen Dreiecksverteilung (Abb. 12.18c). Für Merkmale mit dieser Verteilungsform kann die Streuung nach folgender Gleichung geschätzt werden:

$$\hat{\sigma} = \frac{R}{\sqrt{18}} = 0.236 \cdot R.$$

Gleichseitige Dreiecksverteilung. Bei einem häufig anzutreffenden Verteilungsmodell strebt die Dichte vom Merkmalszentrum aus nach beiden Seiten gegen Null (Abb. 12.18d). Für diese Verteilungsform lässt sich die Streuung in folgender Weise schätzen:

$$\hat{\sigma} = \frac{R}{\sqrt{24}} = 0.204 \cdot R.$$

Normalverteilung. Ist es realistisch, für das untersuchte Merkmal eine Normalverteilung anzunehmen, ermöglicht die folgende Gleichung eine brauchbare Streuungsschätzung:

$$\hat{\sigma} = \frac{R}{5.15} = 0.194 \cdot R.$$

Die Streubreite R entspricht hierbei einem Intervall, in dem sich etwa 99 % aller Werte befinden (Abb. 12.18e).

probe ► Abschn. 9.3.2). Die Praxis lehrt uns jedoch, dass für viele Zielpopulationen derartige Listen nicht existieren bzw. nur mit einem unzumutbaren Aufwand erstellt werden können.

Will man dennoch auf eine probabilistische Stichprobe nicht verzichten (was für die seriöse wissenschaftliche Umfrageforschung, epidemiologische Erhebungen und andere Anwendungsfälle populationsbeschreibender Studien unabdingbar ist), kann auf **weitere probabilistische Stichprobentypen** zurückgegriffen werden: Auf geschichtete/stratifizierte Stichproben, auf Klumpenstichproben sowie auf mehrstufige Stichproben und wiederholte Stichprobenuntersuchungen im Sinne von Paneluntersuchungen (zu Stichprobenarten für quantitative Studien ► Abschn. 9.3).

Diese Stichprobentechniken sind nicht nur oft **forschungsökonomisch günstiger**, sondern haben gegenüber der einfachen Zufallsstichprobe auch den Vorteil einer **genaueren Parameterschätzung**. Dies setzt allerdings voraus, dass man bereits vor der Untersuchung weiß, welche weiteren Variablen mit dem interessierenden Merkmal zusammenhängen bzw. wie dieses Merkmal ungefähr verteilt ist. Wenn derartige Vorkenntnisse geschickt im jeweiligen Stichprobenplan eingesetzt werden, kann sich dieses Wissen in Form von präziseren Parameterschätzungen mehr als bezahlt machen.

Beispiel: Wird für die Schätzung eines Populationsparameters (z. B. finanzielle Ausgaben von Auszubildenden für Genussmittel) anstelle einer einfachen Zufallsstichprobe eine **geschichtete Stichprobe** gezogen, so kann dadurch eine deutliche Reduzierung des Standardfehlers bzw. eine Erhöhung der Schätzgenauigkeit erzielt werden, was gleichzeitig auch den notwendigen optimalen Stichprobenumfang reduziert. Dafür ist es jedoch erforderlich, ein geeignetes Schichtungsmerkmal zu kennen, das in engem Zusammenhang mit der zu schätzenden Variable steht und gleichzeitig im Zuge der Stichprobenkonstruktion auch forschungspraktisch problemlos zu berücksichtigen ist (z. B. Alter der Auszubildenden). Idealerweise wählt man die Aufteilung der Schichten so, dass einerseits die Größe der Schicht in der Population (große Schichten werden auch stärker in der Stichprobe repräs-

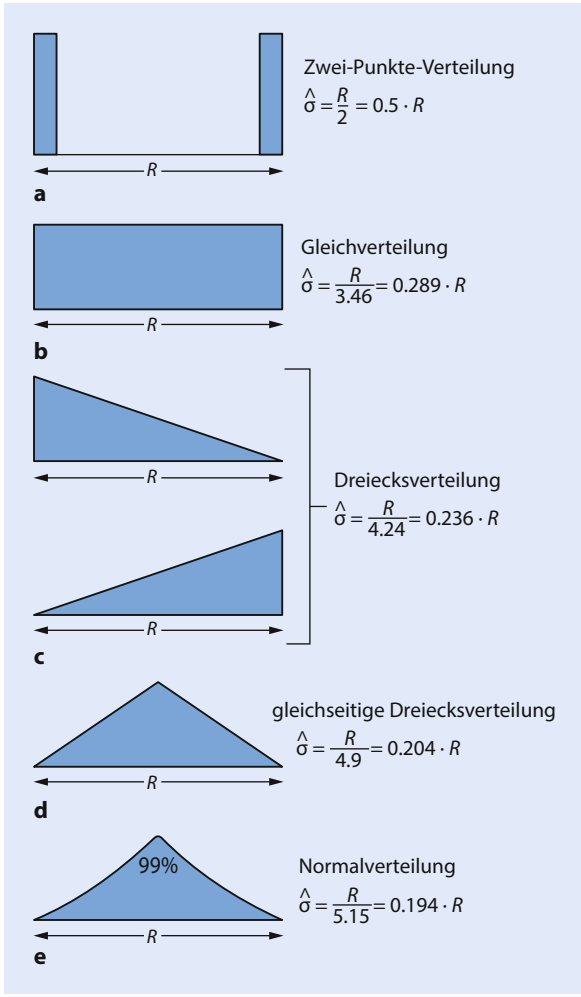


Abb. 12.18 **a–e** Streuungsschätzungen für verschiedene Verteilungsformen. **a** Zwei-Punkte-Verteilung, **b** Gleichverteilung, **c** Dreiecksverteilung, **d** gleichseitige Dreiecksverteilung, **e** Normalverteilung. R = Range bzw. Variationsbreite

sentiert) und andererseits die Populationsstreuung jeder Schicht (bei größerer Streuung wird eine größere Teilstichprobe der jeweiligen Schicht gezogen) berücksichtigt werden. Insgesamt ist auf diese Weise bei vergleichbarer Schätzgenauigkeit für die Gesamtstichprobe der geschichteten Stichprobe ein kleineres n notwendig als bei einer einfachen Zufallsstichprobe. Die bereits eingeführten Formeln für das Konfidenzintervall des arithmetischen Mittels, das Konfidenzintervall eines Populationsanteils sowie die jeweiligen optimalen Stichprobenumfänge sind für den Anwendungsfall der optimal geschichteten Stichprobe entsprechend abzuwandeln.

Eine ausführliche Darstellung der Präzisierung von Parameterschätzungen bei geschichteten und anderen probabilistischen Stichproben – jeweils mit Formeln und Rechenbeispielen – ist dem ▶ Webangebot zu diesem Buch zu entnehmen. Aus Platzgründen wird an dieser Stelle auf Details verzichtet, da probabilistische Stichproben und Studien mit dem Anspruch der Parameterschätzung in der akademischen Forschung im Fach Psychologie sowie in Nachbardisziplinen deutlich seltener vorkommen als hypothesenprüfende Studien, denen wir uns im nächsten Teilkapitel umfassend widmen werden.

Lernquiz

1. Welche Qualitätskriterien sollen Punktschätzer erfüllen?
2. Bei der Körpergröße handelt es sich um eine normalverteilte Zufallsvariable. Wie groß ist die Wahrscheinlichkeit für das Auftreten einer ganz bestimmten Körpergröße bei einer untersuchten Person?
3. Welche Aussagen stimmen?
   a) Je höher der Konfidenzkoeffizient, umso breiter ist das Konfidenzintervall.
   b) Je größer die Populationsstreuung, umso breiter ist das Konfidenzintervall.
   c) Die Varianz einer einfachen Zufallsstichprobe ist ein erwartungstreuer Schätzer der Populationsvarianz.
   d) Der Mittelwert einer einfachen Zufallsstichprobe ist ein erwartungstreuer Schätzer des Populationsmittelwertes.
4. Warum ist es erforderlich, Parameterschätzungen mit einem Konfidenzintervall zu versehen?
5. Welche Methoden der Punktschätzung werden unterschieden?
6. Welche Aussagen stimmen nicht?
   a) die z-Transformation überführt jede Verteilung in eine Verteilung mit Mittelwert 1 und Streuung 0.
   b) die z-Transformation überführt jede Verteilung in eine Verteilung mit Mittelwert 1 und Varianz 0
   c) die z-Transformation überführt jede Verteilung in eine Verteilung mit Mittelwert 0 und Varianz 1
   d) die z-Transformation überführt jede Verteilung in eine Standardnormalverteilung
7. Bei dem sog. Standardfehler handelt es sich um:
   a) die Streuung der Stichprobe,
   b) die Streuung der Population,
   c) die Streuung der Stichprobenkennwerteverteilung,
   d) die Differenz zwischen Schätzwert und Populationsparameter.
8. Für die Berechnung des Konfidenzintervalls des Mittelwertes liegen verschiedene Formeln vor. Warum?

9. Um im Vorfeld einer deskriptiven Studie zu bestimmen, welcher optimale Stichprobenumfang für die Intervallschätzung des interessierenden Populationsparameters notwendig ist, müssen welche Entscheidungen getroffen werden?
10. Eine Lehrerin interessiert sich für die Frage, wie viel Zeit 11-jährige Schulkinder täglich für ihre Hausaufgaben aufwenden. Für ihre Untersuchung nimmt sie in Kauf, dass die wahre durchschnittliche Hausaufgabenzeit um maximal 3 Minuten verschätzt wird. Das zu ermittelnde Konfidenzintervall soll mit einem Konfidenzkoeffizienten von 99 % abgesichert werden. Die Streuung der Hausaufgabenzeit bei den Schulkindern der betrachteten Altersgruppe möge bei $\sigma = 35$ min liegen. Mit welchem optimalen Stichprobenumfang n_{opt} sollte eine entsprechende empirische Studie angelegt sein, um – ohne Ressourcenvergeudung – ein aussagekräftiges Konfidenzintervall berechnen zu können?

Übungsaufgaben

1. Eine repräsentative Befragung unter Profi-Fechter/innen mit Weltranglistenplatzierung habe ergeben, dass 70 von 200 Befragten angaben, sie seien Linkshänder/innen. Führen Sie eine Punktschätzung für den Populationsparameter π durch und geben Sie das dazugehörige 95 %ige und das 99 %ige Konfidenzintervall an.
2. In der o. g. Befragung wurde ebenfalls die Häufigkeit erhoben, mit der sich die Fechter/innen durch die Beobachtung zukünftiger Turnierpartner/innen auf kommende Gefechte vorbereiten. Erhoben auf einer Häufigkeits-Ratingskala von 1 (nie) bis 5 (immer) wiesen die Linkshänder/innen einen Mittelwert $\bar{x} = 4.10$; $\sigma^2 = 0.64$, die Rechtshänder/innen einen Mittelwert $\bar{x} = 2.90$; $\sigma^2 = 0.49$ auf. Ermitteln Sie jeweils für Links- und Rechtshänder/innen die Punktschätzer für die Mittelwerte und die dazugehörigen 95 %igen und 99 %igen Konfidenzintervalle.
3. Ermitteln Sie für die oben genannte Befragung den optimalen Stichprobenumfang für das 95 %ige und das 99 %ige Konfidenzintervall des Mittelwerts. Berücksichtigen Sie dabei, dass der Schätzfehler e nicht mehr als 15 % der Merkmalstreuung betragen soll.

Weiterführende Literatur

- Altman, D., Machin, D., Bryant, T., & Gardner, S. (Eds.). (2000). *Statistics with confidence: confidence intervals and statistical guidelines* (book with disc for Windows 95, 98, NT) (2nd ed.). Bristol: BMJ Books.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer
- Müller-Benedict, V. (2011). Konfidenzintervalle. In V. Müller-Benedict (Hrsg.), *Grundkurs Statistik in den Sozialwissenschaften. Eine leicht verständliche, anwendungsorientierte Einführung in das sozialwissenschaftlich notwendige statistische Wissen* (5. Aufl., S. 166–186). Wiesbaden: VS.
- Projekt Neue Statistik (Producer). (o. A., 05.10.2011) Schätzen. Abgerufen 23. August, 2013, unter http://web.neuestatistik.de/inhalte_web/content/start_schaetzen.html#schaetzen.
- Smithson, M. (Ed.). (2003). *Confidence intervals (quantitative applications in the social sciences)*. Thousand Oaks, California: Sage.

Literatur

Assenmacher, W. (2000). *Induktive Statistik*. Berlin: Springer.
Bortz, J. (2005). *Statistik* (6. Aufl.). Heidelberg: Springer.
Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Heidelberg: Springer
Brandstätter, E. (1999). Konfidenzintervalle als Alternative zu Signifikanztests. *Methods of Psychological Research Online, 4*(2), 1–17.
Daniel, C. & Wood, F. S. (1971). *Fitting equations to data*. New York: Wiley.
Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton: CRC Press.
Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.
Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the royal society of London. Series A, 222*, 309–368.
Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, 22*, 700–725.
Hays, W. L. & Winkler, R. L. (1970). *Statistics*. New York: Holt, Rinehart & Winston.

Helten, E. (1974). Wahrscheinlichkeitsrechnung. In J. v. Koolwijk & M. Wieken-Mayser (Hrsg.), *Techniken der empirischen Sozialforschung* (Vol. 6, S. 9–38). München: Oldenbourg.

Huber, P. & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). Hoboken: Wiley.

Kendall, M. G. & Stuart, A. (1973). *The advanced theory of statistics* (2nd ed.). London: Griffin.

Kreyszig, E. (1973). *Statistische Methoden und ihre Anwendungen*. Göttingen: Vandenhoeck & Ruprecht.

Lampert, T., Sygusch, R., & Schlack, R. (2007). Nutzung elektronischer Medien im Jugendalter. Ergebnisse des Kinder- und Jugendgesundheitssurveys (Kiggs). *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz, 50*, 643–652.

Marronna, R., Martin, D., & Yohai, V. (2006). *Robust statistics. Theory and methods*. Hoboken: Wiley.

Neyman, J. (1937). Outline of *a* theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, 236*(767), 333–380.

Sachs, L. (2002). *Statistische Auswertungsmethoden* (10. Aufl.). Berlin: Springer.

Schwarz, H. (1960). Abschätzung der Streuung bei der Planung von Stichprobenerhebungen. *Statistische Praxis, 15*(5), 111–114.

Schwarz, H. (1966). Über die Abschätzung der Standardabweichung zahlenmäßiger Merkmale durch Annahme bestimmter Verteilungen. *Statistische Praxis, 11*.

Schwarz, H. (1975). *Stichprobenverfahren*. München: Oldenbourg.

Stegmüller, W. (1985). *Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie* (2. Aufl.). Berlin: Springer.

Student. (1908). The probable error of *a* mean. *Biometrika, 6*(1), 1–25.

Wirtz, M. & Nachtigall, C. (1998). *Deskriptive Statistik: Statistische Methoden für Psychologen Teil 1*. Weinheim: Juventa.

Wirtz, M. & Nachtigall, C. (2009). *Deskriptive Statistik. Statistische Methoden für Psychologen Teil 1* (5. Aufl.). Weinheim: Juventa.

Yamane, T. (1976). *Statistik*. Frankfurt/Main: Fischer.

12.5 Quantitative Datenanalyse bei explanativen Studien

Lernziele

- Wissen, was man unter quantitativer Datenanalyse bei explanativen Studien versteht.
- Die Logik des klassischen statistischen Nullhypothesen-Signifikanztests erklären können.
- Hypothesen für Zusammenhangs-, Unterschieds-, Veränderungs- und Einzelfalluntersuchungen formulieren können.
- Statistische Verfahren zur Überprüfung von Zusammenhangs-, Unterschieds-, Veränderungs- und Einzelfallhypothesen kennen und anwenden können.

Ein Großteil wissenschaftlicher Studien in den Sozial- und Humanwissenschaften dient der Hypothesenprüfung. In der quantitativen Forschung erfolgt die Hypothesenprüfung meist mit dem **klassischen Signifikanztest**. Teilweise werden auch Konfidenzintervalle – die ansonsten zur Schätzung von Populationsparametern in deskriptiven Studien zur Anwendung kommen (► Abschn. 12.4.2) – für Hypothesentests genutzt. Weitere Alternativen zum klassischen Signifikanztest sind Resampling-Verfahren und Verfahren der Bayes-Statistik (► Abschn. 12.2.1), Minimum-Effektgrößen-Tests (► Kap. 15) sowie Strukturgleichungsmodelle (► Kap. 17).

Wir werden in diesem Teilkapitel zunächst die Logik und Vorgehensweise des klassischen Nullhypothesen-Signifikanztests vorstellen (► Abschn. 12.5.1), bevor konkrete statistische Auswertungstechniken zur Überprüfung von Zusammenhangshypothesen (► Abschn. 12.5.2), von Unterschiedshypothesen (► Abschn. 12.5.3), von Veränderungshypothesen (► Abschn. 12.5.4) sowie von Einzelfallhypothesen (► Abschn. 12.5.5) ausführlich behandelt werden (zur Übersicht ► Inhaltliche Forschungshypothesen für Zusammenhangs-, Unterschieds-, Veränderungs- und Einzelfalluntersuchungen).

Im Kontext der quantitativen Datenanalyse ist immer zu beachten, dass auch eine noch so ausgefeilte und sorgfältige statistische Hypothesenprüfung nur dann zu brauchbaren wissenschaftlichen Resultaten führt, wenn entsprechend aussagekräftige Daten verarbeitet werden. Andernfalls gilt das **GIGO-(Garbage-in-Garbage-out-) Prinzip**: Dürftiges Datenmaterial kann bei der statistischen Analyse nicht verbessert werden, sondern führt notgedrungen auch zu zweifelhaften Resultaten. Die in diesem Teilkapitel betrachtete statistische Hypothesenprüfung im Zuge der Datenanalyse ist also immer rückzubinden an Fragen nach der Angemessenheit der Hypothesen, des Untersuchungsdesigns, der Operationalisierung der Variablen, der Stichprobenziehung, der Datenerhebung sowie der Datenaufbereitung (zur wissenschaftstheoretischen Reflexion von Hypothesentests ► Abschn. 2.2.9). Insbesondere das **Untersuchungsdesign** entscheidet darüber, welchen Informationsgehalt die Daten in Bezug auf die zu prüfenden Hypothesen haben. So kann z. B. ein Korrelationskoeffizient in Abhängigkeit vom Untersuchungsdesign mehr oder minder großen kausalen Erklärungswert haben (Wilkinson & Task Force on Statistical Inference, 1999). Vorrangig für den folgenden Text ist die Beschreibung der Untersuchungspläne im Zusammenhang mit den jeweils geeigneten Signifikanztests zur Hypothesenprüfung. Dabei widmen wir uns primär der **Indikationsfrage**: Welcher Signifikanztest ist für welche Hypothese und welches Design am besten geeignet? Für Details zur Berechnung der einzelnen statistischen Signifikanztests wird jeweils auf einschlägige Statistik-Lehrbücher verwiesen.

Man beachte, dass die Einteilung von Designs und Hypothesentests gemäß der Prüfung von **Zusammenhängen, Unterschieden und Veränderungen** bei Gruppen oder Einzelfällen (► Inhaltliche Forschungshypothesen für Zusammenhangs-, Unterschieds-, Veränderungs- und Einzelfalluntersuchungen) v. a. aus didaktischen Gründen erfolgt (andere Einteilungen sind möglich und sinnvoll; Hager, 1987, 2004). Zudem können für ein und dieselbe Forschungshypothese mehrere Untersuchungsvarianten infrage kommen: So ließe sich z. B. die Unterschiedshypothese: „Soziale Schichten unterscheiden sich im Erziehungsstil“ auch als Zusammenhangshypothese formulieren: „Zwischen den Merkmalen ‚Soziale Schicht‘ und ‚Erziehungsstil‘ besteht ein Zusammenhang“. Ob hierfür ein Untersuchungsplan zur Überprüfung einer Zusammenhangs- oder Unterschiedshypothese gewählt wird, kann unerheblich sein. (In diesem Sinne äquivalente Pläne sind im folgenden Text entsprechend gekennzeichnet.) In vielen Fällen wird sich jedoch herausstellen, dass scheinbar äquivalente Pläne sich in ihrer Praktikabilität bzw. in der Eindeutigkeit ihrer Ergebnisse unterscheiden.

Es sei darauf hingewiesen, dass sich eine Reihe statistischer Signifikanztests zur Überprüfung von Zusammenhangs-, Unterschieds- und Veränderungshypothesen auf ein allgemeines Auswertungsprinzip, das sog. **Allgemeine Lineare Modell** (ALM) zurückführen lassen (Bortz & Schuster, 2010, Kap. 22). Das ALM macht die hier vorgenommene inhaltliche Unterscheidung von Hypothesenarten also im Prinzip überflüssig. Formale Äquivalenzen zwischen statistischen Hypothesentests (d. h. alternative statistische Berechnungsgänge, die zu identischen Resultaten führen) sind indessen nicht gleichbedeutend mit forschungslogischer Gleichwertigkeit von Untersuchungsergebnissen in Abhängigkeit vom Untersuchungsdesign.

Inhaltliche Forschungshypothesen für Zusammenhangs-, Unterschieds-, Veränderungs- und Einzelfalluntersuchungen

Beispielhafte **Zusammenhangshypothesen** und zugehörige statistische Auswertungsverfahren (▶ Abschn. 12.5.2):

- Zwischen zwei Merkmalen x und y besteht ein Zusammenhang. Beispiel: Zwischen der Verbalisierungsfähigkeit von Schülern und dem Lehrerurteil über die Intelligenz der Schüler besteht ein positiver Zusammenhang **(bivariate Korrelation)**.
- Zwischen zwei Merkmalen x und y besteht auch dann ein Zusammenhang, wenn man den Einfluss eines dritten Merkmals z außer Acht lässt. Beispiel: Zwischen der Kommunikationsstruktur von Gruppen und ihrer Produktivität besteht ungeachtet der Gruppengröße ein Zusammenhang **(Partialkorrelation)**.
- Zwischen mehreren Prädiktorvariablen $(x_1, x_2, \ldots x_p)$ und einer Kriteriumsvariablen (y) besteht ein Zusammenhang. Beispiel: Zwischen dem durch mehrere Merkmale beschriebenen Erziehungsverhalten von Eltern und der Hilfsbereitschaft ihrer Kinder besteht ein Zusammenhang **(multiple Korrelations- und Regressionsanalyse)**.
- Zwischen mehreren Prädiktorvariablen $(x_1, x_2, \ldots x_p)$ und mehreren Kriteriumsvariablen $(y_1, y_2, \ldots y_q)$ besteht ein Zusammenhang. Beispiel: Zwischen dem durch mehrere Merkmale beschriebenen Erziehungsverhalten von Eltern und der durch mehrere Merkmale erfassten Prosozialität ihrer Kinder besteht ein Zusammenhang **(kanonische Korrelation)**.
- Zwischen mehreren manifesten und latenten Variablen bestehen komplexe kausale Zusammenhänge **(Strukturgleichungsmodelle)**.

Beispielhafte **Unterschiedshypothesen** und zugehörige statistische Auswertungsverfahren (▶ Abschn. 12.5.3):

- Eine zweifach gestufte unabhängige Variable (Treatment) hat einen Einfluss auf eine abhängige Variable. Beispiel: Bei einer Werbeanzeige mit sexuellen Schlüsselreizen wird das beworbene Produkt schlechter erinnert als bei einer Anzeige ohne sexuelle Schlüsselreize, geprüft in einer experimentellen Laborstudie **(*t*-Test für unabhängige Stichproben)**.
- Eine mehr als zweifach gestufte unabhängige Variable hat einen Einfluss auf eine abhängige Variable. Beispiel: Vier verschiedene Werbeanzeigen für ein neues Produkt erzeugen unterschiedlich intensive Kaufabsichten **(einfaktorielle Varianzanalyse)**.
- Zwei zwei- oder mehrfach gestufte unabhängige Variablen A und B haben jeweils einen separaten Einfluss und/oder einen spezifischen gemeinsamen Einfluss auf die abhängige Variable. Beispiel: Die Schlafqualität hängt von der Altersgruppe (Jugendliche, Erwachsene, Senioren) und vom Geschlecht ab **(zweifaktorielle Varianzanalyse)**.
- Die Wirkung der unabhängigen Variablen auf die abhängige Variable wird durch Kontrollvariablen beeinflusst. Beispiel: Die Schlafqualität hängt von der Altersgruppe und vom Geschlecht ab, und zwar auch dann, wenn der Einfluss der konsumierten Koffein- und Alkoholmenge statistisch kontrolliert wird **(Kovarianzanalyse)**.
- Es wird die Wirkung der unabhängigen Variablen auf mehrere abhängige Variablen gleichzeitig untersucht. Beispiel einer Nullhypothese als Forschungshypothese: Einzelkinder und Geschwisterkinder unterscheiden sich nicht hinsichtlich Selbstwertgefühl, Narzissmus, Hilfsbereitschaft, sozialer Verträglichkeit und Toleranz **(multivariate Varianzanalyse oder Diskriminanzanalyse)**.

Beispielhafte **Veränderungshypothesen** und zugehörige statistische Auswertungsverfahren (▶ Abschn. 12.5.4):

- Ein Treatment übt eine verändernde Wirkung auf eine abhängige Variable aus und dies wird in einem **experimentellen Design** geprüft, typischerweise einem randomisierten Prä-Post-Kontrollgruppendesign (dabei sind ggf. Pretest- und Sequenzeffekte zu kontrollieren). Beispiel: Das regelmäßige Lesen einer konservativen Tageszeitung verändert die politischen Ansichten ihrer Leserinnen und Leser in Richtung Konservatismus (z. B. **zwei- oder mehrfaktorielle Varianzanalyse mit Messwiederholungen auf einem Faktor**).
- Ein Treatment verändert eine abhängige Variable in einer Population A stärker als in einer Population B, was – wenn die Populationszugehörigkeit nicht experimentell variierbar ist – nur mit **nicht- oder quasi-experimentellen Vorher-Nachher-Designs** zu prüfen ist (dabei ist u. a. der Effekt der Regression zur Mitte zu kontrollieren). Beispiel: Ein Rückengymnastik-Programm reduziert Rückenschmerzen bei Bandscheiben-Patienten stärker als bei Nicht-Patienten (z. B. **zwei- oder mehrfaktorielle Varianzanalyse mit Messwiederholungen auf einem Faktor**).
- Die Veränderung einer abhängigen Variablen hängt von einer Drittvariablen ab. Beispiel: Genesungsfortschritte von Kranken hängen von deren Motivation ab, gesund werden zu wollen (z. B. **bivariate Korrelation der Veränderungswerte mit der Drittvariablen**).
- Bei einer einzelnen Person verändert sich ein Merkmal durch eine individualisierte Intervention. Beispiel: Die sportlichen Ergebnisse einer Leistungssportlerin steigern sich durch ein neues Trainingsprogramm stärker als durch das herkömmliche Training (zur Überprüfung von Veränderungshypothesen zu Einzelfällen sind die **für Einzelfallhypothesen indizierten statistischen Verfahren** einzusetzen).
- Ein Merkmal des Erlebens oder Verhaltens verändert sich mit dem Lebensalter, wobei gleichzeitig Generationen- und Epocheneffekte eine Rolle spielen. Beispiel: Mit zunehmendem Lebensalter sinkt das Interesse an neuen technischen Geräten, d. h. es ist bei Kindern und Jugendlichen deutlich stärker ausgeprägt als bei Erwachsenen und Hochbetagten (untersuchbar mittels Querschnitt-, Längsschnitt- und Zeitwandel-Design und auswertbar mittels **Varianzanalysen mit bzw. ohne Messwiederholungen**).
- Eine Intervention führt zu einer sprunghaften Veränderung einer Zeitreihe. Beispiel: Die Verabschiedung eines neuen Scheidungsgesetzes führt schnell zu einer Verdoppelung der jährlichen Ehescheidungen in einem Staat (**Zeitreihenanalyse**).

Beispielhafte **Einzelfallhypothesen** und zugehörige statistische Auswertungsverfahren (▶ Abschn. 12.5.5):

- Eine Intervention wirkt sich in der Form auf die Ausprägung eines kardinalskalierten Merkmals einer einzelnen Person aus, dass die Merkmalsausprägung bei Intervention anders ausfällt als ohne Intervention. Beispiel: Während eines logopädischen Trainings spricht ein Kind mehr Wörter fehlerfrei als ohne Training **(Randomisierungstest)**.

- Eine Intervention wirkt sich in der Form auf die Ausprägung eines kardinalskalierten Merkmals einer einzelnen Person aus, dass die Merkmalsausprägung bei Intervention anders ausfällt als ohne Intervention und über mehrere Behandlungsphasen hinweg einem Trend folgt. Beispiel: Während eines mehrphasigen logopädischen Trainings spricht das Kind von Phase zu Phase mehr Wörter fehlerfrei aus (monotoner Trend: **Trendtest**; linearer Trend: **Permutationstest**).
- Die zeitliche Abfolge von dichotomen Ereignissen ist nicht zufällig. Beispiel: Die Magenschmerzen von Frau M. hängen mit beruflichen Misserfolgen zusammen, d. h. sie treten bei Misserfolgen auf und bleiben aus, wenn keine Misserfolge erlebt werden (**Iterationshäufigkeitstest**).
- Ein Treatment senkt die Auftretenswahrscheinlichkeit von dichotomen Ereignissen. Beispiel: Das Bettnässen eines Kindes tritt im Verlauf einer verhaltenstherapeutischen Maßnahme von Behandlungsphase zu Behandlungsphase zunehmend seltener auf (**Rangsummentest**).
- Die bei einer Testperson in zwei Untertests einer Testbatterie gemessene Testwertedifferenz ist diagnostisch verwertbar. Beispiel: Das Intelligenzprofil eines Abiturienten weist systematische, nicht durch Zufall erklärbare Schwankungen in der Weise auf, dass seine Leistungsfähigkeit im verbalen Denken die im räumlichen Denken übertrifft (für solche **zufallskritischen Aussagen über einzelne Testwerte** müssen die Kennwerte des betreffenden Tests – z. B. seine Messgenauigkeit – in die statistische Analyse einbezogen werden).

12.5.1 Hypothesenprüfung mittels klassischem Signifikanztest

Lernziele

- Zielsetzung und Logik des statistischen Nullhypothesen-Signifikanztests beschreiben können.
- Wissen, was man unter statistischer Signifikanz versteht.
- Die Bedeutung von Effektgröße und Teststärke bei der Interpretation von Signifikanztest-Ergebnissen kennen.
- Wissen, welche Bedeutung der sogenannte optimale Stichprobenumfang im Kontext der statistischen Hypothesenprüfung hat.
- Verstehen, warum Konfidenzintervalle eine Alternative zum statistischen Signifikanztest darstellen können.

Hypothesen werden in der quantitativen Sozialforschung traditionell mit statistischen Hypothesentests – sog. **statistischen Signifikanztests** – geprüft. Die Frage, ob ein empirisches Forschungsergebnis „statistisch signifikant" geworden ist oder nicht, steht dabei im Mittelpunkt. Was genau sich hinter dem Konzept der „statistischen Signifikanz" verbirgt und welche Interpretationsfehler zu vermeiden sind, soll im Zuge der Lektüre dieses Teilkapitels deutlich werden.

Hinweis Ausführliche Behandlungen der Signifikanztestproblematik findet man z. B. bei Bakan (1966); Bredenkamp (1969, 1972, 1980); Carver (1978); Cohen (1994); Cook, Grader, Hennigan und Flay (1979); Crane (1980); Erdfelder und Bredenkamp (1994); Greenwald (1975); Morrison und Henkel (2007); Harlow, Mulaik und Steiger (1997); Harnatt (1975); Heerden und van Hoogstraten (1978); Kline (2004, Kap. 3); Krause und Metzler (1978); Lane und Dunlap (1978); Lykken (1968); Nickerson (2000); Willmes (1996); Witte (1977); Wottawa (1990).

Die Logik des statistischen Signifikanztests

Der klassische statistische Signifikanztest, der auch als **Nullhypothesen-Signifikanztest** („null hypothesis significance testing", NHST) bezeichnet wird, hat sich historisch als eine Art Mischmodell aus dem reinen Nullhypothesentest nach Ronald Fisher (1925, 1956) und dem Testmodell von Jerzy Neyman und Egon Pearson (1928, 1933) entwickelt, welches neben der Nullhypothese auch eine Alternativhypothese betrachtet.

Die Logik des statistischen Signifikanztests wird hier am **Beispiel einer bivariaten Korrelationshypothese** erläutert. Die Forschungshypothese lautet: Je größer der Stress am Arbeitsplatz ist, desto mehr Fehlzeiten der Beschäftigten treten auf. Anders formuliert: Es besteht ein positiver Zusammenhang zwischen Stress am Arbeitsplatz und Fehlzeiten der Beschäftigten in der Population. Die folgende Darstellung ist nicht mathematisch formalisiert, sondern stellt die Logik des Tests möglichst anschaulich dar (für mathematische Grundlagen s. Bortz & Schuster, 2010, Kap. 10; Eid, Gollwitzer, & Schmitt, 2010, Kap. 8). Dabei werden sieben Aspekte behandelt: 1.) Zielsetzung des Signifikanztests, 2.) Auswahl des passenden Signifikanztests, 3.) inhaltliche Alternativhypothese und Nullhypothese, 4.) das statistische Hypothesenpaar, 5.) das H_0-Modell und die Prüfverteilung, 6.) die Signifikanzentscheidung und 7.) die Ergebnisdarstellung des Signifikanztests.

1. Zielsetzung des Signifikanztests

Zielsetzung des statistischen Signifikanztests ist die **Überprüfung von Hypothesen** anhand quantitativer em-

pirischer Daten. Er stellt eine Regel dar, um anhand von empirischen Daten eine Entscheidung über die Forschungshypothese zu treffen.

Es ist somit nur dann sinnvoll, einen Signifikanztest durchzuführen, wenn man im Vorfeld wohlbegründete Hypothesen (▶ Kap. 5) auf der Basis des aktuellen Forschungsstandes (▶ Kap. 6) formuliert und dann mit einem entsprechenden Untersuchungsdesign (▶ Kap. 7), angemessener Operationalisierung (▶ Kap. 8) und Stichprobe (▶ Kap. 9) sowie passender Datenerhebungsmethode (▶ Kap. 10) quantitative Daten erhoben und einer sorgfältigen Datenaufbereitung (▶ Kap. 11) unterzogen hat.

Bei unserem Beispiel gehen wir davon aus, dass diese Voraussetzungen erfüllt sind: Die Hypothese zum positiven Zusammenhang zwischen Stress und Fehlzeiten wurde aus Theorien zum Absentismus (Abwesenheit vom Arbeitsplatz, die nicht auf Krankheit zurückgeht) abgeleitet. Bei 100 zufällig ausgewählten Erwerbstätigen wurde unter Beachtung der Forschungsethik (▶ Abschn. 4.1) die Stressbelastung am Arbeitsplatz mit einer Stress-Skala erhoben, zudem wurde die Anzahl der im vergangenen Jahr registrierten Fehltage erfasst. Es liegt ein entsprechender anonymisierter und fehlerbereinigter quantitativer Datensatz elektronisch vor.

2. Auswahl des passenden Signifikanztests

Für die zu prüfende Hypothese ist jeweils der passende Signifikanztest aus der großen Vielfalt existierender statistischer Signifikanztests auszuwählen **(Indikationsfrage)**.

Die Auswahl des passenden Signifikanztests für eine inhaltlich sinnvolle Hypothesenprüfung hängt v. a. davon ab,

1. welche Art von Hypothese zu prüfen ist (Zusammenhangs-, Unterschieds-, Veränderungs- oder Einzelfallhypothese),
2. wie viele Variablen zu berücksichtigen sind (z. B. uni-, bi- oder multivariate Analyse),
3. welches Skalenniveau die Variablen haben (nominal-, ordinal-, intervall- oder verhältnisskalierte Daten),
4. welcher Stichprobenumfang vorliegt (z. B. Stichprobenumfang $n \geq 30$ erlaubt Anwendung des Zentralen Grenzwerttheorems) und
5. welche Verteilungseigenschaften die Daten haben (z. B. Normalverteilung bei kleinen Stichprobenumfängen; Varianzhomogenität zwischen Gruppen).

In unserem Beispiel postuliert die Forschungshypothese einen Zusammenhang zwischen zwei Variablen, die mindestens intervallskaliert (bzw. kardinalskaliert/metrisch) erhoben wurden (die Stressbelastung über eine Skala und die Fehlzeiten des letzten Jahres in Tagen über vorliegende Dokumente). Zusammenhangshypothesen für zwei kardinalskalierte Variablen werden über die **bivariate Korrelationsanalyse** (auch: Produkt-Moment-Korrelation, Pearson-Korrelation) getestet (Bortz & Schuster, 2010, S. 153ff.). Dabei setzt der Korrelationstest eine bivariate Normalverteilung der Merkmale voraus. Diese Voraussetzung ist schwer zu prüfen, allerdings reagiert der Test auf Voraussetzungsverletzungen robust (Bortz & Schuster, 2010, S. 162).

3. Alternativhypothese und Nullhypothese

Der klassische Signifikanztest wird auch als **Nullhypothesen-Signifikanztest** bezeichnet. Er bezieht sich auf ein Hypothesenpaar bestehend aus Alternativ- und Nullhypothese.

Die zu prüfende Forschungshypothese ist in der Regel eine **Alternativhypothese (H_1)**, welche die Existenz eines bestimmten Effekts (d. h. eines Zusammenhanges, eines Unterschieds oder einer Veränderung) in der Population postuliert. Die Alternativhypothese kann neben der Existenz des Effektes auch noch seine Richtung (dann handelt es sich um eine **gerichtete Hypothese**) und ggf. sogar noch seine Größe vorhersagen (dann handelt es sich um eine **spezifische Hypothese**).

Wenn die Alternativhypothese feststeht, ist damit indirekt auch die Nullhypothese fixiert. Denn die **Nullhypothese (H_0)** widerspricht der H_1 und behauptet genau das Gegenteil, nämlich dass überhaupt kein Effekt oder sogar ein gegengerichteter Effekt vorliegt. H_1 und H_0 bilden zusammen ein Hypothesenpaar, das alle möglichen empirischen Situationen vollständig abbildet.

Alternativhypothese und Nullhypothese – Im Kontext des statistischen Nullhypothesen-Signifikanztests wird zu jeder **Alternativhypothese** (H_1, „alternative hypothesis“), die theoriebasiert die Existenz (oft auch die Richtung und zuweilen sogar die Größe) eines Populationseffekts postuliert, eine **Nullhypothese** (H_0, „null hypothesis“) aufgestellt, die den betreffenden Effekt negiert. H_1 und H_0 bilden zusammen ein Hypothesenpaar, das alle möglichen Ausprägungen des betrachteten Effekts beinhaltet.

Für unser Beispiel: Die Alternativhypothese postuliert in der Population der Erwerbstätigen einen positiven (gleichgerichteten) Zusammenhang zwischen Stressbelastung am Arbeitsplatz und Fehlzeiten, d. h. je höher die Stressbelastung, umso höher auch die Fehlzeiten. Die Nullhypothese widerspricht und behauptet, dass es in der Population keinen Zusammenhang zwischen Stress und Fehlzeiten gibt (d. h. erhöhter Stress geht unsystematisch mal mit höheren, mal mit niedrigeren, mal mit durchschnittlichen Fehlzeiten einher) oder dass es sogar einen negativen (gegengerichteten) Zusammenhang gibt (d. h. erhöhter Stress geht mit verminderten Fehlzeiten einher). Es ergibt sich somit folgendes Hypothesenpaar:

Tabelle 12.11 Überblick über Hypothesenarten

Hypothesenarten	Kennzeichen	Beispiel
Alternativhypothese (H_1) vs. Nullhypothese (H_0)	Die Alternativhypothese (H_1) postuliert die Existenz (oft auch die Richtung und zuweilen sogar die Größe) eines bestimmten Effektes in der Population. Meist ist die Forschungshypothese eine Alternativhypothese Die Nullhypothese (H_0) negiert den von der Alternativhypothese H_1 postulierten Effekt und behauptet, es gäbe in der Population keinen oder den gegenteiligen Effekt.	H_1: Es besteht bei Erwerbstätigen ein positiver Zusammenhang zwischen Stress am Arbeitsplatz und Fehlzeiten. H_0: Es besteht bei Erwerbstätigen kein Zusammenhang oder sogar ein negativer Zusammenhang zwischen Stress am Arbeitsplatz und Fehlzeiten.
Inhaltlich-verbale Hypothese vs. statistische Hypothese	Hypothesen werden zunächst inhaltlich-verbal formuliert. Die inhaltlich-verbalen Hypothesen sind in statistische Hypothesen zu überführen, welche die interessierenden Populationsparameter (symbolisiert durch die entsprechenden griechischen Buchstaben) enthalten.	H_1: $\rho > 0$ (der Populations-Korrelationskoeffizient rho für Stress und Fehlzeiten hat ein positives Vorzeichen) H_0: $\rho \leq 0$
Gerichtete vs. ungerichtete Hypothese	Alternativhypothesen sollten normalerweise gerichtet sein, d. h. die Richtung des erwarteten Effekts angeben. Ungerichtete Alternativhypothesen haben eine sehr geringe Aussagekraft.	H_1: $\rho > 0$ (gerichtet: positiver Zusammenhang) H_1: $\rho \neq 0$ (ungerichtet: positiver oder negativer Zusammenhang)
Spezifische vs. unspezifische Hypothese	Zusätzlich zur Angabe der Effektrichtung (gerichtete Hypothese) sollte bei einer Alternativhypothese idealerweise auch noch die erwartete Effektgröße angegeben werden (spezifische gerichtete Hypothese). Spezifische Hypothesen sind in der Forschungspraxis bislang selten.	H_1: $\rho > 0$ (gerichtet und unspezifisch: positiver Zusammenhang beliebiger Größe) H_1: $\rho \geq .20$ (gerichtet und spezifisch: positiver Zusammenhang der Mindestgröße .20)

- H_1: In der Population der Erwerbstätigen besteht ein positiver Zusammenhang zwischen Stress am Arbeitsplatz und Fehlzeiten.
- H_0: In der Population der Erwerbstätigen besteht kein Zusammenhang oder ein negativer Zusammenhang zwischen Stress am Arbeitsplatz und Fehlzeiten.

4. Das statistische Hypothesenpaar

Das inhaltliche (in Worten ausformulierte) Hypothesenpaar ist für die statistische Signifikanzprüfung in ein statistisches (in statistischen Symbolen dargestelltes) Hypothesenpaar zu überführen (für Hypothesenarten Tab. 12.11). Da sich Hypothesen definitionsgemäß auf Populationen beziehen, müssen statistische Hypothesen immer die **Symbole für Populationsparameter** enthalten: Dabei handelt es sich konventionell um griechische Buchstaben (z. B. μ – my – für Mittelwert, π – pi – für Prozentwert oder ρ – rho – für Korrelationskoeffizient).

Statistische Hypothesen **(„statistical hypotheses") werden nicht in Worten, sondern in statistischen Symbolen formuliert. Da sich Hypothesen immer auf Populationseffekte beziehen, werden in statistischen Hypothesen griechische Buchstaben zur Symbolisierung der interessierenden Populationsparameter verwendet. Für jeden Signifikanztest ist entsprechend dem inhaltlichen Hypothesenpaar ein statistisches Hypothesenpaar bestehend aus H_1 und H_0 aufzustellen.**

Für unser Beispiel: Das statistische Maß für die Enge und Richtung des Zusammenhangs zwischen zwei kardinalskalierten Variablen ist der **bivariate Korrelationskoeffizient** mit einem Wertebereich von -1 (perfekter negativer Zusammenhang) über 0 (kein Zusammenhang) bis zu $+1$ (perfekter positiver Zusammenhang). Die bivariate Korrelation in einer Stichprobe lässt sich über den **Stichprobenkennwert *r*** berechnen, während die Korrelation in der Population – also der entsprechende **Populationsparameter ρ** (rho) – unbekannt ist.

Wir können also nun das oben formulierte inhaltliche Hypothesenpaar in ein **statistisches Hypothesenpaar** überführen, das lautet:

$$H_1: \rho > 0$$
$$H_0: \rho \leq 0$$

Die H_1 postuliert, dass der Populationskorrelationskoeffizient ρ für Stress und Fehlzeiten ein positives Vorzeichen hat, denn es wird ja ein positiver Zusammenhang erwartet. Die H_0 behauptet das Gegenteil, also entweder keinen Zusammenhang ($\rho = 0$) oder einen negativen statistischen Zusammenhang in der Population ($\rho < 0$).

Kommen wir wieder auf unser Beispiel zurück: Angenommen, anhand unserer Zufallsstichprobe von $n = 100$ Beschäftigten lässt sich ein bivariater Korrelationskoeffizient von $\boldsymbol{r = +.28}$ für den Zusammenhang zwischen Stress und Fehlzeiten berechnen. Es zeigt sich also wie von der Forschungshypothese prognostiziert tatsächlich

eine positive Korrelation. Ist damit unsere H_1 bereits bestätigt? Leider nicht, denn wir haben nur ein einzelnes Stichprobenergebnis vorzuweisen (d. h. ein **deskriptivstatistisches Ergebnis**), die Hypothese bezieht sich indessen auf die Population der Erwerbstätigen, aus der die Stichprobe stammt. Die Frage ist also, wie kann auf der Basis von diesem einzelnen Stichprobenergebnis eine Aussage über den unbekannten Populationseffekt – nämlich den Populationsparameter ρ und sein Vorzeichen – getroffen werden?

5. Das H_0-Modell und die Prüfverteilung

Der statistische Signifikanztest ist ein **inferenzstatistisches Verfahren**, das heißt, auf der Basis von Stichprobendaten werden Aussagen über Populationshypothesen getroffen. Der klassische Signifikanztest konstruiert für diesen Rückschluss ein sog. Nullhypothesen-Modell (kurz: ***H*₀-Modell**), d. h. ein wahrscheinlichkeitstheoretisches Modell, das von der Annahme ausgeht, in der Population gelte die H_0 und vor diesem Hintergrund die Wahrscheinlichkeit des gefundenen Stichprobenergebnisses bestimmt. Ist das Stichprobenergebnis unter Annahme der Gültigkeit der H_0 sehr unwahrscheinlich bzw. passt es nicht gut zum H_0-Modell, so liegt ein statistisch signifikantes Ergebnis vor: Man lehnt dann die H_0 ab und entscheidet sich für die H_1. Diese Argumentationslinie einer indirekten „Bestätigung" der H_1 durch die Zurückweisung des H_0-Modells werden wir nun Schritt für Schritt durchgehen (zur Kritik dieser Logik s. z. B. Cohen, 1994).

Für unser Beispiel: Sicher spricht es eher für den von der H_1 postulierten Effekt, dass die Stichprobenkorrelation positiv ausgefallen ist und eine mittlere Größe hat und nicht nur z. B. $r = +.09$ beträgt – aber bietet diese Stichprobenkorrelation von $r = +.28$ eine genügend starke Datenbasis, um von einer ebenfalls positiven Populationskorrelation (also der Gültigkeit der H_1) auszugehen? Würde man denn mit weiteren Stichproben aus derselben Population jeweils ein ähnliches, ebenfalls hypothesenkonformes Ergebnis erzielen?

Der klassische Signifikanztest arbeitet zur Beantwortung dieser Frage mit dem angesprochenen H_0-Modell. Zur Konstruktion des H_0-Modells geht man zunächst gedanklich von der Annahme aus, dass in der Population die H_0 gilt und somit kein Effekt – in unserem Beispiel: kein Zusammenhang – vorliegt. Man stellt sich weiter vor, was passiert, wenn man nun aus einer Population, in welcher die H_0 gilt (also in unserem Beispiel: $\boldsymbol{\rho = 0}$ gilt) per Zufallsverfahren Stichproben zieht und sich die Stichprobenergebnisse (d. h. die Stichprobenkorrelationskoeffizienten r) anschaut. Es ist davon auszugehen, dass man in diesem Gedankenmodell viele Stichproben zieht, deren Korrelationskoeffizienten **bei $\boldsymbol{r = 0}$** liegen. Denn die Zufallsstichproben stellen ja Miniaturabbilder der Population dar. Es kann aber vereinzelt auch passieren, dass in diesem Gedankenmodell zufällig Stichprobenzusammensetzungen auftreten, die von den Populationsverhältnissen abweichen und als Zufallseffekt bzw. Stichprobenfehler einen positiven oder negativen Stichprobenkorrelationskoeffizienten hervorbringen (z. B. $r = +.56$ oder $r = -.27$), obwohl in der Population $\rho = 0$ gilt.

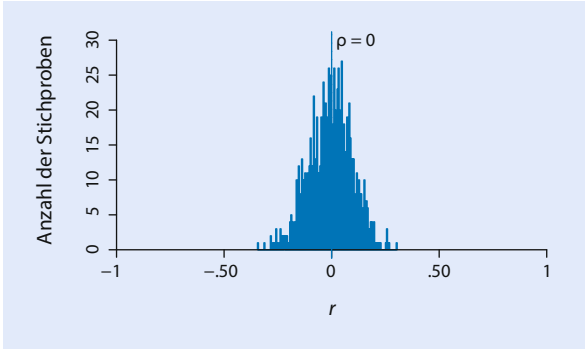


Abb. 12.19 Häufigkeitsverteilung des Stichprobenkennwertes *r*, wenn aus einer Population mit $\rho = 0$ insgesamt 1 000 Stichproben jeweils des Umfangs $n = 100$ gezogen werden (fiktive Daten)

Allgemein ist festzuhalten, dass man in dem Gedankenmodell, bei dem man aus einer Population, in der $\rho = 0$ gilt, mit großer Wahrscheinlichkeit vorwiegend Stichproben zieht, für die $r \approx 0$ gilt. Durch den **Stichprobenfehler** (nicht jede Stichprobe bildet bei einer Zufallsauswahl die Populationsverhältnisse perfekt ab), werden aber in diesem Gedankenmodell auch einige Stichprobenkorrelationen abweichend von 0 auftreten. Starke oder sehr starke Abweichungen von $r = 0$ werden dabei jedoch selten bzw. sehr selten bzw. mit sehr geringer Wahrscheinlichkeit auftreten.

Betrachten wir nun das Ergebnis eines solchen Gedankenexperiments, bei dem wir z. B. 1 000 Stichproben aus der H_0-Population ziehen und für jede Stichprobe den Korrelationskoeffizienten r bestimmen: Wir erhalten also 1 000 Stichprobenkennwerte r mit einer charakteristischen Häufigkeitsverteilung in der Form, dass die allermeisten Stichprobenkorrelationskoeffizienten um 0 liegen und dass Abweichungen nach oben und unten mit wachsender Abweichung von 0 immer seltener auftreten (Abb. 12.19).

Worin liegt nun der Nutzen dieses theoretischen Modells, das eine Häufigkeitsverteilung des Stichprobenkennwertes r darstellt? Es dient als Hilfestellung, um unser empirisches Stichprobenergebnis einordnen zu können. Und zwar fragen wir uns, wie wahrscheinlich es eigentlich ist, dass genau unser empirisches Ergebnis oder

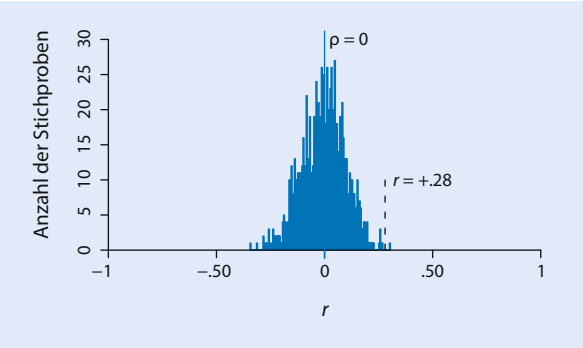


Abb. 12.20 Einordnung unseres empirischen Ergebnisses von $r = +.28$ (basierend auf $n = 100$) in die Stichprobenkennwerteverteilung von r (basierend auf 1 000 Stichproben des Umfanges $n = 100$) unter Annahme der Gültigkeit von $H_0 : \rho = 0$ in der Population (fiktive Daten)

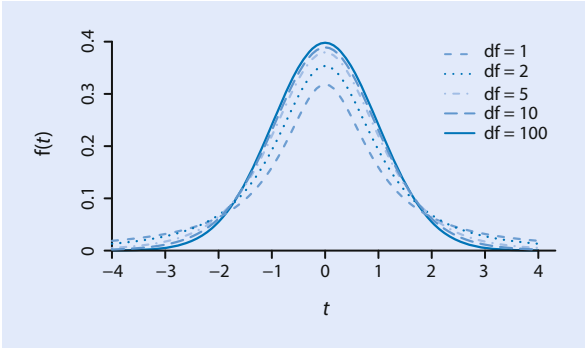


Abb. 12.21 Familie der t-Verteilungen mit unterschiedlichen Freiheitsgraden. df „degrees of freedom"

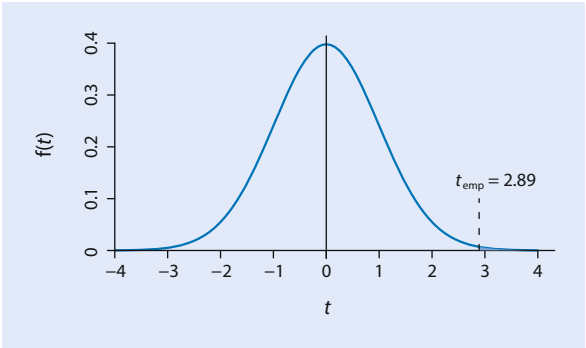


Abb. 12.22 Darstellung des H_0-Modells als t-Verteilung (df = 98) und Position des empirischen Stichprobenergebnisses von $r = +.28$, d. h. äquivalent $t_{emp} = +2.89$ (fiktive Daten)

ein noch extremeres Ergebnis zustande kommt, wenn in der Population die H_0 gilt. Anders gefragt: Passt unser Stichprobenergebnis gut zum H_0-Modell (und spricht somit tendenziell eher für die Gültigkeit der H_0) oder passt es eher nicht so gut zum H_0-Modell, liegt eher im Randbereich der Häufigkeitsverteilung (und spricht somit tendenziell gegen die Gültigkeit der H_0, d. h. indirekt für die H_1; Abb. 12.20)?

Das H_0-Modell als Häufigkeitsverteilung basierend auf 1 000 Stichproben darzustellen, ist nur eine Variante. Ebenso könnte man eine Häufigkeitsverteilung auf der Basis von 5 000 Stichproben konstruieren. In der Forschungsstatistik arbeiten wir mit einem H_0-Modell, das auf theoretisch **unendlich vielen Stichprobenziehungen** aus der Population basiert. Zudem wird anstelle der Verteilung der r-Werte (d. h. der Stichprobenkorrelationen) die Verteilung der t-Werte verwendet. Der ***t*-Wert** ist nichts anderes als eine einfache Umrechnung des r-Wertes gemäß der folgenden Formel:

$$t_{emp} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (12.39)$$

Wenn $r = 0$ ist, dann ist auch $t = 0$, je größer r dem Betrag nach wird, umso größer wird auch t. Während der Wertebereich von r begrenzt ist $(-1 \leq r \leq +1)$ ist der Wertebereich von t unbegrenzt $(-\infty \leq t \leq +\infty)$. Der Vorteil der Umformung der r-Verteilung in eine ***t*-Verteilung** besteht darin, dass die t-Verteilung genau bekannt und austabelliert ist (Tab. 3 im ▶ Webangebot zu diesem Buch). Bei der Korrelationsanalyse basiert der statistische Signifikanztest auf einem H_0-Modell bestehend aus einer t-Verteilung. Die t-Verteilung wird als wahrscheinlichkeitstheoretische **Prüfverteilung** des Korrelationstests verwendet. Dabei ist zu beachten, dass es nicht nur eine einzige t-Verteilung gibt, sondern eine

Familie von *t*-Verteilungen. Jede t-Verteilung ist durch ihre Freiheitsgrade (df = „degrees of freedom") gekennzeichnet, die sich aus dem Stichprobenumfang ableiten. Mit wachsendem Stichprobenumfang bzw. zunehmenden Freiheitsgraden geht die t-Verteilung in eine Normalverteilung über (Abb. 12.21).

Für den Korrelationstest hat die t-Verteilung $n-2$ Freiheitsgrade (df $= n - 2$), d. h. für unser Beispiel mit einer Stichprobe von $n = 100$ wählen wir die t-Verteilung mit df = 98. Unser empirisches Stichprobenergebnis von $r = +.28$ wird gemäß der Formel in einen empirischen t-Wert von $t_{emp} = +2.89$ umgerechnet und kann somit im Kontext des H_0-Modells der t-Verteilung bewertet werden (Abb. 12.22).

Sowohl r-Werte als auch t-Werte sind **stetige Merkmale**, d. h. es gibt beliebig viele feine Abstufungen. Die Wahrscheinlichkeitsfunktion einer stetigen Zufallsvariable nennen wir **Dichtefunktion** (▶ Abschn. 12.4.1). Bei einem stetigen Merkmal geht die Wahrscheinlichkeit für das Auftreten eines einzelnen Wertes z. B. $t = +2.887291667$ gegen Null. Dem einzelnen t-Wert auf der x-Achse der t-Verteilung wird also auf der y-Achse keine Wahrscheinlichkeit, sondern eine Wahrscheinlich-

keitsdichte zugeordnet. Die Dichtefunktion gibt Auskunft über die Wahrscheinlichkeit, dass Werte innerhalb eines bestimmten Intervalls auftreten. Dass irgendein t-Wert zwischen $-\infty$ und $+\infty$ auftritt, ist sicher. Diese Wahrscheinlichkeit ist somit 100 % bzw. 1 und entspricht geometrisch der gesamten Fläche unter der Verteilung (bzw. dem Integral der Dichtefunktion). Interessiert uns nun, wie wahrscheinlich es im H_0-Modell ist, dass unser **empirischer t-Wert oder ein extremerer (d. h. dem Betrag nach größerer) t-Wert auftritt**, so entspricht diese Wahrscheinlichkeit genau der Fläche, die vom empirischen t-Wert von der t-Verteilung rechts abgeschnitten wird (schraffierte Fläche in Abb. 12.22). Es handelt sich dabei um die sog. **Irrtumswahrscheinlichkeit $p(D \mid H_0)$**.

Irrtumswahrscheinlichkeit – Wir bezeichnen die bedingte Wahrscheinlichkeit, dass unter Annahme der Gültigkeit der H_0 das empirische oder ein extremeres Stichprobenergebnis (D: empirische Daten) auftritt, als **Irrtumswahrscheinlichkeit $p(D \mid H_0)$**. Grafisch entspricht diese Wahrscheinlichkeit $p(D \mid H_0)$ der Fläche, die der **empirische Wert** bzw. die empirische Prüfgröße, die aus den Stichprobendaten berechnet wird (z. B. empirischer t-Wert), vom H_0-Modell (z. B. einer t-Verteilung mit entsprechenden Freiheitsgraden) abschneidet.

12

6. Die Signifikanzentscheidung

Die Signifikanzentscheidung wird anhand des H_0-Modells getroffen. Das H_0-Modell ist eine Stichprobenkennwerteverteilung, die sich ergibt, wenn wir aus einer Population, in der die H_0 gilt, theoretisch unendlich viele Stichproben des Umfanges n ziehen und jeweils deren Stichprobenkennwerte bestimmen. Bei einem bivariaten Korrelationstest wird der Korrelationskoeffizient r betrachtet. Er wird für die Bestimmung des H_0-Modells in einen *empirischen* t-Wert transformiert. Als H_0-Modell bzw. Prüfverteilung fungiert die t-Verteilung mit $df = n - 2$.

Das H_0-Modell dient der Einordnung der empirischen Daten bzw. des empirischen Stichprobenergebnisses. Abb. 12.20 deutet ebenso wie Abb. 12.22 per Augenschein darauf hin, dass unser Stichprobenergebnis von $r = +.28$ bzw. von $t_{emp} = +2.89$ nicht besonders gut zum H_0-Modell passt, denn es liegt im Randbereich der Verteilung, schneidet rechts nur eine kleine Fläche ab, ist also unwahrscheinlich. Wenn nun ein Stichprobenergebnis im H_0-Modell sehr unwahrscheinlich ist, lässt dies zwei Interpretationen zu:

- Erste Interpretation: Die H_0: $\rho = 0$ gilt in der Population und die Stichprobe ist zufällig sehr ungewöhnlich zusammengesetzt, deswegen weicht das Stichprobenergebnis stark von $r = 0$ bzw. $t = 0$ ab.
- Zweite Interpretation: Die H_0 gilt nicht, sondern die H_1 gilt in der Population, d. h. es gibt eine positive Populationskorrelation $\rho > 0$, die sich dementsprechend in dem Stichprobenergebnis $r > 0$ bzw. $t > 0$ widerspiegelt.

Beide Interpretationen sind zulässig. Wenn allerdings ein Stichprobenergebnis extrem schlecht zum H_0-Modell passt, dann entscheidet man sich für die zweite Interpretation. Statt am H_0-Modell festzuhalten und zu argumentieren, das Stichprobenergebnis zeige nur deshalb eine substanzielle positive Korrelation bzw. einen positiven t-Wert, weil man zufällig eine vollkommen exotische Stichprobe aus einer Population mit Null-Korrelation gezogen habe, wird man eher annehmen, dass das Stichprobenergebnis Spiegel einer positiven Populationskorrelation ist und somit das H_0-Modell nicht gilt (zur Kritik dieser Argumentation s. Cohen, 1994). Doch wie unwahrscheinlich muss ein Stichprobenergebnis sein, damit man die H_0 ablehnt und die H_1 annimmt? Wo ist hier die Grenze zu ziehen? Diese Grenze nennen wir **Signifikanzniveau** (auch: Signifikanzschwelle, Alpha-Fehler-Niveau).

Konventionell hat sich als Entscheidungsregel in der Forschungsstatistik als **Signifikanzniveau** eine **5 %-Grenze** eingebürgert. Wenn die bedingte Wahrscheinlichkeit, dass unter Annahme der Gültigkeit der H_0 das Stichprobenergebnis oder ein extremeres Ergebnis auftritt, bei maximal 5 % liegt, dann wird die zweite Interpretation präferiert und die H_0 abgelehnt. Ein solches **statistisch signifikantes Stichprobenergebnis**, das nur mit einer Wahrscheinlichkeit von maximal 5 % zum H_0-Modell passt bzw. das ganz am Rand des H_0-Modells liegt, führt zur Ablehnung der H_0 und zur **Annahme der H_1**. Diese Entscheidung ist keine sichere Entscheidung, sondern mit einem Fehler behaftet. Denn rein theoretisch könnte es immer noch sein, dass ein extremes Stichprobenergebnis auf eine exotische Stichprobenzusammensetzung bzw. einen Stichprobenfehler zurückzuführen ist und in Wirklichkeit doch die H_0 gilt. Aber dieser Fall ist eben mit maximal 5 % sehr unwahrscheinlich.

Der Wert, der am Rand einer Prüfverteilung die Signifikanzschwelle markiert, wird als **kritischer Wert** bezeichnet. Bei einer t-Verteilung mit $df = 98$ liegt der kritische Wert bei $t_{krit(\alpha=5\%)} = +1.66$ (Tab. 3 im ▶ Webangebot zu diesem Buch). Die Fläche, die der kritische t-Wert von der t-Verteilung bzw. vom H_0-Modell abschneidet entspricht dem Signifikanzniveau von .05 bzw. 5 %. Man nennt diesen Bereich der Prüfverteilung auch den **Ablehnungsbereich der H_0**, da ein Stichprobenergebnis, das in diesen Bereich fällt, nicht gut zur H_0 passt und zum Anlass genommen wird, die H_0 abzulehnen und somit die H_1 anzunehmen (Abb. 12.23).

Wenn wir nun den kritischen t-Wert, der das 5 %-Signifikanzniveau im H_0-Modell markiert, in das H_0-Modell eintragen, können wir den empirischen t-Wert, der unser empirisches Stichprobenergebnis darstellt, einord-

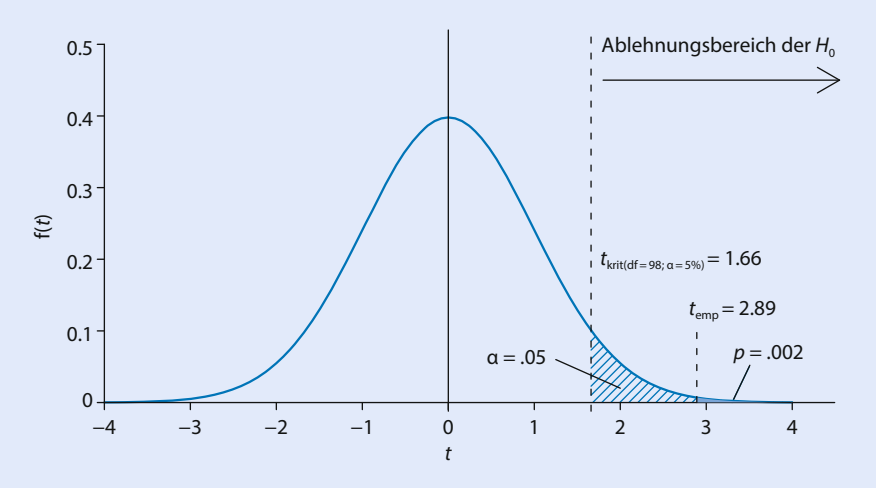


Abb. 12.23 Darstellung des H_0-Modells als t-Verteilung (df = 98) und Position des empirischen Stichprobenergebnisses von $t_{emp} = +2.89$ sowie des kritischen t-Wertes $t_{krit(\alpha=5\%)} = +1.66$, der beim einseitigen Test die Signifikanzschwelle von 5 % repräsentiert (fiktive Daten)

nen. Wir sehen auf den ersten Blick, dass der empirische Wert offensichtlich mit geringerer Wahrscheinlichkeit als 5 % auftritt, denn er ist dem Betrag nach größer als der kritische Wert bzw. liegt rechts vom kritischen Wert und schneidet somit eine noch kleinere Fläche ab. Man kann auch sagen: Der empirische Wert liegt im Ablehnungsbereich der H_0. Wir sprechen von einem statistisch signifikanten Ergebnis, wenn die Irrtumswahrscheinlichkeit kleiner ist als das Signifikanzniveau bzw. wenn der empirische Wert der Prüfverteilung dem Betrag nach über dem kritischen Wert liegt. Bei einem statistisch signifikanten Ergebnis entscheiden wir uns gegen die H_0 und für die H_1.

Signifikanzentscheidung Es handelt sich um ein **statistisch signifikantes Ergebnis**, wenn das Stichprobenergebnis nicht gut zum H_0-Modell passt und somit das H_0-Modell zurückgewiesen und eine Entscheidung zugunsten der H_1 getroffen wird.

Beim Rechnen mit Statistik-Software wird für die Signifikanzentscheidung die Irrtumswahrscheinlichkeit betrachtet und mit dem vorher festgelegten Signifikanzniveau alpha (α) verglichen:

$$\boldsymbol{p(D \mid H_0) \leq \alpha}$$
$$\rightarrow \textbf{ statistisch signifikantes Ergebnis}$$

- $p(D \mid H_0)$: **Irrtumswahrscheinlichkeit** (auch: Alpha-Fehler-Wahrscheinlichkeit); bedingte Wahrscheinlichkeit des Auftretens der Daten D, d. h. des empirischen Stichprobenergebnisses oder eines extremeren Ergebnisses unter Annahme der Gültigkeit der H_0. Die Irrtumswahrscheinlichkeit wird von Statistik-Programmen ausgegeben, sie kann per Hand nicht direkt berechnet werden.
- α: **Signifikanzniveau** (auch: Signifikanzschwelle, Alpha-Fehler-Niveau); Entscheidungskriterium zur Beurteilung der Irrtumswahrscheinlichkeit, das vor jeder Hypothesenprüfung festzulegen ist. Konventionell wird das Signifikanzniveau auf 5 % fixiert, manchmal werden auch 1 % oder 0.1 % verwendet.

Beim Rechnen per Hand (z. B. in Klausuren) wird für die Signifikanzentscheidung der empirische Wert mit dem kritischen Wert verglichen:

$$\textbf{| empirischer Wert | } \geq \textbf{ | kritischer Wert |}$$
$$\rightarrow \textbf{ statistisch signifikantes Ergebnis}$$

- **empirischer Wert**: derjenige Wert aus der Prüfverteilung (z. B. aus der t-, χ^2-, F-Verteilung), der die empirischen Daten repräsentiert; er wird gemäß einer für den jeweiligen Signifikanztest bekannten Formel aus den Stichprobendaten berechnet (per Hand oder per Statistik-Software). Der empirische Wert ist derjenige Wert, der von der Prüfverteilung diejenige Fläche abschneidet, die der Irrtumswahrscheinlichkeit $p(D \mid H_0)$ entspricht.
- **kritischer Wert**: derjenige Wert aus der Prüfverteilung (z. B. aus der t-, χ^2-, F-Verteilung), der das Signifikanzniveau α von 5 %, 1 % oder 0.1 % repräsentiert; er wird aus der austabellierten Prüfverteilung abgelesen. Da beim Rechnen per Hand die direkte Irrtumswahrscheinlichkeit nicht bestimmt werden kann, wird anhand eines Vergleichs des empirischen Wertes mit dem kritischen Wert erschlossen, ob die Irrtumswahrscheinlichkeit kleiner oder gleich dem Signifikanzniveau ist. Dies ist genau dann der Fall, wenn der empirische Wert dem Betrag nach größer oder gleich dem kritischen Wert ist.

Fehlinterpretation des Signifikanzniveaus Das Signifikanzniveau muss immer vor der Hypothesenprüfung (**a priori**) festgelegt werden. Es wird konventionell auf $\alpha = 5\,\%$ fixiert. Zuweilen wird auf einem besonders strengen Niveau von $\alpha = 1\,\%$ oder $\alpha = 0.1\,\%$ getestet, etwa wenn eine fälschlich positive Entscheidung für die H_1 sehr problematisch wäre (zur Begründung der Werte $\alpha = 5\,\%$ bzw. $\alpha = 1\,\%$ als Signifikanzniveau s. Cowles & Davis, 1982). Ebenso wird zuweilen auf einem liberaleren Niveau von $\alpha = 10\,\%$ getestet, etwa bei der Untersuchung eines neuen Phänomens.

Die Wahl des Signifikanzniveaus ist im Vorfeld der Berechnung des Signifikanztests inhaltlich zu begründen. Im Ergebnis kann man dann konstatieren, ob die Irrtumswahrscheinlichkeit die vorher fixierte Signifikanzschwelle überschreitet (statistisch signifikantes Ergebnis) oder nicht (statistisch nicht-signifikantes Ergebnis).

In der Logik des klassischen Signifikanztests ist es dementsprechend **nicht sinnvoll**, nach der Hypothesenprüfung **(a posteriori)** bei sehr kleiner Irrtumswahrscheinlichkeit von einem „hoch signifikanten" ($p < .01$) oder einem „höchst signifikanten" ($p < .001$) Ergebnis zu sprechen.

Bei einem statistisch signifikanten Ergebnis entscheiden wir uns für die Annahme der H_1, was jedoch mit einer Irrtumswahrscheinlichkeit von maximal 5 % verbunden ist. Bei einem **nicht-signifikanten Ergebnis** dürfen wir die H_1 nicht annehmen, allerdings auch nicht unmittelbar die H_0. Was zu beachten ist, wenn man auf der Basis eines nicht-signifikanten Ergebnisses die Nullhypothese als bestätigt ansehen möchte, wird weiter unten sowie in ▶ Abschn. 15.3 genauer erläutert.

7. Die Ergebnisdarstellung des Signifikanztests

In der Fachliteratur werden bei der Präsentation der Hypothesen nur die inhaltlichen Alternativhypothesen vorgestellt (nicht die Nullhypothesen und nicht die statistischen Hypothesen; diese dienen nur als Hintergrundinformation der statistischen Hypothesenprüfung).

Ebenso wird das Signifikanztestergebnis knapp präsentiert, wobei das deskriptivstatistische Ergebnis zu nennen ist (d. h. hier: der Korrelationskoeffizient), zudem sind als inferenzstatistisches Ergebnis der empirische Wert aus der Prüfverteilung (hier: t_{emp}) sowie die Irrtumswahrscheinlichkeit anzugeben (p). Diese Angaben sind dem Output des verwendeten Statistikprogramms zu entnehmen. Der Output selbst sollte nicht 1 : 1 in einen wissenschaftlichen Ergebnisbericht kopiert werden (zur Ergebnisdarstellung ▶ Kap. 13).

Ein Ergebnissatz zu dem oben dargestellten Signifikanztest könnte also lauten:

Hypothesenkonform zeigte sich ein überzufälliger positiver Zusammenhang zwischen Stress am Arbeitsplatz und Fehlzeiten der Erwerbstätigen ($r = +.28$, $t_{emp(df=98)} = +2.89$, $p = .002$).

Bis hierher haben wir die Logik des klassischen Signifikanztests am Beispiel einer gerichteten Zusammenhangshypothese und eines sog. einseitigen statistischen Signifikanztests kennengelernt. Zum vollständigen Verständnis des Signifikanztests und zu dessen sachgerechter Anwendung sind nun noch weitere Konzepte einzuführen: Einseitiger und zweiseitiger Test, Effektgröße, Teststärke und optimaler Stichprobenumfang.

Einseitiger und zweiseitiger Signifikanztest

Forschungshypothesen sollten normalerweise die Richtung des erwarteten Effekts (bzw. das Vorzeichen des interessierenden Populationsparameters) angeben und postulieren, ob es sich um einen positiven oder negativen Zusammenhang handelt, ob das Merkmal in der einen oder anderen Gruppe stärker ausgeprägt ist, ob eine Veränderung im Sinne einer Zunahme oder Abnahme der Variablenausprägung erwartet wird. In unserem Beispiel hatten wir einen positiven Zusammenhang postuliert und somit eine gerichtete Alternativhypothese aufgestellt.

Forschungshypothesen werden normalerweise als *gerichtete Hypothesen* („one-sided hypotheses") formuliert, die angeben, welche Richtung bzw. welches Vorzeichen der postulierte Populationseffekt hat. Ungerichtete Alternativhypothesen („two-sided hypotheses"), die nur „irgendeinen" Zusammenhang, „irgendeinen" Unterschied oder „irgendeine" Veränderung" vorhersagen, ohne Präzisierung der Richtung des Effekts, deuten auf geringe Vorkenntnisse über den Untersuchungsgegenstand hin.

Ungerichtete Forschungshypothesen lassen die Richtung des Effekts offen und sagen somit nur, dass es „irgendeinen" Zusammenhang (positiv oder negativ), „irgendeinen" Gruppenunterschied (zugunsten oder zu ungunsten einer Gruppe) oder „irgendeine" Veränderung (Zunahme oder Abnahme der Merkmalsausprägung) in der Population gibt (Tab. 12.12). Es liegt auf der Hand, dass ungerichtete Hypothesen hinsichtlich ihres theoretischen Erklärungsanspruchs meist unbefriedigend sind. Wenn theoretisch oder aufgrund früherer Studien keine klare Effektrichtung vorhergesagt werden kann, dann deutet dies auf sehr geringe Vorkenntnisse zum Untersuchungsgegenstand hin. Dementsprechend sollte dann in der Regel anstelle einer hypothesen- und theorieprüfend angelegten explanativen Studie eher eine explorative Studie durchgeführt werden, auf deren Basis dann wohlbegründete gerichtete Hypothesen formuliert werden können (▶ Abschn. 7.5).

Tabelle 12.12 Gerichtete und ungerichtete Hypothesenpaare, jeweils bestehend aus Alternativhypothese H_1 und Nullhypothese H_0 (Beispiele für jeweils zwei Variablen)

	Richtung des Effektes	
Art des Effektes	**Gerichtete Hypothesen**	**Ungerichtete Hypothesen**
Zusammenhangshypothese: Zusammenhang zwischen zwei Variablen	H_1: $\rho > 0$ H_0: $\rho \leq 0$	H_1: $\rho \neq 0$ H_0: $\rho = 0$
Unterschiedshypothese: Unterschied zwischen zwei Gruppen	H_1: $\mu_1 > \mu_2$ H_0: $\mu_1 \leq \mu_2$	H_1: $\mu_1 \neq \mu_2$ H_0: $\mu_1 = \mu_2$
Veränderungshypothese: Veränderung zwischen Vorher- und Nachhermessung	H_1: $\mu_{vorher} < \mu_{nachher}$ H_0: $\mu_{vorher} \geq \mu_{nachher}$	H_1: $\mu_{vorher} \neq \mu_{nachher}$ H_0: $\mu_{vorher} = \mu_{nachher}$

Bei einer **gerichteten Forschungshypothese** wird ein **einseitiger Signifikanztest** („one-sided test of significance“) durchgeführt: Bei jedem einseitigen Test wird immer als erstes geprüft, ob das deskriptivstatistische Ergebnis die **vorhergesagte Richtung** anzeigt. Ist die Richtung korrekt, dann wird der empirische Wert mit dem kritischen Wert der Signifikanzgrenze verglichen bzw. es wird geprüft, ob die Irrtumswahrscheinlichkeit bei maximal α bzw. 5 % liegt. Hat man eine gerichtete Hypothese aufgestellt und das deskriptivstatistische Ergebnis weist bereits in die **falsche Richtung**, so erübrigt sich jede weitere Berechnung, die H_1 kann nicht angenommen werden.

Bei einer **ungerichteten Forschungshypothese** wird ein **zweiseitiger Signifikanztest** („two-sided test of significance“) durchgeführt. Bei einer ungerichteten Hypothese muss man sich im Vorfeld nicht auf die Richtung des Effekts festlegen und somit sprechen empirische Ergebnisse in beide Richtungen prinzipiell für die H_1. Soll die ungerichtete H_1 geprüft werden, dass es irgendeinen Zusammenhang zwischen Stress und Absentismus in der untersuchten Population gibt, so sprechen sowohl negative als auch positive Zusammenhänge für die Alternativhypothese und gegen die Nullhypothese. Es muss also an beiden Seiten der Prüfverteilung getestet werden, man spricht deswegen vom zweiseitigen Test. Damit der Test insgesamt weiterhin auf dem Signifikanzniveau α erfolgt, muss beim zweiseitigen Test an jeder Seite der Prüfverteilung die kritische Grenze bei **α/2** gezogen werden (► Studienbeispiel „Zweiseitiger Test bei ungerichteten Hypothesen“).

Nutzt man zur Durchführung des Signifikanztests ein Statistikprogramm, das direkte Irrtumswahrscheinlichkeiten ausgibt, so ist besonders darauf zu achten, a) ob das Programm die **ein- oder zweiseitige Irrtumswahrscheinlichkeit** angibt und b) welche für den eigenen Test gebraucht wird. Häufig wird als Grundeinstellung die zweiseitige Irrtumswahrscheinlichkeit ausgegeben, die für den Test bei ungerichteter Alternativhypothese korrekt ist. Soll indessen ein einseitiger Test durchgeführt werden, so ist die zweiseitige Irrtumswahrscheinlichkeit zu halbieren (denn beim einseitigen Test wird ja – hypothesenkonforme Richtung des geprüften Effektes vorausgesetzt – nur an einer Seite des H_0-Modells die Irrtumswahrscheinlichkeit „abgeschnitten“). Gibt das Programm indessen eine einseitige Irrtumswahrschein-

Studienbeispiel

Zweiseitiger Test bei ungerichteten Hypothesen

Für unser Beispiel zum Zusammenhang von Stress und Fehlzeiten am Arbeitsplatz würde sich bei ungerichteter Alternativhypothese folgendes Vorgehen ergeben:

Inhaltliche Alternativhypothese (ungerichtet): Es besteht ein positiver oder ein negativer Zusammenhang zwischen Stress und Fehlzeiten am Arbeitsplatz.

Statistische Alternativhypothese: H_1: $\rho \neq 0$

Da es sich um eine bivariate Korrelationshypothese handelt und die Variablen auf Kardinalskalenniveau erhoben wurden, ist der Korrelationstest indiziert, der als Prüfverteilung (H_0-Modell) über eine t-Verteilung (df $= n - 2$) erfolgt.

Aus den vorliegenden Stichprobendaten ($n = 100$) wird der Stichprobenkennwert $r = +.28$ berechnet. Er wird in einen empirischen t-Wert von $t_{emp} = +2.89$ umgerechnet.

Für den Signifikanztest muss – beim Rechnen per Hand – der empirische Wert mit einem kritischen Wert verglichen werden. Da eine ungerichtete Alternativhypothese zu prüfen ist, muss der statistische Test zweiseitig erfolgen. Dementsprechend muss bei einem vorab festgelegten Signifikanzniveau von 5 % der **kritische *t*-Wert** nicht für $\alpha = 5\,\%$, sondern für **$\alpha/2 = 2.5\,\%$** aus der Tabelle abgelesen werden. Dieser beträgt laut t-Tabelle bei df $= 98$ $t_{krit(\alpha/2=2.5\%)} = +1.98$. Generell gilt, dass der kritische Wert beim zweiseitigen Test dem Betrag nach höher ist als beim einseitigen Test bzw. dass dasselbe Stichprobenergebnis im zweiseitigen Test schwerer signifikant wird als im einseitigen Test (allerdings muss beim einseitigen Test dafür die Richtung des Effektes im Vorfeld eindeutig festgelegt werden und dann auch empirisch eintreffen).

Im vorliegenden Beispiel ist $t_{emp} = +2.89$ dem Betrag nach größer als $t_{krit(\alpha/2=2.5\%)} = +1.98$ und somit wäre das Ergebnis auch beim zweiseitigen Test statistisch signifikant.

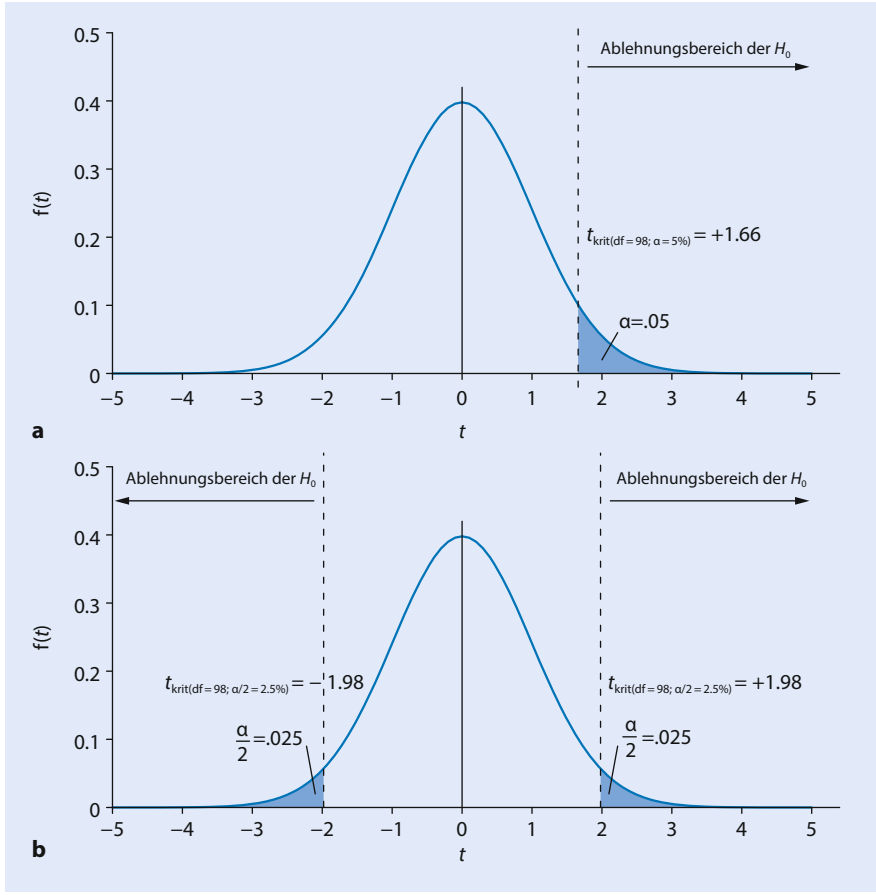


Abb. 12.24 H_0-Modell als t-Verteilung (df = 98) mit Darstellung der Ablehnungsbereiche für H_0 (für ein Signifikanzniveau von 5 %) bei einseitigem Test (**a**) und bei zweiseitigem Test (**b**) für das Absentismus-Beispiel mit $r = +.28$ bzw. $t_{emp} = +2.89$ (fiktive Daten)

lichkeit aus, so muss diese für einen zweiseitigen Test verdoppelt werden. Da man sich in der Hypothese nicht auf eine Richtung des Effekts festlegen konnte, werden beim zweiseitigen Test die Flächen addiert, die der empirische Wert an beiden Seiten der Verteilung abschneidet (Abb. 12.24).

Gerichtete Hypothesen – Gerichtete Hypothesen werden durch einen **einseitigen Signifikanztest** geprüft, indem zunächst die Richtung des Effekts betrachtet und dann die **einseitige Irrtumswahrscheinlichkeit** (Fläche des H_0-Modells, die vom empirischen Wert auf einer Seite der Prüfverteilung abgeschnitten wird) mit dem Signifikanzniveau von 5 % verglichen wird. Alternativ wird beim Rechnen per Hand (z. B. in Klausuren) für die Signifikanzentscheidung der empirische Wert mit dem **kritischen Wert für α = 5 %** verglichen, der auf der einen Seite des H_0-Modells die Grenze des einseitigen Ablehnungsbereichs für die H_0 bildet.

Ungerichtete Hypothesen – Ungerichtete Hypothesen werden durch einen **zweiseitigen Signifikanztest** geprüft, indem die zweiseitige Irrtumswahrscheinlichkeit (Fläche, die vom empirischen Wert an beiden Seiten der Prüfverteilung abgeschnitten wird) mit dem Signifikanzniveau von 5 % verglichen wird. Alternativ wird beim Rechnen per Hand (z. B. in Klausuren) für die Signifikanzentscheidung der empirische Wert mit dem **kritischen Wert für α/2 = 2.5 %** verglichen, der auf beiden Seiten des H_0-Modells die Grenze des zweiseitigen Ablehnungsbereichs für die H_0 bildet.

Signifikantes Ergebnis und Effektgröße

Häufig wird ein statistisch signifikantes Ergebnis automatisch für ein bedeutsames Ergebnis gehalten, insbesondere wenn es auf einer großen Stichprobe basiert. Gerade bei großen Stichproben können jedoch auch sehr kleine, praktisch unbedeutende Effekte statistisch signifikant werden. Denn bei großen Stichproben gewinnt der Signifikanztest an Teststärke („power"), kann also auch kleine Effekte nachweisen. Eine Zurückweisung des H_0-Modells bedeutet schließlich nur, dass der Populationsparameter nicht exakt Null ist. Er kann jedoch winzig klein sein, und – gerade bei großem Stichprobenumfang – dennoch statistisch signifikant werden. (Mit wachsendem Stichprobenumfang reduziert sich nämlich die Streuung der als H_0-Modell verwendeten Stichprobenkennwerteverteilung.)

Deswegen muss bei einem statistisch signifikanten Ergebnis immer die **Effektgröße** (Effektstärke, „effect size", ES) bestimmt und hinsichtlich theoretischer und/oder praktischer Bedeutsamkeit interpretiert werden. Je nach dem, auf welchen Populationsparameter sich die geprüfte Hypothese bezieht, ist das passende unstandardisierte oder standardisierte **Effektgrößenmaß** zu wählen und zu berechnen (▶ Abschn. 14.2). Für die Interpretation der in

einer Studie mit signifikantem Befund berechneten Effektgröße ist dann der **Forschungsstand** zu konsultieren: Ist der gefundene Effekt größer oder kleiner oder vergleichbar mit Effektgrößen, die in anderen Studien des entsprechenden Forschungsfeldes nachgewiesen wurden? Hier muss inhaltlich vergleichend argumentiert werden. Zur Einordnung der verschiedenen Effektgrößenmaße haben sich in der Tradition von Jacob Cohen (1988) als Faustregel auch Normwerte eingebürgert, nach denen Effekte grob als klein, mittel oder groß klassifiziert werden. Eine solche **pauschale Effektgrößen-Klassifikation** ist für die Ergebnisinterpretation aber nur ein vager Anhaltspunkt und ersetzt nicht die inhaltliche Bewertung der berechneten Effektgröße im Licht des Forschungsstandes zum konkreten Forschungsproblem.

Zwar mag man generell davon ausgehen, dass größere Effekte auf größere praktische Bedeutsamkeit hinweisen, für praktische Schlussfolgerungen aus empirischen Befunden sind aber noch weitere Punkte zu bedenken: So kann in bestimmten Praxisfeldern auch ein numerisch sehr kleiner Effekt praktisch hochgradig bedeutsam sein, etwa wenn es um Gesundheit und Menschenleben geht. Umgekehrt kann ein großer Effekt von praktisch geringer Bedeutung sein, etwa weil man empirisch noch nicht genau klären konnte, wie er kausal zustande kommt, oder weil andere Einflussfaktoren als der untersuchte möglicherweise viel bessere Ansatzpunkte für praktische Maßnahmen liefern.

Bei einem statistisch signifikanten Ergebnis muss immer auch die *Effektgröße* angegeben und inhaltlich vor dem Hintergrund des Forschungsstandes hinsichtlich ihrer theoretischen und/oder praktischen Bedeutsamkeit diskutiert werden. Für unterschiedliche Populationsparameter stehen verschiedene Effektgrößenmaße zur Verfügung (► Abschn. 14.2).

Zurück zu unserem Absentismus-Beispiel: Bei der bivariaten Korrelation wird der **Korrelationskoeffizient** direkt als **standardisiertes Effektgrößenmaß** betrachtet. Bei einer Korrelation ab .10 sprechen wir von einem kleinen Effekt, ab .30 von einem mittleren Effekt und ab .50 von einem großen Effekt (Cohen, 1988). Nach dieser groben Effektgrößen-Klassifikation ist der signifikante Korrelationskoeffizient von $r = +.28$ somit als **mittlerer Effekt** einzustufen. Für eine genauere Bewertung der Effektgröße müssten vergleichbare Studien aus der Absentismus-Forschung und die dort gefundenen Korrelationen zwischen Stress und Fehlzeiten herangezogen werden.

Für die anwendungsorientierte Anschlussfrage, ob Anti-Stress-Programme für Erwerbstätige umgesetzt werden sollten, um Fehltage zu reduzieren, reicht die Betrachtung der Effektgröße nicht aus. Hier muss zusätzlich die **Ursache-Wirkungs-Relation** zwischen Stress und Fehlzeiten geklärt werden. Denn der Korrelationskoeffizient aus der hier als Beispiel herangezogenen nicht-experimentellen Querschnittstudie gibt keine Auskunft über die Kausalrichtung. Er besagt nicht, dass Stress die Ursache für die Fehlzeiten ist. Umgekehrt könnten auch erhöhte Fehlzeiten zu verstärktem Stresserleben führen, zudem könnten Drittvariablen wirksam sein. Konzeption und Umsetzung von praktischen Maßnahmen sollten auf statistisch signifikanten wissenschaftlichen Befunden ausreichender Effektgröße basieren, die nicht nur Variablen-Zusammenhänge aufzeigen, sondern auch Ursache-Wirkungs-Ketten belegen. Dies ist vor allem im Rahmen von experimentellen Studien (► Abschn. 7.6) sowie von Längsschnittstudien (► Abschn. 7.8) möglich.

Nicht-signifikantes Ergebnis und Teststärke

Bei einem statistisch nicht-signifikanten Ergebnis dürfen wir uns nicht für die Alternativhypothese entscheiden, denn dafür lässt sich das empirische Ergebnis zu gut mit dem H_0-Modell vereinbaren: Die Irrtums- bzw. **Alpha-Fehler-Wahrscheinlichkeit** (also die Wahrscheinlichkeit, sich fälschlich für die H_1 zu entscheiden, obwohl in der Population die H_0 gilt) ist bei einem nicht-signifikanten Ergebnis definitionsgemäß größer als das Signifikanz- bzw. **Alpha-Fehler-Niveau**: $p(D \mid H_0) > \alpha$.

Dennoch dürfen uns bei einem nicht-signifikanten Ergebnis nicht automatisch für die H_0 entscheiden. Denn es könnte auch sein, dass der gewählte Signifikanztest mangels Teststärke von vorne herein gar nicht signifikant werden konnte. Wenn es uns darauf ankommt, die Nullhypothese anzunehmen, müssen wir die **Beta-Fehler-Wahrscheinlichkeit** (also die Wahrscheinlichkeit, fälschlich die H_0 anzunehmen, obwohl in der Population die H_1 gilt) einbeziehen. Bei einer Entscheidung für die H_0 muss die Beta-Fehler-Wahrscheinlichkeit unter einem festgelegten **Beta-Fehler-Niveau** liegen.

Die beiden möglichen Fehler bei statistischen Entscheidungen veranschaulicht ◘ Tab. 12.13: Noch einmal zusammengefasst: Wir begehen einen **Alpha-Fehler** (α-Fehler, Fehler 1. Art), wenn wir aufgrund einer empirischen Untersuchung zugunsten von H_1 entscheiden, obwohl in Wahrheit (in der Population) die H_0 gilt. Dieser Fehler kann nur auftreten, wenn wir uns bei signifikantem Ergebnis für die H_1 entscheiden. Entscheiden wir uns jedoch bei nicht-signifikantem Ergebnis für die H_0, so kann ein anderer Fehler auftreten, nämlich der **Beta-Fehler** (β-Fehler, Fehler 2. Art). Er besteht darin, dass wir uns fälschlich für die H_0 entscheiden, obwohl in der Population die H_1 gilt.

Es ist wichtig, die drei Konzepte des Alpha-/Beta-**Fehlers**, der entsprechenden Alpha-/Beta-**Fehler-Wahr-**

Tabelle 12.13 Alpha- und Beta-Fehler bei statistischen Entscheidungen (Entscheidungslogik zur Veranschaulichung auch formuliert in Begriffen der Signalentdeckungstheorie sowie juristischer Entscheidungen). (Mod. nach Fiedler, Kutzner, & Krueger, 2012. Mit freundlicher Genehmigung von SAGE Publications)

		In der Population gilt die:	
		H_1	H_0
Entscheidung aufgrund des statistischen Signifikanztests anhand der Stichprobendaten zugunsten der:	H_1	Treffer Ein vorhandener Populationseffekt wird gefunden. Der schuldige Täter wird verurteilt. **Richtige Entscheidung für die H_1** Wahrscheinlichkeit: $1 - \beta$ (Teststärke)	Falsch-positive Entscheidung (falscher Alarm) Ein Populationseffekt wird behauptet, obwohl er gar nicht existiert. Der unschuldige Verdächtige wird verurteilt. **Falsche Entscheidung für die H_1: Alpha-Fehler/ Fehler erster Art** Wahrscheinlichkeit: α-Fehler-Wahrscheinlichkeit $p(D \mid H_0)$
	H_0	Falsch-negative Entscheidung (Verpasser) Ein vorhandener Populationseffekt wird nicht gefunden. Der schuldige Täter wird entlassen. **Falsche Entscheidung für die H_0: Beta-Fehler / Fehler zweiter Art** Wahrscheinlichkeit: β-Fehler-Wahrscheinlichkeit $p(D \mid H_1)$	Korrekte Ablehnung Es wird korrekt festgestellt, dass tatsächlich kein Populationseffekt vorliegt. Der unschuldige Verdächtige wird freigelassen. **Richtige Entscheidung für die H_0** Wahrscheinlichkeit: $1 - \alpha$

12

scheinlichkeit und des zur Bewertung eingesetzten Alpha-/Beta-**Fehler-Niveaus** voneinander zu differenzieren (zur wissenschaftstheoretischen Reflexion des Umgangs mit diesen Fehlern s. Fiedler, Kutzner, & Krueger, 2012).

In Tab. 12.13 wird deutlich, dass das Risiko eines β-Fehlers nur besteht, wenn eine Entscheidung zugunsten von H_0 getroffen wird. Bei einer Entscheidung für die H_0 müssen wir also sichergehen, dass die Beta-Fehler-Wahrscheinlichkeit gering ist. Welche Entscheidungsgrenze wird hier gesetzt? Während sich das **Alpha-Fehler-Niveau von $\alpha = 5\,\%$** etabliert hat, wird das Beta-Fehler-Niveau bislang weniger konventionell gehandhabt. Allerdings zeichnet sich die Tendenz ab, das **Beta-Fehler-Niveau auf $\beta = 20\,\%$** festzulegen. Wenn die Beta-Fehler-Wahrscheinlichkeit maximal 20 % beträgt, dann ist die Teststärke von mindestens 80 % sichergestellt, da die **Teststärke als $1 - \beta$** definiert ist. Die Teststärke gibt an, mit welcher Wahrscheinlichkeit beim Vorliegen eines Populationseffektes ein statistisch signifikantes Ergebnis zustande kommen kann. Ein nicht-signifikantes Ergebnis darf nur dann zum Anlass genommen werden, die H_0 anzunehmen, wenn der Test über genügend Teststärke verfügt, so dass er bei Vorhandensein eines Populationseffektes die Chance gehabt hätte, signifikant zu werden.

Wenn also ein nicht-signifikantes Ergebnis vorliegt, muss in jedem Fall die Teststärke bzw. die β-Fehlerwahrscheinlichkeit bestimmt werden. Ist die Teststärke ausreichend groß ($1 - \beta \geq 80\,\%$), so ist gleichzeitig die Beta-Fehler-Wahrscheinlichkeit gering ($\beta < 20\,\%$) – ein nicht-signifikantes Ergebnis ist nur unter diesen Bedingungen zum Anlass zu nehmen, die H_0 zu akzeptieren. Ist die Teststärke nicht ausreichend, so bedeutet dies, dass der Test nicht genügend Daten zur Verfügung hat, um eine Entscheidung über das aus H_1 und H_0 bestehende Hypothesenpaar zu stützen. In der Praxis lässt sich die Teststärke bei Vorliegen eines nicht-signifikanten Ergebnisses im Nachhinein über eine **Post-hoc-Teststärkenanalyse** bestimmen (▶ Abschn. 14.1.1), z. B. mit dem kostenlosen Statistiktool **G*Power** (Buchner, Erdfelder & Faul, 2012: http://www.gpower.hhu.de).

Bei einem statistisch *nicht-signifikanten Ergebnis* muss immer auch die *Teststärke* bestimmt werden. Nur wenn die Teststärke mindestens 80 % beträgt ($1 - \beta \geq .80$), kann das nicht-signifikante Ergebnis zum Anlass genommen werden, die H_0 anzunehmen. Liegt die Teststärke unter 80 % bzw. umgekehrt, die Beta-Fehler-Wahrscheinlichkeit über 20 % ($\beta > .20$), so ist das Ergebnis nicht eindeutig interpretierbar. Der Signifikanztest basiert dann offensichtlich auf einer zu geringen Datenbasis, um eine Entscheidung H_1 über das Hypothesenpaar zu treffen.

Die Logik der Bestimmung der β-Fehler-Wahrscheinlichkeit bzw. der Teststärke – wie sie z. B. dem Programm g*power zugrunde liegt – sei im Folgenden erläutert. Sie setzt voraus, dass wir in der Lage sind, die in der Alternativhypothese behaupteten Populationsverhältnisse zu präzisieren. Eine gerichtete Alternativhypothese hatte bisher die Form H_1: $\rho > 0$ (erneut sollen die folgenden

Überlegungen exemplarisch an einer bivariaten Korrelationshypothese verdeutlicht werden). Diese Hypothesenart bezeichneten wir als eine unspezifische gerichtete Alternativhypothese. Legen wir fest, dass bei Gültigkeit von H_1 die positive Korrelation einen Mindestbetrag von .20 haben soll (also eine kleine bis mittlere Effektstärke), so resultiert eine spezifische gerichtete Hypothese: $\rho \geq +.20$. Nur eine **spezifische Alternativhypothese** ermöglicht es, die β-Fehler-Wahrscheinlichkeit für eine fälschliche Ablehnung von H_1 zu bestimmen.

Die Beta-Fehler-Wahrscheinlichkeit kann nur bei spezifischer H_1 (Alternativhypothese) anhand eines H_1-Modells bestimmt werden.

Die Bestimmung dieser Wahrscheinlichkeit folgt im Prinzip dem gleichen Gedankengang wie die Berechnung der α-Fehler-Wahrscheinlichkeit. Wenn in der Population die H_1: $\rho \geq +.20$ gilt, resultiert für die Zufallsvariable r bzw. t eine Dichtefunktion, deren mathematischer Aufbau bekannt ist. Mithilfe dieser als $\boldsymbol{H_1}$**-Modell** fungierenden Verteilung lässt sich die (bedingte) Wahrscheinlichkeit ermitteln, mit der empirische Korrelationen bei Gültigkeit von H_1 auftreten können, die mindestens so deutlich (in Richtung H_0) vom H_1-Parameter (hier: $= +.20$) abweichen wie der gefundene Korrelationskoeffizient. Diese bedingte Wahrscheinlichkeit des gefundenen empirischen Ergebnisses (Daten: D) unter Annahme der Gültigkeit der H_1 bezeichnen wir als **β-Fehler-Wahrscheinlichkeit**: $\boldsymbol{p(D \mid H_1)}$. Aus der β-Fehler-Wahrscheinlichkeit können wir die Teststärke $1 - \beta$ berechnen.

Stellt man dem H_0-Modell für unser Absentismus-Beispiel nun jeweils verschiedene H_1-Modelle an die Seite, die auf unterschiedlichen erwarteten Populationseffekten zwischen $\rho = .10$ und $\rho = .40$ basieren, so sieht man deutlich, wie stark die Beta-Fehler-Wahrscheinlichkeit (und somit die Teststärke) von der mit dem H_1-Modell postulierten Effektgröße abhängt (Abb. 12.25). Ein statistisch nicht-signifikantes Ergebnis, das im H_0-Modell **nicht** im Ablehnungsbereich der H_0 liegt (z. B. $t = 1.50$), kann z. B. bei einem H_1-Modell für $\rho = .10$ nicht zum Anlass genommen werden, die H_0 anzunehmen. Denn dieses nicht-signifikante Ergebnis ließe sich angesichts der großen Beta-Fehler-Wahrscheinlichkeit noch sehr gut mit dem H_1-Modell vereinbaren. Anders gesagt: H_0-Modell und H_1-Modell überlappen sich sehr stark und erlauben somit keine klare Entscheidung. Wenn man dagegen einen großen Populationseffekt von $\rho = .40$ zugrunde legt, dann kann bei einem nicht-signifikanten Ergebnis die Nullhypothese beibehalten werden, da das Ergebnis mit einem solchen H_1-Modell nicht vereinbar bzw. die Beta-Fehler-Wahrscheinlichkeit sehr klein ist. Post-hoc-Teststärkebestimmungen hängen also maßgeblich davon ab, wie man die H_1 inhaltlich begründet spezifiziert (▶ Abschn. 14.1.1).

Zu beachten ist, dass die Beta-Fehler-Wahrscheinlichkeit bzw. die Teststärke nicht nur von der Effektgröße, sondern auch von anderen Faktoren abhängt. Allgemein lässt sich sagen, dass die Beta-Fehler-Wahrscheinlichkeit umso geringer und somit die Teststärke umso größer ausfällt, a) je größer die Effektstärke ist, b) je liberaler das Signifikanzniveau α festgelegt ist (z. B. 5 % oder sogar 10 % statt 1 %), c) je geringer die Merkmalsstreuung in der Population ist und d) je größer der Stichprobenumfang ist.

Alpha- und Beta-Fehler verhalten sich gegenläufig: Wird das Alpha-Fehler-Niveau strenger angesetzt bzw. numerisch verkleinert (z. B. von 5 % auf 1 %), so wird damit gleichzeitig die Beta-Fehler-Wahrscheinlichkeit vergrößert (bzw. die Teststärke $1 - \beta$ verringert).

Optimaler Stichprobenumfang

Kaum etwas ist ärgerlicher, als nach Durchführung einer explanativen Studie bei der inferenzstatistischen Auswertung festzustellen, dass man fast nur nicht-signifikante Befunde erhält und dass dies – laut Post-hoc-Teststärkenbestimmung daran liegt, dass der Stichprobenumfang zu gering war, um eine Teststärke von mindestens 80 % zu gewährleisten. In diesem Fall müssen dann alle statistischen Signifikanztests aufgrund mangelnder Datenbasis als nicht aussagekräftig im Hinblick auf die Hypothesen eingeordnet werden. Allenfalls kann dann noch – mit großer Vorsicht – das deskriptivstatistische Ergebnismuster als Hinweis genutzt werden, etwa wenn es der Richtung nach konsistent hypothesenkonform ausfällt.

Um statistisch uneindeutige Ergebnisse zu vermeiden, sollte möglichst bereits bei der Untersuchungsplanung der notwendige Stichprobenumfang für eine Teststärke von 80 % festgelegt und dann realisiert werden. Dieser Mindest-Stichprobenumfang wird zuweilen auch als **„optimaler Stichprobenumfang“** bezeichnet. Die Optimierung bezieht sich auf die Kosten-Nutzen-Bilanz. Denn zwar lässt sich argumentieren, dass ein umso größerer Stichprobenumfang für genauere statistische Berechnungen sorgt. Er kann jedoch letztlich auch eine Ressourcenvergeudung darstellen (unnötige Belastung von Untersuchungspersonen, unnötige Personalkosten für Datenerhebung und Datenanalyse), wenn durch eine übergroße Stichprobe und somit übergroße Teststärke letztlich minimale Effekte statistisch signifikant werden, die praktisch bedeutungslos sind.

Um **a priori** (d. h. vor der Durchführung der Studie) denjenigen „optimalen“ Stichprobenumfang zu iden-

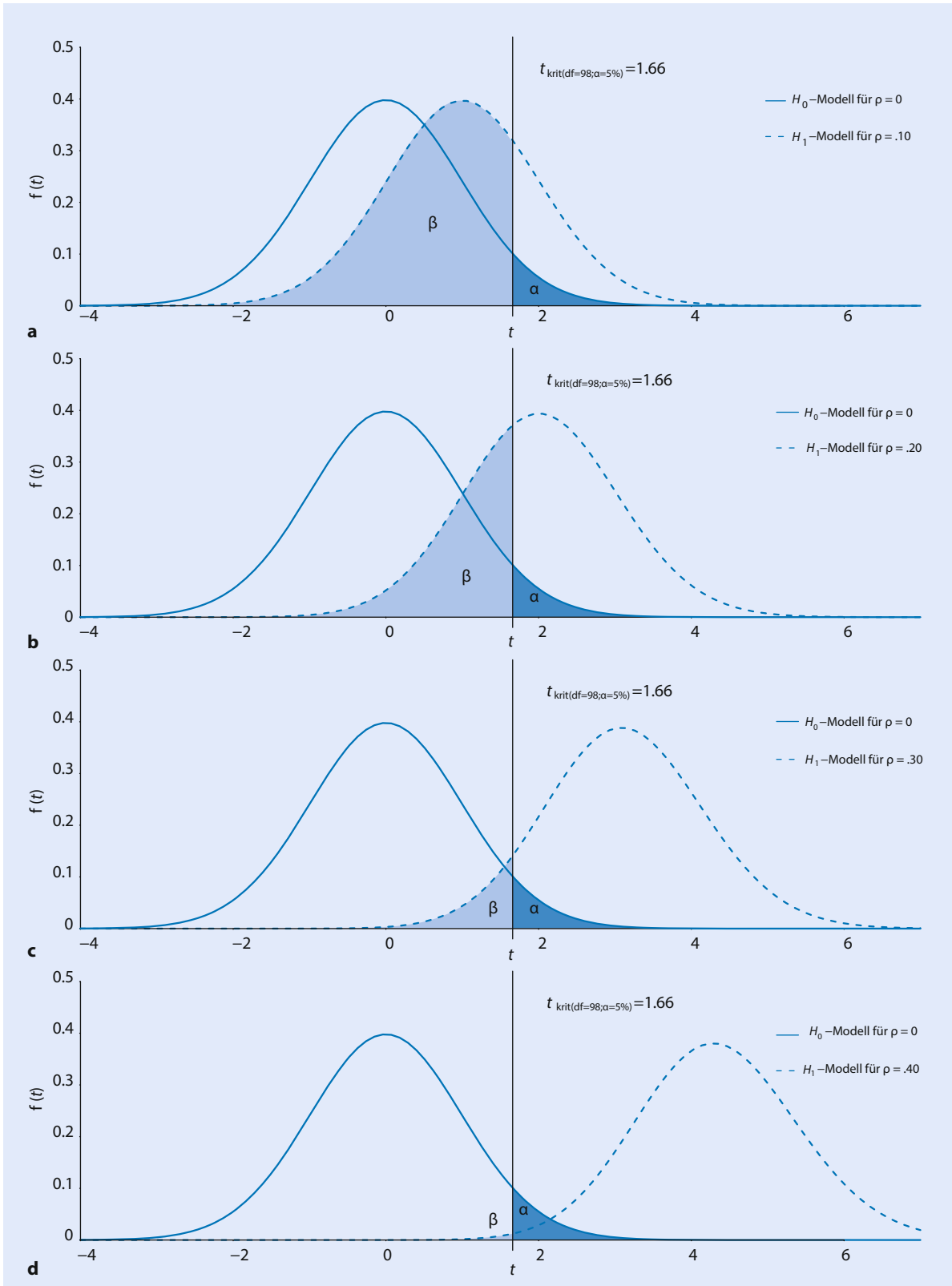


Abb. 12.25 H_0-Modell sowie H_1-Modelle (als *t*-Verteilungen mit df = 98) zu unserem Absentismus-Beispiel für unterschiedliche Effektgrößen von $\rho = .10$ bis $\rho = .40$ mit der jeweiligen Beta-Fehler-Wahrscheinlichkeit

tifizieren, der eine Teststärke von mindestens 80 % sicherstellt, muss gleichzeitig abgeschätzt werden, welche Effektgröße erwartet wird. Generell gilt: Je größer der erwartete Effekt, umso geringer die notwendige Stichprobengröße. Umgekehrt: Je kleiner der erwartete Effekt, umso größer ist der notwendige Stichprobenumfang, um ein aussagekräftiges statistisches Ergebnis über die Hypothesen zu bekommen. Welche Effektgröße man veranschlagt, ist entweder auf der Basis früherer Studien zu entscheiden oder durch eine Vorstudie zu klären.

Optimaler Stichprobenumfang – Um zu vermeiden, dass eine Studie uneindeutige statistische Ergebnisse produziert, sollte bei der Untersuchungsplanung der **optimale Stichprobenumfang** veranschlagt werden. Es handelt sich dabei um den Stichprobenumfang, der mindestens notwendig ist, um einen Effekt bestimmter Größe mit einer Teststärke von mindestens 80 % aufdecken zu können.

Optimale Stichprobenumfänge lassen sich a priori mit dem Statistikprogramm g*power bestimmen und sind für unterschiedliche Signifikanztests auch austabelliert (Bortz & Schuster, 2010). Hätten wir für unser Korrelationsbeispiel etwa einen kleinen Effekt erwartet, so wäre die Studie mit einem optimalen Stichprobenumfang von $n = 1\,073$ anzusetzen gewesen. Da wir es in den Sozialwissenschaften oft mit kleinen Effekten zu tun haben, ist die Problematik kleiner bzw. suboptimaler Stichprobenumfänge ernst zu nehmen, um Studien zu vermeiden, die von vorne herein als nicht aussagekräftig im Hinblick auf die zu prüfenden Hypothesen einzuordnen sind (zur Untersuchungsplanung mit optimalen Stichprobenumfängen ▶ Abschn. 14.3).

Konfidenzintervalle als Alternative zum Signifikanztest

Der statistische Nullhypothesen-Signifikanztest hat sich zwar stark etabliert, ist aber auch umstritten. Ein Hauptkritikpunkt betrifft das zugrunde liegende H_0-Modell auf der Basis einer sog. **Punkt-Nullhypothese**, die unrealistischerweise von einem exakten Nulleffekt ausgeht. Deswegen wurden Minimum-Effektgrößen-Tests vorgeschlagen (▶ Kap. 15), die nicht prüfen, ob ein Effekt von dem Wert Null abweicht (das trifft fast immer zu), sondern ob er einen vorher postulierten, praktisch bedeutsamen Minimaleffekt überschreitet.

Ein weiterer Kritikpunkt am klassischen Nullhypothesen-Signifikanztest betrifft die Argumentationslogik: Der Test soll der Hypothesenprüfung dienen, berechnet aber mit der Irrtumswahrscheinlichkeit $p(D \mid H_0)$ gerade nicht die Wahrscheinlichkeit, dass die Hypothese zutrifft, sondern nur die bedingte Wahrscheinlichkeit des Auftretens des empirischen oder eines extremeren Ergebnisses unter Annahme der Gültigkeit der Nullhypothese. Anhand einer Datenwahrscheinlichkeit eine Entscheidung über Hypothesen zu treffen, halten Kritiker für inkonsistent (Cohen, 1994).

Vor diesem Hintergrund wird das **Konfidenzintervall** als eine Alternative zum klassischen Signifikanztest betrachtet (Brandstätter, 1999; Eid et al., 2010, S. 227). Das Konfidenzintervall als inferenzstatistische Methode der Parameterschätzung gibt für einen interessierenden Populationsparameter auf der Basis des empirischen Stichprobenkennwertes einen Vertrauensbereich an, in dem der unbekannte wahre Populationsparameter mit hoher Wahrscheinlichkeit liegt. Dabei wird üblicherweise ein Konfidenzkoeffizient von 95 % gewählt (▶ Abschn. 12.4.2).

Für unser Beispiel zum Zusammenhang zwischen Stressbelastung und Fehltagen lässt sich folgendes Konfidenzintervall (KI) bestimmen:

$$\text{KI}\,(95\,\%): +.09 \leq \rho \leq +.45.$$

International üblich und vom APA Publication Manual empfohlen ist folgende Schreibweise für Konfidenzintervalle (CI, „confidence interval“), wobei dann aus dem Kontext hervorgehen muss, welcher Populationsparameter dargestellt wird:

$$95\,\%\,\text{CI} \quad [.09, .45]$$

Die Angabe des Konfidenzintervalls zeigt auf den ersten Blick die Effektgröße des Populationseffektes, der gemäß Effektgrößen-Klassifikation nach Cohen (1988) im Bereich klein bis mittel einzustufen ist. Zudem ist erkennbar, dass das Konfidenzintervall nicht den Wert Null einschließt, sondern vollständig im positiven Bereich liegt. Somit ist davon auszugehen, dass auch die Populationskorrelation ein positives Vorzeichen hat bzw. dass in der Population der Erwerbstätigen, aus der die untersuchte Zufallsstichprobe stammt, ein positiver Zusammenhang zwischen Stressbelastung und Fehlzeiten vorliegt. Damit liegen aus Sicht der Befürworter von Konfidenzintervallen alle Informationen vor, die uns im Hinblick auf die Einschätzung der Forschungshypothese interessieren. Die Versuchung, ein Ergebnis allein aufgrund „statistischer Signifikanz“ überzubewerten, sei bei der Konfidenzintervallmethode reduziert, da sowohl ein kleiner Effekt als auch eine unsichere Schätzung – dann ist das Konfidenzintervall sehr breit – sofort ins Auge stechen.

Zudem ist die Interpretationslogik, nach der der interessierende Populationsparameter mit einer subjektiven Wahrscheinlichkeit von 95 % (Konfidenzkoeffizient) im Vertrauensbereich liegt (Brandstätter, 1999, S. 7), konsistenter als beim Signifikanztest, bei dem Aussagen über Hypothesen mit Datenwahrscheinlichkeiten begründet werden. Will man das Konfidenzintervall nicht gemäß subjektiver Wahrscheinlichkeitstheorie, sondern auf der

Basis klassischer frequentistischer Wahrscheinlichkeitstheorie interpretieren, so ist zu konstatieren, dass das Konfidenzintervall eine Zufallsvariable darstellt und dass bei theoretisch unendlich vielen Stichprobenziehungen aus der Population, der wahre Populationsparameter sich in 95 % der Konfidenzintervalle befindet (▶ Abschn. 12.4.2).

Der breiten Verwendung von Konfidenzintervallen anstelle von Signifikanztests steht entgegen, dass Konfidenzintervalle von gängigen Statistikprogrammen nicht direkt ausgegeben werden, sondern anhand des Outputs manuell bestimmt werden müssen und dass sich die Forschungscommunity weitgehend an statistische Nullhypothesen-Signifikanztests gewöhnt hat bzw. diese fordert (z. B. gemäß Autorenhinweisen von Fachzeitschriften; Wilkinson & Task Force on Statistical Inference, 1999).

Zur Prüfung von Hypothesen können anstelle des klassischen statistischen Nullhypothesen-Signifikanztestes auch *Konfidenzintervalle* zum Einsatz kommen. Sie zeigen an, in welchem Bereich der interessierende Populationsparameter mit hoher Wahrscheinlichkeit liegt.

Enthält ein Konfidenzintervall den Wert 0, dann lässt sich über die Richtung des Effekts nichts sagen (dies entspricht einem nicht-signifikanten Ergebnis beim Signifikanztest).

Liegt ein Konfidenzintervall jedoch vollständig unter oder über dem Wert 0, so lässt sich damit die negative bzw. positive Richtung des Populationseffektes mit hoher Wahrscheinlichkeit angeben (dies entspricht einem signifikanten Ergebnis beim Signifikanztest).

Der Vorteil von Konfidenzintervallen besteht darin, dass das Augenmerk direkt auf der Effektgröße (Beträge der Grenzen des Konfidenzintervalls) und der Schätzgenauigkeit (Breite des Konfidenzintervalls) liegt. Zudem ist die Interpretationslogik, nach der der interessierende Populationsparameter mit einer subjektiven Wahrscheinlichkeit von 95 % (Konfidenzkoeffizient) im Vertrauensbereich liegt, konsistenter als beim Signifikanztest, bei dem Aussagen über Hypothesen mit Datenwahrscheinlichkeiten begründet werden.

Lernquiz

1. Welche Aussage ist richtig?
   a) Wenig aussagekräftige Daten können vor der Hypothesenprüfung so aufbereitet werden, dass eine statistische Überprüfung sinnvoll stattfinden kann.
   b) Das Untersuchungsdesign entscheidet über den Informationsgehalt der Daten in Bezug auf die zu überprüfenden Hypothesen.
   c) Ungerichtete Hypothesen haben einen höheren Aussagegehalt als gerichtete Hypothesen.
   d) Hypothesen sollten nach der Betrachtung des empirischen Stichprobenergebnisses an die Datenlage angepasst werden.
   e) In wissenschaftlichen Publikationen werden immer die inhaltlichen und die statistischen Hypothesen genannt.
2. Wovon hängt die Auswahl eines passenden Signifikanztestes für einen inhaltlich sinnvollen statistischen Hypothesentest ab?
3. Warum wird der klassische Signifikanztest auch als Nullhypothesen-Signifikanztest bezeichnet?
4. Warum wird als Entscheidungsregel für die Signifikanz eines Ergebnisses einerseits $[p(D \mid H_0) \leq \alpha]$ und andererseits [|empirischer Wert| $\geq$ |kritischer Wert|] angegeben?
5. Im klassischen Nullhypothesen-Signifikanztest werden bei großem Stichprobenumfang auch kleine Effekte statistisch signifikant, denn die Streuung des H_0-Modells reduziert sich und der kritische Wert wird dem Betrag nach kleiner. Illustrieren Sie diese Aussage, indem Sie den kritischen t-Wert für die 5 %-Signifikanzschwelle zur einseitigen Prüfung der bivariaten Korrelation jeweils für eine Stichprobengröße von $n = 10$, $n = 20$, $n = 30$ und $n = 100$ gegenüberstellen!
6. In einer Veröffentlichung lesen Sie, dass das Ergebnis eines Signifikanztests „höchst signifikant ($p < .001$) ausgefallen“ sei. Diskutieren Sie diese Aussage!
7. „Bei einer Korrelation von $r = -.14$ handelt es sich um einen kleinen und somit praktisch bedeutungslosen Effekt.“ Diskutieren Sie diese Aussage!
8. Welche Aussagen sind falsch?
   a) Ein einseitiger Signifikanztest ist teststärker als ein zweiseitiger Signifikanztest.
   b) Bei einem einseitigen Signifikanztest wird das Signifikanzniveau auf 10 % festgelegt, weil der kritische Wert nur auf einer Seite der Prüfverteilung abgetragen wird.
   c) Ein statistisch signifikantes Ergebnis deutet auf praktische Bedeutsamkeit hin.
   d) Die Verwendung von optimalen Stichprobenumfängen garantiert, dass statistisch signifikante Ergebnisse erzielt werden.
9. Grenzen Sie Alpha- und Beta-Fehler-Wahrscheinlichkeit voneinander ab und geben Sie jeweils das korrekte Symbol an.
10. Welches der folgenden Konfidenzintervalle (95 %) lässt keine Aussage über die Richtung des Effekts zu? Welches Konfidenzintervall schätzt den Populationseffekt am genauesten?
    a) $-.21 \leq \rho \leq -.48$
    b) $+.15 \leq \rho \leq +.23$

c) $-.01 \leq \rho \leq +.34$
d) $+.31 \leq \rho \leq +.42$

Übungsaufgaben

1. Formulieren Sie jeweils sinnvolle inhaltliche und statistische Hypothesenpaare für
   a) eine ungerichtete Zusammenhangshypothese für zwei Variablen
   b) eine gerichtete Unterschiedshypothese für zwei Gruppen
   c) eine spezifische Veränderungshypothese für eine Vorher- und Nachher-Messung.
2. Installieren Sie das Programm g*power und machen Sie sich mit der Oberfläche vertraut.
3. Bestimmen Sie den optimalen Stichprobenumfang mittels g*power für die Überprüfung einer bivariaten Produkt-Moment-Korrelation mit kleinem, mittlerem und großem Effekt gemäß der Effektgrößen-Klassifikation nach Cohen (1988).

Weiterführende Literatur

- Aron, A. & Aron, E. N. (2002). *Statistics for Psychology.* Prentice Hall: Pearson.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer
- Bühner, M. & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler.* München: Pearson Studium.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.

Literatur

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423–437.

Bortz, J. & Schuster, C. (2010). *Statistik für Human– und Sozialwissenschaftler* (7. Ed.). Berlin, Heidelberg: Springer

Brandstätter, E. (1999). Konfidenzintervalle als Alternative zu Signifikanztests. *Methods of Psychological Research Online, 4*(2), 1–17.

Bredenkamp, J. (1969). Über Maße der praktischen Signifikanz. *Zeitschrift für Psychologie, 177*, 310–318.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung*. Frankfurt/Main: Akademische Verlagsanstalt.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente*. Darmstadt: Steinkopff.

Buchner, A., Erdfelder, E., & Faul, F. (2012). G*Power. Abgerufen 27. Februar, 2012, unter http://www.gpower.hhu.de

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378–399.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum.

Cohen, J. (1994). The earth is round (p < 0.05). *American Psychologist, 49*(12), 997–1003.

Cook, T. D., Grader, C. L., Hennigan, K. M., & Flay, B. R. (1979). The history of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin, 86*(4), 662–679.

Cowles, M. & Davis, C. (1982). On the origins of the .05 level of significance. *American Psychologist, 37*(5), 553–558.

Crane, J. A. (1980). Relative likelihood analysis versus significance tests. *Evaluation Review, 4*(6), 824–842.

Eid, M., Gollwitzer, M., & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.

Erdfelder, E. & Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann & W. T. Tack (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich B, Serie I, Methodologische Grundlagen der Psychologie* (Bd. 1, S. 604–648). Göttingen: Hogrefe.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*(6), 661–669.

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. (1956). *Statistical methods and scientific inference*. Edinburgh, UK: Oliver & Boyd.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1–20.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Hrsg.), *Allgemeine experimentelle Psychologie* (S. 43–264). Stuttgart: Fischer.

Hager, W. (2004). *Testplanung zur statistischen Prüfung psychologischer Hypothesen*. Göttingen: Hogrefe.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Hillsdale: Erlbaum.

Harnatt, J. (1975). Der statistische Signifikanztest in kritischer Betrachtung. *Psychologische Beiträge, 17*, 595–612.

Heerden, J. V. & van Hoogstraten, J. (1978). Significance as a determinant of interest in scientific research. *European Journal of Social Psychology, 8*, 141–143.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychological Review, 2*(3), 196–217.

Kline, R. B. (2004). *Beyond significance testing*. Washington: American Psychological Association.

Krause, B. & Metzler, P. (1978). Zur Anwendung der Inferenzstatistik in der psychologischen Forschung. *Zeitschrift für Psychologie, 186*, 244–267.

Lane, D. M. & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology 31*(2), 107–112.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3), 151–157.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research. Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.

Morrison, D. E. & Henkel, R. E. (2007). *The significance test controversy* (new ed.). Chicago: Aldine.

Neyman, J. & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika, 20A*(1/2), 175–294.

Neyman, J. & Pearson, E. (1933). The testing of statistical hypotheses in relation to probability a priori. *Mathematical Proceedings of the Cambridge Philosophical Society, 29*, 492–510.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301.

Wilkinson, L. & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Willmes, K. (1996). Neyman-Pearson-Theorie statistischen Testens. In E. Erdfelder, R. Mausfeld & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 109–122). Weinheim: Psychologie Verlags Union.

Witte, E. H. (1977). Zur Logik und Anwendung der Inferenzstatistik. *Psychologische Beiträge, 19*, 290–303.

Wottawa, H. (1990). Einige Überlegungen zu (Fehl-)Entwicklungen der psychologischen Methodenlehre. *Psychologische Rundschau, 41*(2), 84–107.

12.5.2 Signifikanztests für Zusammenhangshypothesen

Lernziele

- Verschiedene Verfahren zur statistischen Überprüfung von Zusammenhangshypothesen kennen.
- Geeignete statistische Signifikanztests passend zur Hypothese und Datenlage auswählen können.
- Stichprobenfehler bei der Untersuchungsplanung berücksichtigen und vermeiden können.
- Den Unterschied zwischen statistischem Zusammenhang und Kausalität verstehen.
- Ergebnisse der Überprüfung von Zusammenhangshypothesen richtig interpretieren können.
- Das Konzept der Mediator- und der Moderatorvariable erläutern können.

Untersuchungen zur Überprüfung von Hypothesen, die sich auf Zusammenhänge zwischen Merkmalen beziehen, werden oft als **Korrelationsstudien** oder als **Interdependenzanalysen** (Selg, 1971) bezeichnet. Der in einer Interdependenzanalyse gefundene Zusammenhang sagt zunächst nichts über Kausalbeziehungen der untersuchten Merkmale aus. Schlussfolgerungen, die aus Interdependenzanalysen gezogen werden können, beziehen sich primär nur auf die Art und Intensität des miteinander Variierens **(Kovariierens)** zweier oder mehrerer Merkmale. Untersuchungstechnische Vorkehrungen oder inhaltliche Überlegungen können jedoch bestimmte kausale Wirkungsmodelle besonders nahe legen bzw. ausschließen, so dass die Anzahl kausaler Erklärungsalternativen eingeschränkt bzw. die interne Validität der Interdependenzanalyse erhöht wird (zu diesem Problem Jäger, 1974; Köbben, 1970).

Zur inferenzstatistischen Überprüfung von Zusammenhangshypothesen stehen eine Reihe von klassischen Signifikanztests – v. a. die sog. **Korrelations- und Regressionsanalyse** – zur Verfügung (◘ Tab. 12.14). Hier muss in Abhängigkeit von der Anzahl der beteiligten Variablen sowie ihrem Skalenniveau das richtige Verfahren ausgewählt werden. Bei kardinalskalierten Variablen sind **lineare Modelle** einschlägig. Nicht selten stehen Variablen in den Human- und Sozialwissenschaften jedoch auch in einem **nicht-linearen** Zusammenhang zueinander, typisch ist z. B. ein umgekehrt U-förmiger (parabolischer) Zusammenhang. **Komplexe Zusammenhangsmodelle** werden anstelle der multiplen Regressions- und Pfadanalyse (als klassischen Signifikanztests) heute zunehmend über Strukturgleichungsmodelle (▶ Kap. 17) überprüft.

Bei der statistischen Überprüfung von Zusammenhangshypothesen mittels Korrelations- und Regressionsanalysen als klassischen Signifikanztests sind deren Grenzen zu beachten und entsprechende Vorkehrungen zu treffen.

- Statistische Korrelations- und Regressionsanalysen sind an diverse **Voraussetzungen des Datenmaterials** geknüpft, so dass eine gründliche Datenbereinigung (z. B. Ausreißerwerte, Missings) und Voraussetzungsprüfung (Skalenniveau, Normalverteilung etc.) notwendig sind, um zu aussagekräftigen Ergebnissen zu kommen.
- Da im Zuge von Korrelationsstudien oft sehr viele Variablen verarbeitet werden, ist genau zu differenzieren, wo man tatsächlich a priori aufgestellte Hypothesen prüft und wo man nur explorativ nach neuen Zusammenhängen sucht. Forschungslogisch sowie wissenschaftsethisch unbedingt zu vermeiden ist das **Signifikanzenfischen**, bei dem explorativ gefundene Effekte als erfolgreiche Hypothesentests ausgegeben werden.
- Bei nicht-signifikanten Ergebnissen sollte überprüft werden, ob überhaupt eine ausreichende **Teststärke** vorhanden war.
- Bei signifikanten Ergebnissen sollte diskutiert werden, inwiefern die **Effektgröße** praktisch bedeutsam ist, d. h. der gefundene Korrelations-, Regressions- oder Determinationskoeffizient ist seiner Größe entsprechend als kleiner, mittlerer oder großer Effekt einzuordnen und vor allem vor dem Hintergrund des Forschungsstandes in seiner Bedeutsamkeit zu interpretieren.
- Ergänzend (oder sogar alternativ) zum klassischen statistischen Signifikanztest sollten die Effekte – d. h. die interessierenden Populationsparameter – nicht nur als Punkt-, sondern auch als Intervallschätzungen angegeben werden. **Konfidenzintervalle** von Korrelations- und Regressionskoeffizienten sowie Determinationskoeffizienten gelten allgemein als aussagekräftiger als die Betrachtung von Punktschätzungen und Irrtumswahrscheinlichkeiten. Das Berichten von Konfidenzintervallen etabliert sich in der Fachliteratur erst langsam. Nachteilig ist auch, dass viele Statistikprogramme nicht automatisch die Konfidenzintervalle ausgeben, sondern dass diese von den Forschenden separat anhand des Programm-Outputs berechnet werden müssen (für entsprechende Berechnungshinweise s. z. B. Hoyt, Imel, & Chan, 2008).
- Sollen kausale Zusammenhangshypothesen geprüft werden, so ist dies nicht allein mit statistischen Methoden zu bewerkstelligen. Für **kausale Interpretationen** von Zusammenhängen sind Aspekte der Untersuchungsplanung (z. B. aufwändige Designs wie experimentelle Untersuchungen, Längsschnittstudien sowie die Mitberücksichtigung von Moderator- und Mediatorvariablen) ebenso wichtig wie Aspekte der

Tabelle 12.14 Ausgewählte Signifikanztests für Zusammenhangshypothesen

Art der Zusammenhangshypothese	Signifikanztests (Beispiele, Kommentar)
Bivariate Zusammenhangshypothesen	Enge und teilweise auch Richtung des Zusammenhangs zwischen zwei Variablen. Diverse Korrelationskoeffizienten in Abhängigkeit vom Skalenniveau der Variablen
Bivariate Korrelation	– Produkt-Moment-Korrelation, einfache lineare Regression (zwei kardinalskalierte Variablen) – Rangkorrelation (zwei ordinalskalierte Variablen) – Kontingenzkoeffizient, einfache logistische Regression (zwei nominalskalierte Variablen)
Partielle Zusammenhangshypothesen	Enge und Richtung des Zusammenhangs zwischen zwei Variablen, aus dem der Einfluss einer oder mehrerer Kontrollvariablen eliminiert wurde
Partialkorrelation	– Partialkorrelation erster Ordnung: Bivariate Korrelation, bei der Prädiktor und Kriterium um den Einfluss von einer Kontrollvariablen bereinigt wurden – Partialkorrelation zweiter Ordnung: Bivariate Korrelation, bei der Prädiktor und Kriterium um den Einfluss von zwei Kontrollvariablen bereinigt wurden – Semipartialkorrelation: Bivariate Korrelation, bei der eine Variable (Prädiktor oder Kriterium) um den Einfluss einer Kontrollvariablen bereinigt wurde
Multivariate Zusammenhangshypothesen	Zusammenhänge zwischen mehr als zwei Variablen
Multiple Korrelation und Regression	Zusammenhang zwischen mehreren Prädiktoren (dichotom oder kardinalskaliert) und einem kardinalskalierten Kriterium
Kanonische Korrelation	Zusammenhang zwischen mehreren Prädiktoren (dichotom oder kardinalskaliert) und mehreren Kriterien (kardinalskaliert)
Konfigurationsfrequenzanalyse Log-lineare Modelle	Zusammenhang zwischen mehreren nominalskalierten Variablen (z. B. $k \times l \times j$-Kontingenztafel)
Logistische Regression	Zusammenhang zwischen mehreren Prädiktoren (dichotom oder kardinalskaliert) und einem nominalskalierten Kriterium, das entweder dichotom (binäre/binomiale logistische Regression) oder polytom (multinomiale logistische Regression) ausgeprägt sein kann
Hierarchische Zusammenhangshypothesen	
Mehrebenenanalyse/Mehrebenenregression	Zusammenhang zwischen mehreren Prädiktoren (dichotom oder kardinalskaliert) und einem kardinalskalierten Kriterium, wobei die Untersuchungseinheiten nicht voneinander unabhängig sind, sondern natürlichen Gruppen angehören und die Variablen somit auf verschiedenen Ebenen angesiedelt sind (hierarchische Datenstruktur)
Kausale Zusammenhangshypothesen	
Cross-Lagged-Panel-Design	Längsschnittliche Korrelationen zwischen mindestens zwei Variablen, die zu mindestens zwei Zeitpunkten erhoben werden
Pfadanalyse	Komplexe Modelle von Zusammenhängen zwischen Variablen
Strukturgleichungsmodelle	Komplexe Modelle von Zusammenhängen zwischen manifesten und latenten Variablen

angemessenen Datenanalyse und theoriebasierten kritischen Reflexion der Befunde, wobei alle möglichen plausiblen Kausalpfade mitzubedenken sind.

Korrelationsstudien bereiten – etwa im Vergleich zu experimentellen Gruppenvergleichen (► Abschn. 12.5.3) oder Längsschnittstudien (► Abschn. 12.5.4) – vergleichsweise wenig Untersuchungsaufwand. Die „klassische" Interdependenzanalyse ist eine einfache **Querschnittuntersuchung** („cross sectional design"), bei der man zu einem bestimmten Zeitpunkt zwei oder mehr Merkmale an einer möglichst repräsentativen Stichprobe erhebt. Die Korrelationsstudie eignet sich v. a. für Untersuchungen, bei denen man auf eine systematische Kontrolle der Untersuchungsbedingungen weitgehend verzichten muss.

Querschnittliche Interdependenzanalysen haben primär die Aufgabe, vermutete **Gemeinsamkeiten** zwischen Merkmalen statistisch abzusichern. Oft geht es nicht nur

darum, Hypothesen über Zusammenhänge in Populationen zu prüfen, sondern auch Aussagen über die zugrunde liegenden **Ursache-Wirkungs-Prinzipien** (Kausalität) zu treffen. In dieser Logik unterscheidet man im Kontext der Korrelationsanalyse zwischen unabhängigen Variablen bzw. **Prädiktoren** („predictor variables"), die als Ursachenfaktoren betrachtet werden, und abhängigen Variablen bzw. **Kriterien** („criterion variables"), an denen die Wirkungen ablesbar sind. Ein gravierender und sehr verbreiteter Interpretationsfehler im Kontext der Prüfung von Zusammenhangshypothesen besteht darin, Zusammenhänge vorschnell als Belege für bestimmte Kausalmodelle zu interpretieren und dabei alternative Kausalmodelle zu ignorieren.

Die Ergebnisse einer querschnittlichen Interdependenzanalyse lassen sich nur sehr bedingt kausal interpretieren. Da nur zu einem Zeitpunkt untersucht wird, ist die **interne Validität** durch zeitabhängige Störfaktoren (externe zeitliche Einflüsse, Reifungsprozesse, statistische Regressionseffekte und experimentelle Mortalität; ▶ Abschn. 3.2.2) nicht gefährdet. Man beachte jedoch, dass die zeitliche Generalisierbarkeit bei jeder Querschnittstudie zu problematisieren ist, was zu Lasten der **externen Validität** geht. Die interne Validität im Sinne einer Kausalaussage lässt sich erheblich steigern, wenn bivariate oder multivariate Zusammenhangsanalysen längsschnittliche Elemente aufweisen, denn Vergangenes kann niemals die Folge von Zukünftigem sein. Längsschnittlich festgestellte Korrelationen zwischen Prädiktoren und Kriterien (z. B. im Kontext eines in diesem Teilkapitel dargestellten Cross-Lagged-Panel-Designs) haben deshalb einen höheren Aussagegehalt als querschnittliche Korrelationen. Dies setzt allerdings voraus, dass Störfaktoren wie externe zeitliche Einflüsse, Reifungsprozesse der Untersuchungsteilnehmer, Testübung oder experimentelle Mortalität untersuchungstechnisch bzw. statistisch kontrolliert werden.

Die **interne Validität von Interdependenzanalysen** ist meistens gering. Sie lässt sich u. a. durch folgende Maßnahmen erhöhen:

- zeitversetzte Messungen von Prädiktor- und Kriteriumsvariablen (z. B. Betrachtung der Korrelationen im Cross-Lagged-Panel-Design),
- Neutralisierung der Wirkung von Kontroll- oder Störvariablen (z. B. Partialkorrelation),
- Berücksichtigung von Moderator- und Mediatoreffekten (z. B. Moderator- und/oder Mediatoranalyse im Rahmen multipler Regression),
- Detailanalysen von Wirkungspfaden in komplexen Kausalmodellen (z. B. Pfadanalyse, Strukturgleichungsmodelle).

Im Folgenden unterscheiden wir Untersuchungen zur Prüfung **bivariater Zusammenhangshypothesen** (Zusammenhänge zwischen jeweils zwei Variablen) und zur Prüfung **multivariater Zusammenhangshypothesen** (Zusammenhänge zwischen mehr als zwei Variablen), wobei partielle Zusammenhangshypothesen und hierarchische Zusammenhangshypothesen mitbehandelt werden. Der letzte Abschnitt diskutiert Untersuchungsvarianten, die zur Überprüfung **kausaler Zusammenhangshypothesen** entwickelt wurden.

Bivariate Zusammenhangshypothesen

Die formale Struktur einer ungerichteten bivariaten Zusammenhangshypothese lautet: H_1: Zwischen zwei Merkmalen X **(Prädiktor)** und Y **(Kriterium)** besteht ein positiver oder negativer Zusammenhang. Als ungerichtete statistische Alternativhypothese formuliert: H_1: $\rho \neq 0$. Die zugehörige Nullhypothese postuliert, dass kein Zusammenhang in der Population vorliegt bzw. dass der Korrelationskoeffizient in der Population den Wert Null hat: H_0: $\rho = 0$. Eine gerichtete Hypothese legt zusätzlich die Richtung des Zusammenhanges fest, d. h., bei einer gerichteten Alternativhypothese muss entschieden werden, ob sich die Merkmale gleichsinnig (**positiver Zusammenhang**: Je mehr des einen Merkmals, umso mehr des anderen Merkmals) oder gegensinnig verhalten (**negativer Zusammenhang**: Je mehr des einen Merkmals, umso weniger des anderen Merkmals).

Beispiel: Zwischen der Verbalisierungsfähigkeit von Schülern und der Fremdeinschätzung ihrer Intelligenz besteht ein positiver Zusammenhang. Diese gerichtete Hypothese behauptet also, dass höhere Verbalfähigkeiten mit höheren Intelligenzeinschätzungen einhergehen, dass die Merkmale also gleichsinnig kovariieren. Man beachte, dass die Richtung des Zusammenhanges bzw. das Vorzeichen des Korrelationskoeffizienten von der Operationalisierung abhängt: Ein inhaltlich positiver Zusammenhang zwischen der Intelligenz von Kindern und ihrer Schulleistung spiegelt sich in einer negativen Korrelation zwischen Intelligenztestergebnis und Durchschnittsnote wider (je höher der IQ, umso niedriger = besser die Schulnoten). Das **Vorzeichen** des Korrelationskoeffizienten gibt die Richtung des Zusammenhanges an (positiver oder negativer Zusammenhang), der **Betrag** des Korrelationskoeffizienten gibt die Enge des Zusammenhanges an (schwacher oder starker Zusammenhang). Ein **schwacher** Zusammenhang (niedriger Betrag des Korrelationskoeffizienten nahe 0) ist ein unsystematischer Zusammenhang, d. h. bei einer schwachen positiven Korrelation, gehen hohe Werte auf der einen Variablen relativ häufig mit hohen Werten der anderen Variablen einher, es gibt aber auch

viele Abweichungen. Bei einem **starken** Zusammenhang (hoher Betrag des Korrelationskoeffizienten nahe 1) ist das Muster sehr eindeutig: Hohe Werte auf der einen Variablen gehen fast durchgängig mit entsprechend hohen Werten der anderen Variablen einher.

Die Enge und Richtung des Zusammenhanges wird mit einem **Korrelationskoeffizienten** quantifiziert. Der empirisch berechnete Korrelationskoeffizient in der Stichprobe (deskriptiv-statistisches Ergebnis) wird auf **Signifikanz** geprüft, wenn man Hypothesen über den unbekannten Korrelationskoeffizienten in der Population prüfen möchte (inferenzstatistisches Ergebnis der Hypothesenprüfung).

Datenerhebung

Merkmalszusammenhänge können nur berechnet werden, wenn für alle interessierenden Merkmale jeweils Messwerte vorliegen, die einander zugeordnet sind. Im bivariaten Fall werden also die Variablen X und Y in einer Weise erhoben, die sinnvolle Messwertpaare erzeugt. Dies geschieht oft, indem bei allen Untersuchungsobjekten jeweils beide interessierenden Merkmale erhoben werden (z. B. bei jedem Arbeitnehmer in der Stichprobe werden sowohl der Stress am Arbeitsplatz als auch die Fehlzeiten erfasst). Die Berechnung einer Korrelation setzt allerdings nicht voraus, dass jedem Untersuchungsobjekt zwei Messwerte zugeordnet sind. Entscheidend ist, dass einem Messwert ein anderer Messwert eindeutig zugeordnet ist. Diese Forderung wäre auch erfüllt, wenn z. B. Zusammenhänge zwischen den Neurotizismuswerten von Eheleuten, zwischen dem Körpergewicht von Hundebesitzern und dem Gewicht ihrer Hunde, dem Einkommen von Autobesitzern und der PS-Zahl ihrer Autos etc. untersucht werden.

Die hier skizzierte Vorgehensweise bereitet Probleme, wenn ein Merkmal an Paaren erhoben wurde, deren Paarlinge austauschbar sind. Man denke hierbei etwa an die Überprüfung des Zusammenhanges der Intelligenz zweieiiger Zwillinge, bei der nicht entschieden werden kann, welcher Zwilling zum „Merkmal X“ und welcher zum „Merkmal Y“ gehört. Für Fragestellungen dieser Art verwendet man statt des üblichen Korrelationskoeffizienten den Intraklassen-Tau-Koeffizienten (zum Stichwort „Zwillingskorrelation“ s. Bortz & Lienert, 2003, Abschn. 5.2.6) oder den in ▶ Abschn. 10.1.3 „Beobachterübereinstimmung“ behandelten Intra-Klassen-Korrelationskoeffizienten.

Diese Beispiele verdeutlichen für Untersuchungen zur Überprüfung von Zusammenhangshypothesen folgende Leitlinien:

- Es muss zweifelsfrei geklärt sein, welche Messwerte der untersuchten Variablen Messwertpaare bilden.
- Soll der Zusammenhang zweier Merkmale für eine Stichprobe von Untersuchungsobjekten bestimmt werden, darf pro Untersuchungsobjekt nur ein Messwertpaar in die Korrelationsberechnung eingehen.
- Interessiert der Zusammenhang zwischen zwei Stichproben in Bezug auf ein Merkmal, müssen die Untersuchungsobjekte beider Stichproben Paare bilden, deren Paarlinge nicht austauschbar sein dürfen (z. B. Korrelation zwischen Merkmal des Therapeuten und Merkmal des Patienten).
- Sind die Paarlinge der aus zwei Stichproben gebildeten Paare prinzipiell austauschbar („Zwillingskorrelation“), ist die Zusammenhangshypothese über den Intra-Klassen-Korrelationskoeffizienten zu prüfen.

Bivariate Korrelationen

Das Skalenniveau (▶ Abschn. 8.4) der in einer Zusammenhangshypothese genannten Merkmale bestimmt die Korrelationsart, mit der die Hypothese statistisch geprüft wird. In ◘ Tab. 12.15 wird dargestellt, welche Assoziationsmaße bzw. Korrelationsmaße welchen Skalenkombinationen zugeordnet sind. Für manche Skalenkombinationen existieren mehrere verschiedene Assoziationsmaße, die den Zusammenhang zwischen zwei Variablen jeweils etwas unterschiedlich quantifizieren. Weitere spezielle Korrelationstechniken findet man z. B. bei Benninghaus (1989, 1998) oder bei Kubinger (1990). Zur Berechnung der in ◘ Tab. 12.15 genannten und weiteren Korrelationsmaßen wird auf Bortz, Lienert, & Boehnke (2000) bzw. Bortz und Schuster (2010, Kap. 10) sowie Eid, Gollwitzer und Schmitt (2010, Kap. 15) verwiesen. Man beachte, dass die Berechnung eines Korrelationskoeffizienten aus den Stichprobendaten zunächst nur ein deskriptivstatistisches Ergebnis darstellt. Um Hypothesen über **Korrelationskoeffizienten** in der Population zu prüfen (z. B. die gerichtete Forschungshypothese H_1: $\rho > 0$ oder die ungerichtete Forschungshypothese H_1: $\rho \neq 0$), muss der jeweilige Korrelationskoeffizient einem **Signifikanztest** unterzogen werden. Die Berechnung der linearen bzw. Produkt-Moment-Korrelation, die kardinalskalierte Variablen voraussetzt, wird oft begleitet durch eine entsprechende **lineare Regressionsanalyse**, mit deren Hilfe auf der Basis der Prädiktorwerte X die Kriteriumswerte Y vorhergesagt werden können (Bortz & Schuster, 2010, Kap. 11; Eid et al., 2010, Kap. 16).

Bivariate Korrelation – Die bivariate Korrelation („bivariate correlation“) bestimmt über einen Korrelationskoeffizienten die Enge des Zusammenhangs (schwacher oder starker Zusammenhang) sowie die Richtung des Zusammenhangs (positiver oder negativer Zusammenhang) zwischen zwei Merkmalen. Für Variablen unterschiedlichen Skalenniveaus existieren verschiedene bivariate Assoziationsmaße. Der bivariate Korrelationskoeffizient hat einen Wertebereich von −1 (perfek-

Tabelle 12.15 Übersicht bivariater Assoziationsmaße. (Mod. nach Eid et al., 2010, S. 539. Mit freundlicher Genehmigung von Beltz)

Merkmal *y*	Merkmal *x*					
	Kardinalskala	Ordinalskala singuläre Variable	Ordinalskala Rangklassen	Künstliche Dichotomie	Natürliche Dichotomie	Nominalskala
Kardinalskala (Intervall- oder Verhältnisskala)	Produkt-Moment-Korrelation	Spearmans Rangkorrelation ρ [rho] Kendalls τ [tau] Wilsons *e*	Polyseriale Korrelation	Biseriale Korrelation	Punktbiseriale Korrelation	Kontingenzkoeffizient *C* Koeffizient η [eta]
Ordinalskala singuläre Variable	–	Spearmans Rangkorrelation ρ [rho] Kendalls τ [tau] Wilsons *e*	Somers d_{YX} Kims $d_{Y.X}$	Biseriale Rangkorrelation Somers d_{YX} Kims $d_{Y.X}$	Biseriale Rangkorrelation Somers d_{YX} Kims $d_{Y.X}$	(Aufteilung der singulären Variable in Rangklassen)
Ordinalskala Rangklassen	–	–	γ-Koeffizient [gamma]	γ-Koeffizient	γ-Koeffizient	Kontingenzkoeffizient *C* Cramérs *V*
Künstliche Dichotomie	–	–	–	φ-Koeffizient [phi] Yules *Q* Tetrachorische Korrelation	φ-Koeffizient Yules *Q* Korrelationskoeffizient ν [ny]	Kontingenzkoeffizient *C* Cramérs *V*
Natürliche Dichotomie	–	–	–	–	φ-Koeffizient Yules *Q*	Kontingenzkoeffizient *C* Cramérs *V*
Nominalskala	–	–	–	–	–	Kontingenzkoeffizient *C* Cramérs *V*

ter negativer = gegensinniger Zusammenhang) bis +1 (perfekter positiver = gleichsinniger Zusammenhang). Bei Werten um Null existiert kein systematischer Zusammenhang.

Dichotome (binäre) Merkmale sind zweifach gestufte Merkmale. Wir sprechen von einer **künstlichen Dichotomie**, wenn ein eigentlich kontinuierlich verteiltes Merkmal auf zwei Stufen reduziert wird (z. B. Prüfungsleistung: bestanden – nicht bestanden) und von einer **natürlichen Dichotomie**, wenn das Merkmal tatsächlich nur zwei Ausprägungen hat (Schwangerschaft: schwanger – nicht schwanger).

Bei ordinalskalierten Merkmalen ist zu differenzieren, ob es sich um **singuläre Variablen** handelt, bei denen jeder Untersuchungseinheit ein separater ordinaler Messwert zugeordnet wird (z. B. Rangreihe des Eintreffens beim Wettlauf), oder ob es sich bei den ordinalen Messwerten um **Rangklassen** handelt (z. B. Einordnung von Personen als Nichtraucher, Gelegenheitsraucher, regelmäßiger Raucher, Kettenraucher; Eid et al., 2010, S. 107; zur Ordinalskala ▶ Abschn. 8.4.3). Zudem ist bei Assoziationsmaßen für Ordinaldaten zu beachten, wie **Rangbindungen** („ties“) jeweils zu behandeln sind (manche Rangkorrelationskoeffizienten sind nur auf Daten ohne Rangbindungen anwendbar).

Nicht jeder Skalenkombination ist eine eigene Korrelationsart zugeordnet. Existiert bei Merkmalen mit unterschiedlichem Skalenniveau kein spezielles Korrelationsmaß, wird das Merkmal mit dem höheren Skalenniveau auf das Skalenniveau des Vergleichsmerkmals **transformiert**. Generell lässt sich ein höheres Skalenniveau immer auf ein niedrigeres heruntertransformieren, allerdings geht dies mit Informationsverlust einher. Will man z. B. das Alter von Untersuchungspersonen (kardinalskaliert) mit ihren Farbpräferenzen (nominalskaliert) in Beziehung setzen, ist es erforderlich, das eigentlich kontinuierlich verteilte Altersmerkmal auf einige wenige Alterskategorien zu reduzieren (zur Kategorienbildung bei kontinuierlichen Merkmalen ▶ Abschn. 8.4.6). Der Kontingenzkoeffizient *C* sowie Cramérs *V* (Bortz & Lienert, 2003, Abschn. 5.1.3) behandeln dann beide Merkmale wie Nominalskalen. Es besteht ferner die Möglichkeit, diese Fragestellung als eine multivariate Zusammenhangshypothese bzw. als eine Unterschiedshypothese (▶ Abschn. 12.5.3) aufzufassen und zu überprüfen.

Für die Kombination „natürliche Dichotomie“ mit einem künstlich dichotomen Merkmal haben Ulrich und Wirtz (2004) einen **Korrelationskoeffizienten *ν*** (griechisches n; sprich: ny) vorgeschlagen, der dem in Tab. 12.15 aufgeführten Phi-Koeffizienten in verschiedener

Hinsicht überlegen ist. Ein besonderes Merkmal dieses Korrelationskoeffizienten ist darin zu sehen, dass er – anders als Phi – unabhängig ist vom Cut-Off-Point, der das kontinuierliche Merkmal auf eine künstliche Dichotomie reduziert. Hinweise dazu, unter welchen Bedingungen der Phi-Koeffizient oder **Yules *Q*** (Odds-Ratio, der auf den Wertebereich −1 bis +1 normiert ist) als Zusammenhangsmaße für dichotome Merkmale zu verwenden sind, finden sich bei Eid et al. (2010, S. 525).

Üblicherweise interessieren wir uns bei kardinalskalierten Merkmalen für die Enge des **linearen** Zusammenhanges. Es lassen sich jedoch auch **nicht-lineare Zusammenhänge** mit Verfahren quantifizieren, die z. B. Draper und Smith (1998, Kap. 24), von Eye und Schuster (1998, Kap. 7, Kap. 9) oder Lehmann (1980) beschreiben. Allerdings ist hierbei darauf zu achten, dass die Zusammenhangshypothese die Art des nichtlinearen Zusammenhanges (exponentiell, logarithmisch etc.) spezifiziert.

Merkmalsprofile

Nach unserem bisherigen Verständnis betreffen Zusammenhangshypothesen typischerweise die Beziehung zweier Merkmale. Es lassen sich jedoch auch Zusammenhänge (oder besser: Ähnlichkeiten) zweier oder mehrerer Personen (Untersuchungsobjekte) analysieren. Hierbei geht man davon aus, dass für jede der zu vergleichenden Personen bezüglich mehrerer Variablen Messungen vorliegen, die zusammengenommen individuelle Merkmalsprofile ergeben. Maße der Ähnlichkeit von Merkmalsprofilen werden z. B. bei Schlosser (1976) oder allgemein in der Literatur zur Clusteranalyse behandelt. So wird z. B. zur Prüfung der Übereinstimmung zwischen zwei Profilen im semantischen Differenzial (▶ Abschn. 8.6.4) die sog. Q-Korrelation eingesetzt (vgl. Schäfer, 1983).

Stichprobenfehler

Bei der korrelationsstatistischen Überprüfung von Zusammenhangshypothesen ist darauf zu achten, dass die Stichprobe tatsächlich die gesamte Population, für die das Untersuchungsergebnis gelten soll, repräsentiert. Zu welchen Verzerrungen der Zusammenhangsschätzung es bei Stichprobenfehlern kommen kann, zeigt Stelzl (1982) anhand einiger Beispiele, die in Abb. 12.26 wiedergegeben sind.

Es geht um die Überprüfung des Zusammenhanges zwischen schulischer Leistung (Y) und Intelligenz (X), wobei beide Merkmale kardinalskaliert operationalisiert seien. Für die Population aller Schüler möge eine Produkt-Moment-Korrelation von $\rho = .71$ zutreffend sein (die Schulleistung ist hier nicht über Noten, sondern über Punkte operationalisiert, so dass ein positiver Korrelationskoeffizient auftritt). In Abb. 12.26a wird gezeigt, wie sich der Zusammenhang dieser Merkmale ändert bzw. verkleinert, wenn nur Schüler mit einem IQ über 85 bzw. unter 85 untersucht werden. Sie beträgt im ersten Fall (IQ $>$ 85) $r = .63$ und im zweiten Fall (IQ $<$ 85) $r = .42$.

Ähnliches gilt für die in Abb. 12.26b vorgenommene Selektion. Hier beträgt die Korrelation in der Teilstichprobe IQ $<$ 115 $r = .59$ und in Teilstichprobe IQ $>$ 115 $r = .48$. Eine **kleine homogene Teilstichprobe**, die nicht die gesamte Merkmalsvariation in der Population abbildet, kann also den wahren Zusammenhang unterschätzen.

Diesem zur Unterschätzung des Gesamtzusammenhanges führenden Stichprobenfehler steht ein anderer gegenüber, der eine Überschätzung des Gesamtzusammenhanges bedingt: die **Extremgruppenselektion**. Durch Weglassen von Schülern mit mittleren Intelligenzquotienten erhöht sich die Korrelation auf $r = .81$ (Abb. 12.26c). Verzichtet man zusätzlich auf die Einbeziehung von Schülern mit durchschnittlichen Schulleistungen, erhöht sich der Zusammenhang weiter auf $r = .91$ (Abb. 12.26d). Über die statistischen Probleme, die durch Extremgruppenselektion entstehen, berichten ausführlich Alf und Abrahams (1975) sowie Preacher, Rucker, MacCullum, & Nicewander (2005).

Stelzl (1982, Abschn. 5.2) erläutert **Artefakte** bei der Überprüfung von Zusammenhangshypothesen, die durch mathematische Abhängigkeit der untersuchten Merkmale entstehen (Beispiel: Wenn $X + Y =$ konstant ist, resultiert zwangsläufig zwischen X und Y eine negative Korrelation).

Weiterhin ist zu beachten, dass Korrelationskoeffizienten durch einzelne **Ausreißerwerte** („outliers“; Extremwerte) in der Stichprobe stark verzerrt werden können. Vor der inferenzstatistischen Prüfung von Zusammenhangshypothesen sollte also eine Dateninspektion erfolgen, um Ausreißer zu erkennen und sachgerecht zu behandeln. Ausreißerwerte sind v. a. dann ein Problem, wenn der Wertebereich einer Variablen nicht – wie z. B. bei Ratingskalen – beschränkt ist (z. B. offene Erfassung des Einkommens, der Freundeszahl, der Telefonate in der letzten Woche etc.).

Für die Verallgemeinerung einer Korrelation auf eine Grundgesamtheit ist zu fordern, dass die untersuchte Stichprobe tatsächlich zufällig gezogen wurde und keine absichtliche oder unabsichtliche systematische Selektion darstellt. So kann Extremgruppenselektion z. B. zu einer dramatischen Überschätzung von Korrelationen führen.

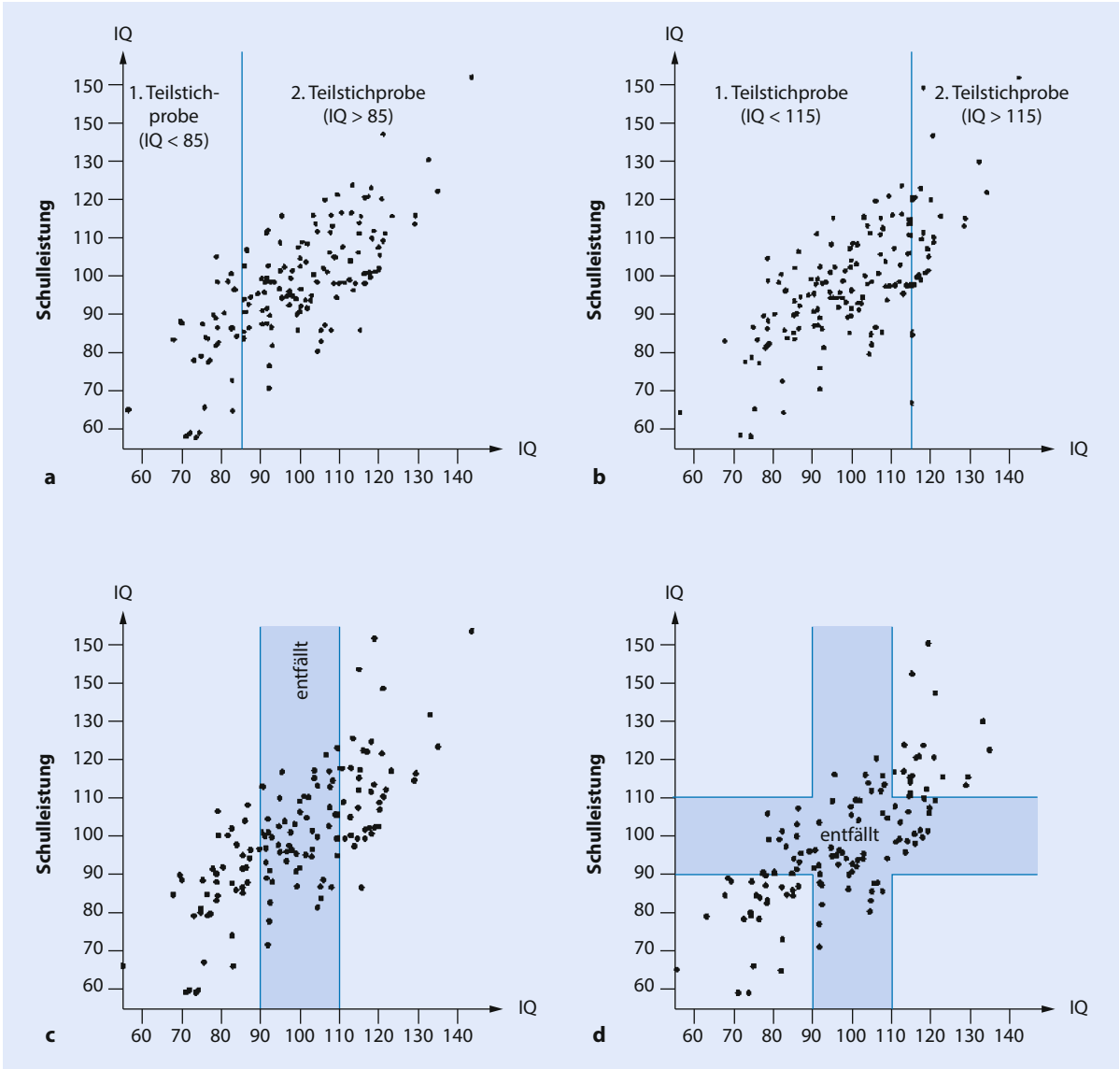


Abb. 12.26 **a–d** Verzerrung von Korrelationen durch Stichprobenselektion. **a** Zwei nach dem IQ selegierte Teilstichproben mit einem Grenzwert des IQs von 85, **b** zwei nach dem IQ selegierte Teilstichproben mit einem Grenzwert des IQs von 115, **c** Extremgruppenselektion nach dem IQ, **d** Extremgruppenselektion nach der Schulleistung und dem IQ

Multivariate Zusammenhangshypothesen

Sind mehr als zwei Variablen beteiligt, so sprechen wir von multivariaten Zusammenhangshypothesen. Hier sind v. a. drei Konstellationen zu unterscheiden:

1. Bei der **Partialkorrelation** wird der bivariate Zusammenhang zwischen zwei Merkmalen um den Einfluss einer oder mehrerer Drittvariablen Z ($Z_1, Z_2, \ldots$) bereinigt.
2. Bei der **multiplen Korrelation** wird der Zusammenhang zwischen mehreren Prädiktorvariablen X ($X_1, X_2, \ldots$) und einer Kriteriumsvariable Y bestimmt.
3. Bei der **kanonischen Korrelation** wird der Zusammenhang zwischen mehreren Prädiktorvariablen X ($X_1, X_2, \ldots$) und mehreren Kriteriumsvariablen Y ($Y_1, Y_2, \ldots$) berechnet.

Bei der statistischen Überprüfung von Hypothesen zu partiellen Zusammenhängen, multiplen Zusammenhängen und kanonischen Zusammenhängen geht es jeweils um die Berechnung des entsprechenden Korrelationskoeffizienten (Stichprobenkennwert), der dann im Zuge eines Signifikanztests auf Überzufälligkeit geprüft wird. Die Forschungshypothese lautet dabei, dass ein entsprechender Zusammenhang in der Population vorliegt (H_1: $\rho \neq 0$), während die H_0 diesen Effekt negiert (H_0: $\rho = 0$). Man beachte, dass bei einem multiplen oder

kanonischen Zusammenhang der Korrelationskoeffizient nur die Stärke des Zusammenhanges (0: kein Zusammenhang, 1: perfekter Zusammenhang), nicht jedoch die Richtung des Zusammenhanges angeben kann. Denn bei mehreren Prädiktoren und/oder mehreren Kriterien können einige in positiver und andere in negativer Richtung kovariieren. Liegt ein überzufälliger multipler oder kanonischer Gesamtzusammenhang vor, so wird man im Zuge der statistischen Datenanalyse noch prüfen, welche Prädiktoren bzw. Kriterien wie stark und in welcher Richtung überzufällig an dem Gesamtzusammenhang beteiligt sind. Eine entsprechende Hypothesenprüfung für die einzelnen kardinalskalierten Variablen eines multivariaten Zusammenhanges erfolgt bei multiplen Zusammenhängen über die multiple lineare Regressionsanalyse: Jeder Prädiktor erhält ein standardisiertes Beta-Gewicht mit dem Wertebereich -1 bis $+1$, das Stärke und Richtung des Beitrags des Prädiktors ausdrückt und auf Signifikanz geprüft wird (Bortz & Schuster, 2010, Kap. 21; Eid et al., 2010, Kap. 18). Bei einem kanonischen Zusammenhang werden die Einzelbeiträge der Prädiktoren und Kriterien über Faktorladungen bestimmt (Bortz & Schuster, 2010, Kap. 28).

12

▪ Partielle Zusammenhänge

Partialkorrelationen sind eigentlich Verfahren zur Überprüfung bivariater Zusammenhangshypothesen. Dass wir sie dennoch unter der Rubrik „Multivariate Zusammenhangshypothesen" behandeln, wird damit begründet, dass diese Verfahren die Beziehungen mehrerer Merkmale simultan berücksichtigen. Zielsetzung der Partialkorrelation ist es zu prüfen, inwiefern ein bivariater Zusammenhang zwischen X und Y statistisch durch eine oder mehrere Drittvariablen Z beeinflusst wird. Dazu werden die interessierenden Variablen X und Y um den Einfluss von Z statistisch bereinigt, bzw. aus dem bivariaten Zusammenhang wird der Einfluss von Z herauspartialisiert (für ein ▶ Studienbeispiel „Partialkorrelation: Zusammenhang zwischen Regierungsqualität und Zufriedenheit der Bevölkerung unter Kontrolle des Wohlstandes").

Der Nachweis eines statistisch gesicherten Zusammenhanges zweier Merkmale X und Y verlangt Überlegungen, wie dieser Zusammenhang zu erklären ist. Hierfür bietet sich häufig eine dritte Variable Z an, von der sowohl X als auch Y abhängen. Besteht zwischen X und Z sowie zwischen Y und Z jeweils ein enger Zusammenhang (was nicht bedeuten muss, dass Z die Merkmale X und Y kausal beeinflusst), erwarten wir zwangsläufig auch zwischen X und Y einen engen Zusammenhang. Wir könnten nun danach fragen, wie eng der Zusammenhang zwischen X und Y wäre, wenn wir die Gemeinsamkeiten des Merkmals Z mit den Merkmalen X und Y außer Acht lassen. Der Erklärungswert dieses Ansatzes wird anschaulicher, wenn sich die Annahme, Z beeinflusse X und Y kausal, inhaltlich begründen lässt. Wir fragen dann, wie X und Y zusammenhängen, wenn der Einfluss von Z auf X und Y ausgeschaltet wird.

Sollte es so sein, dass der Zusammenhang zwischen X und Y durch das Ausschalten von Z verschwindet, müsste die Korrelation zwischen X und Y als **Scheinkorrelation** („spurious correlation") bezeichnet werden. Dies ist eine Korrelation, die einen direkten Zusammenhang zwischen X und Y lediglich „vortäuscht", die jedoch bedeutungslos wird, wenn man eine dritte Variable Z beachtet. Beispiel: Die positive Korrelation zwischen Schuhgröße (X) und Lesbarkeit der Handschrift (Y) von Kindern verschwindet, wenn man das Alter (Z) der Kinder kontrolliert. Oder: Heuschnupfen (X) und Weizenpreis (Y) korrelieren negativ, weil gute Ernten mit vielen Weizenpollenallergien und niedrigen Weizenpreisen einhergehen und schlechte Ernten mit wenig Allergien, aber guten Preisen verbunden sind. Drittvariable ist hier also das Wetter (Z) (weitere Beispiele für Scheinkorrelationen findet man bei Beck-Bornholdt & Dubben, 2001, S. 142ff.; Krämer, 1995, Kap. 14).

Mit diesen Überlegungen erweitern wir eine einfache bivariate Zusammenhangshypothese zu einer partiellen, bivariaten Zusammenhangshypothese: Zwischen zwei Merkmalen X und Y besteht auch dann ein Zusammenhang, wenn der „Einfluss" einer dritten Variablen Z ausgeschaltet wird. (Wir setzen das Wort „Einfluss" in Anführungszeichen, weil die Beziehung zwischen Z und den Merkmalen X und Y nicht kausal – wie die Bezeichnung „Einfluss" suggeriert – sein muss.) Ein Beispiel: Zwischen der Produktivität von Gruppen und ihrer Kommunikationsintensität besteht auch dann ein Zusammenhang, wenn man den Einfluss der Gruppengröße ausschaltet.

Die Überprüfung dieser erweiterten bivariaten Zusammenhangshypothese erfolgt mit der **Partialkorrelation** (Bortz & Schuster, 2010, Kap. 21; Eid et al., 2010, Kap. 17). Sie lässt sich berechnen, wenn von einer Stichprobe Messungen auf allen drei Variablen vorliegen und setzt in der Regel voraus, dass alle drei untersuchten Merkmale kardinalskaliert sind. (Über spezielle Verfahren zur Überprüfung partieller Zusammenhänge bei nicht intervallskalierten Merkmalen berichten z. B. Bortz, Lienert, & Boehnke, 2008, Abschn. 8.2.4.) Will man eine polytom-nominalskalierte Variable Z kontrollieren, muss diese zuvor in dichotome Indikatorvariablen umkodiert werden; ▶ Kodierung eines polytom-nominalen Merkmals durch dichotome Indikatorvariablen. Man beachte, dass das „Ausschalten" einer Kontrollvariablen statistisch erfolgt und nicht untersuchungstechnisch (wie z. B. durch das Konstanthalten der Kontrollvariablen). Den Vorgang der „Bereinigung" der Merkmale X und Y um diejenigen

Studienbeispiel

Partialkorrelation: Zusammenhang zwischen Regierungsqualität und Zufriedenheit der Bevölkerung unter Kontrolle des Wohlstandes

Eine politikwissenschaftliche Sekundäranalyse untersuchte für 130 Länder, über die entsprechende Daten vorlagen (Gelegenheitsstichprobe aus der Population der 194 Staaten dieser Welt), ob ein positiver Zusammenhang zwischen Regierungsqualität einerseits und Lebenszufriedenheit der Bevölkerung andererseits vorliegt H_1: $\rho > 0$ (Ott, 2011). Prädiktor- und Kriteriumsvariable wurden kardinalskaliert operationalisiert: Die subjektive Lebenszufriedenheit wurde als Einzelindikator mit einer Ratingskala (von 0: maximal unzufrieden bis 10: maximal zufrieden) gemessen und die Regierungsqualität über einen Index zur sog. technischen Regierungsqualität (z-standardisierte Werte von -2: schlechte Qualität bis $+2$: gute Qualität). In den Index flossen als Indikatoren z. B. Effizienz der öffentlichen Verwaltung und Grad der Korruptionskontrolle ein (die Erfassung der Indikatoren erfolgte im Auftrag der Weltbank durch jeweils 60 unabhängige Expertenurteile).

Es zeigte sich ein sehr enger positiver linearer **bivariater** Zusammenhang von $\boldsymbol{r = +.75}$, der statistisch signifikant ist ($t_{emp} = 12.79$, $p < .001$, zur Berechnung des Signifikanztests des bivariaten Korrelationskoeffizienten über die t-Verteilung s. Bortz & Schuster, 2010, S. 163). Wir können also die Alternativhypothese annehmen, dass bessere Regierungsqualität systematisch mit höherer Lebenszufriedenheit der Bevölkerung einhergeht bzw. dass geringere Regierungsqualität mit geringerer Zufriedenheit kovariiert. Der Scatterplot (Streudiagramm) mit den entsprechenden Messwertpaaren und der Regressionsgeraden ist Abb. 12.27 zu entnehmen. Der zum bivariaten Korrelationskoeffizienten $r = +.75$ gehörende Determinationskoeffizient $\boldsymbol{r^2 = 56\,\%}$ zeigt, dass mehr als die Hälfte der Unterschiedlichkeit der Zufriedenheitswerte auf Unterschiede in der technischen Regierungsqualität zurückzuführen sind.

Man kann sich nun fragen, ob dieser bivariate Zusammenhang möglicherweise nur eine **Scheinkorrelation** darstellt, die in Wirklichkeit durch eine Drittvariable zustande kommt. Inhaltlich könnte man vermuten, dass der Wohlstand eines Landes (operationalisiert über das durchschnittliche Pro-Kopf-Einkommen) sowohl dazu führt, dass ein besser funktionierendes Politiksystem existiert als auch, dass die Bevölkerung zufriedener ist. Zu berechnen ist somit die **Partialkorrelation** (partielle Korrelation), also der um den Einfluss des Wohlstandes bereinigte Zusammenhang zwischen technischer Regierungsqualität und Zufriedenheit. Im Vergleich zu der sehr großen bivariaten Korrelation $r_{xy} = +.75$ ($r^2 = 56\,\%$) sank die Partialkorrelation auf $\boldsymbol{r_{xy.z} = +.14}$ ($r^2 = 2\,\%$) deutlich ab, und verfehlte bei einseitigem Test knapp das 5 %-Niveau ($z = 1.58$, $p = .057$; zur Berechnung des Signifikanztests des partiellen Korrelationskoeffizienten über die Standardnormalverteilung siehe Bortz & Schuster, 2010, S. 341).

Im Kontext der Studie (Ott, 2011) wurden neben der technischen Regierungsqualität auch andere Aspekte der Regierungsqualität mit der Zufriedenheit der Bevölkerung korreliert. Das Ergebnismuster deutete im Gesamtbild auf einen positiven Zusammenhang zwischen Regierungsqualität und Zufriedenheit hin.

Wie lässt sich dieser Zusammenhang nun kausal interpretieren? Der Autor (Ott, 2011) diskutiert drei verschiedene **kausale Interpretationsmöglichkeiten**:

1. **Kein kausaler** Zusammenhang: Der Zusammenhang stellt nur eine durch Drittvariablen erzeugte Scheinkorrelation dar. Dieses Argument wird als unplausibel zurückgewiesen, da sich nach Herauspartialisieren von Drittvariablen wie Wohlstand noch immer ein positiver Zusammenhang an der Signifikanzgrenze zeigt.
2. **Zufriedenheit verursacht Regierungsqualität**: Zufriedene Bürgerinnen und Bürger verhalten sich staatstreuer, engagieren sich stärker politisch und tragen somit zur positiven Regierungsqualität bei. Diese Möglichkeit hält der Autor für plausibel.
3. **Regierungsqualität verursacht Zufriedenheit**: Als entscheidenden Verursachungszusammenhang betrachtet der Autor den Einfluss der Regierungsqualität auf die Zufriedenheit, wobei er direkte und indirekte Effekte der Regierung auf die Zufriedenheit der Bevölkerung differenziert:
   a) **Direkter Kausaleffekt**: Es stimmt die Bevölkerung zufriedener, wenn sie gut regiert wird.
   b) **Indirekter Kausaleffekt**: Eine hohe Regierungsqualität fördert Lebensbedingungen (z. B. Wohlstand, gutes Gesundheitssystem, Geschlechtergleichberechtigung), die Zufriedenheit in der Bevölkerung steigern.

Wann immer bivariate, partielle oder multiple Zusammenhänge kardinalskalierter Variablen analysiert werden, wird zunächst vom einfachsten Zusammenhangsmodell – dem **linearen** Zusammenhang ausgegangen (Abb. 12.27). Eine gründliche grafische Inspektion des Datenmaterials sowie daran anschließende statistische Berechnungen können jedoch auch **nicht-lineare Zusammenhänge** nahelegen.

So zeigte die hier vorgestellte Studie einen linearen Zusammenhang zwischen technischer Regierungsqualität und der Intensität der Zufriedenheit, gleichzeitig aber auch einen nicht-linearen **quadratischen** Zusammenhang zwischen technischer Regierungsqualität und Heterogenität der Zufriedenheit in der Bevölkerung (operationalisiert über die Standardabweichung SD der subjektiven Zufriedenheitswerte der Bevölkerung eines Landes): Bei sehr geringer Regierungsqualität ist auch die Heterogenität der Zufriedenheit gering: Die Bevölkerungsmehrheit ist relativ einheitlich unzufrieden. Wächst die Regierungsqualität, so steigt die Zufriedenheit, aber in unterschiedlicher Weise in unterschiedlichen Bevölkerungsgruppen: Bei mittlerer Regierungsqualität herrscht also größere Ungleichheit in der Zufriedenheit. Steigt die Regierungsqualität noch weiter an, so werden benachteiligte und unzufriedene Bevölkerungsgruppen besser integriert und unterstützt, so dass sich die Zufriedenheitswerte auf höherem und einheitlicherem Niveau befinden (Abb. 12.28 für die Punktewolke der Messwertpaare und die Parabel der quadratischen Regressionsgleichung). Würde man für die Daten fälschlich eine lineare Korrelation berechnen, so würde $r = -.18$ ($r^2 = 3\,\%$) resultieren, während der quadratische Zusammenhang in $\boldsymbol{r = +.45}$ ($\boldsymbol{r^2 = 20\,\%}$, $p < .05$) resultiert. Der quadratische Zusammenhang lässt sich aufteilen in zwei lineare Zusammenhänge: Für Länder mit geringer Regierungsqualität ($z < 0$), ist der lineare Zusammenhang positiv ($r = +.29$), für Länder mit hoher Regierungsqualität ist der lineare Zusammenhang negativ ($r = -.64$; Ott, 2011, S. 11).

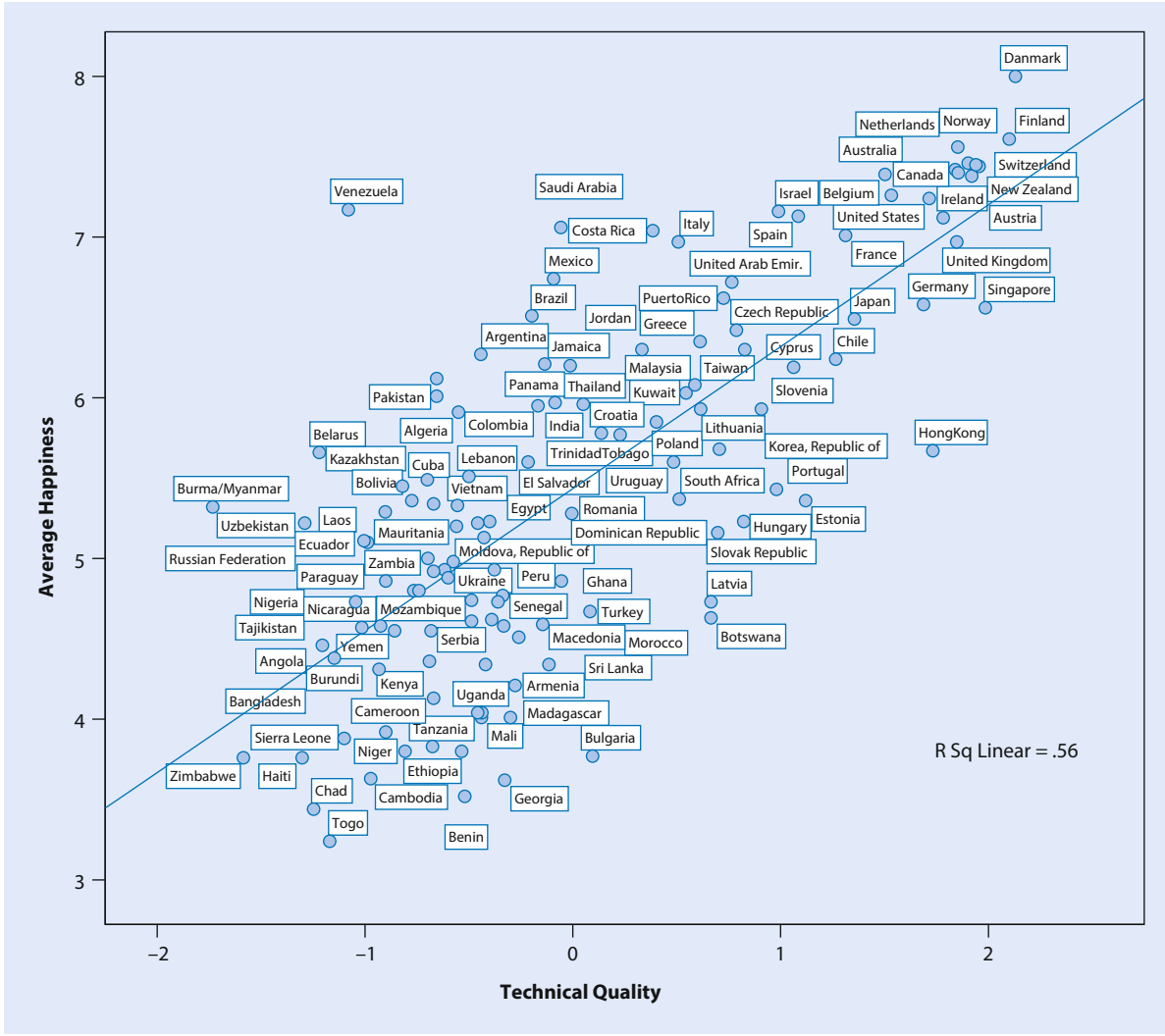


Abb. 12.27 Technische Regierungsqualität und durchschnittliche Lebenszufriedenheit im Jahr 2006. (Aus Ott, 2011, S. 10)

Anteile, die auf eine Kontrollvariable Z zurückgehen, bezeichnet man als **Herauspartialisieren** von Z.

Partialkorrelation – Die Partialkorrelation bestimmt über einen Partialkorrelationskoeffizienten („partial correlation coeffizient") die Enge und Richtung des Zusammenhang zwischen zwei Merkmalen, wobei der Einfluss von einer oder mehreren Drittvariablen kontrolliert (herauspartialisiert) wird.

Manchmal lassen sich nicht nur eine Kontrollvariable Z, sondern mehrere Kontrollvariablen Z_1, $Z_2 \ldots Z_p$ benennen, von denen man annimmt, sie üben auf den Zusammenhang von X und Y einen „Einfluss" aus. Die Hypothese „ Zwischen zwei Merkmalen X und Y besteht ein Zusammenhang, auch wenn der Einfluss mehrerer Kontrollvariablen außer Acht bleibt" wird mit **Partialkorrelationen höherer Ordnung** überprüft. Beispiel: Zwischen Zigarettenkonsum (X) und Krebsrisiko (Y) besteht auch dann ein Zusammenhang, wenn man den „Einfluss" von Staub am Arbeitsplatz (Z_1) und sportlichen Aktivitäten (Z_2) eliminiert.

Ob der Zusammenhang zweier Merkmale X und Y „echt" ist oder durch ein Drittmerkmal Z erklärt werden kann (Scheinkorrelation), erfährt man über die Partialkorrelation.

Das Konzept der Bereinigung von Merkmalen lässt sich in vielfältiger Weise nutzen. Es gestattet z. B. auch, Hypothesen zu überprüfen, die behaupten, dass zwischen zwei Merkmalen X und Y ein Zusammenhang besteht, wenn ein Kontrollmerkmal Z nur aus einer der beiden Variablen herauspartialisiert wird. Eine solche Hypothese könnte etwa besagen, dass zwischen den Merkmalen

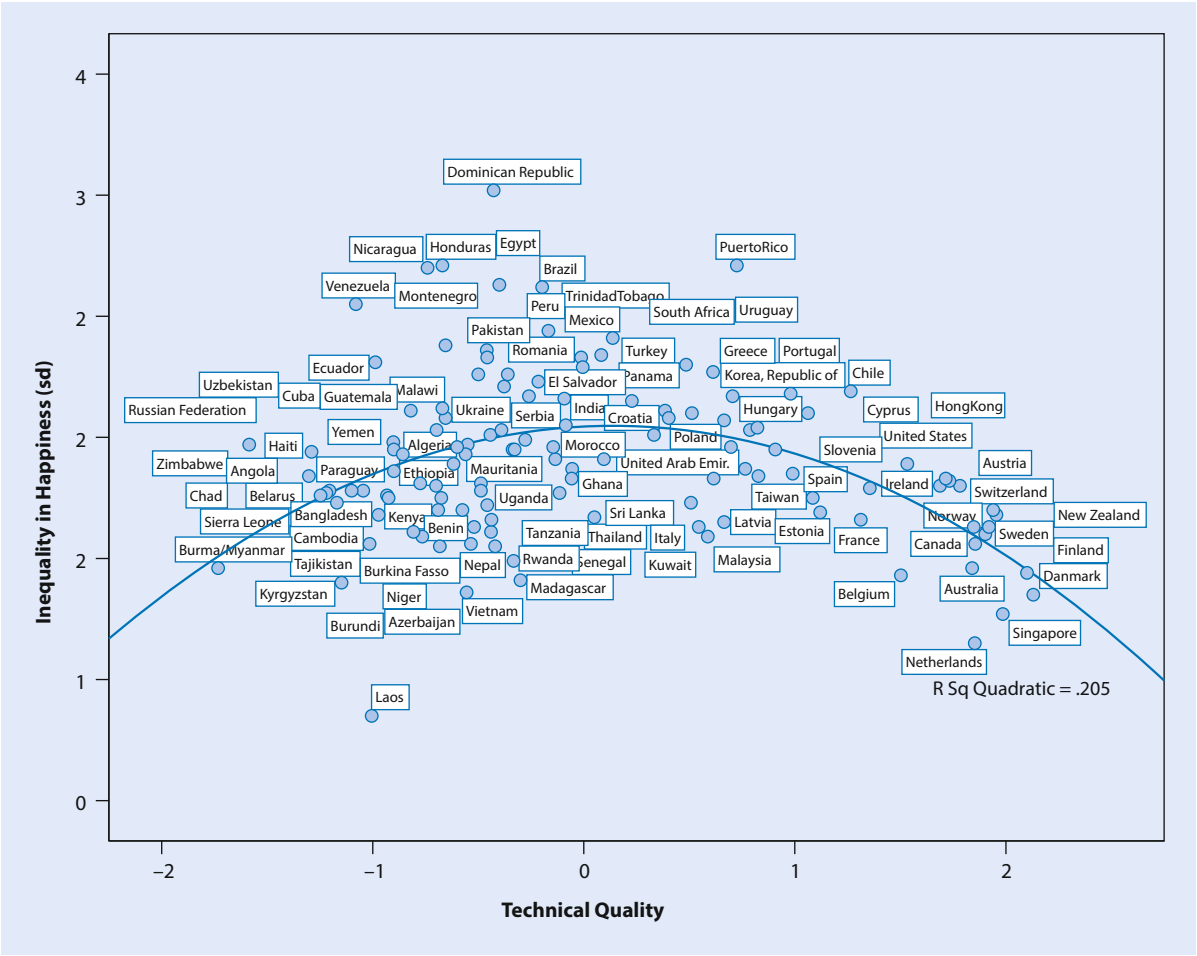


Abb. 12.28 Technische Regierungsqualität und Ungleichheit von Lebenszufriedenheit im Jahr 2006. (Aus Ott, 2011, S. 12)

„Prüfungsleistung“ und „beruflicher Erfolg“ ein Zusammenhang besteht, wenn das Merkmal „Prüfungsangst“ aus dem Merkmal „ Prüfungsleistung“ herauspartialisiert wird, wenn also die Prüfungsleistungen bzgl. der Prüfungsangst „bereinigt“ werden. Das Verfahren, mit dem derartige Hypothesen überprüft werden, heißt **Semipartialkorrelation**.

Multiple Zusammenhänge

Multiple Zusammenhangshypothesen betreffen Beziehungen zwischen einem Merkmalskomplex mit den Merkmalen X_1, $X_2 \ldots X_p$ und einem Merkmal Y. Lässt sich inhaltlich die Richtung eines möglichen kausalen Einflusses begründen, bezeichnet man diese Variablen auch als **Prädiktoren bzw. Prädiktorvariablen** (entspricht der unabhängigen/ursächlichen Variablen, die bestimmte Wirkungen erzeugt) und als **Kriterium bzw. Kriteriumsvariable** (entspricht der abhängigen Variablen, auf der sich bestimmte Wirkungen zeigen). Die Zusammenhangshypothese lautet dann: Zwischen mehreren Prädiktorvariablen und einer Kriteriumsvariablen besteht ein Zusammenhang.

Multiple Zusammenhangshypothese – Eine multiple Zusammenhangshypothese behauptet, dass zwischen mehreren Prädiktorvariablen und einer Kriteriumsvariablen ein Zusammenhang besteht. Sie wird mit der multiplen Korrelations- und Regressionsanalyse überprüft.

Multipler Korrelationskoeffizient R – Der multiple Korrelationskoeffizient R („multiple correlation coefficient“) quantifiziert die Enge (nicht jedoch die Richtung) des Gesamtzusammenhangs zwischen dem Prädiktorvariablensatz und dem Kriterium. Er wird auf Signifikanz geprüft.

Multiple Regressionsanalyse – Die multiple Regressionsanalyse bestimmt die Regressionskoeffizienten (Beta-Gewichte) der Prädiktoren, die Auskunft über die Enge und Richtung des Einflusses der einzelnen Prädiktoren im Kontext des Gesamtzusammenhangs aller betrachteten Prädiktorvariablen geben. Auch die Beta-Gewichte werden auf Signifikanz geprüft.

Viele Zusammenhangshypothesen lassen sich sinnvoll nur als multiple Zusammenhangshypothesen formulieren. Dies trifft umso eher zu, je komplexer die zu untersu-

Kodierung eines polytom-nominalen Merkmals durch dichotome Indikatorvariablen

Es wird die Hypothese untersucht, dass zwischen der Art der Berufsausübung (als Arbeiter, Angestellter, Beamter oder als Selbstständiger) und der Anzahl der jährlichen Urlaubstage ein Zusammenhang besteht. (Vereinfachend gehen wir davon aus, dass die Art der Berufsausübung hier in eindeutige, trennscharfe und erschöpfende Kategorien eingeteilt wurde. Zu Anforderungen an nominalskalierte Variablen ► Abschn. 8.4.2.) Diese Hypothese kann 1.) als bivariate Zusammenhangshypothese über den Kontingenzkoeffizienten bzw. Cramérs Phi, 2.) als Unterschiedshypothese über die Varianzanalyse oder 3.) als multivariate Zusammenhangshypothese über die multiple Korrelation geprüft werden.

Alle drei Auswertungstechniken erfordern dasselbe Untersuchungsmaterial, nämlich Angaben über die Art der Berufsausübung und die Anzahl jährlicher Urlaubstage einer möglichst repräsentativen Stichprobe berufstätiger Personen. Im Ergebnis unterscheidet sich die erste Auswertungsart geringfügig von der zweiten bzw. dritten Auswertungsart. (Durch die Zusammenfassung des Merkmals „Anzahl der jährlichen Urlaubstage“ in einzelne Kategorien, die für die Anwendung des Kontingenzkoeffizienten erforderlich ist, gehen Informationen verloren.) Die zweite und dritte Auswertungsart sind genauer und führen zu identischen Resultaten.

Hier soll nur demonstriert werden, wie das Untersuchungsmaterial für eine Auswertung über eine multiple Korrelation vorbereitet wird. Wir nehmen einfachheitshalber an, es seien in jeder Berufskategorie lediglich 3 Personen befragt worden:

Berufskategorie	Urlaubstage
Arbeiter	26, 30, 24
Angestellte	28, 25, 25
Beamte	26, 32, 30
Selbstständige	30, 16, 26

Wir definieren eine sog. **Indikatorvariable** d_1, auf der alle Personen der ersten Berufsgruppe eine 1 und die Personen der zweiten und dritten Gruppe eine 0 erhalten. Mit einer zweiten Indikatorvariablen d_2 wird entschieden, welche Personen zur zweiten Berufsgruppe gehören. Diese erhalten hier eine 1 und die Personen der ersten und dritten Gruppe eine 0. Entsprechend verfahren wir mit einer dritten Indikatorvariablen d_3: Personen der dritten Berufskategorie werden hier mit 1 und die der ersten und zweiten Kategorie mit 0 verschlüsselt. Diese dichotomen Indikatorvariablen mit den Ausprägungen 0 und 1 werden auch als **Dummy-Variablen** und die Kodierung dementsprechend als **Dummy-Kodierung** bezeichnet (Bortz & Schuster, 2010, S. 364).

Bei dieser Vorgehensweise bleibt offen, wie mit Personen der vierten Berufskategorie umzugehen ist. Führen wir das Kodierungsprinzip logisch weiter, benötigen wir eine vierte Indikatorvariable d_4, die darüber entscheidet, ob eine Person zur vierten Kategorie gehört ($d_4 = 1$) oder nicht ($d_4 = 0$). Diese vierte (oder allgemein bei k Kategorien die k-te) Indikatorvariable ist jedoch überflüssig. Ordnen wir der vierten Kategorie auf allen 3 Indikatorvariablen (d_1, d_2, d_3) eine 0 zu, resultieren vier verschiedene Kodierungsmuster, die eindeutig zwischen den 4 Kategorien differenzieren: Kategorie 1: 1, 0, 0; Kategorie 2: 0, 1, 0; Kategorie 3: 0, 0, 1 und Kategorie 4: 0, 0, 0. Die multiple Korrelation wird damit über folgende Datenmatrix berechnet:

Prädiktoren			Kriterium
d_1	d_2	d_3	y
1	0	0	26
1	0	0	30
1	0	0	24
0	1	0	28
0	1	0	25
0	1	0	25
0	0	1	26
0	0	1	32
0	0	1	30
0	0	0	30
0	0	0	16
0	0	0	26

Über die Theorie dieser Vorgehensweise sowie weitere Ansätze zur Verschlüsselung nominaler Merkmale als Indikatorvariablen berichten u. a. Bortz & Schuster (2010, S. 363ff.); Gaensslen und Schubö (1973, Kap. 12.1); Moosbrugger (1978, 2002); Overall und Klett (1972); Rochel (1983); Sievers (1977); von Eye und Schuster (1998, Kap. 4); Werner (1997, Kap. 4) sowie Wolf und Cartwright (1974)

chenden Variablen sind. Eine Sportpsychologin hätte sicherlich keinerlei Schwierigkeiten, die Kriteriumsvariable „Weitsprungleistung“ zu messen. Interessiert sie jedoch der Zusammenhang dieses Merkmals mit dem Prädiktor „Trainingsmotivation“, steht sie vor der weitaus schwierigeren Aufgabe, diesen komplexen Prädiktor zu operationalisieren und zu quantifizieren. Es erscheint zweifelhaft, dass sich dieses Merkmal durch nur einen Wert eines jeden Untersuchungsteilnehmers – was der Vorgehensweise für die Überprüfung einer bivariaten Zusammenhangshypothese entspräche – vollständig abbilden lässt. Zufriedenstellender wären hier mehrere operationale Indikatoren der Trainingsmotivation, wie z. B. die Anzahl freiwillig absolvierter Trainingsstunden, die Konzentration während des Trainings, die Intensität des Trainings, die Anzahl der Pausen etc., also Indikatoren, die verschiedene, wichtig erscheinende Teilaspekte des untersuchten Merkmals erfassen. Die multiple Zusammenhangshypothese würde dann lauten: Zwischen den Indikatorvariablen X_1, $X_2 \ldots X_p$ des Merkmals „Trainingsmotivation“ (sie fungieren als Prädiktoren) und dem Merkmal „Weitsprungleistung“ (Y), das hier das Kriterium darstellt, besteht ein Zusammenhang.

Eine sehr interessante Veranschaulichung multipler Zusammenhänge findet man bei Dawes, Faust, & Meehl (1993). Hier geht es um den Vergleich klinischer Vorhersagen von Krankheitsverläufen durch erfahrene Mediziner („clinical prediction“) mit statistischen Vorhersagen aufgrund mehrerer Prädiktorvariablen („statistical prediction“; s. hierzu auch den Klassiker „Clinical vs. Statistical

Prediction“ von Meehl, 1954). Es wird gezeigt, dass bei identischen Ausgangsinformationen die statistischen Vorhersagen den klinischen Vorhersagen überlegen sind. Einige Studien belegen darüber hinaus, dass sogar Kombinationen von klinischen und statistischen Vorhersagen schlechter abschneiden als statistische Vorhersagen allein.

Die Überprüfung einer multiplen Zusammenhangshypothese, die eine Beziehung zwischen mehreren (Prädiktor-)Variablen und einer (Kriteriums-)Variablen behauptet, erfolgt über die **multiple Korrelations- und Regressionsanalyse** (Bortz & Schuster, 2010, Kap. 21; Eid et al., 2010, Kap. 18). Der multiple Korrelationskoeffizient R quantifiziert die Höhe des **Gesamtzusammenhangs** und ist auf Signifikanz zu prüfen. Die multiple Regressionsanalyse berechnet die **Regressionskoeffizienten** für die einzelnen Prädiktoren **(Beta-Gewichte)**, denen zu entnehmen ist, ob ein Prädiktor im Kontext des betrachteten Prädiktorvariablensatzes substanziell (signifikant) mit dem Kriterium korreliert und welche Stärke sowie welche Richtung dieser Zusammenhang zwischen dem einzelnen Prädiktor und dem Kriterium hat (Betrag und Vorzeichen des β-Gewichtes). Wir möchten noch einmal betonen, dass die Regressionskoeffizienten eines multiplen Zusammenhangs sehr deutlich von den einzelnen bivariaten Korrelationen zwischen jedem Prädiktor und dem Kriterium abweichen können. (Regressionskoeffizienten und bivariate Prädiktor-Kriteriums-Korrelationen sind nur dann identisch, wenn alle Prädiktoren untereinander zu $r = 0$ korrelieren.) Denn die Regressionsgewichte geben Auskunft über den Erklärungsbeitrag des einzelnen Prädiktors im Zusammenhang mit dem gesamten Prädiktorvariablensatz. Das β-Gewicht eines Prädiktor ändert sich dementsprechend, wenn – bei gleichbleibendem Kriterium – weitere oder andere Prädiktoren in die multiple Regression aufgenommen oder Variablen eliminiert werden. Das ▶ Studienbeispiel „Multiple Regression: Der Zusammenhang zwischen acht Persönlichkeitseigenschaften und akademischem Betrug“ illustriert die Prüfung multipler Zusammenhangshypothesen.

Untersuchungstechnisch bereitet die Prüfung multipler Zusammenhangshypothesen relativ wenig Aufwand. Alle Merkmale, d. h., die Prädiktoren und das Kriterium, werden an einer möglichst repräsentativen Stichprobe im Zuge einer Querschnittstudie erhoben (ein Messzeitpunkt). Die multiple Korrelation ist berechenbar, wenn zumindest die Kriteriumsvariable intervallskaliert ist. Die Prädiktoren können auch dichotom- oder polytom-nominalskaliert sein.

■ Nominalskalierte Prädiktoren

Die Kategorien eines dichotomen Merkmals kodiert man einfachheitshalber mit den Werten 0 und 1. Verwendet man z. B. das Ergebnis eines Leistungstests als Prädiktor, erhalten z. B. alle Personen, die den Test nicht bestanden haben, den Wert 0 und alle, die ihn bestanden haben, den Wert 1. Wie man mit einem nominalen Merkmal mit mehr als zwei Kategorien verfährt, zeigt ▶ Kodierung eines polytom-nominalen Merkmals durch dichotome Indikatorvariablen.

Es wird deutlich, dass eine bivariate Zusammenhangshypothese zwischen einem nominalskalierten und einem intervallskalierten Merkmal, für deren Überprüfung in ◘ Tab. 12.15 (nach Reduktion des intervallskalierten Merkmals auf einzelne Kategorien) der Kontingenzkoeffizient vorgeschlagen wurde, auch mit Hilfe der multiplen Korrelation geprüft werden kann. Diese verwendet die Kodierungsvariablen (dichotome **Indikatorvariablen** mit den Ausprägungen 0 und 1 bzw. Dummy-Variablen) als Prädiktoren und die intervallskalierte Variable als Kriterium. Neben dem durch Indikatorvariablen kodierten nominalen Merkmal können in einer multiplen Korrelation gleichzeitig weitere nominal- und/oder intervallskalierte Prädiktorvariablen berücksichtigt werden.

Hierbei ist allerdings zu beachten, dass die Anzahl der erforderlichen dichotomen Indikatorvariablen bei einer großen Zahl an polytom-nominalen Variablen (die jeweils womöglich noch viele Ausprägungen haben) sehr groß werden kann, was die Interpretation der multiplen Korrelation erschwert.

Ersatzweise könnte deshalb eine unter der Bezeichnung „Optimal Scaling“ bekannt gewordene Technik eingesetzt werden. Das **Optimal Scaling** basiert auf der Idee, für die k Kategorien eines nominalen Merkmals metrische Werte zu schätzen, sodass man statt $k - 1$ Dummy-Variablen nur eine Variable benötigt. Die Kategorienwerte werden so geschätzt, dass die bivariate Korrelation zwischen dem optimal skalierten nominalen (Prädiktor-)Merkmal und einer Kriteriumsvariablen genau so hoch ist wie die multiple Korrelation zwischen den $k - 1$ Dummy-Variablen und der Kriteriumsvariablen (man beachte allerdings die in diesem Zusammenhang auftretenden inferenzstatistischen Probleme). Ausführliche Informationen zu diesem Verfahren findet man z. B. bei Gifi (1990), Meulman (1992) oder Young (1981). Im SPSS-Programmpaket ist dieses Verfahren unter der Bezeichnung CATREG integriert und im SAS-Paket unter PROC.TRANSREG (MORALS). Als Anwendungsbeispiel sei eine Arbeit von Weber (2000) empfohlen, in der das Optimal Scaling im Kontext der Quantifizierung von Determinanten der Fernsehnutzung eingesetzt wird.

Die multiple Korrelation als Verfahren zur Überprüfung multivariater Zusammenhangshypothesen ist natürlich nicht nur einsetzbar, wenn eine komplexe Prädiktorvariable in Form mehrerer Teilindikatoren untersucht wird, sondern auch dann, wenn die Beziehung mehrerer inhaltlich heterogener Prädiktorvariablen zu einer

Studienbeispiel

Multiple Regression: Der Zusammenhang zwischen acht Persönlichkeitseigenschaften und akademischem Betrug

Eine Befragungsstudie unter Studierenden (n = 249 Psychologiestudierende aus den USA) widmete sich dem multiplen Zusammenhang zwischen Persönlichkeitsmerkmalen einerseits und akademischem Betrug andererseits. Unter anderem wurden folgende Hypothesen auf der Basis des Forschungsstandes formuliert:

H_1: Akademischer Betrug korreliert positiv mit allen drei „Dark-Triad-Variablen", wobei Psychopathologie der stärkste Prädiktor ist.

H_2: Akademischer Betrug korreliert unter den fünf „Big-Five-Dimensionen" v. a. negativ mit Verträglichkeit und Gewissenhaftigkeit.

Der **akademische Betrug** wurde kardinalskaliert operationalisiert über einen Häufigkeitsindex (berichtete Häufigkeit, mit der bei Tests in der Schule abgeschrieben und bei Hausaufgaben plagiiert wurde). Die Persönlichkeitseigenschaften wurden kardinalskaliert mit entsprechenden Persönlichkeitstests (▶ Abschn. 10.4.3) gemessen. Berücksichtigt wurden gemäß dem **Big-Five-Modell der Persönlichkeit** (zur Entwicklung dieses Modells durch unterschiedlicher Forschergruppen s. Digman, 1990) die fünf grundlegenden Persönlichkeitsdimensionen Extraversion, Verträglichkeit, Gewissenhaftigkeit, emotionale Stabilität bzw. Neurotizismus und Offenheit für neue Erfahrungen. Zudem wurde die sog. **Dunkle Triade der Persönlichkeit** einbezogen (Paulhus & Williams, 2002), nämlich die drei sozial problematischen Eigenschaften Narzissmus, Machiavellismus und Psychopathie ◘ Tab. 12.16).

Das Muster der bivariaten Korrelationen fiel für die Dark Trias und für Big Five jeweils exakt hypothesenkonform aus. Eine Betrachtung der multiplen Regression, bei der alle acht Persönlichkeitsdimensionen zusammen als Prädiktorensatz zur Vorhersage des akademischen Betrugs genutzt wurde, zeigte, dass aufgrund der starken Überlappungen der Persönlichkeitseigenschaften letztlich nur ein einziger der acht Prädiktoren substanziell mit dem Kriterium zusammenhing: Personen mit höherer Ausprägung in Psychopathie begingen überzufällig häufiger einen akademischen Betrug, wobei es sich um einen starken Effekt handelt. Wenn die Psychopathie als Prädiktor verwendet wird, haben die sieben anderen Prädiktoren keinen zusätzlichen Informationsnutzen und tragen nicht zu einer verbesserten Vorhersage des akademischen Betrugs bei.

Dieses Studienbeispiel illustriert, wie irreführend es sein kann, multiple Zusammenhänge nur anhand einer Betrachtung der bivariaten Korrelationen zu beurteilen, weil hierbei z. B. redundante Informationen bzw. inhaltlich sich überschneidende Konstrukte nicht deutlich werden.

Bei der **Berechnung multipler Regressionen** unterscheidet man drei Vorgehensweisen:

1. **Einschlussmethode** („enter regression"): Alle Prädiktoren werden gemeinsam in einem Schritt in die Regressionsgleichung einbezogen bzw. es wird **ein vollständiges Regressionsmodell** mit allen Prädiktoren gerechnet. Die Regressionsgleichung enthält dementsprechend sowohl signifikante als auch nicht-signifikante Prädiktoren. Diese Methode ist zu präferieren, wenn die Regressionskoeffizienten aller Prädiktoren für die Hypothesenprüfung von Interesse sind.
2. **Induktive schrittweise Methode** („stepwise regression"): Die Prädiktoren werden schrittweise induktiv – gemäß ihrer empirisch vorgefundenen Interkorrelationen – einbezogen, so dass nacheinander **verschiedene induktiv vorgefundene Regressionsmodelle** mit unterschiedlichen Prädiktorvariablensätzen berechnet und verglichen werden können. Bei der üblichen **Vorwärtstechnik** („forward selection method") werden die Prädiktoren schrittweise vom stärksten bzw. nützlichsten bis zum schwächsten Prädiktor einbezogen. In unserem Beispiel zum akademischen Betrug wurde im Zuge der Vorwärtstechnik ein Modell mit nur einem einzigen Prädiktor ausgewählt, alle anderen Prädiktoren wurden statistisch als wenig nützlich bzw. redundant klassifiziert und nicht einbezogen. Bei der seltener verwendeten **Rückwärtstechnik** („backward elimination method") beginnt man mit einem Regressionsmodell mit allen Prädiktoren und entfernt nacheinander diejenigen Prädiktoren, die den geringsten Erklärungsbeitrag leisten bzw. die geringste Nützlichkeit besitzen (Bortz & Schuster, 2010, S. 358).
3. **Theorieprüfende hierarchische Methode** (hierarchische Regressionsanalyse, „ hierarchical regression analysis", HRA): Die Prädiktoren werden nacheinander – auf der Basis ihrer theoretisch erwarteten Priorität – einbezogen, so dass nacheinander **verschiedene theoretisch begründete Regressionsmodelle** mit unterschiedlichen Prädiktorvariablensätzen berechnet und verglichen werden können. Im Zuge explanativer Studien ist dieser Ansatz die Methode der Wahl (Hoyt et al., 2008).

Kriteriumsvariablen simultan erfasst werden soll. Eine Untersuchung von Silbereisen (1977) überprüfte z. B. die Hypothese, dass die Rollenübernahmefähigkeit von Kindern mit Merkmalen wie „Betreuung an Werktagen" „Erziehungsstil der Mutter", „ Erwerbstätigkeit der Mutter", „Kindergartenbesuch", „Geschlecht des Kindes" etc. zusammenhängt. Natürlich hätte diese Hypothese auch in einzelne bivariate Zusammenhangshypothesen zerlegt und geprüft werden können. Abgesehen von inferenzstatistischen Schwierigkeiten (die wiederholte Durchführung von Signifikanztests erschwert die Kalkulation der Irrtumswahrscheinlichkeiten; Bortz et al., 2008), übersieht dieser Ansatz, dass viele bivariate Zusammenhänge nicht den gleichen Informationswert haben wie ein entsprechender multivariater Zusammenhang. Die multiple Korrelations- und Regressionsanalyse berücksichtigt auch **Kombinationswirkungen der Prädiktoren** und ist deshalb sehr viel aussagekräftiger als die separate Betrachtung einzelner bivariater Zusammenhänge.

▪ Kanonische Zusammenhänge

Gelegentlich ist es sinnvoll oder erforderlich, **zwei Variablenkomplexe**, also **mehrere Prädiktorvariablen** und **mehrere Kriteriumsvariablen**, gleichzeitig miteinander in Beziehung zu setzen. Hypothesen über die Beziehungen zwischen zwei Variablensätzen werden „kanonische

Tabelle 12.16 Multiple Regression (beta-Koeffizienten) zwischen Persönlichkeitseigenschaften und akademischem Betrug im Vergleich zu bivariaten Korrelationskoeffizienten. (Daten aus Williams, Nathanson, & Paulhus, 2010, S. 296)

	Bivariate Korrelationen mit akademischem Betrug	Multiple Regression mit akademischem Betrug
Dark Triad-Variablen:		
1. Psychopathie	.58	.50
2. Machiavellismus	.39	–
3. Narzissmus	.20	–
Big Five-Dimensionen:		
4. Extraversion	.09	–
5. Verträglichkeit	−.23	–
6. Gewissenhaftigkeit	−.28	–
7. Emotionale Stabilität	.08	–
8. Offenheit für neue Erfahrungen	.07	–

Zusammenhangshypothesen“ genannt. Soll z. B. die Hypothese geprüft werden, das Wetter beeinflusse die Befindlichkeit des Menschen, wäre einer Studie, die nur ein Merkmal des Wetters (z. B. die Temperatur) mit einem Merkmal der Befindlichkeit (z. B. die Einschätzung der eigenen Leistungsfähigkeit) in Beziehung setzt (bivariate Zusammenhangshypothese), von vornherein wenig Erfolg beschieden. Das Wetter ist nur mit einem Merkmalskomplex sinnvoll beschreibbar, der seinerseits mit vielen, sich wechselseitig beeinflussenden Merkmalen der persönlichen Befindlichkeit zusammenhängen könnte.

Die Überprüfung dieses kanonischen Zusammenhanges (mehrere Prädiktoren und mehrere Kriterien) erfolgt mit der **kanonischen Korrelation** (Bortz & Schuster, 2010, Kap. 28) Untersuchungstechnisch erfordert die Überprüfung dieser Hypothese die (multivariate) Erfassung von Witterungsbedingungen und die (multivariate) Erfassung der Befindlichkeit von Personen, die diesen Witterungsbedingungen ausgesetzt sind. Generell ist die kanonische Korrelation als Auswertungstechnik indiziert, wenn von einer Stichprobe von Merkmalsträgern (z. B. Personen) Messungen auf mehreren Prädiktorvariablen und mehreren Kriteriumsvariablen vorliegen. Der kanonische Zusammenhang wird über globale Zusammenhangsmaße – **kanonische Korrelationskoeffizienten** – quantifiziert und auf Signifikanz geprüft, zudem wird der Einfluss der einzelnen Prädiktoren und Kriterien im Kontext des kanonischen Gesamtzusammenhanges betrachtet.

Kanonische Zusammenhangshypothese – Eine kanonische Zusammenhangshypothese behauptet, dass zwischen mehreren Prädiktorvariablen einerseits und mehreren Kriteriumsvariablen andererseits ein systematischer Zusammenhang besteht. Sie wird mit der kanonischen Korrelationsanalyse überprüft. Dabei werden die Prädiktoren sowie die Kriterien jeweils zu Faktoren gebündelt (pro Prädiktor- und Kriteriumsvariablensatz entspricht die Zahl der gebildeten Faktoren jeweils der Zahl der Variablen).

Kanonische Korrelationskoeffizienten CR – Kanonische Korrelationskoeffizienten CR („canonical correlation coefficients“) sind die paarweisen Korrelationen zwischen den Prädiktor- und Kriteriumsvariablenfaktoren, sie quantifizieren die Enge (nicht jedoch die Richtung) der Zusammenhänge der Variablensätze. In einer kanonischen Korrelationsanalyse können maximal so viele kanonische Korrelationskoeffizienten berechnet werden, wie sich Variablen im kleineren Variablensatz befinden. Die kanonischen Korrelationskoeffizienten werden auf Signifikanz geprüft. Enge und Richtung des Einflusses der einzelnen Prädiktoren und Kriterien auf den Gesamtzusammenhang sind an deren Faktorladungen ablesbar.

Nominalskalierte bzw. kategoriale Kriterien Während kategoriale unabhängige Variablen bzw. Prädiktoren durch Transformation in binäre Dummy-Variablen in den meisten regressionsanalytischen Techniken zulässig sind, stellen **kategoriale abhängige Variablen bzw. Kriterien** einen Sonderfall dar. Unter den Techniken zur Analyse kategorialer Daten spielt die **logistische Regression** eine wichtige Rolle. Mit ihrer Hilfe kann ein binäres oder polytom-nominalskaliertes Kriterium durch mehrere Prädiktoren vorhergesagt werden (Agresti, 2007; Eid et al., 2010, Kap. 21).

Handelt es sich um kategoriale Kriterien und kategoriale Prädiktoren, so kommen die **Konfigurationsfrequenzanalyse** (KFA; Krauth, 1993; Krauth & Lienert, 1973; von Eye, 1990) oder **Log-lineare Modelle** zum Einsatz (Andreß, Hagenaars, & Kühnnel 1997; Arminger, 1982; Bishop, Fienberg, & Holland 1975; Langeheine,

a

		Aufgabe 1	
		+	–
		Aufgabe 2	
		+	–
Aufgabe 3	+	50	0
	–	0	50

		Aufgabe 1	
		+	–
Aufgabe 2	+	50	50
	–	50	50

b

		Aufgabe 1	
		+	–
Aufgabe 3	+	50	50
	–	50	50

c

		Aufgabe 1	
		+	–
Aufgabe 3	+	50	50
	–	50	50

d

		Aufgabe 2	
		+	–
Aufgabe 3	+	50	50
	–	50	50

Abb. 12.29 Kombinations-(Interaktions-)Wirkung von Variablen. Das Meehl'sche Paradoxon

1980; Agresti, 2007; Eid et al., 2010, Kap. 21). Eine Zusammenhangshypothese für kategoriale Variablen könnte z. B. behaupten, dass die Diagnose einer psychischen Krankheit (Schizophrenie, Depression, Paranoia etc.) von der sozialen Schicht der Patienten und Patientinnen (Unterschicht, Mittelschicht, Oberschicht) sowie der Wohngegend (städtisch vs. ländlich) abhängt. Logistische Regressionen und Log-lineare Modelle können mit gängigen Statistik-Programmpaketen wie z. B. SPSS berechnet werden.

Bivariate und multivariate Zusammenhänge im Vergleich

Es wurde bereits erwähnt, dass multivariate Zusammenhänge die Bedeutung von Merkmalskombinationen mitberücksichtigen, die bei isolierter Betrachtung bivariater Zusammenhänge verloren gehen. Sie sagen damit mehr aus als die einzelnen bivariaten Zusammenhänge. Ein besonders eindrucksvolles, auf Nominaldaten bezogenes Beispiel hierfür stellt das sog. **Meehl'sche Paradoxon** dar (Meehl, 1950). Zur Veranschaulichung dieses klassischen Beispiels einer Kombinations-(Interaktions-) Wirkung nehmen wir an, 200 Personen hätten drei Aufgaben zu lösen. Die Aufgaben können gelöst werden (+) oder nicht gelöst werden (–). In Abb. 12.29 sieht man die (fiktiven) Ergebnisse dieser Untersuchung.

Alle Personen, die Aufgabe 1 und 2 lösen, haben auch Aufgabe 3 gelöst. Aufgabe 3 wird aber auch von denjenigen gelöst, die die Aufgaben 1 und 2 nicht lösen. Umgekehrt hat jede Person, die nur eine Aufgabe löst (entweder Aufgabe 1 oder Aufgabe 2), Aufgabe 3 nicht gelöst. Damit lässt sich die Lösung oder Nichtlösung von Aufgabe 3 exakt vorhersagen: Personen, die von den Aufgaben 1 und 2 beide oder keine lösen, finden für die dritte Aufgabe die richtige Lösung. Wird von den Aufgaben 1 und 2 jedoch nur eine Aufgabe gelöst, bleibt die dritte Aufgabe ungelöst. Es besteht ein perfekter, multivariater Zusammenhang.

Betrachten wir hingegen nur jeweils zwei Aufgaben (Abb. 12.29b-d), so müssen wir feststellen, dass hier überhaupt keine Zusammenhänge bestehen. Die Tatsache, dass jemand z. B. Aufgabe 1 gelöst hat, sagt nichts über die Lösung oder Nichtlösung von Aufgabe 2 aus. Entsprechendes gilt für die übrigen Zweierkombinationen von Aufgaben. Alle bivariaten Zusammenhänge sind Null bzw. nicht vorhanden. Hätte man also ausschließlich die bivariaten Zusammenhänge betrachtet, wäre man fälschlich zum dem Schluss gekommen, dass die drei Merkmale wechselseitig voneinander unabhängig sind, obwohl bei multivariater Betrachtung zwei der Merkmale das dritte perfekt vorhersagen.

Ein weiteres (fiktives) Beispiel: Es geht um die Evaluation eines Rehabilitationsprogrammes für entlassene Strafgefangene, mit dem Rückfälle bzw. neue Delikte vermieden oder doch zumindest reduziert werden sollen. Von 2 000 Strafgefangenen nahmen 1 000 am Reha-Programm teilt. Tab. 12.17a zeigt, dass von diesen 1 000 Strafgefangenen 500, also 50 %, rückfällig wurden, während von den 1 000 nicht-teilnehmenden Strafgefangenen lediglich 400, also 40 %, rückfällig wurden. Offenbar wirkt das Programm kontraproduktiv: Durch die Teilnahme am Programm kommt es zu mehr Rückfällen als durch Verzicht

Tabelle 12.17 Zur Wirksamkeit eines Rehabilitationsprogramms: das „Simpson-Paradox"

Rückfällig	Teilnahme		Gesamt
	Nein	Ja	
	a) Alle		
Ja	400 (40 %)	500 (50 %)	900
Nein	600	500	1 100
(Gesamt)	1 000	1 000	2 000
	b) nur Männer		
Ja	175 (70 %)	450 (60 %)	625
Nein	75	300	375
(Gesamt)	250	750	1 000
	c) nur Frauen		
Ja	225 (30 %)	50 (20 %)	275
Nein	525	200	725
(Gesamt)	750	250	1 000

bzw. Verweigerung der Teilnahme. Man sollte das Reha-Programm also einstellen.

Nun wollen wir die bivariate Zusammenhangsanalyse durch Einführung eines dritten Merkmals (Geschlecht der Strafgefangenen) zu einer **trivariaten** Zusammenhang**sanalyse** erweitern. Der Einfachheit halber nehmen wir an, dass sich die Gesamtstichprobe zu gleichen Teilen aus Männern und Frauen zusammensetzt.

In Tab. 12.17b) wird der Zusammenhang von Teilnahme und Rückfälligkeit für die männlichen Strafgefangenen verdeutlicht: Von den 750 teilnehmenden Männern wurden 450 rückfällig: Das sind 60 %. Nahmen die Männer nicht am Programm teil (250), wurden 175 rückfällig. Das sind 175 von 250 oder 70 %. Verzicht auf das Programm erhöht bei Männern offensichtlich das Risiko eines Rückfalls, oder: Die Teilnahme am Programm wird dringend empfohlen.

Betrachten wir schließlich die weiblichen Strafgefangenen in Tab. 12.17c). Auch hier – wie bei den Männern – war das Programm erfolgreich, denn nur 20 % der teilnehmenden Frauen wurden rückfällig gegenüber 30 % der nicht-teilnehmenden Frauen.

Die geschlechtsspezifischen Analysen zeigen also, dass das Reha-Programm das Rückfallrisiko sowohl bei Männern als auch bei Frauen um 10 Prozentpunkte reduziert. Fassen wir jedoch Tab. 12.17b und c zusammen, resultiert Tab. 12.17a („Alle") mit dem paradoxen Ergebnis, dass das Reha-Programm das Rückfallrisiko erhöht statt es zu senken. Wie kann man diesen Widerspruch bzw. das sog. **Simpson-Paradox** erklären?

Die Erklärung ist darin zu sehen, dass die relativ hohe männliche Rückfallquote (60 %) auf vielen teilnehmenden Männern basiert (750) und die relativ geringe weibliche Rückfallquote (20 %) auf wenigen teilnehmenden Frauen (250). Die Merkmale „Geschlecht" und „Teilnahme" sind nicht voneinander unabhängig. Sie sind **„konfundiert"**. In der Addition der männlichen und weiblichen Teilnehmenden erhält die Rückfallquote der Männer ein 3faches Gewicht ($750:250 = 3$), was zu der gesamten Rückfälligkeitsquote von 50 % führt: $(3 \times 60\,\% + 1 \times 20\,\%)/4 = 50\,\%$.

Umgekehrt erhalten bei den nicht-teilnehmenden Personen die Frauen ein 3faches Gewicht, d. h., deren relativ geringe Rückfallquote von 30 % ist mit 3 und die hohe Rückfallquote der Männer mit 1 zu gewichten: $(1 \times 70\,\% + 3 \times 30\,\%)/4 = 40\,\%$. Dies führt in der Gesamtbilanz zum schlechten Abschneiden des Reha-Programms.

Man hätte natürlich auch argumentieren können, dass die teilnehmenden Frauen genauso zu gewichten seien wie die teilnehmenden Männer (etwa weil die Entwicklungskosten für das Reha-Programm pro teilnehmende Frau 3-mal so hoch sind wie die Entwicklungskosten pro teilnehmendem Mann). In diesem Falle hätten die teilnehmenden Personen eine Rückfallquote von $(60\,\% + 20\,\%)/2 = 40\,\%$ und die nicht-teilnehmenden Personen von $(70\,\% + 30\,\%)/2 = 50\,\%$. Jetzt würde also auch das Gesamtergebnis für das Reha-Programm sprechen.

Dieses Ergebnis hätte man auch erzielt, wenn aus ethischen und untersuchungstechnischen Gründen eine zufällige Zuweisung (Randomisierung) der Männer und Frauen auf die Untersuchungsbedingungen: „Teilnahme nein/ja" möglich gewesen wäre. Mit 500 teilnehmenden Männern und 500 teilnehmenden Frauen wären die Merkmale Geschlecht und Teilnahme dann voneinander unabhängig bzw. nicht konfundiert, so dass sich der Gesamteffekt additiv aus den geschlechtsspezifischen Effekten ergibt.

Übernehmen wir die Rückfallquoten aus Tab. 12.17b und c, resultieren nun für die teilnehmenden Personen 300 rückfällige Männer (60 % von 500) und 100 rückfällige Frauen (20 % von 500) bzw. insgesamt 400 rückfällige Personen bzw. 40 %. Für die nicht-teilnehmenden Personen lauten die entsprechenden Zahlen: 350 rückfällige Männer (70 % von 500), 150 rückfällige Frauen (30 % von 500), also insgesamt 500 Rückfälle bzw. 50 %. Auch diese Vorgehensweise hätte also den Erfolg des Reha-Programmes bestätigt.

Weitere Informationen und Literatur zum Simpson-Paradox findet man bei Yarnold (1996). Die Beispiele zei-

Tabelle 12.18 Hypothesenmatrizen für eine konfirmatorische Faktorenanalyse

	Matrix a			Matrix b			Matrix c		
	FI	FII	FIII	FI	FII	FIII	FI	FII	FIII
1. Ich kann vor lauter Sorgen nicht schlafen.	+1	0	0	?	0	0	.55	−.13	.17
2. Ich kann schwer einschlafen.	+1	0	0	?	0	0	.79	−.16	−.12
3. Ich leide unter Schlaflosigkeit.	+1	0	0	?	0	0	.99	.15	−.05
4. Ich lasse nichts auf mir sitzen.	0	+1	0	0	?	0	.35	.94	.02
5. Ich bin immer schnell mit einer Antwort zur Hand.	0	+1	0	0	?	0	.03	.38	.15
6. Aus Streitgesprächen halte ich mich raus.	0	−1	0	0	?	0	−.01	−.82	.17
7. Ich bin ein rundum glücklicher Mensch.	0	0	+1	0	0	?	−.05	.02	.91
8. Ich führe ein sorgloses Dasein.	0	0	+1	0	0	?	−.15	.14	.82
9. Ich liebe Spontanentschlüsse.	0	0	+1	0	0	?	−.03	.13	.46

gen, dass bei der Überprüfung multivariater Hypothesen in Form einzelner bivariater Hypothesen entscheidende Informationen verloren gehen können – ein Befund, der nicht nur für nominale Merkmale gilt. Geeignete statistische Auswertungsverfahren, um Scheinkorrelationen und maskierte Zusammenhänge wie das Simpson-Paradox bei kategorialen Variablen aufzudecken, sind die Log-linearen Modelle und Logit-Modelle (Eid et al., 2010, Kap. 20).

Wenn man sich für die Zusammenhänge zwischen mehr als zwei Variablen interessiert, sollte stets auf *multivariate bzw. multiple Korrelations- und Regressionsanalysen* zurückgegriffen werden. Die Überprüfung einer multivariaten Zusammenhangshypothese durch mehrere bivariate Korrelationen führt oft zu Fehlinterpretationen.

Faktorielle Zusammenhänge

Eine weitere multivariate Zusammenhangshypothese behauptet, dass die wechselseitigen Zusammenhänge vieler Merkmale durch wenige, in der Regel voneinander unabhängige (orthogonale) **Dimensionen** oder **Faktoren** erklärbar sind. Ein Beispiel (nach Mulaik, 1975) soll diese Hypothesenart erläutern.

Untersucht wird ein aus 9 Items bestehender Selbsteinschätzungsfragebogen (Tab. 12.18), der von einer Stichprobe beantwortet wurde. Es wird die Hypothese geprüft, dass diese Items die von Eysenck (1969) berichteten, voneinander unabhängigen Faktoren **„Schlafschwierigkeiten", „Schlagfertigkeit" und „Sorglosigkeit"** erfassen, bzw. dass die wechselseitigen Zusammenhänge (Interkorrelationen) der 9 Items auf diese 3 Faktoren zurückzuführen sind.

Die faktoriellen Hypothesen besagen, dass die ersten 3 Items eindeutig und ausschließlich einem Faktor FI (Schlafschwierigkeiten), die Items 4–6 einem Faktor FII (Schlagfertigkeit) und die 3 letzten Items einem Faktor FIII (Sorglosigkeit) zugeordnet sind. Die Hypothesenmatrix a in Tab. 12.18 fasst diese faktoriellen Hypothesen symbolisch zusammen. Die hier wiedergegebenen Zahlenwerte sind Korrelationen der Items (Variablen) mit den Faktoren, die man auch als **Faktorladungen** bezeichnet. Es werden perfekt positive (+1) bzw. perfekt negative (−1) Korrelationen der Items mit dem jeweiligen Faktor angenommen, was in der Praxis unrealistisch ist.

Gegenüber der Hypothesenmatrix a ist die Hypothesenmatrix b weniger restriktiv und somit realistischer. Hier wird nur behauptet, dass die Items mit denjenigen Faktoren, mit denen sie theoretisch nichts zu tun haben sollten, in keinem Zusammenhang stehen bzw. zu Null korrelieren. Über die Höhe des Zusammenhanges der Items mit den ihnen zugeordneten Faktoren wird keine Aussage gemacht.

Hypothesenmatrix c gibt eine konkrete, empirisch ermittelte Faktorenstruktur vor, die für männliche Personen errechnet wurde. Man könnte nun eine Hypothese formulieren, die behauptet, dass zwischen dieser, für männliche Personen ermittelten Faktorstruktur und der Faktorstruktur weiblicher Personen ein Zusammenhang besteht. Diese Zusammenhangshypothese wird mit einem **Faktorstrukturvergleich** überprüft (Bortz & Schuster, 2010, S. 424f.).

Weitere Hypothesen dieser Art beziehen sich auf die Anzahl der Faktoren eines Variablensatzes bzw. darauf, welche Faktoren voneinander unabhängig (orthogonal) bzw. abhängig (oblique) sind. Die Überprüfung derartiger Hypothesen erfolgt mit der **konfirmatorischen Faktorenanalyse** bzw. mit **Strukturgleichungsmodellen** (▶ Kap. 17).

Faktorenanalyse – Die **Faktorenanalyse** bündelt die Variablen gemäß ihrer Interkorrelationen zu Faktoren. Man unterscheidet explorative Faktorenanalysen („exploratory factor analysis", EFA), die ohne Vorannahmen durchgeführt werden und keine Signifikanzprüfung enthalten, von konfirmatorischen Faktorenanalysen („confirmatory factor analysis", CFA), bei denen ein Faktorladungsmuster als Hypothese vorgegeben und inferenzstatistisch geprüft wird.

Ausgangsmaterial für eine Faktorenanalyse ist typischerweise eine Matrix der Variableninterkorrelationen oder -kovarianzen. Häufig jedoch sind es ausschließlich nominale Merkmale, die im Blickpunkt des Interesses stehen. Die Auszählung dieser Merkmale führt zu Häufigkeiten bzw. Kontigenztafeln, deren „faktorielle Struktur" mittels der sog. **multiplen Korrespondenzanalyse** (MCA, auch „Dual Scaling" oder „Additive Scoring" genannt) überprüft werden kann. Ziel der Korrespondenzanalyse ist es, die Kategorien von zwei oder mehr Merkmalen als Punkte in einem „Faktorenraum" mit möglichst wenigen Dimensionen abzubilden. Wenn man so will, ist die Korrespondenzanalyse also die „Faktorenanalyse" für kategoriale Daten (ausführlicher hierzu Clausen, 1998; Greenacre, 1993).

▪ Hierarchische Zusammenhangshypothesen

Hierarchische Datenstrukturen („nested data") entstehen durch mehrstufige Stichprobenziehung in Form von Klumpenstichproben (▶ Abschn. 9.3.2). Die Untersuchungseinheiten sind dadurch nicht voneinander unabhängig, sondern innerhalb eines Clusters (z. B. Schulklasse, Betrieb, Krankenhaus, Wahlkreis) einander in der Regel ähnlicher als zwischen verschiedenen Clustern. Dadurch ist die **statistische Unabhängigkeits-Voraussetzung** der linearen multiplen Regression verletzt, so dass die Mehrebenenregression bzw. **Mehrebenenanalyse** („multilevel analysis") zum Einsatz kommen muss (Hox, 2002; Hox & Roberts, 2010; Snijders & Bosker, 2011; Eid et al., 2010, Kap. 19).

Im einfachsten Fall einer hierarchischen Datenstruktur liegen zwei Ebenen vor (Individualebene und Gruppenebene). Je nach Reichweite der Studie sind auch Datensätze und Hypothesen mit einer 3-Ebenen-Struktur möglich: Untersuchte Patienten können z. B. aus verschiedenen Kliniken und diese wiederum aus verschiedenen Ländern stammen; untersuchte Schüler können aus verschiedenen Schulen aus verschiedenen Bundesländern stammen. Die Berücksichtigung der hierarchischen Datenstruktur verhindert nicht nur Fehler bei der statistischen Parameterschätzung, sondern auch inhaltlich einen sog. **ökologischen Fehlschluss**, der zustande kommt, wenn man einen Effekt auf Gruppenebene fälschlich direkt auf die Individualebene überträgt. Im Kontext einer Mehrebenenanalyse werden die das untersuchte Kriterium vorhersagenden Prädiktoren den unterschiedlichen Ebenen (Level 1: Individualebene, Level 2: Gruppenebene, Level 3: übergeordnete Gruppenebene) ausdrücklich zugeordnet und in ihrem separaten und gemeinsamen Einfluss auf das Kriterium inferenzstatistisch geprüft, wozu eine Reihe von Modellvarianten zur Verfügung stehen.

So wurde z. B. in einer sozialpsychologischen Studie in der Schweiz untersucht, wie gut sich negative Einstellungen der Schweizer Bevölkerung gegenüber der religiösen Kopfbedeckung von Muslima aus Level-1-Variablen (individuelle Werthaltungen zu Sicherheit, Tradition und Konservatismus) und/oder Level-2-Variablen (Immigranten-Anteil und politisches Klima in der Wohngemeinde) erklären lassen. Die Zwei-Ebenen-Analyse zeigte, dass Prädiktoren beider Ebenen einen Erklärungsbeitrag liefern und sich auch substantielle Interaktionseffekte zeigen (Fasel, Green, & Sarrasin, 2013).

Als Faustregel für den minimalen Stichprobenumfang wird von mindestens 30 Untersuchungseinheiten pro Ebene ausgegangen. Zur Durchführung von Mehrebenenanalysen sowie für die Bestimmung ihrer Teststärken und optimalen Stichprobenumfänge liegen **spezielle Statistik-Programme** bzw. Makros vor (Eid et al., 2010, Kap. 19), z. B. **HLM** (Hierarchical Linear Modeling; http://www.ssicentral.com/hlm/), **MPlus** (ermöglicht auch die Integration latenter Variablen; http://www.statmodel.com) sowie **PinT** (Power in two-level designs; http://www.stats.ox.ac.uk/~snijders/multilevel.htm; Snijders & Bosker, 1993).

Kausale Zusammenhangshypothesen

Die Formulierung von Zusammenhangshypothesen leidet unter der Schwierigkeit, dass das deutschsprachige Vokabular (und nicht nur dieses) wenig Ausdrücke enthält, die einen schlichten Zusammenhang zweier oder mehrerer Merkmale, d. h. das Faktum, dass sich bei Veränderung eines Merkmals ein anderes Merkmal der Tendenz nach gleichsinnig oder gegenläufig verändert, treffend beschreiben. So liest man häufig im Zusammenhang mit der Interpretation von Korrelationen, dass ein Merkmal ein anderes „determiniert", „erklärt", „bedingt", „beeinflusst", dass ein Merkmal von einem anderen „abhängt" oder für ein anderes Merkmal „Bedeutung hat", dass sich ein Merkmal auf ein anderes „auswirkt" etc. Gegen den Gebrauch dieser Redewendungen ist nichts einzuwenden, wenn dabei aus dem Kontext ersichtlich wird, dass Kor-

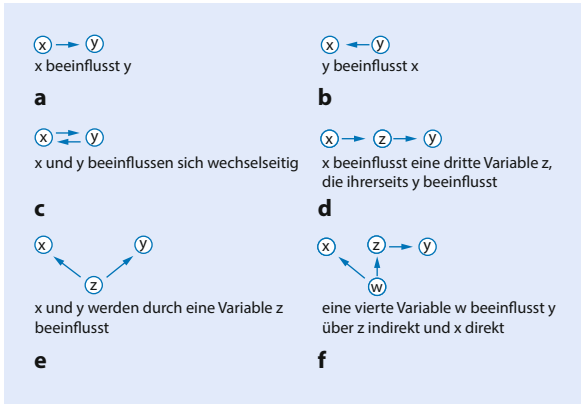


Abb. 12.30 Kausalmodelle und ihre Stützung durch eine Korrelation

relationen nicht fälschlicherweise – wie die Ausdrücke es nahelegen – als kausale Zusammenhänge interpretiert werden.

Korrelationen geben Auskunft über die Richtung und Enge eines Zusammenhanges, nicht jedoch über seine Ursachen. Die Prüfung *kausaler* Zusammenhang*shypothesen* kann – unter Anwendung diverser Zusatztechniken – stets nur annäherungsweise erfolgen.

Kausalmodelle

Eine Korrelation sagt für sich genommen nichts über Ursache-Wirkungs-Verhältnisse aus. In Abb. 12.30 findet sich eine Auswahl von Kausalmodellen, die alle durch den Nachweis einer bivariaten Korrelation r_{xy} „bestätigt" werden. Es erscheint jedoch fraglich, ob eines der hier dargestellten Kausalmodelle empirisch Bestand hat (s. unten); reale Ursachen-Wirkungs-Gefüge dürften in den Sozial- und Humanwissenschaften weitaus komplizierter sein, als Kausalbeziehungen, die sich mit drei oder vier Merkmalen theoretisch konstruieren lassen.

Kausale Hypothesen können allerdings durch nicht vorhandene Korrelationen widerlegt werden. (Diese theoretische Behauptung setzt praktisch voraus, dass das Ausbleiben einer Korrelation nicht durch Stichprobenselektionsfehler oder Messfehler erklärt werden kann. Zudem meint „Korrelation" hier Zusammenhänge beliebiger Art, d. h. auch Korrelationen, die ein nicht-lineares Ursache-Wirkungs-Gefüge modellieren.)

Wenn z. B. behauptet wird, übermäßiger Alkoholgenuss (X) reduziere die Lebenserwartung (Y) (**Modell a** in Abb. 12.30), ist diese Kausalhypothese widerlegt, wenn sich zwischen diesen Merkmalen keine irgendwie geartete Korrelation nachweisen lässt. Im umgekehrten Falle, wenn also Lebensdauer und Alkoholkonsum zusammenhängen, spricht dieses Ergebnis nicht gegen das behauptete Kausalmodell; es unterstützt aber gleichzeitig auch andere Kausalmodelle. Verwenden wir als Beispiele die in Abb. 12.30 veranschaulichten formalen Kausalmodelle b bis f, ließen sich diese wie folgt konkretisieren:

- **Modell b:** Eine geringe Lebenserwartung verursacht erhöhten Alkoholkonsum.
- **Modell c:** Erhöhter Alkoholkonsum macht depressiv und verdunkelt damit die Lebensperspektive. Diese Lebensunlust lässt erneut zur Flasche greifen.
- **Modell d:** Durch höheren Alkoholkonsum wird man arbeitsunfähig und damit arm. Armut (Z) bedingt schlechte Ernährung, die das Leben verkürzt. Armut wäre hier eine **Mediatorvariable** (▶ Moderator- und Mediatoreffekte).
- **Modell e:** Eine angeborene „Ich-Schwäche" (Z) erhöht die Anfälligkeit für lebensbedrohende Krankheiten und für Alkohol.
- **Modell f:** Stress (W) verursacht Trinken und Rauchen (Z). Lebensverkürzend wirkt aber nur das Rauchen.

Die Beispiele sind bewusst unterschiedlich „glaubwürdig" gehalten. Ihre subjektive Glaubwürdigkeit resultiert aber nicht aus der Korrelation; diese unterstützt alle Kausalannahmen gleichermaßen. Es sind vielmehr subjektive Überzeugungen und Hintergrundwissen, die das eine oder andere Kausalmodell als plausibler erscheinen lassen. Eine Korrelationsstudie alleine (hier die Bestimmung der Korrelation zwischen Lebensdauer und Alkoholkonsum) differenziert diese Kausalmodelle nicht. Eine Technik, mit der z. B. die Modelle d und e bei kategorialen Merkmalen differenziert werden können, wurde von von Eye und Schuster (1998) vorgeschlagen.

Korrelationen sind nicht geeignet, die Gültigkeit eines Kausalmodells nachzuweisen. Allerdings ist es möglich, durch Nullkorrelationen Kausalmodelle zu falsifizieren, da Kausalrelationen Korrelationen implizieren.

Korrelationsstudien haben damit nur eine geringe **interne Validität** und sind experimentellen und z. T. auch quasi-experimentellen Plänen unterlegen (in derartigen Studien werden typischerweise Unterschiedshypothesen statistisch geprüft ▶ Abschn. 12.5.3). Dennoch haben sie in der empirischen Forschung eine wichtige Funktion: Sie gestatten es, ohne besonderen Untersuchungsaufwand bestimmte Kausalhypothesen von vornherein als äußerst unwahrscheinlich auszuschließen.

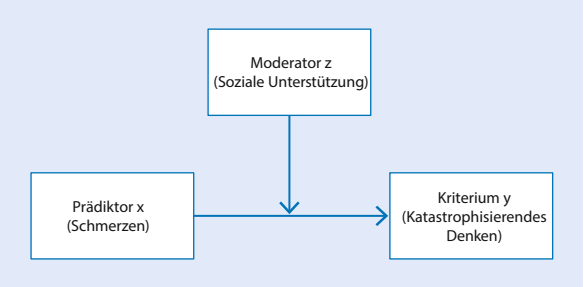
Kausalinterpretationen von Korrelationen sind – wenn überhaupt – nur inhaltlich bzw. logisch zu begründen. Ließe sich die Hypothese „Zwischen Witterungsbedingungen

Moderator- und Mediatoreffekte

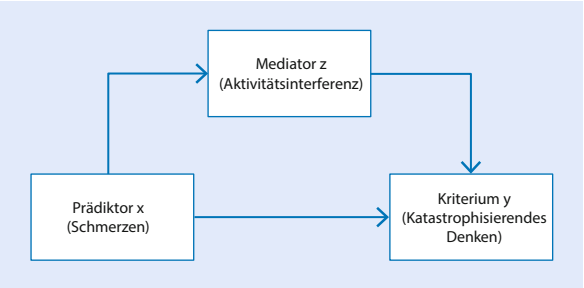
Wenn wir komplexere Zusammenhangsmodelle und darin enthaltene Wirkbeziehungen betrachten, so sind neben direkten Einflüssen auch sog. indirekte Einflüsse, v. a. Moderator- und Mediatorvariablen von Interesse (Baron & Kenny, 1986; zur statistischen Auswertung von Moderator- und Mediatoreffekten illustriert an Beispielen aus der Rehabilitationspsychologie s. Hoyt et al., 2008).

Eine **Moderatorvariable** („moderator variable“) ist eine Variable, welche die Enge und/oder Richtung des Zusammenhanges zwischen zwei Variablen beeinflusst (◘ Abb. 12.31). Hypothesen über Moderatoreffekte treffen Aussagen darüber, unter welchen Bedingungen bzw. für welche Teilgruppen der betrachtete Zusammenhang besonders stark oder schwach ausgeprägt ist. Beispiel: Zwischen Schmerzen (Prädiktor) und katastrophisierendem Denken (Kriterium) besteht ein positiver Zusammenhang, wer mehr Schmerzen hat, denkt eher über katastrophale Folgen (weitere Schmerzsteigerung, dauerhafte gesundheitliche Schäden etc.) nach als eine Person, die weniger Schmerzen hat. Dieser Zusammenhang wird durch die Menge verfügbarer sozialer Unterstützung moderiert. Bei Personen mit viel sozialer Unterstützung ist der Zusammenhang zwischen Schmerz und katastrophisierendem Denken geringer, sie werden z. B. durch ihre Sozialkontakte abgelenkt und aufgemuntert. Bei Personen mit geringer sozialer Unterstützung, ist der Zusammenhang zwischen Schmerz und Katastrophendenken enger (Hoyt et al., 2008). Zusammengefasst: Eine Moderatorvariable gibt an, unter welchen Bedingungen bzw. bei welchen Personengruppen ein Effekt **stärker oder schwächer ausgeprägt** ist. **Hypothesen über Moderatoreffekte** können u. a. im Rahmen der multiplen Regressionsanalyse geprüft werden (Eid et al., 2010, S. 644ff.).

Eine **Mediatorvariable** („mediator variable“) ist eine intervenierende Variable. Sie wird kausal von der unabhängigen bzw. Prädiktorvariable beeinflusst und beeinflusst dann ihrerseits kausal die abhängige bzw. Kriteriumsvariable (◘ Abb. 12.32). Die Mediatorvariable ist also ein notwendiges Bindeglied in der Kausalkette. Durch das Einfügen von Mediatorvariablen in Kausalmodelle werden die kausalen Wirkprozesse theoretisch und empirisch detaillierter erfasst und untersucht. Beispiel: Patienten, die unter chronischen Schmerzen leiden, entwickeln oft eine Neigung zum katastrophisierenden Denken, das um negative Entwicklungen der eigenen Gesundheit kreist. Eine mögliche Mediatorvariable könnte die Aktivitätsinterferenz der Schmerzen sein: Durch Schmerzen können die Patienten nicht mehr ihren gewohnten Arbeits- und/oder Freizeitaktivitäten nachgehen (Aktivitätsinterferenz), infolgedessen haben sie weniger positive Erlebnisse und mehr Zeit über die Schmerzen und die Krankheit nachzudenken, was zu katastrophisierenden Denkmustern führt (vgl. Hoyt et al., 2008). Zusammengefasst: Eine Mediatorvariable stellt ein notwendiges **Bindeglied in einer Kausalkette** dar, sie sorgt mit dafür, dass der interessierende Effekt überhaupt zustande kommt. **Hypothesen über Mediatoreffekte** können im Rahmen von multiplen Regressionsanalysen (Baron & Kenny, 1986), Pfadanalysen und Strukturgleichungsmodellen sowie auf der Basis der bias-korrigierten Bootstrapping-Methode (MacKinnon, 2008) geprüft werden.



◘ Abb. 12.31 Moderatoreffekt der Variablen *z* (z. B. *soziale Unterstützung*) auf den kausalen Zusammenhang zwischen dem *Prädiktor x* (z. B. *Schmerzen*) auf das *Kriterium y* (z. B. *katastrophisierendes Denken*)



◘ Abb. 12.32 Mediatoreffekt der Variable *z* (z. B. *Aktivitätsinterferenz*) als Bindeglied in der Kausalkette zwischen *Prädiktor x* (z. B. *Schmerzen*) und *Kriterium y* (z. B. *katastrophisierendes Denken*)

und Befindlichkeit besteht ein Zusammenhang“ korrelationsstatistisch bestätigen, würde wohl niemand auf die Idee kommen, damit das Kausalmodell „Die Befindlichkeit beeinflusst das Wetter“ als bestätigt zu sehen. Unser Wissen über die Entstehung von Wetterverhältnissen lässt als Erklärung dieser Korrelation nur das Kausalmodell „Wetter beeinflusst Befindlichkeit“ zu oder bestenfalls Modelle vom Typus d in ◘ Abb. 12.30, nach denen das Wetter die Befindlichkeit indirekt beeinflusst. Wir favorisieren dieses Kausalmodell jedoch nicht wegen der Korrelation, sondern weil wir (mehr oder weniger) genau wissen, wie das Wetter entsteht bzw. weil wir sicher wissen, dass die menschliche Befindlichkeit das Wetter nicht beeinflusst.

▪ Längsschnittstudien

Die Anzahl konkurrierender Kausalmodelle wird erheblich eingeschränkt, wenn die zu korrelierenden Merkmale zu unterschiedlichen Zeitpunkten erhoben werden, weil man sicher sein kann, dass das später erhobene Merkmal das früher erhobene Merkmal nicht beeinflusst.

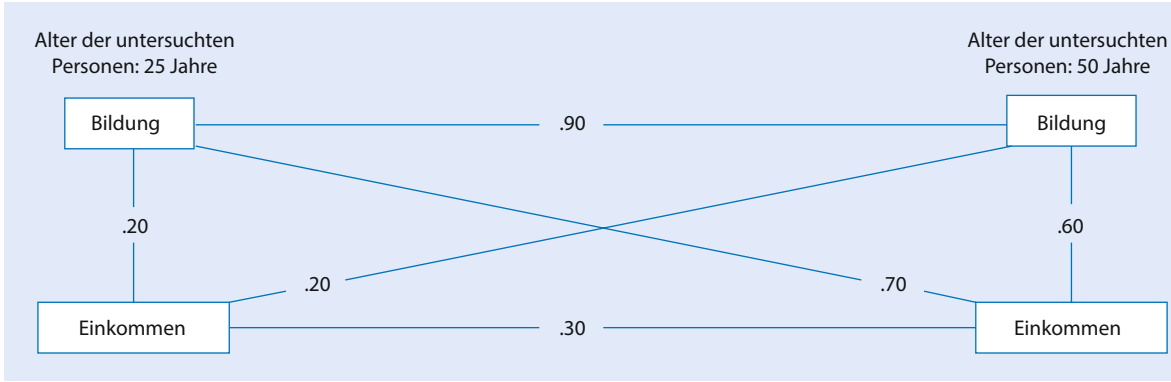


Abb. 12.33 Cross-Lagged-Panel-Design. (Mod. nach Spector, 1981. Republished with permission of SAGE Publications, from Research Designs, Spector, P.E., Copyright 2014; permission conveyed through Copyright Clearance Center, Inc.)

Korrelieren die an einer Stichprobe von Vorschulkindern erhobenen Testwerte eines Schulreifetests mit den späteren schulischen Leistungen derselben Kinder, scheidet die kausale Erklärungsalternative „Die schulischen Leistungen beeinflussen die Ergebnisse im Vorschultest" aus. Der umgekehrte Erklärungsansatz, die schulischen Leistungen hingen von der Schulreife ab, ist mit dieser Korrelation jedoch keineswegs gesichert.

Die gleiche Korrelation wäre auch zu erwarten, wenn der Zusammenhang beider Merkmale auf ein drittes Merkmal (z. B. kognitive und sprachliche Förderung durch die Eltern im Vorschulalter und im Schulalter) zurückgeht (Abb. 12.30, Modell e) oder wenn sich die Vorschultestergebnisse nur indirekt auf die schulischen Leistungen auswirken (z. B. über die Erwartungshaltungen der Lehrer, die Kinder mit guten Testergebnissen mehr fördern als Kinder mit schlechten Testergebnissen; Abb. 12.30, Modell d). Dennoch kann man davon ausgehen, dass die interne Validität von Korrelationsstudien über zeitlich versetzt erhobene Merkmale (Längsschnittuntersuchung) in der Regel höher ist als die interne Validität von Korrelationsstudien, die dieselben Merkmale zu einem Zeitpunkt prüfen (Querschnittuntersuchung).

In der zeitreihenanalytischen Forschung (▶ Abschn. 12.5.4) wird der Kausalitätsbegriff häufig durch das Konzept der Prognostizierbarkeit im Sinne der sog. **Wiener-Granger-Kausalität** ersetzt (Schmitz, 1989). Diese wird unterstützt, wenn die zukünftige Ausprägung einer Variablen $Y(y_{t+1})$ umso besser vorhergesagt werden kann, je mehr frühere Ausprägungen einer Variablen $X(x_t, x_{t-1} \ldots)$ berücksichtigt werden.

Cross-Lagged-Panel-Design

Die Idee, dass konkurrierende Kausalmodelle in korrelativen Längsschnittuntersuchungen unterschiedliche Plausibilität aufweisen, wurde von Campbell (1963) bzw. Pelz und Andrews (1964) aufgegriffen und zu einem eigenständigen Untersuchungstyp ausgebaut: dem Cross-Lagged-Panel-Design. In Abb. 12.33 (nach Spector, 1981) wird der Aufbau dieses Korrelationsstudientyps an einem Beispiel verdeutlicht.

In diesem Beispiel konkurrieren die Kausalhypothesen „Die Bildung beeinflusst das Einkommen" und „Das Einkommen beeinflusst die Bildung". Hierzu wurde eine Stichprobe wiederholt bzgl. der Merkmale „Bildung" und „Einkommen" untersucht: einmal im Alter von 25 Jahren und ein zweites Mal im Alter von 50 Jahren. Damit ergeben sich **sechs mögliche Korrelationen**: 2 Korrelationen eines jeden Merkmals mit sich selbst, gemessen zu zwei Zeitpunkten **(Autokorrelationen)**, 2 Korrelationen zwischen den zwei verschiedenen, zeitversetzt gemessenen Merkmalen **(verzögerte Kreuzkorrelationen)** und 2 Korrelationen zwischen zwei verschiedenen, gleichzeitig gemessenen Merkmalen **(synchrone Korrelationen)**. Die vier zuletzt genannten Korrelationen sind für die Entscheidung, welcher der beiden Kausalhypothesen der Vorzug zu geben sei, besonders wichtig.

Vertritt man die Hypothese, dass die Bildung das Einkommen bestimmt, das Einkommen jedoch die Bildung nur schwach beeinflusst, würde man bei den verzögerten Kreuzkorrelationen zwischen der Bildung mit 25 Jahren und dem Einkommen mit 50 Jahren eine hohe und zwischen dem Einkommen mit 25 Jahren und der Bildung mit 50 Jahren eine niedrige Korrelation erwarten. Gleichzeitig müssten bei den synchronen Korrelationen die Merkmale Bildung und Einkommen mit 50 Jahren höher korrelieren als mit 25 Jahren. Mit 25 Jahren konnte die Bildung das Einkommen erst wenig beeinflussen. Mit 50 Jahren hingegen ist das Einkommen „bildungsgemäß".

Diese Zahlenverhältnisse sind in Abb. 12.33 wiedergegeben. Die Untersuchung favorisiert also die Hypothese „Bildung beeinflusst das Einkommen" gegenüber der Hypothese „Das Einkommen beeinflusst die Bildung." Es muss jedoch betont werden, dass auch diese Untersu-

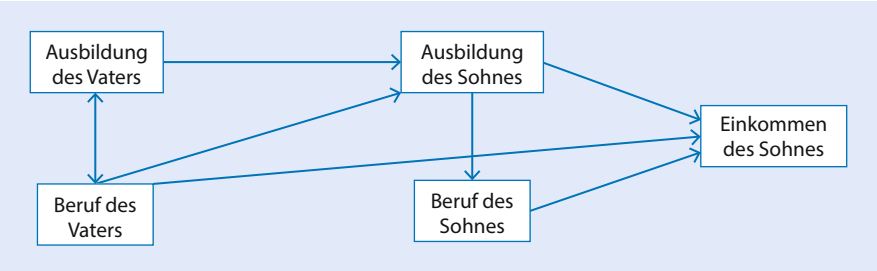


Abb. 12.34 Beispiel für ein pfadanalytisches Kausalmodell. (Mod. nach Spaeth, 1975. Bildrechte: McCatchan, Berkeley, CA)

chungsart weitere kausale Erklärungen nicht ausschließt. Sie entscheidet „lediglich" über die relative Plausibilität von zwei konkurrierenden Kausalhypothesen.

Die interne Validität eines Cross-Lagged-Panel-Designs lässt sich durch die Einbeziehung von mehr als zwei Messzeitpunkten erhöhen. Hierüber und über weitere Modifikationen dieses Untersuchungstyps berichten z. B. Cook und Campbell (1976) sowie Kenny und Harackiewicz (1979; zur Kritik s. Rogosa, 1980, 1995).

Mit dem auf Stichprobenuntersuchungen zugeschnittenen Cross-Lagged-Panel-Design nicht zu verwechseln ist die Prüfung von Einzelfallhypothesen mittels „Lag Sequential Analysis". Hierbei handelt es sich um eine Analyse dyadischer Interaktionsprozesse mit dem Ziel, die wechselseitige Bestimmtheit zweier aufeinander bezogener Verhaltenssequenzen zu quantifizieren (z. B. Analyse von Schüler-Lehrer-Interaktionen, um herauszufinden, welches Lehrerverhalten störendes Schülerverhalten begünstigt oder unterbindet; Gunter, Jack, Shores, Carrell, & Flowers, 1993). Literatur und Auswertungstechniken zu diesem Verfahren findet man bei Faraone und Dorfman (1987) oder Schmitz, Kruse, & Tasche (1985).

Pfadanalyse

Die Widerlegung komplexer Kausalmodelle ermöglicht ein Verfahren, dessen Grundzüge auf Wright (1921) zurückgehen und das heute unter der Bezeichnung Pfadanalyse („path analysis") bekannt ist (Bentler, 1980; Blalock, 1971; Weede, 1970). Ein Kausalmodell, das Gegenstand einer pfadanalytischen Untersuchung sein könnte, zeigt Abb. 12.34 (nach Spaeth, 1975).

Die hier wiedergegebenen Kausalhypothesen lassen sich verkürzt folgendermaßen formulieren: Es geht darum, die Höhe des Einkommens männlicher Personen zu erklären. Es wird behauptet, dieses sei von der Ausbildung und dem Beruf der Person sowie dem Beruf des Vaters abhängig. Die Ausbildung des Sohnes, die ihrerseits von der Ausbildung und dem Beruf des Vaters abhängt, beeinflusst den Beruf des Sohnes etc.

Wir wollen auf die einzelnen Schritte einer pfadanalytischen Überprüfung dieses Modells verzichten. Es sei jedoch darauf hingewiesen, dass im **Pfadmodell** („path model", „path analytic model") die Enge und Richtung der Pfade durch sog. **Pfadkoeffizienten** ausgedrückt werden, die im Zuge der Prüfung entsprechender kausaler Zusammenhangshypothesen auf Signifikanz zu testen sind, wobei – wie bei jedem Signifikanztest – für ausreichende Teststärke zu sorgen ist. Man beachte, dass Pfadkoeffizienten nicht identisch sind mit Regressionskoeffizienten (wohl aber aus ihnen berechnet werden können; ausführlicher hierzu Bortz & Schuster, 2010, Kap. 24; Eid et al., 2010, Kap. 25). Keinesfalls dürfen in ein grafisches Pfadmodell wie Abb. 12.34 bivariate Regressions- oder gar Korrelationskoeffizienten eingetragen werden, da diese von den eigentlichen Pfadkoeffizienten, die das Gesamtgefüge der Korrelationen der beteiligten Variablen mitberücksichtigen, stark abweichen (für eine Pfadanalyse der Zusammenhänge zwischen Geschlecht und Verhalten von Schülern sowie ihrer Leistungsfähigkeit einerseits und ihren Schulnoten bzw. Beurteilungen durch Lehrer andererseits s. Bennett, Gottesman, Rock, & Cerullo, 1993).

Die Logik, nach der durch Korrelations- bzw. Pfadanalysen kausale Zusammenhangsmodelle **widerlegt** werden können, sei noch einmal anhand der in Abb. 12.30 dargestellten Modelle verdeutlicht: Trivialerweise sind alle in Abb. 12.30 aufgeführten Modelle widerlegt, wenn die Korrelation der Merkmale X und Y unbedeutend ist. Korrelieren diese Merkmale jedoch substanziell, können alle 6 Modelle (und weitere, in Abb. 12.30 nicht wiedergegebene Modelle) richtig sein. Eine Differenzierung zwischen den Modellen allein aufgrund einer substanziellen (d. h. statistisch signifikanten) Korrelation r_{xy} ist nicht möglich.

Bei substanzieller Korrelation r_{xy} scheiden jedoch die Modelle d und e aus, wenn die Partialkorrelation $r_{xy\cdot z}$ (die Korrelation zwischen X und Y, aus der Z herauspartialisiert wurde) gegenüber der Korrelation r_{xy} praktisch unverändert ist. Sie gelten als vorläufig bestätigt, wenn $r_{xy\cdot z}$ unbedeutend wird. Dies heißt gleichzeitig, dass dann auch die Modelle a, b und c ausscheiden.

Im Widerspruch zu Modell f steht entweder eine bedeutende Partialkorrelation $r_{xz\cdot w}$ (bei gleichzeitig hoher Korrelation r_{xz}) und/oder eine bedeutende Partialkorrelation $r_{xy\cdot z}$. Dieses Modell wird unterstützt, wenn sowohl

das Herauspartialisieren von W aus r_{xz} als auch das Herauspartialisieren von Z aus r_{xy} die Korrelation r_{xz} bzw. r_{xy} nicht verändern. Man bedenke jedoch, dass diese Korrelationskonstellation z. B. auch ein Modell bestätigen würde, bei dem Z durch Y und W durch Z kausal beeinflusst werden, also ein Modell, bei dem in Abb. 12.30 (Modell f) die Pfeilrichtungen zwischen Z und Y bzw. W und Z umgekehrt sind.

Erneut zeigt sich also, dass Kausalhypothesen korrelations- bzw. regressionsanalytisch (und damit auch pfadanalytisch) zu widerlegen, aber nicht eindeutig zu bestätigen sind. Stehen die empirischen Korrelationen zu einem Kausalmodell nicht im Widerspruch, heißt dies nicht, dass dieses Kausalmodell tatsächlich der Realität entspricht. Dieser Schluss wäre nur zulässig, wenn sich die korrelativen Zusammenhänge durch keine weiteren Kausalmodelle erklären ließen. Weitere Rahmenbedingungen für Kausalität erörtern Edwards und Bagozzi (2000). Wie man sich jedoch leicht anhand von Abb. 12.30 überzeugen kann (indem man z. B. die Pfeilrichtungen ändert oder neue Pfeile einfügt), lassen sich zu einem Korrelationsgefüge mühelos mehrere Kausalmodelle konstruieren, über deren relative Plausibilität die Korrelationen allein nichts aussagen. (Ein eindrucksvolles Beispiel für eine Fehlinterpretation eines pfadanalytischen Ergebnisses findet man bei Stelzl, 1982, Kap. 9.3.3.).

Neben der Tatsache, dass auch die Widerlegung eines Kausalmodells Erkenntnisfortschritt bedeutet, verbindet sich mit pfadanalytischen Ansätzen der Vorteil, dass sie – anders als einfache Korrelationsstudien – die Forschenden dazu zwingen, sich über Ursache-Wirkungs-Verhältnisse Gedanken zu machen bzw. kausale Modelle zu konstruieren. Prüfungstechnisch kann man die Pfadanalyse als einen Vortest ansehen, der relativ einfach durchzuführen ist und den man häufig einsetzt, bevor man – wenn dies möglich ist – eine Kausalhypothese gezielt mit Untersuchungen überprüft, die eine höhere interne Validität aufweisen als Zusammenhangsanalysen.

Eine Zusammenfassung der Kontroversen zum Thema „Pfadanalyse“ findet man bei Meehl und Waller (2002). Die Autoren haben zudem eine Technik vorgeschlagen, mit der rivalisierende Pfadmodelle erzeugt und bzgl. ihrer Kompatibilität mit einer vorgegebenen Korrelationsmatrix geprüft werden können („Delete One, Add One Rule“; D1-A1-Rule). Arbeiten, die sich mit diesem Ansatz kritisch auseinandersetzen, werden bei Waller und Meehl (2002) kommentiert.

Lineare Strukturgleichungsmodelle

Weiterentwicklungen der Pfadanalyse überprüfen nicht nur Annahmen bzgl. der wechselseitigen Kausalbeziehungen der untersuchten Merkmale, sondern zusätzlich Hypothesen, die sich auf latente, nicht direkt beobachtbare Merkmale bzw. deren Beziehungen untereinander und zu den untersuchten Merkmalen beziehen. Derartige Strukturgleichungsmodelle (**„structural equation modelling“**, SEM) werden in Kap. 17 genauer dargestellt. Wie die Pfadanalyse erfordern auch lineare Strukturgleichungsmodelle, dass sich die Forschenden vor Untersuchungsbeginn sehr genau überlegen, zwischen welchen Variablen bzw. Konstrukten kausale Beziehungen oder kausale Wirkketten bestehen könnten. In der Forschungsliteratur sind Auswertungen mit Strukturgleichungsmodellen mittlerweile sehr viel häufiger zu finden als klassische Pfadanalysen.

Kausale Mikromediatoren

Wenn eine Untersuchung zeigt, dass eine Maßnahme oder ein Treatment wirkt, ist damit keineswegs geklärt, was die tatsächlichen Wirkmechanismen waren, die beim Individuum das erwartete Verhalten auslösten. Die Dekomposition globaler Wirkprozesse in „kausale Mikromediatoren“ kann hier Abhilfe schaffen und die Überprüfung kausaler Hypothesen erheblich präzisieren. Man versucht hierbei – z. B. durch qualitative Interviews (▶ Abschn. 10.2.2) – die eigentlichen, vom Individuum erlebten Ursachen des (vermeintlich) durch das Treatment ausgelösten Verhaltens zu ergründen. Hat man derartige Mikromediatoren ausfindig gemacht, können Fragen nach der Generalisierbarkeit des kausalen Effekts (unter welchen Umständen ist damit zu rechnen, dass das Individuum ähnlich reagiert wie in der konkreten Untersuchung?) sehr viel leichter beantwortet werden.

Beispiel: Cook und Shadish (1994) berichten in diesem Zusammenhang über einen öffentlich begangenen Mord, der von vielen Schaulustigen tatenlos hingenommen wurde; sie erklären dieses Phänomen mit dem von Latane und Darley (1970) eingeführten Konzept der Verantwortungsdiffusion („diffusion of responsibility“): Keiner hilft, weil alle denken, andere müssten helfen. Diese kausale Erklärung, die aus einer genauen Analyse dessen hervorging, was die Schaulustigen beim Anblick der Tat dachten und erlebten, erwies sich als kausaler Mikromediator für unterlassene Hilfeleistungen in vielen vergleichbaren Situationen als sehr tragfähig. Weitere Informationen zu dieser Thematik findet man bei Shadish (1996).

Kausalmodelle in der Evaluationsforschung

Bei der Bewertung von Praxismaßnahmen im Zuge der Evaluationsforschung geht es nicht nur darum nachzuprüfen, **ob** die Maßnahmen die erwarteten Wirkungen zeigen (und keine unerwünschten Nebenwirkungen haben), sondern auch darum, die Wirkmechanismen im Detail zu verstehen und ggf. zu optimieren. Evaluationsansätze,

die sich detailliert mit den Ursache-Wirkungs-Ketten von Maßnahmen befassen, sie konzeptuell modellieren und empirisch prüfen, werden in die Gruppe der **theorieorientierten Evaluationsansätze** eingeordnet (▶ Abschn. 18.2.3). Hierbei kommen Zusammenhangsanalysen, aber auch andere hypothesenprüfende Verfahren zum Einsatz.

▪ Metaanalyse

Für die Stützung einer Kausalhypothese bzw. zur Klärung der Frage, inwieweit eine kausale Beziehung generalisierbar ist, sind wiederholte Prüfungen der gleichen Kausalhypothese von großem Wert. Da Replikationen eine prototypische Untersuchung niemals deckungsgleich nachstellen können (andere Untersuchungspersonen, andere Untersuchungszeitpunkte, ggf. modifizierte Untersuchungsbedingungen und Messinstrumente etc.), erfährt man aus den Ergebnissen vergleichbarer Untersuchungen, wie stark der geprüfte Kausaleffekt ist, bei welchen Subgruppen oder Untersuchungsbedingungen er auftritt bzw. auch, in welcher Weise seine Generalisierung eingeschränkt ist. Liegen also eine ganze Reihe von Studien vor, die eine kausale Zusammenhangshypothese mit intern validen Designs überprüfen und deren korrelations-/regressionsanalytische Auswertungen zu statistisch signifikanten Ergebnissen führen, so ist dies wesentlich überzeugender als ein einzelnes Studienergebnis. Entsprechenden Überblick vermitteln Metaanalyen, welche die statistischen Ergebnisse vergleichbarer Studien zu einem Gesamtwert und einer Gesamtinterpretation zusammenfassen (zu Metaanalysen ▶ Kap. 16).

Hinweis: Weitere Informationen zur Überprüfung kausaler Hypothesen findet man z. B. bei van Koolwijk und Wieken-Mayser (1986) sowie McKim und Turner (1997). Über formale Randbedingungen, die erfüllt sein müssen, um Regressionsmodelle kausal interpretieren zu können, berichtet Steyer (1992, 2003, Teil III).

Lernquiz

1. Welche interne Validität weisen Interdependenzanalysen auf und wie ist diese im Hinblick auf kausale Interpretationen zu bewerten?
2. Lässt sich durch einen Korrelationskoeffizienten von $r = .58$ auf einen starken Zusammenhang in der Population schließen?
3. Die multiplen Regressionskoeffizienten (d. h. die Beta-Gewichte einer multiplen Regression) entsprechen den bivariaten Korrelationskoeffizienten zwischen Prädiktoren und Kriterium. Diskutieren Sie diese Aussage!
4. Was stimmt nicht?
   a) Das Ausschließen von Ausreißerwerten vor der Datenanalyse ist nicht zwingend notwendig, da Korrelationskoeffizienten durch diese kaum verzerrt werden.
   b) Eine kleine homogene Teilstichprobe aus der Population kann den wahren Zusammenhang unterschätzen.
   c) Die Extremgruppenselektion führt zu einer Unterschätzung des Zusammenhanges.
   d) Zur Prüfung von Zusammenhangshypothesen sollten Stichproben gezogen werden, die die Population möglichst gut repräsentieren.
5. Grenzen Sie die Partialkorrelation von der Semipartialkorrelation ab!
6. Was trifft zu?
   a) Multiple Zusammenhangshypothesen beziehen sich auf Zusammenhänge zwischen zwei kardinalskalierten Variablen.
   b) Multiple Zusammenhangshypothesen spielen eine untergeordnete Rolle in der Hypothesenprüfung.
   c) Die Überprüfung multipler Zusammenhangshypothesen erfolgt in der Regel über die multiple Korrelations- und Regressionsanalyse.
   d) In der sozialwissenschaftlichen Forschung sind lediglich lineare Zusammenhänge von Interesse.
   e) Multiple Korrelation und kanonische Korrelation sind Synonyme.
   f) Multiple Zusammenhänge sollten mittels multivariater Verfahren überprüft werden und nicht über eine Mehrfachberechnung von bivariaten Verfahren.
7. Wie ist ein Cross-Lagged-Panel-Design aufgebaut und wozu braucht man es?
8. Was versteht man unter einer Pfadanalyse?
9. Grenzen Sie Moderator- und Mediatorvariable voneinander ab!
10. Welche Rolle spielen Metaanalysen bei der Überprüfung von kausalen Zusammenhangshypothesen?

Übungsaufgaben

1. Sie möchten überprüfen, ob sich die Studienrichtung auf den Berufserfolg (operationalisiert über das monatliche Einkommen in Euro) auswirkt. Der Einfachheit halber nehmen wir an, pro Studienrichtung wurden 5 Personen befragt. Führen Sie für den nachfolgend dargestellten Datensatz eine Dummy-Kodierung durch: Wirtschaftswissenschaften (25 000; 1 500; 500; 2 800; 3 120); Sozialwissenschaften (3 200; 1 600; 8 000; 1 200; 1 000); Ingenieurwissenschaften (2 500; 2 300; 27 000; 2 500; 7 600); Medizin (1 500; 2 350; 10 115; 9 800; 3 210).

2. Welches Verfahren würden Sie zur Auswertung wählen? Welche Variablenarten sind im Datensatz vorhanden? Welche weiteren Variablen sind Ihrer Meinung nach sinnvoll, um der oben genannten Fragestellung detailliert nachzugehen?
3. Erstellen Sie grafisch ein Pfadmodell, das die Zusammenhänge zwischen Studienrichtung, Studienabschlussnote, Einkommen und Reputation des Studienortes darstellt.

Weiterführende Literatur

- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.
- Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont: Wadsworth, Cengage Learning.
- Kuckartz, U., Rädiker, S., Ebert, T. & Schehl, J. (2010). *Statistik. Eine verständliche Einführung*. Wiesbaden: VS.
- Nachtigall, C. & Wirtz, M. (2009). *Wahrscheinlichkeitsrechnung und Inferenzstatistik. Statistische Methoden für Psychologen* – Band 2 (5. Aufl.). Weinheim: Juventa.

Literatur

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Alf, E. F., Jr. & Abrahams, N. M. (1975). The use of extreme groups in assessing relationships. *Psychometrika, 40*(4), 563–572.

Andreß, H. J., Hagenaars, J. A., & Kühnnel, S. (1997). *Analyse von Tabellen und kategorialen Daten*. Heidelberg: Springer.

Arminger, G. (1982). *Klassische Anwendungen verallgemeinerter linearer Modelle in der empirischen Sozialforschung. ZUMA Arbeitsberichte, Nr. 1982/03*. Mannheim.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical consideration. *Journal of Personality and Social Psychology, 51*(6), 1173–1182.

Beck-Bornholdt, H. P. & Dubben, H. H. (2001). *Der Hund, der Eier legt. Erkennen von Fehlinformationen durch Querdenken*. Reinbeck: Rowohlt.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*(2), 347–356.

Benninghaus, H. (1989). *Deskriptive Statistik. Statistik für Soziologen, Bd. 1* (6. Aufl.). Stuttgart: Teubner.

Benninghaus, H. (1998). *Einführung in die sozialwissenschaftliche Datenanalyse* (5. Aufl.). München: Oldenbourg.

Bentler, P. M. (1980). Multivariate analysis with variables: Causal modeling. *Annual Review of Psychology, 31*, 419–456.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge: MIT Press.

Blalock, H. M. (Ed.). (1971). *Causal models in the social sciences*. London: MacMillan.

Bortz, J. & Lienert, G. A. (2003). *Kurzgefaßte Statistik für die klinische Forschung. Ein praktischer Leitfaden für die Analyse kleiner Stichproben* (2. Aufl.). Heidelberg: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2000). *Verteilungsfreie Methoden in der Biostatistik* (2. Aufl.). Heidelbeberg: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer

Campbell, D. T. (1963). From description to experimentation: Interpreting trends as quasi-experiements. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 212-254). Madison: University of Wisconsin Press.

Clausen, S. E. (1998). *Applied correspondence analysis. An introduction*. Thousand Oaks: Sage.

Cook, T. D. & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), *Handbook of industrial and organizational research* (pp. 223–326). Chicago: Rand McNally.

Cook, T. D. & Shadish, W. R. (1994). Social experiments: Some developments over the past fiffteen years. *Annual Review of Psychology, 45*, 548–580.

Dawes, R. M., Faust, D., & Meehl, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues* (pp. 351–367). Hillsdale: Erlbaum.

Digman, J. M. (1990). „Personaliy structure: Emergence of the five-factor model". *Annual Review of Psychology, 41*, 417–440.

Draper, N. & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Edwards, J. R. & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–174.

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.

Eye, A. von (1990). *Introduction to configural frequency analysis: The search for types and antitypes in cross-classification*. Cambridge: Cambridge University Press.

Eye, A. von & Schuster, C. (1998). *Regression analysis for social sciences*. San Diego: Academic Press.

Eysenck, H. J. (1969). *Personality structure and measurement*. London: Routledge & Paul.

Faraone, S. V. & Dorfman, D. D. (1987). Lag sequential analysis: Robust statistical methods. *Psychological Bulletin, 101*(2), 312–323.

Fasel, N., Green, E. G. T., & Sarrasin, O. (2013). Unveiling naturalization: A multilevel study on minority proportion, conservative ideologies, and attitudes toward the Muslim veil. *Zeitschrift für Psychologie, 221*(4), 242–251.

Gaensslen, H. & Schubö, W. (1973). *Einfache und komplexe statisische Analyse*. München: Reinhardt.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.

Greenacre, M. J. (1993). *Correspondence analysis in practice*. London: Academic Press.

Gunter, P. L., Jack, S. L., Shores, R. E., Carrell, D. E., & Flowers, J. (1993). Lag sequential analysis as a tool for functional analysis of student disruptive behavior in classrooms. *Journal of Emotional and Behavioral Disorders, 1*(3), 138–148.

Hox, J. J. (2002). *Multilevel analysis. Techniques and Applications*. Mahwah, NJ: Earlbaum.
Hox, J. J. & Roberts, J. K. (Eds.). (2010). *Handbook of advanced multilevel analysis*. New York, NY: Routledge.
Hoyt, W. T., Imel, Z. E., & Chan, F. (2008). Multiple regression and correlation techniques: Recent Controversies and best practices. *Rehabilitation Psychology, 53*(3), 321–339.
Jäger, R. (1974). Zur Gültigkeit von Aussagen, die auf korrelationsstatistischen Verfahren beruhen. *Archiv für Psychologie, 126*(2–4), 253–264.
Kenny, D. A. & Harackiewicz, J. M. (1979). Cross-lagged panel correlation. Practice and promise. *Journal of Applied Psychology, 64*(4), 372–379.
Van Koolwijk, J. & Wieken-Mayser, M. (Hrsg.). (1986). *Techniken der empirischen Sozialforschung. Bd. 8 Kausalanalyse.* München: Oldenbourg.
Köbben, A. (1970). Cause and intention. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology* (pp. 89–98). Garden City: Natural History Press.
Krämer, W. (1995). *So lügt man mit Statistik*. Frankfurt: Campus.
Krauth, J. (1993). *Einführung in die Konfigurationsfrequenzanalyse (KFA)*. Weinheim: Beltz.
Krauth, J. & Lienert, G. A. (1973). *Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendungen in Psychologie und Medizin*. Freiburg: Alber.
Kubinger, K. D. (1990). Übersicht und Interpretation der verschiedenen Assoziationsmaße. *Psychologische Beiträge, 22*, 290–346.
Langeheine, R. (1980). Multivariate Hypothesentestung bei qualitativen Daten. *Zeitschrift für Sozialpsychologie, 11*, 140–151.
Latane, B. & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton Crofts.
Lehmann, G. (1980). Nichtlineare „Kausal"- bzw. Dominanz-Analysen in psychologischen Variablensystemen. *Zeitschrift für experimentelle und angewandte Psychologie, 27*(2), 257–276.
MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Earlbaum.
McKim, V. R. & Turner, S. P. (Eds.). (1997). *Counseling in crisis? Statistical methods in the search for causal knowledge in the social sciences*. Notre Dame: University of Notre Dame Press.
Meehl, P. E. (1950). Configural scoring. *Journal of Consulting Psychology, 14*(3), 165–171.
Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneanapolis: University of Minnesota Press.
Meehl, P. E. & Waller, N. G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods, 7*(3), 283–300.
Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika, 57*(4), 539–565.
Moosbrugger, H. (1978). *Multivariate statistische Analyseverfahren*. Stuttgart: Kohlhammer.
Moosbrugger, H. (2002). *Lineare Modelle. Regressions- und Varianzanalysen* (3. Aufl.). Bern: Huber.
Mulaik, S. A. (1975). Confirmatory Factor Analysis. In D. J. Amick & H. J. Walberg (Eds.), *Introductory multivariate analysis* (pp. 170–207). Berkeley: McCutchan.
Ott, J. C. (2011). Government and happiness in 130 nations: Good governance fosters higher level and more equality of happiness. *Social Indicators Research, 102*(1), 3–22.
Overall, J. E. & Klett, C. J. (1972). *Applied multivariate analysis*. New York: McGraw Hill.
Paulhus, D. L. & Williams, K. M. (2002). The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality, 36*(6), 556–563.
Pelz, D. C. & Andrews, F. M. (1964). Detecting causal priorities in panel study data. *American Sociological Review, 29*(6), 836–848.
Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178–192.
Rochel, H. (1983). *Planung und Auswertung von Untersuchungen im Rahmen des allgemeinen linearen Modells*. Heidelberg: Springer.
Rogosa, D. R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88*(2), 245–258.
Rogosa, D. R. (1995). Myths and methods: „Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah: Erlbaum.
Schäfer, B. (1983). Semantische Differentialtechnik. In H. Feger & J. Bredenkamp (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich B, Serie I, Bd. 2, Datenerhebung* (S. 154–221). Göttingen: Hogrefe.
Schlosser, O. (1976). *Einführung in die sozialwissenschaftliche* Zusammenhang*sanalyse*. Reinbek: Rowohlt.
Schmitz, B. (1989). *Einführung in die Zeitreihenanalyse. Modelle, Softwarebeschreibung, Anwendung*. Bern: Huber.
Schmitz, B., Kruse, F. O., & Tasche, K. G. (1985). Anwendung zeitreihenanalytischer Verfahren bei prozeßorientierter Paardiagnostik. In H. Appelt & B. Strauß (Hrsg.), *Ergebnisse einzelfallstatistischer Untersuchungen* (S. 84–113). Berlin: Springer.
Selg, H. (1971). *Einführung in die experimentelle Psychologie*. Stuttgart: Kohlhammer.
Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes. A primer of examples, methods and issues. *Psychological Methods, 1*(1), 47–65.
Sievers, W. (1977). Über Dummy-Variablen-Kodierung in der Varianzanalyse. *Psychologische Beiträge, 19*, 454–462.
Silbereisen, R. K. (1977). Prädiktoren der Rollenübernahme bei Kindern. *Psychologie in Erziehung und Unterricht, 24*, 86–92.
Snijders, T. A. B. & Bosker, R. (1993). Standard errors and sample sizes for two-level research. [doi:10.2307/1165134]. *Journal of Educational Statistics, 18*(3), 237–259.
Snijders, T. A. B. & Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.
Spaeth, J. L. (1975). Path analysis. In D. J. Amick & H. J. Walberg (Eds.), *Introductory multivariate analysis* (pp. 53–89). Berkeley: McCutchan.
Spector, P. E. (1981). *Research designs*. Thousand Oaks: Sage.
Stelzl, I. (1982). *Fehler und Fallen der Statistik*. Bern: Huber.
Steyer, R. (1992). *Theorie kausaler Regressionsmodelle*. Stuttgart: Fischer.
Steyer, R. (2003). *Wahrscheinlichkeit und Regression*. Heidelberg: Springer.
Ulrich, R. & Wirtz, M. (2004). On the correlation of a naturally and an artificially dichotomized variable. *British Journal of Mathematical and Statistical Psychology, 57*(2), 235–251.
Waller, N. G. & Meehl, P. E. (2002). Risky tests, versimilitude, and path analysis. *Psychological Methods, 7*(3), 323–337.
Weber, R. (2000). *Prognosemodelle zur Vorhersage der Fernsehnutzung. Neuronale Netze, Tree-Modelle und klassische Statistik im Vergleich*. München: Fischer.
Weede, E. (1970). Zur Methodik der kausalen Abhängigkeitsanalyse (Pfadanalyse) der nichtexperimentellen Forschung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 22*(3), 532–550.
Werner, J. (1997). *Lineare Statistik. Allgemeines lineares Modell*. Weinheim: Psychologie Verlags Union.
Williams, K. M., Nathanson, C., & Paulhus, D. L. (2010). Identifying and profiling scholastic cheaters: Their personality, cognitive ability,

and motivation. *Journal of Experimental Psychology: Applied, 16*(3), 293–307.
Wolf, G. & Cartwright, B. (1974). Rules for coding dummy variables in multiple regression. *Psychological Bulletin, 81*(3), 173–179.
Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*(7), 557–585.
Yarnold, P. R. (1996). Characterizing and cirumventing simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement, 56*(3), 430–442.
Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46*(4), 357–387.

12.5.3 Signifikanztests für Unterschiedshypothesen

Lernziele

- Verschiedene Verfahren zur statistischen Überprüfung von Unterschiedshypothesen kennen.
- Geeignete statistische Auswertungsverfahren für experimentelle, quasi-experimentelle und nichtexperimentelle Designs auswählen können.
- Ergebnisse der statistischen Überprüfung von Unterschiedshypothesen richtig interpretieren können.
- Das Konzept von Interaktionseffekten verstehen und Interaktionseffekte interpretieren können.
- Das Konzept von Kontrollvariablen verstehen und anwenden können.

Fragen wie „Hat diese Maßnahme eine Wirkung?“ oder „Welchen Effekt löst diese Behandlung aus?“ werden in der Grundlagen- und Evaluationsforschung häufig gestellt. Eine Strategie, derartige Fragen zu untersuchen, wäre genauso naheliegend wie falsch: Man führt die Maßnahme ein bzw. die Behandlung durch (wir wollen im Folgenden die hierfür übliche englischsprachige Bezeichnung **Treatment** übernehmen) und dokumentiert den Zustand der untersuchten Personen im Anschluss an das Treatment. Diese Vorgehensweise (bzw. Varianten hiervon) führt jedoch zu uneindeutigen Ergebnissen, denn man kann niemals sicher sein, ob die registrierten Effekte nicht auf andere Ursachen als das Treatment zurückgehen bzw. ob die vermeintliche Treatmentwirkung auch ohne das eigentliche Treatment eingetreten wäre. Allgemein formuliert lassen die Ergebnisse viele Interpretationen zu, d. h., derartige Eingruppenpläne („one shot case studies“) sind durch eine geringe interne Validität (▶ Abschn. 3.2.2) gekennzeichnet.

Dieses Teilkapitel befasst sich mit der statistischen Datenanalyse für Untersuchungspläne, welche die eingangs gestellten Forschungsfragen bzw. Forschungshypothesen (denn in der Regel wird eine konkrete Maßnahmenwirkung im Rahmen einer Unterschiedsphypothese prognostiziert) präziser bzw. eindeutiger beantworten. Charakteristisch für diese Untersuchungspläne ist der Vergleich zweier (oder mehrerer) Stichproben, die sich in Bezug auf eine (oder mehrere) unabhängige Variable(n) unterscheiden. Im einfachsten Fall, der auf viele Varianten derartiger Forschungsfragen anwendbar ist, hat man nur eine unabhängige Variable mit zwei Stufen: behandelt vs. nicht behandelt. Die Untersuchung liefe damit auf den Vergleich zweier Gruppen, nämlich einer behandelten „Treatment-Gruppe“ und einer nicht behandelten „Kontrollgruppe“ hinaus. Unterscheiden sich diese beiden Gruppen in Bezug auf eine abhängige Variable im Anschluss an das Treatment, ist damit eine Treatmentwirkung sehr viel besser belegt als mit einem Eingruppenplan.

Hypothesen, die sich auf die Wirksamkeit einer Maßnahme oder eines Treatments beziehen, sollten als *Unterschiedshypothesen* oder – wenn Messwiederholungen an denselben Untersuchungsobjekten möglich sind – als Veränderungshypothesen formuliert und statistisch geprüft werden.

Unterschiedshypothesen können auf vielfältige Weise geprüft werden, wobei es sich stets um Gruppenvergleiche handelt, dabei werden die Gruppen durch die Ausprägungen einer oder mehrerer nominalskalierter unabhängiger Variablen gebildet. Die Maßnahmenwirkung wird dagegen in der Regel auf kardinalskalierten abhängigen Variablen gemessen. Die höchste interne Validität und Aussagekraft besitzen experimentelle Studien, bei denen die unterschiedlich behandelten Untersuchungsgruppen per Randomisierung zusammengestellt (Kontrolle personenbezogener Störvariablen) und zudem untersuchungsbedingte Störvariablen strikt kontrolliert werden (▶ Abschn. 7.6). Eine etwas geringere interne Validität besitzen quasi-experimentelle Designs, bei denen die unterschiedlich behandelten Untersuchungsgruppen nicht zufällig zusammengestellt wurden (▶ Abschn. 7.6.2). Das Untersuchungsdesign hat wesentlichen Einfluss darauf, wie sich ein statistisches Ergebnis (sei es signifikant oder nicht signifikant) inhaltlich hinsichtlich der zu prüfenden Unterschiedshypothese interpretieren lässt.

Wir behandeln im Folgenden Zweigruppenpläne, Mehrgruppenpläne, faktorielle Pläne, hierarchische Pläne, quadratische Pläne, Pläne mit Kontrollvariablen sowie multivariate Pläne (Tab. 12.19). Bezüglich der statistischen Auswertung werden wir uns auch in diesem Teilkapitel mit Hinweisen auf einschlägige Signifikanztests begnügen (Indikationsfrage), für Berechnungsdetails wird auf Statistiklehrbücher verwiesen.

Alle genannten Untersuchungspläne zur Prüfung von Unterschiedshypothesen lassen sich durch das Hinzunehmen eines Zeit- bzw. Messwiederholungsfaktors (z. B. Vorher-Nachher-Messung) erweitern. Was beim Faktor Zeit zu beachten ist, wird im Zusammenhang mit Signifikanztests für **Veränderungshypothesen** (▶ Abschn. 12.5.4) behandelt. Insbesondere in der Evaluationsforschung wird beim statistischen Nachweis von Maßnahmenwirkungen meist eine Berücksichtigung des Zeitfaktors verlangt.

In der Darstellung dieses Kapitels dominiert die Varianzanalyse in ihren verschiedenen Varianten. Es sei jedoch noch einmal darauf verwiesen, dass sowohl Hypothesenarten als auch statistische Auswertungsverfahren, die hier aus didaktischen Gründen getrennt behandelt

Tabelle 12.19 Ausgewählte Signifikanztests für Unterschiedshypothesen (UV: unabhängige Variable; AV: abhängige Variable)

Art der Untersuchungspläne	Signifikanztests (Beispiele, Kommentar)
Zweigruppenpläne	Unterschiedshypothesen, die sich auf **eine zweifach gestufte (dichotom-nominalskalierte) unabhängige Variable** (entweder ein Treatment und eine Kontrollgruppe oder zwei Treatments) und mindestens eine kardinalskalierte abhängige Variable beziehen
Varianten und statistische Auswertungsverfahren	Einfaktorieller (1 UV) univariater (1 AV) Versuchsplan: t-Test für unabhängige Stichproben; bei Matching-Prozedur in Quasi-Experimenten t-Test für abhängige Stichproben Bei Verletzung der Voraussetzungen für parametrische Verfahren werden verteilungsfreie (nonparametrische) Verfahren eingesetzt (z. B. U-Test statt t-Test) Bei mehr als einer AV s. unten multivariate Pläne
Mehrgruppenpläne	Unterschiedshypothesen, die sich auf mehr als zwei Gruppen beziehen. Es wird mit **einer mehrfach gestuften (polytom-nominalskalierten) unabhängigen Variablen** und mindestens einer kardinalskalierten abhängigen Variablen gearbeitet
Varianten und statistische Auswertungsverfahren	Einfaktorieller (1 UV) univariater (1 AV) Versuchsplan: Einfaktorielle univariate Varianzanalyse mit einem festen oder zufälligen Faktor; A-priori-Kontraste; A-posteriori-Einzelvergleiche (Post-hoc-Tests); Trendtests; bei Matching-Prozedur in Quasi-Experimenten: Varianzanalyse für abhängige Stichproben Bei mehr als einer AV s. unten multivariate Pläne
(Mehr)Faktorielle Pläne	Unterschiedshypothesen, die sich auf **mindestens zwei unabhängige Variablen (dichotom- und/oder polytom-nominalskaliert)** und eine kardinalskalierte abhängige Variable beziehen
Varianten und statistische Auswertungsverfahren	Bei 2 UV: Zweifaktorielle Varianzanalyse Bei 3 UV: Dreifaktorielle Varianzanalyse A-priori-Vergleiche mittels Kontrasten; A-posteriori-Vergleiche mit Post-hoc-Tests; Trendtests Bei faktoriellen Plänen ist die Betrachtung der Interaktionseffekte besonders wichtig Bei mehr als einer AV s. unten multivariate Pläne
Hierarchische Pläne	**Unvollständige Untersuchungspläne** mit mindestens zwei unabhängigen Variablen, die nur einen Teil der möglichen Faktorenstufenkombinationen berücksichtigen, so dass ein Faktor (eine UV) unter eine andere UV geschachtelt ist
Varianten und statistische Auswertungsverfahren	Differenzierung der Pläne nach Art der Schachtelung: lateinische Quadrate, griechisch-lateinische Quadrate, teilhierarchische Pläne Auswertung mittels Varianzanalyse; Post-hoc-Tests; A-priori-Kontrasten; Trendtests Keine vollständige Überprüfung von Interaktionseffekten möglich Bei mehr als einer AV s. unten multivariate Pläne
Pläne mit Kontrollvariablen	Alle Pläne für Unterschiedshypothesen, bei denen **mindestens eine kardinalskalierte Kontrollvariable** berücksichtigt wird
Varianten und statistische Auswertungsverfahren	Kovarianzanalyse, bei der die Werte der abhängigen Variablen um den Einfluss einer oder mehrerer Kontrollvariablen bereinigt werden
Multivariate Pläne	Alle Pläne für Unterschiedshypothesen, die **mindestens zwei abhängige Variablen** enthalten
Varianten und statistische Auswertungsverfahren	Ein- und mehrfaktorielle multivariate Varianzanalyse Ein- und mehrfaktorielle multivariate Kovarianzanalyse Ein- und mehrfaktorielle Diskriminanzanalyse

werden, formal teilweise ineinander überführbar sind. Die Unterschiedshypothese, dass Personen, die regelmäßig meditieren, bessere Konzentrationsleistungen aufweisen als Personen, die nicht meditieren, lässt sich auch als **Zusammenhangshypothese** formulieren und korrelations- bzw. regressionsanalytisch auswerten: Es besteht ein positiver Zusammenhang zwischen Meditation und Konzentration (► Abschn. 12.5.2).

Des Weiteren wird bei varianzanalytischen Techniken von kontinuierlichen bzw. mindestens intervallskalierten abhängigen Variablen ausgegangen. Interessiert man sich für **kategoriale abhängige bzw. Outcome-Variablen** (z. B. Genesung/keine Genesung; Berufserfolg/kein Berufserfolg), so ist eine statistische Analyse mittels Diskriminanzanalyse möglich, wobei hier die Prädiktoren kontinuierlich sein sollten. Sind abhängige und unab-

hängige Variablen kategorial, so lassen sich loglineare Modelle einsetzen. Mittels logistischer Regression lassen sich kategoriale Variablen vorhersagen, wobei die unabhängigen Variablen bzw. Prädiktoren dann sowohl intervallskaliert als auch kategorial (Voraussetzung: Dummy-Kodierung) sein können (zur statistischen Analyse kategorialer Daten s. Agresti, 2007). Gruppenpläne, die ausschließlich durch nominalskalierte Variablen zustande kommen und bei denen die Zellen keine Messwerte, sondern die Anzahl der Untersuchungsobjekte enthalten, werden mit **Chi-Quadrat-Verfahren** ausgewertet (Bortz & Schuster, 2010).

Treatmentgruppe	Kontrollgruppe
Treatmentgruppe 1	Treatmentgruppe 2
S_1	S_2

Abb. 12.35 Untersuchungsschema eines Zweigruppenplans mit einer Treatment- und einer Kontrollgruppe oder zwei Treatmentgruppen

Zweigruppenpläne

Einfache Effekthypothesen der Art „Treatment X hat einen Einfluss auf die abhängige Variable Y" sollten als Unterschiedshypothesen (bzw. bei einer Messwiederholung als Veränderungshypothesen; ▶ Abschn. 12.5.4) geprüft werden. Die entsprechende Unterschiedshypothese lautet: „Die mit einem Treatment X behandelte Population unterscheidet sich bezüglich Y von einer nicht behandelten Population" (zur Formulierung einer gerichteten Unterschiedshypothese ▶ Abschn. 12.5.1).

Bei einem Zweigruppenplan („two groups design") arbeitet man mit einer zweifach gestuften (dichotom-nominalskalierten) unabhängigen Variablen und einer kardinalskalierten/metrischen abhängigen Variablen.

Ein solcher Zweigruppenplan ist die einfachste Variante eines einfaktoriellen (1 UV) univariaten (1 AV) Untersuchungsplans. Typischerweise erfolgt die statistische Hypothesenprüfung über einen t-Test für unabhängige Stichproben.

Experimentelle Untersuchungen

Die Durchführung beginnt mit der Ziehung einer Stichprobe des Umfanges n aus derjenigen Population, für die die Untersuchungsergebnisse gelten sollen (zur Art der Stichprobe ▶ Abschn. 9.3 und zur optimale Größe der Stichprobe ▶ Abschn. 14.3). In der Experimentalforschung wird in der Praxis oft nicht mit Zufallsstichproben, sondern mit nicht-probabilistischen Stichproben gearbeitet (z. B. Gelegenheitsstichproben, Quotenstichproben). Die n Untersuchungsteilnehmenden werden nach einem Zufallsverfahren **(Randomisierung)** in 2 Gruppen bzw. 2 **Teilstichproben S_1 und S_2** mit den Umfängen n_1 und n_2 aufgeteilt, wobei n_1 und n_2 nach Möglichkeit gleich groß sein sollten. Beachte: Die Frage der Stichprobenauswahl (im Experiment oft keine Zufallsstichprobe) ist abzugrenzen von der Frage der Gruppenzuteilung (im Experiment definitionsgemäß immer zufällig im Sinne einer Randomisierung; ▶ Abschn. 7.6.1). Die Stichprobenart beeinflusst vor allem das Qualitätskriterium der externen Validität, die Gruppenzuteilung das Kriterium der internen Validität (▶ Abschn. 3.2.2).

Die eine Untersuchungsgruppe erhält das Treatment (Treatment- oder Experimentalgruppe), die andere bleibt unbehandelt oder wird konventionell behandelt (Kontrollgruppe). Es resultiert das in Abb. 12.35 wiedergegebene Untersuchungsschema. Die Untersuchung endet mit der Erhebung der abhängigen Variablen in beiden Gruppen bzw. mit der Überprüfung des Unterschiedes der beiden Stichprobenmittelwerte auf statistische Signifikanz.

Bei großen und eher homogenen Stichproben (z. B. Psychologiestudierende im Grundstudium) soll und kann die Randomisierung personengebundene Störvariablen weitgehend ausschalten: In beiden zufällig gebildeten Gruppen sind im Durchschnitt gleiche Ausgangsbedingungen zu finden. Ein nach der Behandlung gefundener Gruppenunterschied ist dann eindeutig auf die Behandlung zurückzuführen. Die interne Validität ist jedoch gefährdet, wenn die Voraussetzung gleicher Ausgangsbedingungen der Gruppen verletzt ist. Dies kann geschehen, wenn bei der Randomisierung Fehler unterlaufen oder wenn die Stichprobe zu klein und/oder zu heterogen ist, um per Zufallszuteilung eine Vergleichbarkeit der Gruppen herzustellen. Deswegen ist es generell empfehlenswert und bei kleinen bzw. heterogenen Gruppen zwingend notwendig, die Vergleichbarkeit der Gruppen vor der Behandlung in einem **Vortest (Pretest, Prätest)** zu erheben. Damit kann nicht nur die Vergleichbarkeit belegt, sondern – indem die Ausgangswerte als Kontrollvariablen/Kovariaten – verwendet werden, im Falle von Abweichungen die Vergleichbarkeit auch statistisch verbessert werden. Durch Einbeziehung einer Vorher-Messung wird der Zeitfaktor in das Design eingeführt, so dass die Studie dann den Veränderungsplänen zugeordnet werden kann (▶ Abschn. 12.5.4).

Das „Treatment" in experimentellen Studien ist jedoch nicht mit einer **Behandlungsmaßnahme** (z. B. Therapie-

oder Trainingsmaßnahme), die auf kurz- oder langfristige Veränderungen der Personen abzielt und somit am besten als Veränderungshypothese zu prüfen ist (▶ Abschn. 12.5.4), gleichzusetzen. Ein „Treatment" kann z. B. auch aus einem **Stimulus** bestehen, bei dem man sich fragt, ob seine Variation zu unterschiedlichen Effekten führt: Wird ein Text anders wahrgenommen, wenn er nur das generische Maskulinum nutzt („Liebe Studenten") oder männliche und weibliche grammatische Formen verwendet („Liebe Studentinnen und Studenten")? Wirkt eine Unternehmens-Website auf Jobsuchende attraktiver, wenn die ethnische Vielfalt in Teams betont wird, als wenn dies nicht der Fall ist? Kommen Werbeanzeigen mit natürlich aussehenden Models besser an als Werbeanzeigen mit stark idealisierten Models? Bei derartigen Fragestellungen konfrontiert man die Untersuchungspersonen der einen Gruppe mit der einen Stimulusvariante und die der anderen Gruppe mit der anderen Variante und betrachtet den durchschnittlichen Gruppenunterschied auf der oder den relevanten abhängigen Variablen. Bei Designs, die zwei Stimulusvarianten kontrastieren, ist es oft nicht sinnvoll, von Treatment- und Kontrollgruppe zu sprechen, vielmehr werden zwei Treatment- bzw. zwei Stimulusgruppen verglichen (◘ Abb. 12.35). Die Begriffe „Treatment" und „Stimulus" verweisen auf Untersuchungsbedingungen, die von den Forschenden gezielt vorgegeben werden, was in experimentellen und quasi-experimentellen Studien der Fall ist. In nicht-experimentellen Studien, in denen vorgefundene Gruppen verglichen werden (z. B. Raucher vs. Nichtraucher; Computerspieler vs. Nichtspieler), die oft selbstgewählt und über lange Zeit hinweg mit unterschiedlichen Bedingungen konfrontiert sind, spricht man anstelle von Treatmentgruppen eher von Exposuregruppen.

Für die statistische Überprüfung von Unterschiedshypothesen, die in experimentellen Zweigruppenplänen untersucht werden, steht der ***t*-Test für unabhängige Stichproben** zur Verfügung. Sind die Voraussetzungen des *t*-Tests verletzt, ist als Signifikanztest ein Verfahren aus der Klasse der verteilungsfreien Methoden zu wählen (z. B. der *U*-Test; Bortz, Lienert, & Boehnke, 2008, S. 200).

Quasi-experimentelle Untersuchungen

Zweigruppenpläne sind auch für die Durchführung von quasi-experimentellen Untersuchungen geeignet, d. h. für Untersuchungen, bei denen die Untersuchungsgruppen zwar wie im Experiment gezielt unterschiedlich behandelt werden, allerdings nicht randomisiert gebildet werden konnten. In diesem Sinne quasi-experimentell wäre z. B. eine Untersuchung, die die Wirkung von Tablet-gestütztem Unterricht auf die Schülerleistungen untersucht, dabei aber die Schüler nicht zufällig den Klassen (= Untersuchungsgruppen) zuordnet, sondern eine vorgefundene Klasse als Tablet-Klasse (Experimentalbedingung) und eine andere vorgefundene Klasse als Kontrollgruppe verwendet.

Bei quasi-experimentellen Untersuchungen besteht die Gefahr, dass die unabhängige Variable mit anderen, für die abhängige Variable bedeutsamen Variablen konfundiert ist. Diese **Störvariablen** bzw. konfundierende Variablen („confounder") können personengebunden oder untersuchungsbedingt sein. Maßnahmen zur Kontrolle derartiger Störvariablen wurden in ▶ Abschn. 3.2.2 diskutiert. Wählt man als Kontrolltechnik eine **Matching-Prozedur**, ist die Unterschiedshypothese mit einem ***t*-Test für abhängige Stichproben** zu überprüfen.

Extremgruppenvergleiche

Eine spezielle Variante des Zweigruppenplans stellt der sog. Extremgruppenvergleich dar. Hierbei werden nur Untersuchungspersonen berücksichtigt, die bzgl. einer mehrfach gestuften oder kontinuierlichen unabhängigen Variablen besonders hohe oder besonders niedrige Ausprägungen aufweisen (z. B. Personen mit sehr hoher und sehr niedriger Pflegestufe; Personen mit sehr guter und sehr schlechter Bildung; Personen mit sehr hohem oder sehr geringem Einkommen etc.).

Zur Hypothesenprüfung im engeren Sinne sollten Extremgruppenvergleiche nicht herangezogen werden, da sie den interessierenden Effekt überschätzen. Extremgruppenvergleiche stehen auf der gleichen Stufe wie die in ▶ Abschn. 12.5.2 kritisierten Korrelationsstudien, die den mittleren Bereich einer Variablen außer Acht lassen (◘ Abb. 12.26c). Sie können aber im Zuge der Exploration dazu dienen, den potenziellen Erklärungswert einer unabhängigen Variable auf eine abhängige Variable zu prüfen. Wenn der **Signifikanztest auf Probe** (▶ Abschn. 12.3.2) beim Extremgruppenvergleich einen überzufälligen Unterschied aufzeigt (z. B. signifikantes Ergebnis im *t*-Test oder *U*-Test), können Folgestudien in Erwägung gezogen werden, die das gesamte Merkmalsspektrum der UV – also neben den Gruppen mit extremer Merkmalsausprägung auch die Gruppen mit mittleren Merkmalsausprägungen abdecken. Eine ausführliche Erörterung der Probleme von Extremgruppenvergleichen findet man bei Preacher, Rucker, MacCallum und Nicewander (2005).

Mehrgruppenpläne

Unterschiedshypothesen, die sich nicht nur auf zwei, sondern auf mehr als 2 Treatments (allgemein: p Treatments A_1, A_2 … A_p) ggf. inklusive Kontrollbedingung(en) beziehen, die alle als Ausprägungen *einer* polytom-nominalskalierten unabhängigen Variable zu verstehen sind, werden mit einem einfaktoriellen Mehrgruppenplan untersucht. Die experimentelle Vorgehensweise entspricht

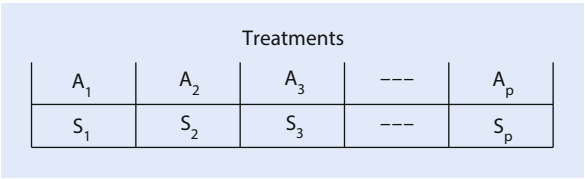


Abb. 12.36 Untersuchungsschema eines Mehrgruppenplans

der eines Zweigruppenplans: Man zieht eine Stichprobe des Umfanges n und teilt diese zufällig in p Gruppen bzw. Teilstichproben S_1, S_2 … S_p mit den Umfängen n_1, n_2 … n_p auf. Hierbei ist es von Vorteil, wenn alle Gruppen gleich groß sind. Jeder Gruppe wird dann ein Treatment zugeordnet. Es resultiert das in Abb. 12.36 wiedergegebene Untersuchungsschema. Ebenso wie bei Zweigruppen- gilt auch bei Mehrgruppenplänen, dass Treatments, die auf Merkmalsveränderungen bei Personen abzielen, am besten mit Vorher-Nachher-Messungen und somit einem Messwiederholungsdesign zu prüfen sind (▶ Abschn. 12.5.4), während Effekte von Stimuli oft durch eine einzige Messung (nämlich jeweils nach Konfrontation mit dem spezifischen Stimulus) geprüft werden. So können z. B. Effekte der sprachlichen Gestaltung von Texten experimentell durch 4 Stimulus-Varianten geprüft werden, in denen nur männliche Formen („liebe Studenten"), männliche und weibliche Formen („liebe Studentinnen und Studenten", nur weibliche Formen („liebe Studentinnen") und neutrale Formen („liebe Studierende") verwendet werden. Werden die Textvarianten jeweils durch Randomisierung vergleichbaren Gruppen von Untersuchungspersonen vorgelegt, so können Mittelwertsunterschiede in der Textbewertung auf die Textgestaltung zurückgeführt werden.

Bei einem *Mehrgruppenplan* („multiple groups design") arbeitet man mit einer mehrfach gestuften (polytom-nominalskalierten) unabhängigen Variablen und einer kardinalskalierten/metrischen abhängigen Variablen. Ein solcher Mehrgruppenplan ist ein einfaktorieller (1 UV) univariater (1 AV) Untersuchungsplan.

Typischerweise erfolgt die statistische Hypothesenprüfung beim Mehrgruppenplan über eine einfaktorielle Varianzanalyse („one-way analysis of variance", „one-way ANOVA") bzw. über A-priori- oder A-posteriori-Einzelvergleiche und ggf. auch Trendtests. Man unterscheidet im Rahmen der Varianzanalyse feste Faktoren („fixed factor"), bei denen nur Aussagen über die realisierten Faktorstufen getroffen werden sollen, von zufälligen Faktoren („random factor"), bei denen die realisierten Faktorenstufen eine Zufallsauswahl möglicher Faktorstufen darstellen.

Ein weiteres Beispiel: Es wird die Hypothese überprüft, dass die Reproduktion eines Textes von der Art der Informationsaufnahme abhängt. Eine Untersuchungsgruppe muss sich einen Text durch leises Lesen einprägen (Treatment A_1), eine zweite denselben Text durch lautes Lesen (Treatment A_2) und einer dritten Gruppe wird derselbe Text vorgelesen (A_3). Diese genau definierten Faktorstufen machen die UV zu einem **festen Faktor**. Würde man dagegen untersuchen wollen, ob die Informationsaufnahme beim Vorlesen von der Person abhängt, die einen Text vorliest und würde man dafür 5 Vorleser engagieren, so wären diese 5 Stufen zufällige Ausprägungen des Faktors (denn es gibt noch zahllose weitere Vorleser), der Faktor wäre also ein sog. **zufälliger Faktor** („random factor"). Als abhängige Variable werden die Fehler gezählt, die die Untersuchungsteilnehmenden bei einem abschließenden Wissenstest über den Text machen.

Die statistische Überprüfung dieser Unterschiedshypothese erfolgt mithilfe der **einfaktoriellen Varianzanalyse**, der zunächst im Sinne einer **Overall-Signifikanz** die Information zu entnehmen ist, ob sich mindestens zwei der drei untersuchten Gruppen überzufällig voneinander unterscheiden. Zusätzlich kann man mithilfe sog. Einzelvergleiche überprüfen, ob sich bestimmte Treatments signifikant voneinander unterscheiden (im Beispiel: Haben lautes und leises Lesen unterschiedliche Wirkung auf die Behaltensleistung?). Hierbei werden **A-priori-Einzelvergleiche** (Kontraste), die die Formulierung gezielter Einzelvergleichshypothesen vor der Untersuchung voraussetzen (und somit die Overall-Signifikanzprüfung überflüssig werden lassen), und **A-posteriori-Einzelvergleiche** (Post-hoc-Tests) unterschieden, mit denen man im Nachhinein – sofern ein Overall-Effekt gefunden wurde – feststellt, welche Treatments sich signifikant voneinander unterscheiden. Bei Post-hoc-Tests werden alle Gruppen paarweise miteinander verglichen, bei A-priori-Kontrasten nur genau die vorhergesagten (meist gerichteten) Gruppenunterschiede geprüft.

Einzelvergleichsverfahren verwendet man auch, um **Kombinationen einzelner Treatments** mit anderen Treatments zu vergleichen. Diese Auswertungsvariante ist besonders vorteilhaft, wenn neben mehreren Treatment-Gruppen eine oder mehrere Kontrollgruppen untersucht werden und man an der Hypothese interessiert ist, dass sich die behandelten Untersuchungspersonen der verschiedenen Behandlungsformen zusammengenommen von den nichtbehandelten Untersuchungspersonen unterscheiden. Ein typisches Beispiel hierfür ist der Vergleich verschiedener Medikamente mit einem Placebo (einer chemisch wirkungslosen Substanz), bei dem zunächst die Frage interessiert, ob die Wirkung der Medikamente

zusammengenommen überhaupt einer möglichen Placebowirkung überlegen ist.

Weitere Zusatzauswertungen sind möglich, wenn nicht nur die abhängige Variable, sondern auch die unabhängige Variable intervallskaliert (oder zumindest ordinalskaliert) ist (z. B. Reaktionszeiten in Abhängigkeit von verschiedenen Alkoholmengen). Mit sog. statistischen **Trendtests** kann dann z. B. die Hypothese geprüft werden, ob die Werte der abhängigen Variablen linear (oder einem anderen Trend folgend) zur unabhängigen Variablen ausgeprägt sind (im Beispiel: Die Reaktionszeit ist jeweils umso höher, je höher auch die Alkoholmenge ist). Oder: Eine Mode-Werbeanzeige wird umso ästhetischer eingeschätzt, je idealisierter und perfekter die dargestellten Models sind.

Wie bereits mehrfach erwähnt, sind **Quasi-Experimente** mangels Randomisierung weniger aussagekräftig als experimentelle Untersuchungen. Die interne Validität lässt sich jedoch auch hier durch die in ▶ Abschn. 3.2.2 genannten Kontrolltechniken erhöhen. Auf die Matching-Prozedur wird man bei Mehrstichprobenplänen realistischer Weise nur zurückgreifen, wenn höchstens drei oder vier Gruppen kleineren Umfanges zu vergleichen sind, da sonst der Aufwand für das Matching zu groß wird. Die Auswertung erfolgt bei Matched Samples mit einer **Varianzanalyse für abhängige Stichproben** (Bortz & Schuster, 2010; Eid, Gollwitzer, & Schmitt, 2010).

(Mehr-)Faktorielle Pläne

Bisher waren die Treatments bzw. die zu vergleichenden Populationen Stufen **einer** unabhängigen Variablen, was untersuchungstechnisch zu Zwei- oder Mehrgruppenplänen führte (einfaktorielle Pläne). Für viele Forschungsfragen ist es jedoch realistisch, davon auszugehen, dass mehrere unabhängige Variablen simultan wirksam sind. Lassen sich diese hypothetisch benennen, empfiehlt sich eine Untersuchung nach den Regeln (mehr-)faktorieller Pläne. Soll ein Zeitfaktor berücksichtigt werden (z. B. pro Untersuchungsgruppe erfolgt eine Vorher- und eine Nachher-Messung), so sei wiederum auf die Prüfung von Veränderungshypothesen verwiesen (▶ Abschn. 12.5.4).

Bei einem *(mehr)faktoriellen Plan* („factorial design", „full factorial design") arbeitet man mit mehr als einer nominalskalierten bzw. diskreten unabhängigen Variablen und einer abhängigen Variablen. Enthält ein faktorieller Plan zwei unabhängige Variablen, spricht man von einem zweifaktoriellen Plan; enthält er drei unabhängige Variablen, spricht man von einem dreifaktoriellen Plan etc. Sofern nur eine abhängige Variable beteiligt ist, handelt es sich um mehrfaktorielle (mehrere UV) univariate (1 AV) Untersuchungspläne.

	B_1	B_2	B_3	----	B_q
A_1	S_{11}	S_{12}	S_{13}	----	S_{1q}
A_2	S_{21}	S_{22}	S_{23}	----	S_{2q}
A_3	S_{31}	S_{32}	S_{33}	----	S_{3q}
⋮	⋮	⋮	⋮	−+−	⋮
A_p	S_{p1}	S_{p2}	S_{p3}	----	S_{pq}

Abb. 12.37 Untersuchungsschema eines zweifaktoriellen Plans

Die statistische Hypothesenprüfung erfolgt bei faktoriellen Plänen typischerweise über mehrfaktorielle Varianzanalysen sowie A-priori- und A-posteriori-Einzelvergleiche und ggf. Trendtests. Besonders wichtig bei mehrfaktoriellen Varianzanalysen ist neben der Betrachtung der Haupteffekte auch die Analyse der Interaktionseffekte, d. h. der Art der Wechselwirkungen der unabhängigen Variablen in ihrer Wirkung auf die abhängige Variable.

Zweifaktorielle Pläne

Die einfachste Variante mehrfaktorieller Pläne, der zweifaktorielle Plan, kontrolliert gleichzeitig die Bedeutung von zwei unabhängigen Variablen (Faktoren) für eine abhängige Variable. Zusätzlich informiert dieser Plan über die Art der Kombinationswirkung (**Interaktion** oder Wechselwirkung) der beiden unabhängigen Variablen.

Nehmen wir an, die erste unabhängige Variable (Faktor A) sei p-fach und die zweite unabhängige Variable (Faktor B) q-fach gestuft. Es ergeben sich damit insgesamt $p \times q$ Faktorstufenkombinationen. In einem vollständigen experimentellen Plan („full factorial design") werden jeder Faktorstufenkombination per Zufall n Untersuchungsobjekte zugeordnet, d. h., wir benötigen insgesamt $p \times q$ Gruppen bzw. Teilstichproben (S_{11}, $S_{12} \ldots S_{pq}$) bzw. $p \times q \times n$ Untersuchungsobjekte, für die jeweils eine Messung der abhängigen Variablen erhoben wird (zu ungleich großen Stichproben s. Bortz, 2005, ▶ Abschn. 8.4). In ◘ Abb. 12.37 findet sich das Grundschema eines zweifaktoriellen Plans.

Beispiel: Überprüft werden die Hypothesen, dass die Ablesegenauigkeit für Anzeigegeräte (z. B. für Tachometer) von der Form des Gerätes (Faktor A mit den Stufen: A_1 = oval, A_2 = viereckig, A_3 = rund) und von der Art der Zahlendarstellung (Faktor B mit den Stufen: B_1 = analog, B_2 = digital) abhängt. Insgesamt resultieren also $3 \times 2 = 6$ Faktorstufenkombinationen (Arten von Anzeigegeräten). Für jede Faktorstufenkombination erhält eine Stichprobe von Untersuchungspersonen die Aufgabe, in

mehreren Durchgängen in einer vorgegebenen Zeit die angezeigte Zahl zu nennen. Die Anzahl falscher Reaktionen der Untersuchungspersonen sind die Messungen der abhängigen Variablen.

Die nach diesem Schema erhobenen Daten werden mit einer zweifaktoriellen Varianzanalyse („two-way ANOVA“) statistisch ausgewertet. Diese Auswertung entspricht nicht – wie man meinen könnte – der zweifachen Anwendung einer einfaktoriellen Varianzanalyse, denn es wird zusätzlich überprüft, ob auch die Kombinationswirkungen der untersuchten Faktorstufen statistisch bedeutsam sind. Diese könnten z. B. darin bestehen, dass die runde Form in Kombination mit der analogen Zahlendarstellung besonders gut, aber in Kombination mit der digitalen Darstellung besonders schlecht abschneidet.

Wegen ihrer großen forschungslogischen Bedeutung wollen wir uns im Folgenden dem Konzept der statistischen Interaktion etwas ausführlicher zuwenden.

Ein zweifaktorieller Plan wird mit einer zweifaktoriellen Varianzanalyse („two-way analysis of variance“, „two-way ANOVA“) inferenzstatistisch ausgewertet. Dabei kann man Unterschiedshypothesen über drei Effekte prüfen: Haupteffekt A, Haupteffekt B und die Interaktion erster Ordnung $A \times B$.

Haupteffekte und Interaktionen Betrachten wir den einfachsten Fall einer zweifaktoriellen Varianzanalyse mit je nur 2 Stufen für Faktor A und Faktor B. Ein solcher Plan entsteht z. B., wenn man in einem Experiment die Wirksamkeit eines Placebos (A_1) im Vergleich zu einem herkömmlichen Beruhigungsmittel (A_2) testen will und dabei auch mögliche Geschlechtseffekte (B_1: männlich, B_2: weiblich) mit einbezieht. Als abhängige Variable wäre eine Skala zur subjektiven Befindlichkeit (z. B. mit Items wie „Ich fühle mich nervös und angespannt“: stimmt gar nicht – wenig – teils/teils – ziemlich – völlig) oder auch ein physiologisches Maß (z. B. ein physiologischer Erregungsindikator, ▶ Abschn. 10.5) denkbar.

Bei der (fiktiven) Untersuchung von 5 Personen pro Untersuchungsgruppe könnte sich das in Tab. 12.20 dargestellte Datenschema ergeben.

Mit diesem zweifaktoriellen varianzanalytischen Design können drei Forschungshypothesen geprüft werden: zwei **Haupteffekthypothesen** und eine **Interaktionshypothese**. Für das Beispiel formulieren wir:

1. Das Placebo und das Beruhigungsmittel wirken unabhängig vom Geschlecht der behandelten Personen unterschiedlich (Haupteffekt A). (Bei entsprechenden Vorkenntnissen könnte auch eine gerichtete Hypothese formuliert werden.)

Tabelle 12.20 Datenschema einer 2 × 2-Varianzanalyse. Wertebereich der abhängigen Variablen: 0 = geringste Anspannung bis 20 = höchste Anspannung

	A_1 Placebo	A_2 Beruhigungsmittel	$\bar{B}_j$
B_1 Männer	9	6	
	10	7	
	8	6	7.60
	8	6	
	9	7	
B_2 Frauen	7	8	
	8	8	
	6	7	7.10
	8	6	
	7	6	
$\bar{A}_i$	8.00	6.70	7.35

2. Männer und Frauen reagieren insgesamt, d. h. in Bezug auf beide Medikamente, unterschiedlich (Haupteffekt B).
3. Es kommt zu einer differenziellen Wirkung der Medikamente, z. B. von der Art, dass Frauen auf das Placebo stärker reagieren als Männer, dass aber für das Beruhigungsmittel keine geschlechtsspezifischen Wirkunterschiede nachweisbar sind (Interaktion $A \times B$).

Auch ohne statistische Analyse kann man zunächst durch Inspektion der Stichprobenergebnisse Vermutungen darüber anstellen, ob Haupteffekte oder ein Interaktionseffekt vorliegen könnten. Je größer die Unterschiede zwischen den Spaltenmittelwerten (hier: $\bar{A}_1$ und $\bar{A}_2$) bzw. zwischen den Zeilenmittelwerten (hier: $\bar{B}_1$ und $\bar{B}_2$), desto eher spricht dies für einen signifikanten Haupteffekt A bzw. Haupteffekt B. Ob möglicherweise ein signifikanter Interaktionseffekt vorliegt, erkennt man durch Betrachtung der Zellenmittelwerte ($\overline{AB}_{ij}$), die man ergänzend zum obigen Datenschema meist separat in einer kleinen Tabelle einträgt (Tab. 12.21).

Für das (aus didaktischen Gründen extrem kleine) Zahlenbeispiel stellen wir auf deskriptiver Ebene fest, dass der Haupteffekt A mit $8.00 - 6.70 = 1.30$ größer ist als der Haupteffekt B ($7.60 - 7.10 = 0.50$). Einen Interaktionseffekt erkennt man daran, dass sich die Differenzen der $\overline{AB}_{ij}$-Werte zeilen-(oder spalten-)weise unterscheiden. Dies ist im Beispiel der Fall: der Geschlechtsunter-

Tabelle 12.21 Zellenmittelwerte $\overline{AB}_{ij}$ für eine 2 × 2 Varianzanalyse

	A_1 Placebo	A_2 Beruhigungsmittel	$\bar{B}_j$
B_1 Männer	8.80	6.40	7.60
B_2 Frauen	7.20	7.00	7.10
$\bar{A}_i$	8.00	6.70	7.35

schied beträgt beim Placebo $8.80 - 7.20 = 1.60$ und beim Beruhigungsmittel $6.40 - 7.00 = -0.60$. Charakteristisch für eine Interaktion ist, dass die Wirkung eines Faktors auf die abhängige Variable von der Ausprägung des anderen Faktors abhängt Eine Varianzanalyse über die Daten der Tab. 12.20 bestätigt den Haupteffekt A und die Interaktion $A \times B$ ($\alpha = .01$), aber nicht den Haupteffekt B.

Man beachte, dass mit dem Vorliegen eines Interaktionseffektes *nicht* gemeint ist, **dass** zwei Faktoren zusammenwirken bzw. dass sie gemeinsam einen Effekt erzeugen, sondern es geht darum, **auf welche Weise** die Faktorstufen zusammenwirken! (Verstärken sie sich? Schwächen sie sich ab? Kommen beim Zusammenspiel einiger Faktorstufen überraschende Wirkungen zustande, die über ein rein additives Zusammenwirken hinausgehen?) Ein additives Zusammenwirken beider Faktoren ist als interaktionsfreier „Normalfall" definiert; nur **überzufällige Abweichungen von der Additivität** werden als „Interaktionseffekt" bezeichnet. Dazu noch einmal das Beispiel der Anzeigegeräte: Wenn eine digitale Anzeige für sich genommen im Durchschnitt zu geringen Fehlerraten führt und ein rechteckiges Display ebenfalls besonders günstige Fehlerraten hat, muss bei additivem Zusammenwirken der beiden Faktorstufen im Sinne von Additivität ein rechteckiges digitales Anzeigegerät besonders geringe Fehlerdurchschnitte erzeugen (kein Interaktionseffekt). Das Vorliegen eines Interaktionseffektes wäre daran erkennbar, dass trotz guter separater Einzelergebnisse von digitaler Anzeige und von rechteckigem Display in der Kombination plötzlich überraschend schlechtere Werte zustande kommen oder aber die Fehlerraten plötzlich in ganz extremem Maße verringert sind. Interaktionseffekte können deskriptiv anhand der Zellenmittelwerte betrachtet werden, müssen aber jeweils ihrerseits auf Signifikanz geprüft werden. Eine zweifaktorielle Varianzanalyse gibt dementsprechend drei Irrtumswahrscheinlichkeiten für die drei Signifikanzprüfungen an: Für Haupteffekt A, Haupteffekt B und den Interaktionseffekt $A \times B$.

Interaktionseffekt – Ein signifikanter Interaktionseffekt $A \times B$ in der zweifaktoriellen Varianzanalyse besagt, dass beide Faktoren nicht einfach ganz normal additiv, sondern in anderer, nicht-additiver Weise zusammenwirken. Anlässlich eines signifikanten Interaktionseffekts in der zweifaktoriellen Varianzanalyse kann eine ungerichtete Interaktionshypothese angenommen werden. Gerichtete Interaktionshypothesen werden über Interaktionskontraste inferenzstatistisch abgesichert.

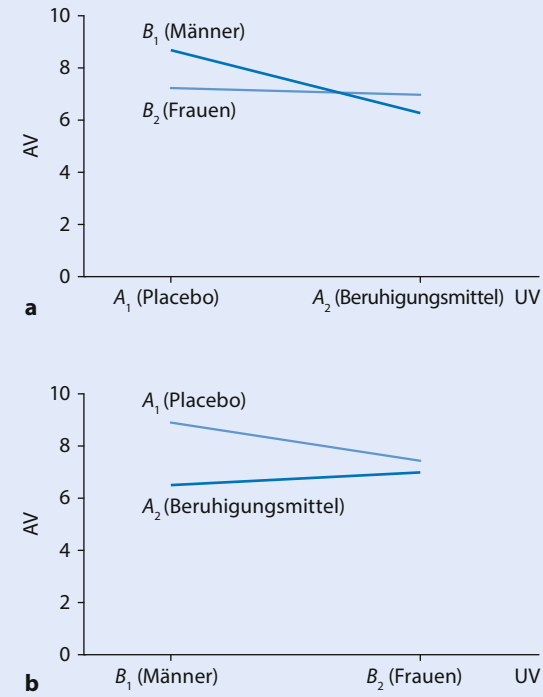


Abb. 12.38 Interaktionsdiagramme für Faktor *A* (Behandlung: Placebo vs. Beruhigungsmittel) (**a**) und Faktor *B* (Geschlecht: Männer vs. Frauen) (**b**) *AV* abhängige Variable: Befindlichkeitsskala. (geringste Anspannung 0 bis 20 höchste Anspannung), *UV* unabhängige Variable

Um die Art des Zusammenwirkens zweier Faktoren sichtbar zu machen, fertigt man ergänzend zur Tabelle der Zellenmittelwerte (Tab. 12.21) sog. **Interaktionsdiagramme** an, in die jeweils alle Zellenmittelwerte (in Tab. 12.21 sind es vier) einzutragen sind. Im Interaktionsdiagramm werden die Werte der abhängigen Variablen AV (hier: die Werte für die Befindlichkeit) auf der Ordinate (y-Achse) und die Stufen einer der beiden Faktoren (z. B. Faktor A) auf der Abszisse (x-Achse) abgetragen. Für jede Stufe des anderen Faktors (hier B) wird ein Linienzug angefertigt, der die Mittelwerte der entsprechenden Faktorstufenkombinationen verbindet. Damit erhält man das Interaktionsdiagramm für Faktor A. Um ein Interaktionsdiagramm für Faktor B zu erstellen, trägt man die Stufen von Faktor B auf der x-Achse ab und zeichnet für die Stufen von Faktor A jeweils einen Grafen. In den Interaktionsdiagrammen für A und B werden somit dieselben Zellenmittelwerte dargestellt, nur jeweils anders gruppiert. Für unser Zahlenbeispiel ergeben sich die Interaktionsdiagramme in Abb. 12.38.

Wenn keine Interaktion vorliegt und die Faktoren nur ganz normal additiv zusammenwirken, sind die im Interaktionsdiagramm abgetragenen Grafen parallel. Je stärker sie von der Parallelität abweichen, desto eher spricht dies für das Vorliegen eines Interaktionseffektes.

Wenn eine Interaktion vorliegt, lassen sich drei Typen von Interaktionen unterscheiden (Leigh & Kinnear, 1980), die wir im Folgenden für einen 2 × 3-Plan verdeutlichen:

- Die **ordinale Interaktion** („ordinal interaction“) ist dadurch gekennzeichnet, dass die Grafen in beiden Interaktionsdiagrammen zwar nicht parallel, aber doch gleichsinnig verlaufen (z. B. beide aufsteigend, beide abfallend, Abb. 12.39a).
- Bei der **hybriden Interaktion** („hybrid interaction“; „semi-disordinal interaction“) dagegen verlaufen die Grafen nur in einem Interaktionsdiagramm gleichsinnig, im anderen nicht (Abb. 12.39b).
- Wenn in beiden Interaktionsdiagrammen die Grafen nicht gleichsinnig verlaufen, spricht man von **disordinaler Interaktion** („disordinal interaction“) Abb. 12.39c.

Man beachte, dass die Frage, ob Grafen gleichsinnig verlaufen, nichts damit zu tun hat, ob sie sich durchkreuzen. Zwei Grafen können z. B. beide aufsteigend sein und sich durchkreuzen, während gegenläufige Grafen ohne Schnittpunkte auftreten können. Statistische Tests, mit denen man überprüfen kann, welcher Interaktionstyp vorliegt, wurden z. B. von Bredenkamp (1982) entwickelt.

Mit diesem Wissen können wir nun die im Beispiel aufgetretene Interaktion (Abb. 12.38) zumindest deskriptiv als hybride Interaktion kennzeichnen.

Die Frage, welcher Interaktionstyp vorliegt, ist für die Interpretation der signifikanten Haupteffekte von Belang. Wenn keine Interaktion oder eine ordinale Interaktion vorliegt, darf man **beide signifikanten Haupteffekte global interpretieren** und dabei über die Stufen des anderen Faktors hinweg generalisieren. Im vorliegenden Beispiel würde man dann sagen, das Beruhigungsmittel sei wirksamer als das Placebo und – falls der Haupteffekt *B* signifikant wäre – Männer seien im Durchschnitt angespannter als Frauen.

Eine solche globale Interpretation ist immer problematisch, wenn eine hybride Interaktion vorliegt. Bei der hybriden Interaktion kann nämlich nur **ein Faktor global interpretiert** werden, wie im Beispiel Faktor *A*: Sowohl bei den Frauen als auch bei den Männern ist der Erregungsgrad mit Placebo (A_1) höher als bei Einnahme des Beruhigungsmittels (A_2), d. h., das Beruhigungsmittel ist generell wirkungsvoller als das Placebo. Faktor *B* ist dagegen nicht global interpretierbar, denn man kann nicht pauschal sagen, dass Männer nervöser und angespannter sind als Frauen. Hier muss man gemäß der signifikanten Interaktion differenziert zum Ausdruck bringen, dass die durchschnittliche Erregung bei den Männern in der Placebo-Bedingung höher ist als bei den Frauen, dass die Männer aber unter der Beruhigungsmittel-Bedingung einen niedrigeren durchschnittlichen Erregungswert als die Frauen aufweisen.

Bei einer disordinalen Interaktion kann **keiner der beiden Faktoren global interpretiert werden**; stattdessen muss eine differenzierte Betrachtung der einzelnen Zellenmittelwerte erfolgen.

Im Kontext der Varianzanalyse spricht man von einem *Interaktionseffekt* („interaction effect“), wenn die unabhängigen Variablen nicht einfach additiv, sondern in anderer Weise zusammenwirken. Die Art der Interaktion – ordinal, hybrid oder disordinal – entscheidet über die Interpretierbarkeit der Haupteffekte („main effects“).

Diese Interaktionen werden wie folgt unterschieden:

- Bei einer **ordinalen** Interaktion können beide Haupteffekte global interpretiert werden.
- Bei einer **hybriden** Interaktion kann nur ein Haupteffekt global interpretiert werden.
- Bei einer **disordinalen** Interaktion kann keiner der beiden Haupteffekte global interpretiert werden, stattdessen müssen beide unabhängigen Variablen in ihren Wechselwirkungen anhand der einzelnen Zellen betrachtet werden.

Interaktionen bilden Sachverhalte ab, die für viele human- und sozialwissenschaftliche Fragen realistischer sind als Haupteffekte. Mit einem Haupteffekt überprüfen wir eine Unterschiedshypothese, die sich auf die gesamte Zielpopulation bezieht, die also behauptet, dass die durch verschiedene Treatments oder Stimuli ausgelösten Effekte einheitlich für die gesamte untersuchte Population gelten. Interaktionshypothesen hingegen beziehen sich auf die differenzielle Wirkung der Treatments, d. h. auf Treatments, deren Wirkung von der Art der untersuchten Subpopulationen abhängt, also etwa von deren Personenmerkmalen oder von weiteren Treatment- bzw. Stimulus-Einflüssen. Interaktionseffekte basieren auf einer nicht-additiven Wirkung von Treatment-Kombinationen.

In Abb. 12.39 werden einige typische, aber keineswegs alle möglichen Muster für eine Interaktion verdeutlicht. Die Gesamtzahl aller Interaktionsmuster nimmt rasch zu, wenn man pro Faktor mehr Stufen untersucht als in den Beispielen. Es ist deshalb durchaus denkbar, dass sich ein empirisch gefundenes Interaktionsmuster als statistisch bedeutsam erweist, obwohl man – wenn überhaupt – ein anderes erwartet hat.

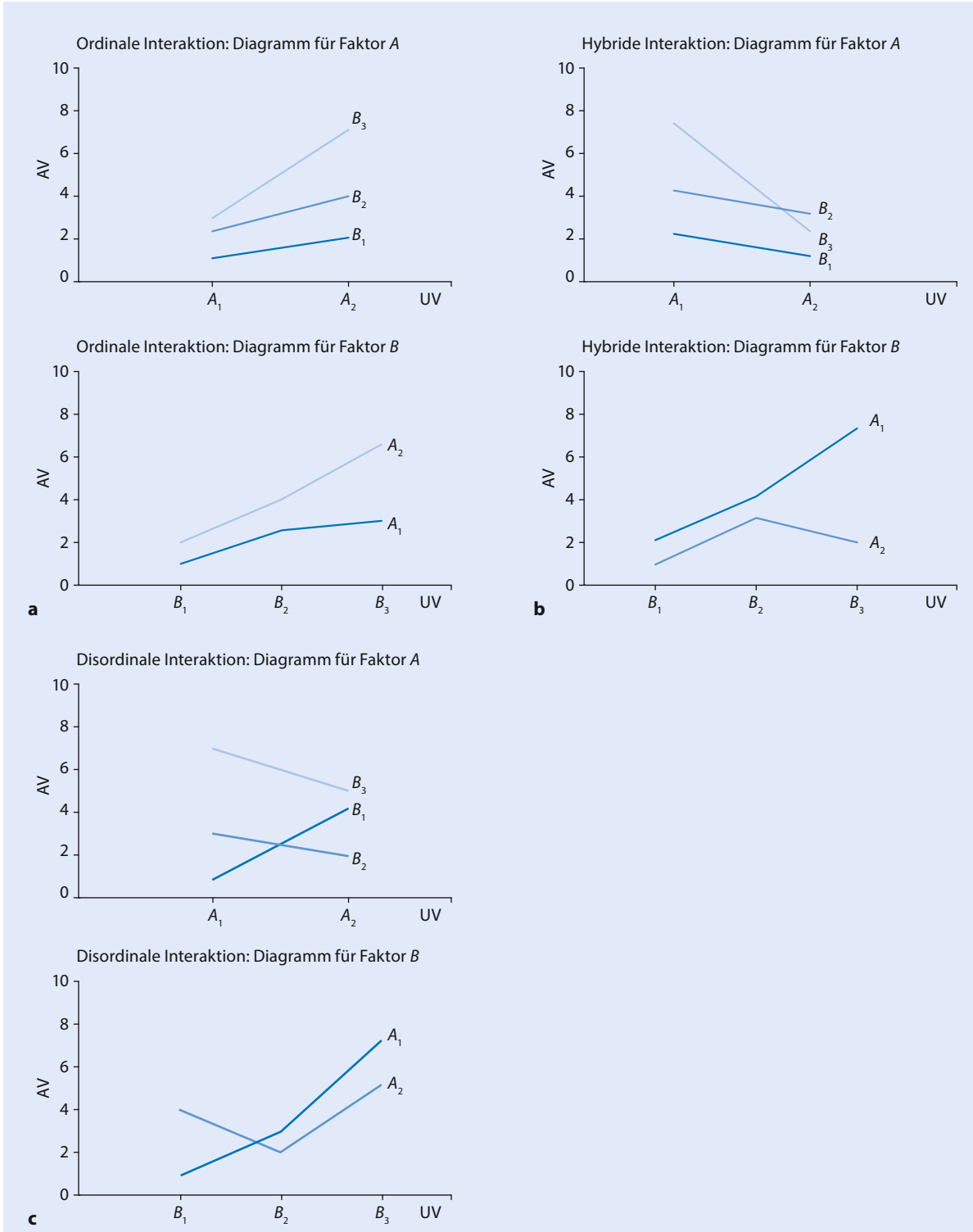


Abb. 12.39 **a** Ordinale Interaktion, **b** hybride Interaktion, **c** disordinale Interaktion. *AV* abhängige Variable, *UV* unabhängige Variable

Der allgemeinen Leitlinie hypothesenprüfender Untersuchungen folgend, ist auch in Bezug auf die Prüfung von Interaktionen zu fordern, dass ihr eine möglichst gezielte **Interaktionshypothese** voranzustellen ist. Statistisch signifikante Interaktionen, die nicht durch eine Hypothese vorhergesagt wurden, haben letztlich nur explorativen Wert. Die Varianzanalyse prüft zunächst nur eine ungerichtete Interaktionshypothese. Will man mit einer gerichteten Interaktionshypothese das Interaktionsmuster genauer vorhersagen und diese Hypothesen inferenzstatistisch absichern, so sind **A-priori-Einzelvergleiche** – speziell **Interaktionskontraste** – indiziert (Bortz & Schuster, 2010, S. 260).

Experimentelle, quasi-experimentelle und nicht-experimentelle zweifaktorielle Pläne Zweifaktorielle Pläne, die mit zweifaktoriellen Varianz- oder Kovarianzanalysen auszuwerten sind, können je nach Untersuchungsdesign ausgerichtet sein als:

- **nicht-experimentelle Studie** (z. B. Kombination von zwei Personenfaktoren: Faktor *A*: Ernährungsform: Vegetarier/Nichtvegetarier; Faktor *B*: Wohnort: Stadt/Land; die Ausprägungen beider Faktoren können nicht von den Forschenden beeinflusst werden, dementsprechend ist auch keine Zuordnung – weder zufällig noch nicht-zufällig – der Untersuchungspersonen zu den Bedingungen möglich),
- **quasi-experimentelle Studie** (z. B. Faktor *A*: Unterrichtsform: herkömmlicher Unterricht/Tablet-gestützter Unterricht; Faktor *B*: erfahrene Lehrkraft/unerfahrene Lehrkraft, beide Faktoren können zwar von den Forschenden gesetzt werden, die Zuordnung der Versuchspersonen erfolgt im Rahmen einer Feldstudie jedoch nicht individuell randomisiert, sondern systematisch durch Zuweisung vorgefundener Schulklassen zu den vier Untersuchungsbedingungen; durch sog. **Cluster-Randomisierung** können aber auch Studien mit vorgefundenen Gruppen in einem echten experimentellen Design untersucht werden) oder
- **experimentelle Studie** (z. B. Faktor *A*: Lernprogramm: Computerprogramm/Smartphone App; Faktor *B*: Lernstoff: Mathe/Englisch/Chemie; die Ausprägungen beider Faktoren werden von den Forschenden gesetzt, eine zufällige Zuordnung der einzelnen Versuchspersonen zu den 6 Untersuchungsbedingungen wird im Rahmen einer Laborstudie realisiert).

Das Design hat insofern einen Einfluss auf die statistische Datenanalyse, als bei quasi- und nicht-experimentellen Studien, die mangels Randomisierung fehlende designtechnische Kontrolle von Störvariablen soweit wie möglich über deren Messung und statistische Kontrolle als Kovariaten (im Rahmen einer **Kovarianzanalyse**) erfolgen sollte, um die interne Validität zu steigern. Entsprechende Kovariaten sind z. B. Personenvariablen oder das Ausgangsniveau der interessierenden abhängigen Variablen (Prä-Werte), deren Einfluss aus den Daten herausgerechnet wird, bevor man den Gruppenunterschied testet. Auch spezielle statistische Techniken wie die **Propensity-Score-Analyse**, die nicht-äquivalente Gruppen hinsichtlich der gemessenen Kovariaten analysiert und vergleichbar macht (Rosenbaum & Rubin, 1983), kommen in der Grundlagen- und Evaluationsforschung in den letzten Jahren verstärkt zum Einsatz (Luellen, Shadish, & Clark, 2005).

Kontrollfaktoren in zweifaktoriellen Plänen Zweifaktorielle Untersuchungspläne überprüfen simultan drei verschiedene Unterschiedshypothesen: zwei Haupteffekthypothesen und eine Interaktionshypothese. Diese drei Hypothesen müssen jedoch nicht immer explizit formuliert sein. Häufig steht nur eine Hypothese im Vordergrund (z. B. eine Hypothese über die unterschiedliche Wirkung verschiedener Treatments) und der zweite Faktor wird nur zu Kontrollzwecken eingeführt.

So wurde im Kontext der in ▶ Abschn. 3.2.2 berichteten Kontrolltechniken darauf hingewiesen, dass die gruppenkonstituierende unabhängige Variable durch andere Merkmale (Confounder: konfundierende Variablen) überlagert sein kann, die als Erklärung der gefundenen Gruppenunterschiede ebenfalls infrage kommen. Lässt sich hierbei ein Merkmal benennen, das mit hoher Wahrscheinlichkeit mit der unabhängigen Variablen konfundiert ist, kann dieses als Kontrollfaktor in die Untersuchung aufgenommen werden, obwohl sich die Forschungshypothese auf den anderen Faktor, die eigentlich interessierende unabhängige Variable bezieht.

Angenommen, es sollen Physiker und Informatiker hinsichtlich ihrer Abstraktionsfähigkeit verglichen werden. Als kritische Störvariable betrachten wir die Berufserfahrung der Untersuchungsteilnehmenden. Hier könnte es sinnvoll sein, neben dem Faktor „Beruf“ (Physiker vs. Informatiker) einen zweiten (Kontroll-)Faktor zu berücksichtigen, der die Untersuchungsteilnehmenden nach Maßgabe ihrer Berufserfahrung in homogene Teilgruppen (Blöcke) einteilt (z. B. wenig, mittel, viel Berufserfahrung). Damit ist derjenige Varianzanteil der abhängigen Variablen, der auf die Berufserfahrung bzw. die Interaktion der beiden Faktoren zurückgeht, varianzanalytisch bestimmbar und die zwischen den Berufsgruppen registrierten Unterschiede sind von der Berufserfahrung unabhängig.

		mit Humor (A_1)		ohne Humor (A_2)		← Humor
		Schwer (B_1)	Leicht (B_2)	Schwer (B_1)	Leicht (B_2)	← Unterrichtsstoff
Reputation	Hoch (C_1)	S_{111}	S_{121}	S_{211}	S_{221}	
	Gering (C_2)	S_{112}	S_{122}	S_{212}	S_{222}	

Abb. 12.40 Untersuchungsschema eines dreifaktoriellen 2 × 2 × 2-Plans

Drei- und mehrfaktorielle Pläne

In faktoriellen Untersuchungsplänen können nicht nur zwei, sondern auch drei oder mehr Faktoren (unabhängige Variablen) sowie deren Interaktionen simultan untersucht werden. Bei vollständigen mehrfaktoriellen Plänen ist darauf zu achten, dass die Stufen eines jeden Faktors mit den Stufen aller anderen Faktoren kombiniert werden und dass unter jeder Faktorstufenkombination eine Teilstichprobe des Umfanges n untersucht wird (bei ungleich großen Stichproben spricht man von nichtorthogonalen Varianzanalysen; Bortz & Schuster, 2010, S. 274). Allerdings nimmt die Anzahl der benötigten Untersuchungsteilnehmenden mit wachsender Faktorzahl exponentiell zu: Ein dreifaktorieller Plan mit jeweils 2 Stufen pro Faktor benötigt $(2 \times 2 \times 2) \times n$ Untersuchungsobjekte, ein vierfaktorieller Plan $2^4 \times n$ Untersuchungsobjekte etc. Für einen dreifaktoriellen Plan mit beliebigen Faktorstufenzahlen p, q und r benötigt man insgesamt $p \times q \times r \times n$ Untersuchungsobjekte. Allgemein werden faktorielle Pläne über die Faktorstufenzahl der einzelnen Faktoren benannt, man spricht also z. B. von einem „2 × 2 × 2 × 4-Design“ oder einem „3 × 2 × 2 × 5-Design“.

Als Beispiel für Hypothesen, die mit einem 2 × 2 × 2-Plan prüfbar sind, wählen wir eine Untersuchung von Perry, Abrami, Leventhal und Check (1979; zit. nach Spector, 1981). Diese Untersuchung überprüft die Hypothesen, dass die Bewertung des Unterrichtes eines Dozenten von seinem Humor, dem Schwierigkeitsgrad des Unterrichtsstoffs und der Reputation des Dozenten abhängt. Die Autoren fertigten acht Videoaufnahmen des Unterrichtes eines Dozenten an, die sich in Bezug auf folgende drei Faktoren unterschieden:

- **Faktor *A*:** Humor des Dozenten (A_1: mit Humor, A_2: ohne Humor),
- **Faktor *B*:** Schwierigkeitsgrad des Unterrichtsstoffes (B_1: schwer, B_2: leicht),
- **Faktor *C*:** Reputation des Dozenten (C_1: Dozent wird als Person mit hoher Reputation vorgestellt, C_2: Dozent wird als Person mit geringer Reputation vorgestellt).

In Abb. 12.40 wird dieser Untersuchungsplan grafisch veranschaulicht. Jeder Faktorstufenkombination wird eine Stichprobe S des Umfanges n zugewiesen, d. h., jede Videoaufnahme wird von n Untersuchungspersonen (hier: Studierenden) beurteilt. Die Teilstichproben der 8 Gruppen addieren sich zum Gesamtstichprobenumfang.

Dreifaktorielle Pläne werden ebenfalls varianzanalytisch ausgewertet. Eine **dreifaktorielle Varianzanalyse** („three-way ANOVA“) überprüft sieben voneinander unabhängige Unterschiedshypothesen: drei Haupteffekte (A, B, C), drei Interaktionen erster Ordnung ($A \times B$, $A \times C$, $B \times C$) und eine **Interaktion zweiter Ordnung** ($A \times B \times C$). Üblicherweise ist man jedoch nicht in der Lage, alle sieben Hypothesen vor Untersuchungsbeginn genau zu begründen, sondern nur einige. Werden dennoch alle sieben Effekte geprüft, sind signifikante Effekte, zu denen keine theoretisch fundierten Hypothesen formuliert wurden, nur deskriptiv zu verwerten. Zu beachten ist allerdings, dass die Teststärke varianzanalytischer Auswertungen in hohem Maße davon abhängt, welche bzw. wie viele Hypothesen geprüft werden sollen (Maxwell, 2004; ► Abschn. 14.1).

Interaktionen zweiter Ordnung Interaktionen zweiter oder höherer Ordnung sind meistens schwer interpretierbar. Ihre Bedeutung wird leichter erkennbar, wenn man sie grafisch illustriert (Abb. 12.41).

Zur Veranschaulichung wählen wir erneut das oben genannte Beispiel zur Unterrichtsbewertung (allerdings mit fiktiven Daten). Die beiden linken Abbildungen verdeutlichen zusammengenommen, dass die Interaktion zweiter Ordnung unbedeutend ist. Die Tatsache, dass der Unterrichtsstil humorvoll oder nicht humorvoll war, spielte bei einem schweren Unterrichtsstoff kaum eine Rolle. War der Unterrichtsstoff hingegen leicht, wurde ein Unterricht mit Humor weitaus positiver bewertet als ein Unterricht ohne Humor (Interaktion $A \times B$). Diese Interaktion war – wie die beiden linken Abbildungen zeigen – von der Reputation des Dozenten unabhängig (keine $A \times B \times C$-Interaktion).

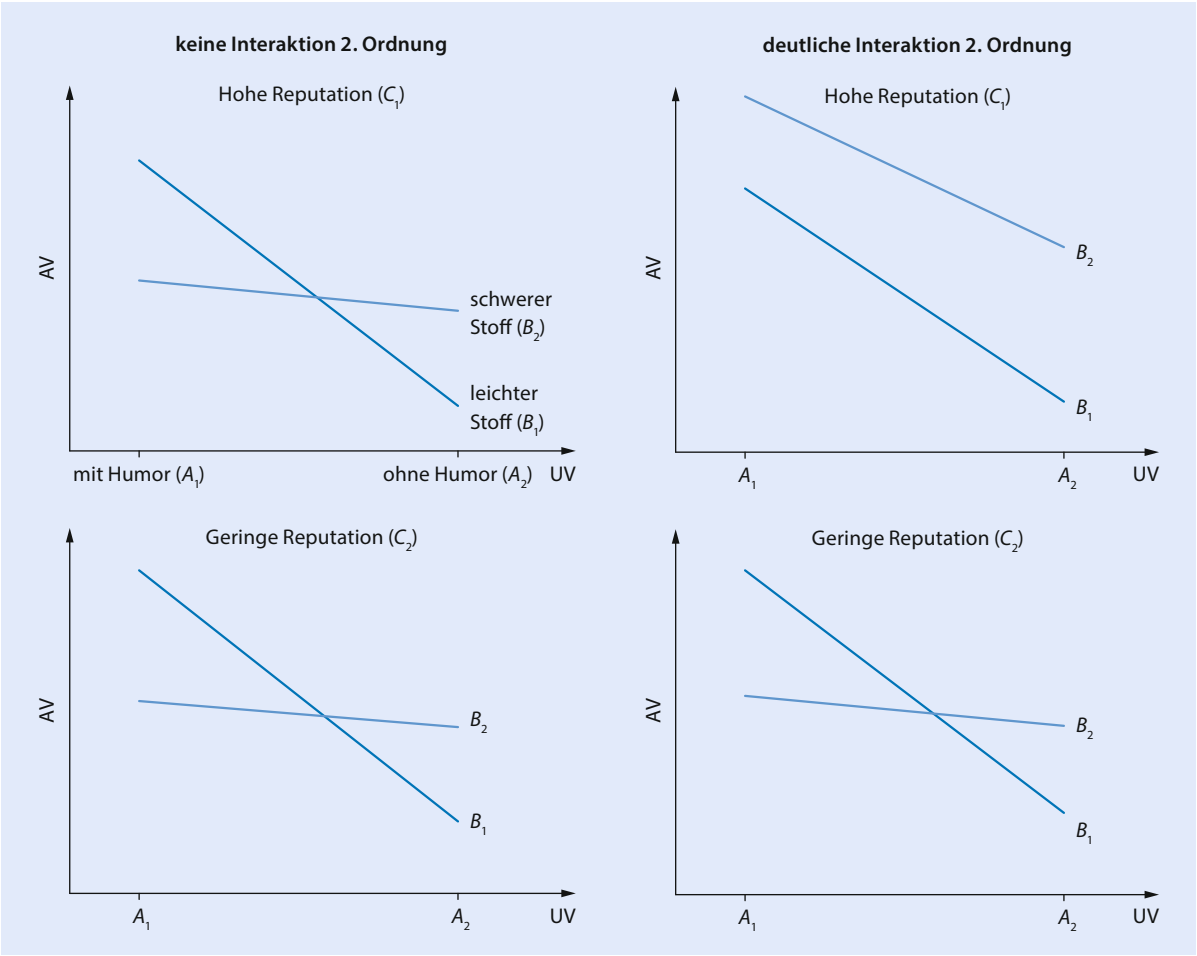


Abb. 12.41 Grafische Darstellung einer Interaktion zweiter Ordnung. *AV* abhängige Variable, *UV* unabhängige Variable

Im Unterschied hierzu verdeutlichen die beiden rechten Abbildungen eine deutliche $A \times B \times C$-Interaktion: Für Dozenten mit geringer Reputation galt die oben beschriebene $A \times B$-Interaktion praktisch unverändert. Verfügte ein Dozent jedoch über eine hohe Reputation, wurde ein schwieriger Unterrichtsstoff unabhängig davon, ob der Unterrichtsstil humorvoll war oder nicht, positiver bewertet als ein leichter Unterrichtsstoff. Zusätzlich wurde auch hier ein humorvoller Unterricht besser beurteilt als ein humorfreier Unterricht.

Allgemein gilt: Ist das Muster der $A \times B$-Interaktion auf allen Stufen des Faktors C ungefähr gleich, besteht keine Interaktion zweiter Ordnung. Unterscheiden sich die Muster der $A \times B$-Interaktion für verschiedene C-Stufen, ist dies als deskriptiver Hinweis auf eine Interaktion zweiter Ordnung zu werten.

Statt die $A \times B$-Interaktion für die Stufen des Faktors C darzustellen, hätte man auch die $A \times C$-Interaktion für die Stufen des Faktors B bzw. die $B \times C$-Interaktion für die Stufen des Faktors A grafisch veranschaulichen können. Grundsätzlich sollte diejenige Darstellungsart gewählt werden, die die inhaltliche Bedeutung der Interaktion möglichst einfach und treffend beschreibt.

Interaktion zweiter Ordnung – Wir sprechen von einer Interaktion zweiter Ordnung ($A \times B \times C$-Interaktion, Tripel-Interaktion, „second order interaction"), wenn die Art der Interaktion zwischen zwei Faktoren (Interaktion erster Ordnung, z. B. $A \times B$) von den Stufen eines dritten Faktors (z. B. Faktor C) abhängt.

Hierarchische Pläne

Nur selten werden alle Hypothesen, die ein vollständiger mehrfaktorieller Plan prüft, tatsächlich vor Untersuchungsbeginn explizit formuliert. Meistens sind es Interaktionen höherer Ordnung, über die man keine Hypothesen formulieren kann oder will, weil sie nicht interessieren. Dennoch werden viele Fragestellungen mit vollständigen, mehrfaktoriellen Untersuchungen geprüft, obwohl dieser Untersuchungsplan mehr Fragen beantwortet als ursprünglich gestellt wurden.

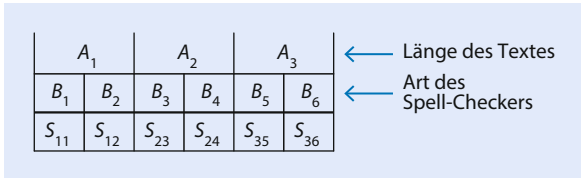


Abb. 12.42 Zweifaktorieller hierarchischer Plan

Dieser „Luxus“ erfordert einen untersuchungstechnischen Aufwand, der sich reduzieren lässt, wenn man statt vollständiger, mehrfaktorieller Pläne **unvollständige Pläne** einsetzen kann, die nur einige der möglichen Faktorstufenkombinationen berücksichtigen. Zu diesen unvollständigen Plänen gehören die hierarchischen und teilhierarchischen sowie die im nächsten Abschnitt zu behandelnden quadratischen Pläne. Beide Planvarianten können experimentell, quasi-experimentell oder nicht-experimentell realisiert sein.

Faktorielle Versuchspläne, bei denen durch die Schachtelung des einen Faktors unter den anderen Faktor (geschachtelter Faktor, „nested factor“) eine Hierarchie der Faktoren entsteht, bezeichnet man als *hierarchische Pläne* („hierarchical design“, „nested designs“). Stehen nicht alle, sondern nur einige Faktoren in hierarchischer Beziehung, so spricht man von einem *teilhierarchischen Plan* („semi-nested designs“).

Während bei vollständigen Plänen („full factorial designs“, „fully-crossed factorial designs“) alle Haupteffekte und Interaktionen geprüft werden können, lassen sich bei hierarchischen Plänen keine Interaktionseffekte und bei teilhierarchischen Plänen nur Interaktionen zwischen den vollständig miteinander kombinierten Faktoren testen. Die inferenzstatistische Auswertung erfordert einige Modifikationen im Vergleich zur Varianzanalyse mit vollständigen Plänen.

Zweifaktorielle hierarchische Pläne

In Abb. 12.42 findet sich ein zweifaktorieller hierarchischer Versuchsplan. Hier werden nicht alle Stufen des Faktors B mit allen Stufen von A kombiniert, sondern zwei Stufen von B mit A_1, zwei weitere Stufen von B mit A_2 und die letzten beiden der sechs B-Stufen mit A_3. Allgemein: Jede der p Stufen eines Faktors A ist mit anderen q Stufen eines Faktors B kombiniert. Wir sagen: Die Stufen des Faktors B sind unter die Stufen des Faktors A „**geschachtelt**“ („nested“).

Mit diesem hierarchischen bzw. geschachtelten Plan lassen sich z. B. die Unterschiedshypothesen überprüfen, dass die Rechtschreibleistung am Computer (operationalisiert als Anzahl der Rechtschreibfehler pro getippter Seite) von der Länge des Textes (Faktor A: 1 Seite, 3 Seiten, 15 Seiten) und vom verwendeten Rechtschreibkorrekturprogramm (Faktor B: 6 verschiedene Spell-Checkers) abhängt, wobei jeweils zwei Rechtschreibkorrekturprogramme den drei Textvarianten zugeordnet sind.

Der Vorteil dieser Untersuchungsanlage liegt auf der Hand. Statt der 18 Stichproben, die ein vollständiger zweifaktorieller Plan zur Überprüfung der genannten Hypothesen erfordert, kommt der hierarchische Plan mit nur 6 Stichproben bzw. Untersuchungsgruppen aus. Hierarchische Pläne erfordern also weniger Untersuchungseinheiten als vollständige Pläne.

Diesem Vorteil steht jedoch ein gravierender Nachteil gegenüber: Unterschiede zwischen den Stufen des Faktors A sind nur in Verbindung mit den jeweiligen Stufen des Faktors B, die unter die entsprechenden Stufen des Faktors A geschachtelt sind, interpretierbar. Der Effekt, den die Länge des Textes auslöst, gilt nur für die Rechtschreibkorrekturprogramme, mit denen der jeweilige Text geschrieben wurde. In ähnlicher Weise können auch Unterschiede zwischen den Stufen von B durch Effekte des Faktors A überlagert sein. Die Faktoren A und B sind nur dann eindeutig interpretierbar, wenn Texteffekte von der Art der Spell-Checkers und Spell-Checker-Effekte von der Art der Texte unabhängig sind oder kurz: wenn zwischen den Faktoren keine Interaktion besteht.

Interaktionen sind jedoch in hierarchischen Plänen statistisch nicht überprüfbar. Die Interpretierbarkeit eines hierarchischen Versuchsplans hängt deshalb davon ab, ob sich theoretisch rechtfertigen oder durch andere Untersuchungen belegen lässt, dass Interaktionen höchst unwahrscheinlich sind oder die Interpretation der Haupteffekte von möglichen Interaktionseffekten nicht bedeutsam beeinträchtigt wird.

Angesichts dieser Einschränkungen könnte man den praktischen Wert hierarchischer Versuchspläne bezweifeln. Dem ist entgegenzuhalten, dass manche Fragestellungen überhaupt nur mit hierarchischen Plänen überprüfbar sind, weil die vollständige Kombination der Faktorstufen unsinnig oder unmöglich wäre. Eine Untersuchung, die – etwa im Rahmen einer **multizentrischen Studie** („multi-site study“, d. h. einer Studie, an der mehrere Forschungszentren beteiligt sind) – die Wirksamkeit verschiedener therapeutischer Techniken überprüft, kann darauf angewiesen sein, Kliniken zu finden, die sich auf die zu vergleichenden Therapien spezialisiert haben. Bei dieser Fragestellung wäre es unrealistisch, davon auszugehen, dass jede Klinik jede der zu untersuchenden Therapien praktiziert. Man könnte vielmehr feststellen, dass Therapie A_1 in den Kliniken B_1, B_2 und B_3, The-

A_1			A_2			A_3												← Therapien
B_1			B_2			B_3			B_4			B_5			B_6			← Kliniken
C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	← Ärzte
S_{111}	S_{112}	S_{113}	S_{124}	S_{125}	S_{126}	S_{237}	S_{238}	S_{239}	S_{2410}	S_{2411}	S_{2412}	S_{3513}	S_{3514}	S_{3515}	S_{3516}	S_{3617}	S_{3618}	

Abb. 12.43 Dreifaktorieller vollständig hierarchischer Plan

rapie A_2 in den Kliniken B_4, B_5 und B_6 etc. zum Einsatz gelangen, d. h., für die Untersuchung kommt prinzipiell nur ein hierarchischer Plan infrage.

Erweist sich der Haupteffekt „Art der Therapie" als statistisch signifikant, ist dieses Ergebnis nur in Verbindung mit denjenigen Kliniken, die die untersuchten Therapien praktizieren, interpretierbar. Umgekehrt muss bei bedeutsamen Klinikunterschieden in Rechnung gestellt werden, dass die Kliniken verschiedene Therapien einsetzen. Diese interpretativen Vorbehalte entfallen, wenn man davon ausgehen kann, dass zwischen den Faktoren „Art der Therapie" und „Kliniken" keine Interaktion besteht, dass also die Wirksamkeit einer Therapie nicht davon abhängt, in welcher Klinik sie durchgeführt wird.

Besonderheiten dieses Plans v. a. im Hinblick auf die Kalkulation von Effektgrößen, findet man bei Wampold und Serlin (2000). Mit Teststärkeüberlegungen in multizentrischen Untersuchungen befassen sich Raudenbusch und Liu (2000).

Dreifaktorielle hierarchische Pläne

Von einem vollständig hierarchischen dreifaktoriellen Plan spricht man, wenn nicht nur die Stufen des Faktors B unter die Stufen des Faktors A, sondern auch die Stufen eines dritten Faktors C unter die Stufen von B geschachtelt sind. Im zuletzt genannten Beispiel könnte man sich zusätzlich dafür interessieren, ob der Therapieerfolg auch vom behandelnden Arzt (Therapeuten) abhängt. Erneut müssen wir davon ausgehen, dass ein Arzt bzw. eine Ärztin nicht alle untersuchten Therapien beherrscht und schon gar nicht gleichzeitig in allen untersuchten Kliniken praktiziert, d. h., auch dieser Faktor lässt sich nicht vollständig mit allen Stufen der beiden übrigen Faktoren kombinieren. Man wird deshalb pro Klinik verschiedene Ärztinnen und Ärzte auswählen und erhält damit den in Abb. 12.43 dargestellten Untersuchungsplan (mit $p = 3$ verschiedenen Therapien, $q = 2$ Kliniken pro Therapie und $r = 3$ Ärzten pro Klinik).

Gegenüber dem entsprechenden vollständigen dreifaktoriellen Plan mit $3 \times 6 \times 18 = 324$ Untersuchungsgruppen bzw. Patientenstichproben des Umfanges n benötigt der hierarchische Plan nur 18 Stichproben. Überprüfbar sind mit diesem Plan nur die drei Haupteffekte, deren Interpretation den gleichen Restriktionen unterliegt wie die Haupteffekte eines zweifaktoriellen hierarchischen Plans.

	A_1		A_2		A_3		← Therapien
	B_1	B_2	B_3	B_4	B_5	B_6	← Kliniken
C_1(weiblich)	S_{111}	S_{121}	S_{231}	S_{241}	S_{351}	S_{361}	
C_2(männlich)	S_{112}	S_{122}	S_{232}	S_{242}	S_{352}	S_{362}	

Abb. 12.44 Dreifaktorieller teilhierarchischer Plan

Teilhierarchische Pläne

Lässt sich im Beispiel die Hypothese begründen, dass der Therapieerfolg zusätzlich auch vom Geschlecht der Patienten (oder einem anderen Merkmal) abhängt, ist es ratsam, die Untersuchung nach einem mehrfaktoriellen teilhierarchischen Plan anzulegen. Der Plan heißt deshalb teilhierarchisch, weil der Geschlechtsfaktor mit allen 3 Faktoren kombinierbar ist. Verzichten wir auf den Ärztefaktor, resultiert der in Abb. 12.44 wiedergegebene dreifaktorielle teilhierarchische Plan.

Mit diesem Plan überprüft man nicht nur die drei Haupteffekte, sondern zusätzlich auch die Interaktionen zwischen denjenigen Faktoren, die vollständig miteinander kombiniert sind (im Beispiel: $A \times C$ und $B \times C$).

Quadratische Pläne

Das Untersuchungsschema eines zweifaktoriellen Plans, der zwei Faktoren mit gleicher Stufenzahl kontrolliert, lässt sich als ein Quadrat darstellen. Wenn jeder Faktor p Stufen aufweist, erfordert dieser Plan p^2 Stichproben des Umfanges n, wobei die Gruppengrößen möglichst ähnlich sein sollten, aber nicht identisch sein müssen. Er überprüft zwei Haupteffekthypothesen und eine Interaktionshypothese. Mit dem gleichen Aufwand an Untersuchungspersonen lassen sich jedoch auch drei Haupteffekthypothesen testen, vorausgesetzt, alle Faktoren haben die gleiche Stufenzahl.

Quadratische Pläne – Quadratische Pläne sind eine Sonderform der hierarchischen bzw. unvollständigen Pläne, bei denen alle Faktoren die

gleiche Stufenzahl aufweisen. Mit Plänen dieser Art können nur Haupteffekte überprüft werden. Varianten quadratischer Pläne sind lateinische sowie griechisch-lateinische Quadrate.

Das für einen zweifaktoriellen quadratischen Plan erforderliche Untersuchungsschema ist in Abb. 12.45 wiedergegeben.

Lateinische Quadrate

Der in Abb. 12.45 dargestellte Plan geht davon aus, dass jeder Faktor 4 Stufen hat, d. h., insgesamt benötigt der Plan $4 \times 4 = 16$ Stichproben. Die erste Stichprobe wird der Faktorstufenkombination $A_1 \times B_1 \times C_1$ zugewiesen, die zweite Stichprobe der Kombination $A_1 \times B_2 \times C_2$, die dritte Stichprobe der Kombination $A_1 \times B_3 \times C_3$ etc. Wie man dem Untersuchungsschema leicht entnehmen kann, ist jede Stufe eines jeden Faktors mit allen Stufen der beiden übrigen Faktoren vollständig kombiniert. Man sagt, der Plan ist in Bezug auf die Haupteffekte **ausbalanciert**. Pläne dieser Art heißen lateinische Quadrate („latin square designs").

Als Beispiel für ein lateinisches Quadrat (mit $p = 3$) wählen wir folgende quasi-experimentelle Untersuchung: Es soll überprüft werden, ob die Einstellung zur Kernenergie abhängt von

- der in einem Haushalt zur Wärmeerzeugung genutzten Energieart (Faktor A: Kohle, Öl, elektrischer Strom),
- dem Lebensalter (Jugend-, Erwachsenen-, Seniorenalter) und/oder
- der Wohngegend (ländlich, kleinstädtisch, großstädtisch).

Man benötigt damit Stichproben aus folgenden Populationen:

Stichprobe 1:	Kohle – Jugendalter – ländlich
Stichprobe 2:	Kohle – Erwachsenenalter – kleinstädtisch
Stichprobe 3:	Kohle – Seniorenalter – großstädtisch
Stichprobe 4:	Öl – Jugendalter – kleinstädtisch
...	...
Stichprobe 9:	Strom – Seniorenalter – kleinstädtisch

Als Konstruktionsprinzip für die Erstellung eines lateinischen Quadrates wählt man einfachheitshalber die sog. **zyklische Permutation**. Hierbei enthält die erste Zeile des lateinischen Quadrates die C-Stufen in natürlicher Abfolge. Die zweite Zeile bilden wir, indem zu den Indizes der ersten Zeile der Wert 1 addiert und von dem Index, der durch die Addition den Wert $p + 1$ erhält, p abgezogen wird. Wird die zweite Zeile in gleicher Weise geändert, resultiert die dritte Zeile etc. Man erhält so eine Anordnung, die als Standardform eines lateinischen Quadrates bezeichnet wird. Für ein lateinisches Quadrat mit $p = 5$ Stufen ergibt sich folgende Standardform für die

	A_1	A_2	A_3	A_4
B_1	C_1	C_2	C_3	C_4
B_2	C_2	C_3	C_4	C_1
B_3	C_3	C_4	C_1	C_2
B_4	C_4	C_1	C_2	C_3

Abb. 12.45 Lateinisches Quadrat ($p = 4$)

Abfolge der C-Stufen:

$$C_1 \times C_2 \times C_3 \times C_4 \times C_5$$
$$C_2 \times C_3 \times C_4 \times C_5 \times C_1$$
$$C_3 \times C_4 \times C_5 \times C_1 \times C_2$$
$$C_4 \times C_5 \times C_1 \times C_2 \times C_3$$
$$C_5 \times C_1 \times C_2 \times C_3 \times C_4$$

(Auf die Eintragung der Faktoren A und B wurde verzichtet.)

Vollständige mehrfaktorielle Pläne sind nicht nur in Bezug auf die Haupteffekte, sondern auch in Bezug auf die Interaktionen ausbalanciert. Letzteres gilt nicht für lateinische Quadrate. Wie man Abb. 12.45 leicht entnehmen kann, ist z. B. die Stufe C_1 nur mit $A_1 \times B_1$, $A_4 \times B_2$, $A_3 \times B_3$ und $A_2 \times B_4$ kombiniert. Die verbleibenden 12 $A \times B$-Kombinationen sind mit anderen C-Stufen verbunden. Dies hat nicht nur zur Folge, dass in lateinischen Quadraten keine Interaktionshypothesen geprüft werden können; zusätzlich sind die Haupteffekte nur dann eindeutig interpretierbar, wenn die Interaktionen zwischen den Faktoren zu vernachlässigen sind (zur inferenzstatistischen Auswertung entsprechender Pläne s. z. B. Bortz & Schuster, 2010, Kap. 20).

Griechisch-lateinische Quadrate

Neben der Standardform gibt es weitere Anordnungen, die ebenfalls den Anforderungen eines lateinischen Quadrates genügen, d. h. in Bezug auf die Haupteffekte ausbalanciert sind. Ein „griechisch-lateinisches Quadrat" („graeco-latin square design") entsteht, wenn man zwei lateinische Quadrate kombiniert und diese zueinander orthogonal sind. Zwei lateinische Quadrate sind orthogonal, wenn deren Kombination zu einer neuen Anordnung führt, in der jede Zweierkombination der Faktorstufen genau einmal vorkommt. Der Unterschied zwischen orthogonalen und nicht-orthogonalen lateinischen Quadraten wird in Abb. 12.46 verdeutlicht.

Die beiden aufgeführten lateinischen Quadrate bezeichnet man als nicht-orthogonal, weil deren Kombi-

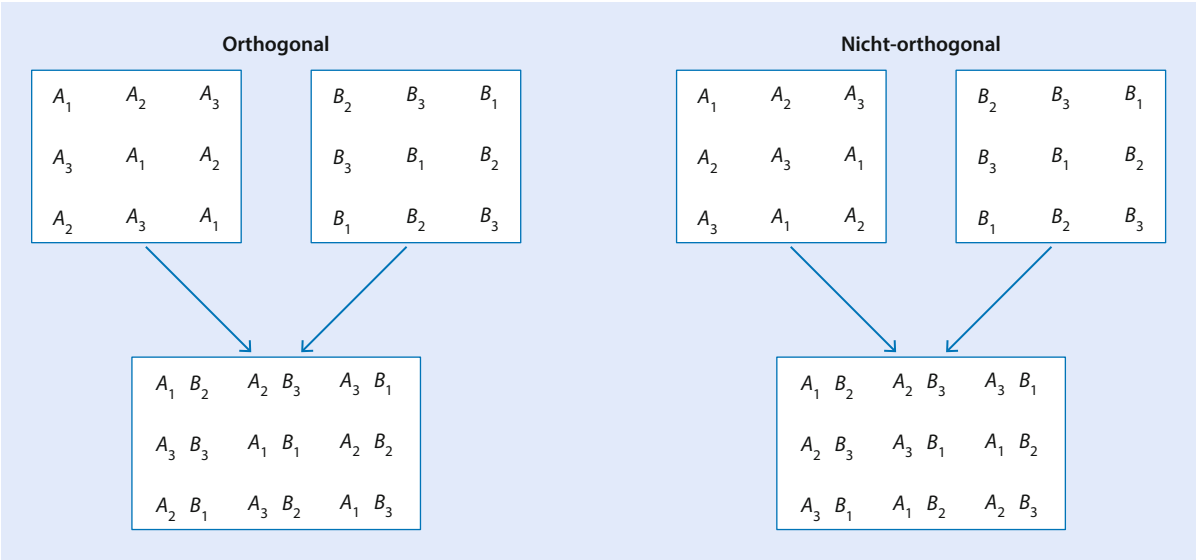


Abb. 12.46 Orthogonale und nicht-orthogonale lateinische Quadrate

	A_1	A_2	A_3	A_4
B_1	C_1 D_1	C_2 D_3	C_3 D_4	C_4 D_2
B_2	C_2 D_2	C_1 D_4	C_4 D_3	C_3 D_1
B_3	C_3 D_3	C_4 D_1	C_1 D_2	C_2 D_4
B_4	C_4 D_4	C_3 D_2	C_2 D_1	C_1 D_3

Abb. 12.47 Griechisch-lateinisches Quadrat ($p = 4$)

nation zu einer Anordnung führt, in der sich die Faktorstufenpaare $A_1 \times B_2$, $A_2 \times B_3$ und $A_3 \times B_1$ jeweils dreimal wiederholen. Die beiden linken lateinischen Quadrate hingegen sind orthogonal, denn deren Kombination enthält alle Faktorstufenpaare.

Mit griechisch-lateinischen Quadraten können in einer Untersuchung 4 Faktoren kontrolliert werden. Will man z. B. experimentell überprüfen, wie sich vier verschiedene Lärmbedingungen (Faktor A), vier Temperaturbedingungen (Faktor B), vier Beleuchtungsbedingungen (Faktor C) und vier Luftfeuchtigkeitsbedingungen (Faktor D) auf die Arbeitszufriedenheit von Fließbandarbeitern auswirken, kann statt eines vollständigen vierfaktoriellen Plans das in Abb. 12.47 dargestellte, weniger aufwändige griechisch-lateinische Quadrat eingesetzt werden.

Statt der $4^4 = 256$ Stichproben des Umfanges n im vollständigen Plan kommt das griechisch-lateinische Quadrat mit nur 16 Stichproben aus. Mit jeder dieser 16 Stichproben wird eine andere Kombination der vier Faktoren untersucht. Die Kombinationen sind so zusammengestellt, dass die Stufen eines jeden Faktors mit allen Stufen der verbleibenden drei Faktoren genau einmal verbunden sind, d. h., auch dieser Plan ist in Bezug auf die Haupteffekte ausbalanciert. Interaktionen sind erneut nicht prüfbar und sollten für eine bessere Interpretierbarkeit der Haupteffekte zu vernachlässigen sein.

Untersuchungen nach dem Schema eines griechisch-lateinischen Quadrates sind durchführbar, wenn die **Faktorstufenzahl aller Faktoren gleich** ist und die Konstruktion zweier orthogonaler lateinischer Quadrate zulässt. Dies ist nur der Fall, wenn die **Faktorstufenzahl als ganzzahlige Potenz einer Primzahl** darstellbar ist (z. B. $p = 3 = 3^1$, $p = 4 = 2^2$, $p = 5 = 5^1$ etc.). Für $p = 6$ und $p = 10$ lassen sich z. B. keine griechisch-lateinischen Quadrate konstruieren. (Näheres hierzu vgl. z. B. Cochran & Cox, 1966, S. 146ff.).

Über die statistische Auswertung lateinischer bzw. griechisch-lateinischer Quadrate wird z. B. bei Bortz & Schuster (2010, Kap. 20) berichtet.

Pläne mit Kontrollvariablen

Um Unterschiede zwischen Treatment- und Kontrollgruppen in der bzw. den abhängigen Variablen eindeutig auf das Treatment zurückführen zu können, sind vergleichbare Ausgangsbedingungen in den Gruppen notwendig. Diese können am besten in einem Laborexperiment hergestellt werden. Viele Forschungsfragen – insbesondere auch in der Evaluationsforschung – werden jedoch im Rahmen von quasi-experimentellen Feldstudien untersucht. Hier kann die Äquivalenz der Gruppen hinsichtlich personen- und untersuchungsbedingter Störvariablen de-

signtechnisch nicht sichergestellt werden. Stattdessen gilt es, mögliche Störvariablen im Vorfeld zu identifizieren und im Zuge der Untersuchung zu messen, so dass sie als Kontrollvariablen in die statistische Analyse eingehen können. Die Identifikation von relevanten Kontrollvariablen erfolgt auf der Basis von Theorien sowie des aktuellen Forschungsstandes, ggf. auch eigener Vorstudien. Das Arbeiten mit Kontrollvariablen ist umso erfolgreicher, je vollständiger die relevanten Kontrollvariablen erfasst werden und je besser sie operationalisiert sind. Kontrollvariablen werden teilweise auch in echten Experimenten erhoben und sind für quasi- und nicht-experimentelle Studien unverzichtbar.

Eine verbreitete Auswertungstechnik, um den Einfluss von Kontrollvariablen auf die abhängige Variable zu eliminieren, ist die **Kovarianzanalyse**. Die Eliminierung des Einflusses einer Kontrollvariablen auf die abhängige Variable geschieht hierbei regressionstechnisch mittels der **Partialkorrelation** (► Abschn. 12.5.2).Will man z. B. den Behandlungserfolg verschiedener psychotherapeutischer Techniken evaluieren und hält es für wahrscheinlich, dass die Heilerfolge auch von der Verbalisierungsfähigkeit der Patienten abhängen, wird diese als Kontrollvariable miterhoben. Die kovarianzanalytische Auswertung der Untersuchung führt zu Ergebnissen, die die Wirkung der psychotherapeutischen Methoden unabhängig von den Verbalfähigkeiten der untersuchten Patienten widerspiegeln. Hier hat die kovarianzanalytische Berücksichtigung einer Kontrollvariablen in erster Linie die Funktion, die Fehlervarianz der abhängigen Variablen zu reduzieren (Näheres hierzu bei Bortz & Schuster, 2010, Kap. 19; für ein regressionsanalytisches Pendant zur klassischen Kovarianzanalyse s. Eid, Schuster, & Schmidt, 2010, S. 664ff.).

Eine Kontrollvariable („control variable“, „confounder“) ist eine Störvariable, deren Ausprägung bei jedem Untersuchungsobjekt gemessen wurde und deren Einfluss aus der abhängigen Variablen herausgerechnet (herauspartialisiert) wird.

Im Kontext der Prüfung von Unterschiedshypothesen auf der Basis von Daten aus Gruppenplänen wird anstelle der Varianzanalyse die Kovarianzanalyse („analysis of covariance“, ANCOVA) genutzt, um den Einfluss der unabhängigen Variablen auf die – um den Einfluss der von Kontrollvariablen bereinigte – abhängige Variable zu bestimmen.

Grundsätzlich besteht auch die Möglichkeit, in einer Untersuchung mehrere Kontrollvariablen zu berücksichtigen. Das kovarianzanalytische Auswertungsmodell entspricht dann dem einer Partialkorrelation höherer Ordnung. **Nominalskalierte Kontrollvariablen** können ebenfalls kovarianzanalytisch verarbeitet werden, wenn sie zuvor als Dummy-Variablen kodiert wurden (vgl. ► Abschn. 12.5.2, ► „Kodierung eines polytom-nominalen Merkmals durch dichotome Indikatorvariablen“). Die Kovarianzanalyse ist an einige Voraussetzungen geknüpft, die die Breite ihrer Einsatzmöglichkeiten einschränken (Bortz & Schuster, 2010, Kap. 19). Erweist sich das kovarianzanalytische Auswertungsmodell für eine konkrete Untersuchung als unangemessen, sollte das Blockbildungsverfahren („randomized block design“, ► Abschn. 12.5.4) vorgezogen werden (ausführliche Informationen zum Vergleich von Kovarianzanalyse und Randomized-Block-Design findet man z. B. bei Feldt, 1958; vgl. zu dieser Thematik auch Little, An, Johanns, & Giordani, 2000).

Um Unterschiede zwischen Treatment- und Kontrollgruppen in der bzw. den abhängigen Variablen eindeutig auf das Treatment zurückführen zu können, sind vergleichbare Ausgangsbedingungen in den Gruppen notwendig. Diese werden oft im Rahmen einer Vorher-Messung erhoben, so dass die Vorher-Messungen bzw. **Prä-Werte als Kontrollvariablen** genutzt werden können, um die Post-Werte zu bereinigen – in diesem Fall liegt durch die Messwiederholung eine Veränderungsmessung vor (► Abschn. 12.5.4).

Die **Propensity-Score-Analyse** ist ein fortgeschrittenes Verfahren, mit dem eine statistische Vergleichbarkeit nicht-äquivalenter Gruppen bei Quasi- und Nicht-Experimenten hergestellt werden soll. Bereits in den 1980er-Jahren entwickelt (Rosenbaum & Rubin, 1983), hat die Propensity-Score-Analyse erst in den letzten Jahren breitere Anwendung gefunden, z. B. in der biomedizinischen Forschung sowie in der Evaluationsforschung. Im Rahmen der Propensity-Score-Analyse wird anhand aller betrachteten Kontrollvariablen die Wahrscheinlichkeit berechnet, mit der eine Untersuchungsperson der jeweiligen Untersuchungsgruppe angehört. Für die Berechnung der Propensity-Scores stehen unterschiedliche Ansätze zur Verfügung. Die Anwendung der Propensity-Scores ist unterschiedlich. Sie können z. B. designtechnisch für ein Matching von Untersuchungsgruppen genutzt werden oder auch als eine Kovariate aus der abhängigen Variablen herauspartialisiert werden (Guo & Fraser, 2010; Harder, Stuart, & Anthony, 2010; Pruzek, 2011; Rubin, 1997).

Man beachte, dass die Kontrollvariablen normalerweise mit der abhängigen und nicht mit der unabhängigen Variablen korrelieren sollten. Eine Korrelation zwischen der unabhängigen Variablen und einer Kontrollvariablen kann den eigentlich interessierenden Effekt reduzieren oder gar zum Verschwinden bringen (Lieberson, 1985).

In Sonderfällen kann es jedoch erwünscht sein, einen Effekt durch das Herauspartialisieren einer Kontrollvariablen „zum Verschwinden“ zu bringen. Dies ist immer dann der Fall, wenn die **Forschungshypothese der Null-**

hypothese entspricht und sich ein unerwarteter Effekt einstellt, den man aufgrund des theoretischen Vorwissens zunächst nicht zum Anlass nehmen will, die Nullhypothese zu falsifizieren.

Beispiel: In einer wissenschaftssoziologischen Untersuchung werden Publikationsregeln und Publikationsverhalten in unterschiedlichen Fachdisziplinen verglichen. Dabei könnte sich herausstellen, dass die in der Studie erfassten medizinischen Forschungsinstitute im Durchschnitt signifikant mehr Publikationen aufweisen als die pharmakologischen. Bevor dieser Effekt interpretiert wird (etwa anhand von Umfang und Anspruch der Studien in den beiden Fächern etc.), versucht man zu klären, ob es sich bei dem Effekt möglicherweise um ein Artefakt handelt. So wäre es z. B. möglich, dass die untersuchten medizinischen Forschungseinrichtungen besser mit Hilfskräften ausgestattet sind und es sich somit nicht um einen fachspezifischen Effekt, sondern allein um einen personellen Ausstattungsfaktor handelt. Ein Herauspartialisieren der Anzahl der den Forschenden zugeordneten Hilfskraftstellen müsste im Falle eines Artefakts die Gruppenunterschiede nivellieren.

Multivariate Pläne

Alle bisher behandelten Pläne zur Überprüfung von Unterschiedshypothesen gingen davon aus, dass jeweils nur ***eine*** **abhängige Variable** untersucht wird. Man bezeichnet sie deshalb auch als **univariate** Pläne, wobei es unerheblich ist, ob nur eine unabhängige Variable (einfaktorieller Plan) oder mehrere unabhängige Variablen (mehrfaktorieller Plan) geprüft werden. Ein Plan heißt **multivariat**, wenn man in einer Untersuchung simultan **mehrere abhängige Variablen** überprüft.

An einem mehrfaktoriellen Plan sind mehrere unabhängige Variablen (Faktoren, UVs), an einem multivariaten Plan mehrere abhängige Variablen (AV) beteiligt.

Multivariate Pläne – seien sie ein- oder mehrfaktoriell – werden zur Prüfung von Unterschiedshypothesen oft mit der **multivariaten Varianzanalyse** („multivariate analysis of variance", MANOVA), der **multivariaten Kovarianzanalyse** („multivariate analysis of covariance", MANCOVA) bzw. mit der **Diskriminanzanalyse** („discriminant analysis") ausgewertet.

Viele Unterschiedshypothesen können angemessen nur multivariat formuliert werden. Der Vorteil des multivariaten Ansatzes gegenüber dem univariaten Ansatz ist darin zu sehen, dass er die wechselseitigen Beziehungen der abhängigen Variablen untereinander berücksichtigt und aufdeckt. Dies kann besonders wichtig sein, wenn die abhängige Variable komplex ist und sich sinnvoll nur durch mehrere operationale Indikatoren erfassen lässt. Arbeitsleistung, Therapieerfolg, Einstellungen etc. sind Beispiele für komplexe Konstrukte, die sich mit einem einzigen operationalen Indikator nur sehr ungenau beschreiben lassen. Hier ist es wünschenswert, verschiedene Subdimensionen des Konstrukts in die Analyse einzubeziehen.

Hat man für eine komplexe abhängige Variable mehrere operationale Indikatoren definiert (z. B. Arbeitsmenge und Anzahl der Fehler als Indikatoren von Arbeitsleistung), könnte man daran denken, die Unterschiedshypothese (z. B.: H_1: die Arbeitsleistungen in drei Abteilungen eines Betriebes sind unterschiedlich) in mehreren separaten univariaten Analysen zu überprüfen. Neben dem bereits erwähnten Nachteil, dass bei dieser Vorgehensweise die Beziehungen der abhängigen Variablen untereinander unentdeckt bleiben, führt die wiederholte Durchführung univariater Analysen zur Überprüfung **einer** Hypothese zu gravierenden inferenzstatistischen Problemen. In ▶ Abschn. 12.5.1 wurde berichtet, dass die Alternativhypothese beim klassischen Nullhypothesen-Signifikanztest üblicherweise angenommen wird, wenn die Irrtumswahrscheinlichkeit kleiner als 5 % (1 %) ist. Diese lässt sich jedoch nur schwer kalkulieren, wenn über eine Hypothese aufgrund mehrerer Signifikanztests entschieden wird und es somit zur α-Fehler-Kumulation kommt.

Wenn z. B. 100 Signifikanztests auf dem Signifikanzniveau von 5 % durchgeführt werden, erwarten wir bei Gültigkeit der Nullhypothese, dass ungefähr 5 Signifikanztests zufällig signifikant werden. Führen nun die separaten Analysen von zehn abhängigen Variablen zu signifikanten Resultaten, kann nicht mehr entschieden werden, welche dieser Signifikanzen „zufällig" und welche „echt" sind, es sei denn, man korrigiert das Signifikanzniveau (Näheres zur **α-Fehler-Adjustierung** bzw. Bonferroni-Korrektur bei Bortz & Schuster, 2010, S. 232). Diese Schwierigkeiten lassen sich vermeiden, wenn statt vieler separater univariater Analysen zur selben Hypothese lieber eine einzelne multivariate Analyse durchgeführt wird.

Sämtliche hier besprochenen Pläne zur Überprüfung von Unterschiedshypothesen lassen sich zu multivariaten Plänen erweitern. Die aufgeführten Beispiele gelten somit auch für multivariate Pläne, wenn statt einer mehrere abhängige Variablen untersucht werden. Im Übrigen gelten die Argumente, die in ▶ Abschn. 12.5.2 bei der Gegenüberstellung bivariater und multivariater Zusammenhangshypothesen genannt wurden, für Unterschiedshypothesen analog: Die „Zusammenschau" mehrerer univariater Analysen liefert in der Regel weniger Erkenntnisse als eine multivariate Analyse.

Zu den hypothesenprüfenden Auswertungstechniken für multivariate Pläne zählen u. a. die **multivariate Varianzanalyse**, die **multivariate Kovarianzanalyse** so-

wie die **Diskriminanzanalyse** (Bortz & Schuster, 2010, Kap. 26 und 27). Weitere hypothesenprüfende multivariate Verfahren sind u. a. die multiple lineare Korrelations- und Regressionsanalyse und die logistische Regressionsanalyse (zur Prüfung von Zusammenhangshypothesen ► Abschn. 12.5.2).

Lernquiz

1. Ein experimenteller und ein quasi-experimenteller Viergruppenplan sollen jeweils statistisch ausgewertet werden. Welche der beiden Studien geht mit höherem Aufwand bei der statistischen Datenanalyse einher (Begründung)?
2. In einem Seniorenwohnheim wird ein Jahr lang eine wöchentliche Kochgruppe durchgeführt. Die 25 Mitglieder berichteten nach dem Ende der Kochgruppe auf einem standardisierten Ernährungsfragebogen (1–10 Punkte) eine überdurchschnittlich gesundheitsbewusste Ernährung ($M = 7.69$). Eine Kontrollgruppe von 25 Senioren, die nicht an der Kochgruppe teilgenommen hatten, zeigte signifikant (t-Test für unabhängige Stichproben) weniger gesundheitsbewusstes Ernährungsverhalten ($M = 3.45$). Ist damit belegt, dass die Kochgruppe gesundheitsbewusste Ernährung bei Senioren fördert? (Begründung) Unter welchen Bedingungen kann eine solche Studie Auskunft über die Wirksamkeit der Kochgruppe geben? Welches statistische Auswertungsverfahren ist indiziert?
3. Eine groß angelegte experimentelle Studie zum Vergleich zwischen herkömmlichen Frontalvorlesungen, Blended-Learning-Veranstaltungen (in denen Unterrichtsmaterialien begleitend auf einer E-Learning-Plattform bereitgestellt werden) und Online-Kursen (in denen alle Inhalte auf einer E-Learning-Plattform inklusive interaktiver Lernmaterialien bereitgestellt werden) habe ergeben, dass sich die Lernleistung signifikant zwischen den verschiedenen Unterrichtsformen unterscheidet. Welche Aussagen können Sie über die Effektivität der jeweiligen Lernformen treffen?
4. Grenzen Sie mehrfaktorielle und multivariate Varianzanalyse voneinander ab.
5. Welche Aussage ist richtig?
   a) Wenn in einer zweifaktoriellen Varianzanalyse beide Haupteffekte signifikant werden, wird auch der Interaktionseffekt signifikant.
   b) Ein signifikanter Interaktionseffekt $A \times B$ besagt, dass beide Faktoren additiv zusammenwirken.
   c) Ordinale Interaktion und hybride Interaktion besagen das gleiche.
   d) Bei einer disordinalen Interaktion erster Ordnung kann keiner der beiden Haupteffekte global interpretiert werden.
   e) Interaktionseffekte sind in der Forschung kaum von Interesse.
6. Welche Möglichkeiten der Kontrolle von personengebundenen Störvariablen werden bei der Prüfung von Unterschiedshypothesen eingesetzt?
7. Was ist eine Interaktion zweiter Ordnung und wie benennt man sie?
8. „Hierarchische und teilhierarchische bzw. allgemein: unvollständige Pläne bieten die Möglichkeit, trotz fehlender Werte im Datensatz eine aussagekräftige statistische Analyse der Haupteffekte durchzuführen.“ Diskutieren Sie diese Aussage!
9. Was stimmt nicht?
   a) Einflussreiche Kontrollvariablen korrelieren mit der abhängigen Variablen.
   b) Durch den Einfluss von Störvariablen werden Gruppenunterschiede generell abgeschwächt.
   c) Nominalskalierte Kontrollvariablen können bei der varianzanalytischen Auswertung von Unterschiedshypothesen nicht berücksichtigt werden.
   d) Eine Kovarianzanalyse kann nur in univariaten Versuchsplänen eingesetzt werden.
10. Im Rahmen eines Experiments soll überprüft werden, inwiefern die Publikumsbewertung einer Zeichnung von der Person des Künstlers abhängt. Dazu wurde dieselbe Zeichnung 4 Gruppen von Untersuchungspersonen zur Bewertung vorgelegt. Die Gruppen erhielten jeweils eine unterschiedliche Information dazu, von wem die Zeichnung stammt: a) von einem Kind, b) von einem Patienten einer Kunsttherapiegruppe, c) von einem Hobbymaler, d) von einem renommierten Künstler. Die Untersuchungsteilnehmenden wurden gebeten, jeweils auf Ratingskalen anzugeben, wie schön sie die Zeichnung finden und wie gut sie ihnen insgesamt gefällt. Was für ein Untersuchungsplan liegt vor und wie muss diese Studie statistisch ausgewertet werden?

Übungsaufgaben

Im Zusammenhang mit der Erforschung der Mensch-Roboter-Interaktion spielt die Persönlichkeit von sozialen Robotern eine wichtige Rolle. So könnte z. B. eine „extravertierte Persönlichkeit“ des Roboters (z. B. dargestellt durch lebhafte Mimik und entsprechende sprachliche Äußerungen) Einfluss darauf haben, wie Menschen den Roboter wahrnehmen und mit ihm interagieren. In einer experimentellen Studie wurde den studentischen Versuchspersonen entweder

eine introvertiert oder eine extravertiert gestaltete Version desselben sozialen Roboters (iCat von Philips) im Video präsentiert. Die Versuchspersonen sollten anschließend den Roboter auf verschiedenen Dimensionen bewerten. Ein entsprechender Datensatz mit $n = 174$ vollständigen Fällen befindet sich im ▶ Webangebot zu diesem Buch.

1. Prüfen Sie zunächst die gerichtete Unterschiedshypothese, dass der extravertiert gestaltete Roboter von den Versuchspersonen auch als extravertierter wahrgenommen wurde als der introvertiert gestaltete Roboter. Berechnen Sie den t-Test für unabhängige Stichproben sowie eine einfaktorielle univariate Varianzanalyse, vergleichen und interpretieren Sie die Ergebnisse. Sind beide Verfahren auf den Datensatz gleichermaßen sinnvoll anwendbar oder werden Voraussetzungen verletzt?
2. Prüfen Sie nun die dieselbe gerichtete Unterschiedshypothese und prüfen Sie dabei gleichzeitig, ob der Effekt unabhängig davon auftritt, ob den Versuchspersonen das Roboter-Video im Labor oder Online gezeigt wurde und ob es sich um weibliche oder männliche Versuchspersonen handelt. Führen Sie eine dreifaktorielle univariate Varianzanalyse durch und interpretieren Sie die Ergebnisse für alle sieben getesteten Effekte bzw. Hypothesen. Welche Aussagen sind hinsichtlich Effektgröße und Teststärke zu treffen? Ist die dreifaktorielle Varianzanalyse auf den Datensatz sinnvoll anwendbar oder werden Voraussetzungen verletzt?
3. Prüfen Sie abschließend, inwiefern sich der Extraversions-Effekt bei der Roboterwahrnehmung durch die Technikbegeisterung der Versuchspersonen als Kontrollvariable erklären lässt. Führen Sie dazu eine dreifaktorielle univariate Kovarianzanalyse durch und interpretieren Sie das Ergebnis! Betrachten Sie auch die bivariate Korrelation zwischen Kontrollvariable und abhängiger Variable. Ist die Kovarianzanalyse auf den Datensatz sinnvoll anwendbar, oder werden Voraussetzungen verletzt?

Weiterführende Literatur

- Bittrich, K. & Blankenberger, S. (2011). *Experimentelle Psychologie: Experimente planen, realisieren, präsentieren*. Landsberg: Beltz.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.
- Hussy, W. & Jain, A. (2002). *Experimentelle Hypothesenprüfung in der Psychologie*. Göttingen: Hogrefe.
- Kantowitz, B., Roediger, H. & Elmes, D. (2008). *Experimental psychology, International edition* (9th ed.). Andover: Cengage Learning Emea.

Literatur

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Hoboken: Wiley.

Bittrich, K. & Blankenberger, S. (2011). *Experimentelle Psychologie: Experimente planen, realisieren, präsenteiren*. Landsberg: Beltz.

Bortz, J. (2005). *Statistik* (6. Aufl.). Berlin: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer.

Bredenkamp, J. (1982). Verfahren zur Ermittlung des Typs der statistischen Wechselwirkung. *Psychologische Beiträge, 24*(1), 56–75.

Cochran, W. G. & Cox, G. M. (1966). *Experimental Designs*. New York: Wiley.

Eid, M., Gollwitzer, M., & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.

Feldt, L. S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika, 23*, 335–354

Guo, S. & Fraser, M. W. (2010). *Propensity score analysis. Statistical methods and applications. Advanced quantitative techniques in the social sciences*. Thousand Oaks: Sage.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234–249.

Leigh, J. H. & Kinnear, T. C. (1980). On interaction classification. *Educational and Psychological Measurement, 40*(4), 841–843.

Lieberson, S. (1985). *Making it count. The improvement of social research and theory*. Berkeley: University of California Press.

Little, R. J., An, H., Johanns, J., & Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods, 5*(4), 459–476.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*(6), 530–558.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.

Perry, R. B., Abrami, P. C., Leventhal, L., & Check, J. (1979). Instructor reputation: An expectancy relationship involving student ratings and achievement. *Journal of Educational Psychology, 71*(6), 776–787.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178–192.

Pruzek, R. M. (2011). Introduction to the special issue on propensity score methods in behavioral research. *Multivariate Behavioral Research, 46*(3), 389–398.

Raudenbusch, S. W. & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(8), 757–763.

Spector, P. E. (1981). *Research designs*. Thousand Oaks: Sage.

Wampold, B. E. & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5*(4), 425–433.

12.5.4 Signifikanztests für Veränderungshypothesen

Lernziele

- Vor- und Nachteile von Querschnitt- und Längsschnittstudien abwägen können.
- Verschiedene Verfahren zur statistischen Überprüfung von Veränderungshypothesen kennen.
- Methodische Probleme von Längsschnittstudien bei der Untersuchungsplanung berücksichtigen und vermeiden können.
- Einen Überblick über statistische Zeitreihenanalysen gewinnen.
- Ergebnisse der Überprüfung von Veränderungshypothesen richtig interpretieren können.

Die empirische Untersuchung von **individuellen Veränderungen** ist in der Regel nur möglich, wenn bei jeder untersuchten Person (bzw. bei jedem Untersuchungsobjekt) dieselbe(n) Variable(n) in mehr oder minder großem zeitlichen Abstand bzw. unter ähnlichen oder unterschiedlichen Bedingungen mehrfach hintereinander erfasst werden. In den Human- und Sozialwissenschaften spielen im Zusammenhang mit der Prüfung von Veränderungshypothesen v. a. fünf Arten von Studien eine Rolle:

1. **Experimentelle Gruppenuntersuchungen mit Messwiederholungen** dienen zur Prüfung von **Hypothesen über Veränderungswirkungen von Interventionen** (z. B. Therapie- oder Trainingswirkungen). Im typischen Fall werden eine Vorher- und eine Nachher-Messung durchgeführt und die behandelte(n) Experimentalgruppe(n) mit einer unbehandelten oder herkömmlich behandelten Kontrollgruppe verglichen (**randomisiertes Prä-Post-Kontrollgruppen-Design**). Zur statistischen Auswertung dienen meist Varianzanalysen mit Messwiederholungen, bei denen der Messzeitpunkt als weitere unabhängige Variable fungiert oder Kovarianzanalysen, bei denen die Prätest-Messwerte als Kontrollvariablen aus den Posttestmesswerten herauspartialisiert werden. Erfolgt bei einem echten Experiment eine Randomisierung großer Stichproben, so dass von äquivalenten Untersuchungsgruppen auszugehen ist, so kann eine kausale Veränderungswirkung ausnahmsweise auch ohne Messwiederholung, d. h. ohne Prätest-Messung nur aus einem signifikanten Posttestvergleich zwischen Experimental- und Kontrollgruppe geschlossen werden.
2. **Quasi-experimentelle Gruppenuntersuchungen mit Messwiederholungen** dienen ebenfalls der Prüfung von **Hypothesen über Wirkungen von Interventionen**. Allerdings ist die interne Validität dieser Studien im Vergleich zum Experiment mangels Randomisierung deutlich eingeschränkt, so dass entsprechende Vorkehrungen zu treffen sind, um eine aussagekräftige Hypothesenprüfung zu ermöglichen. Auch bei der Auswertung von nicht-randomisierten Prä-Post-Kontrollgruppen-Designs sind Varianz- bzw. Kovarianzanalysen mit Messwiederholungen einschlägig.
3. Überprüfung von **Interventionen im Rahmen von Einzelfallstudien**, die mit Resampling-Verfahren statistisch auszuwerten sind (► Abschn. 12.5.5).
4. **Längsschnittstudien** dienen der Prüfung von **Entwicklungshypothesen**, bei denen Alters-, Kohorten- und Epocheneffekte eine Rolle spielen (z. B. Veränderung von Intelligenz, körperlicher Fitness oder Selbstkonzept über die Lebensspanne). Hierbei werden Untersuchungsgruppen in großen zeitlichen Abständen (z. B. alle 10 Jahre) und somit über lange Zeitspannen hinweg (z. B. drei Messzeitpunkte über 30 Jahre hinweg) untersucht. Die statistische Auswertung erfolgt mittels Varianzanalysen mit Messwiederholungen oder mittels Strukturgleichungsmodellen, speziell latenten Wachstumsmodellen.
5. **Zeitreihenanalysen** zur Prüfung von **Zeitreihenhypothesen**. Hierbei werden Messwerte in regelmäßigen und oft auch kurzen Abständen erhoben, nicht selten kommen dadurch relativ lange Zeitreihen mit fünfzig oder mehr Messzeitpunkten zustande (z. B. stündliche Logfile-Daten eines Webservers, tägliche Börsenkurse einer Aktie, jährliche Suizidraten eines Landes etc.). Die statistische Auswertung erfolgt hier mittels Zeitreihenanalyse als Sammelbegriff für eine Gruppe von einschlägigen Verfahren (Thome, 2005).

Darüber hinaus existieren noch weitere statistische Methoden der Prüfung von Veränderungshypothesen auf der Basis quantitativer Daten. Nicht behandelt wird hier z. B. die „**Survival Analysis**“, die sich der Frage widmet, ob bzw. wann bestimmte Ereignisse im Lebenslauf einer Zielpopulation eintreten (z. B. erste Anzeichen für Lungenkrebs bei Rauchern, das erste Kind etc.). Eine Einführung in diese Thematik und weiterführende Literatur findet man bei Singer und Willett (1991). Mit einer weiteren Thematik befasst sich die sog. **Ereignisanalyse**, bei der die Zeitintervalle zwischen aufeinander folgenden Ereignissen untersucht werden (z. B. die Frage, mit welcher Wahrscheinlichkeit einzelne Individuen innerhalb eines festgelegten Zeitraumes den Beruf wechseln). Über diese Technik informieren z. B. Blossfeld, Hamerle, & Mayer (1986).

Experimentelle Untersuchungen mit und ohne Messwiederholungen

In den Human- und Sozialwissenschaften geht es sehr oft darum, Hypothesen über die Wirkungen von bestimmten Interventionen (z. B. Psychotherapiemethode, Medikament, Lernprogramm, Fernsehsendung etc.) zu prüfen. Jede **Wirkungshypothese zu Interventionen** ist im Grunde eine Veränderungshypothese, denn sie impliziert, dass sich der Ausgangszustand der Untersuchungsobjekte durch die Intervention in bestimmter – positiver und/oder negativer Weise – verändert hat. Wirkungshypothesen lassen sich mit der höchsten internen Validität durch experimentelle Untersuchungen prüfen, wobei meist ein **Messwiederholungsdesign** gewählt wird: Dabei werden die interessierenden abhängigen Variablen vor der Intervention **(Pretest, Prästest)** und nach der Intervention **(Posttest)** gemessen, wobei die Experimentalgruppe(n) möglichst noch mit einer Kontrollgruppe (die unbehandelt bleibt oder nach herkömmlicher Methode behandelt wird) verglichen wird **(Kontrollgruppendesign)**. Aus der Kombination von Messwiederholungs- und Kontrollgruppendesign ergibt sich der **randomisierte Prä-Post-Kontrollgruppenvergleich** („randomized controlled trial", RCT) als typisches experimentelles Design. Damit Differenzen zwischen Experimental- und Kontrollgruppen wirklich auf die Intervention – und nicht auf andere Einflüsse – zurückführbar sind, müssen sich die Personen in den Experimental- und Kontrollgruppen vor der Behandlung der Experimentalgruppe(n) sehr stark ähneln. Dies wird bei großen Untersuchungsgruppen durch Randomisierung erreicht, bei kleinen Gruppen muss die Äquivalenz von Experimental- und Kontrollgruppen im Vortest geprüft und ggf. aktiv hergestellt werden. Damit ist auch die experimentelle Untersuchung anfällig für die meisten in ► Abschn. 3.2.2 genannten Gefährdungen der internen Validität. Beste Kontrollmöglichkeiten dieser Gefährdungen bietet der **Solomon-Viergruppenplan**. Wir werden im Folgenden experimentelle Prä-Post-Kontrollgruppendesigns sowie den Solomon-Viergruppenplan vorstellen und vorab auf den Ausnahmefall eingehen, in dem experimentelle Veränderungshypothesen auch ohne Messwiederholungen geprüft werden.

Experimentelle Prüfung von Veränderungshypothesen ohne Messwiederholungen

Um kausale Veränderungshypothesen zu prüfen, die sich auf die Wirkung von Treatments beziehen, ist zunächst der Vergleich zwischen Experimentalgruppen, in denen das Treatment bzw. mehrere Treatment-Varianten wirken, und Kontrollgruppen, die nicht oder konventionell behandelt werden, einschlägig. Man erwartet, dass sich bei Wirksamkeit des Treatments die abhängigen Variablen in erwarteter Form ausprägen (z. B. deutlicher Rückgang der Depression in der mit einer Depressionstherapie behandelten Experimentalgruppe) und von der Kontrollgruppe unterscheiden (kein oder allenfalls ein viel geringerer Rückgang der Depression in einer unbehandelten Kontrollgruppe). Um nachzuweisen, dass sich die Depression verringert hat, ist neben der Posttest-Messung normalerweise auch eine Pretest-Messung notwendig.

Nur wenn man genau wüsste, dass sich die Personen in der Experimental- und die Personen in der Kontrollgruppe vor der Behandlung nicht nennenswert in ihrer Depressivität unterschieden haben, könnte man geringere Posttest-Depressionswerte in der Experimentalgruppe im Vergleich zur Kontrollgruppe als indirekten Hinweis auf die erwartete Veränderung in der Behandlungsgruppe werten. Wenn man mit relativ **großen homogenen Stichproben pro Untersuchungsgruppe** arbeitet und diese **randomisiert** bildet, so sollten damit vergleichbare Ausgangsbedingungen in allen Untersuchungsgruppen resultieren, so dass Veränderungshypothesen über Treatment-Wirkungen dann allein auf der Basis empirischer Posttest-Werte möglich sind. Zur statistischen Prüfung von Veränderungshypothesen sind in einem Posttest-Experimentaldesign dieselben Verfahren indiziert, die zur Prüfung von Unterschiedshypothesen dienen, etwa der ***t*-Test für unabhängige Stichproben** für einen Zweigruppen-Plan sowie **ein- oder mehrfaktorielle Varianz- oder Kovarianzanalysen ohne Messwiederholungen**. In der Praxis bevorzugt man jedoch experimentelle Designs, in denen die Ausgangswerte aller Untersuchungspersonen im Pretest erfasst werden.

In experimentellen Untersuchungen mit großen Stichproben ist durch die Randomisierung Äquivalenz der zu vergleichenden Gruppen gewährleistet. Man kann deshalb theoretisch auf Pretest-Messungen verzichten und hypothesenkonforme *Posttest-Unterschiede* als Bestätigung der Veränderungshypothese interpretieren. Die Posttest-Effekte werden mit dem *t*-Test für unabhängige Stichproben (Zweigruppenplan) oder mit Varianzanalysen (Mehrgruppenplan, faktorielle Pläne) ausgewertet.

In der Praxis der experimentellen Grundlagen- und Evaluationsforschung haben sich jedoch Pretest-Posttest-Designs etabliert, d. h. es erfolgt eine Veränderungsmessung durch eine Erfassung der interessierenden abhängigen Variable(n) vor und nach der experimentellen Manipulation.

Pretests sind zudem immer dann zwingend erforderlich, wenn Zweifel an der korrekten Durchführung der Randomisierungsprozedur bestehen oder die Stichproben zu klein sind, um dem zufälligen Ausgleich personenbe-

	Pretest-Messung	1. Posttest-Messung	2. Posttest-Messung	…	Letzte Posttest-Messung
Experimentalgruppe	S_1	S_1	S_1	…	S_1
Kontrollgruppe	S_2	S_2	S_2	…	S_2

Abb. 12.48 Zweifaktorieller Messwiederholungsplan mit Experimentalgruppe und Kontrollgruppe

dingter Störvariablen in Experimental- und Kontrollgruppe trauen zu können (Mittring & Hussy, 2004). Wann immer man befürchtet, dass der statistische Fehlerausgleich per Randomisierung nicht sichergestellt ist, dass also die Experimental- und Kontrollgruppe in Bezug auf die abhängige Variable vor Applikation des Treatments nicht äquivalent sind, sollten die Vergleichsgruppen wie „natürliche" Gruppen behandelt und nach den Richtlinien quasi-experimenteller Untersuchungen ausgewertet werden. Wie man feststellen kann, ob der Unterschied zwischen den Pretest-Werten von Experimental- und Kontrollgruppe genügend klein ist, um von äquivalenten Vergleichsgruppen sprechen zu können, wird bei Klemmert (2004), Rogers, Howard, & Vessey (1993) bzw. Wellek (1994) beschrieben **(„equivalence testing")**.

Experimentelle Prüfung von Veränderungshypothesen mit Messwiederholungen

Im Normalfall erfolgt die experimentelle Prüfung von Veränderungshypothesen in einem Messwiederholungsdesign. Damit wird der Messzeitpunkt (z. B. in den Ausprägungen Vorher-, Nachher-Messung) neben dem Treatment als weitere unabhängige Variable in das Design aufgenommen. Für den einfachen Vergleich einer randomisierten Experimentalgruppe (S_1) mit einer randomisierten Kontrollgruppe (S_2) resultiert dann das in Abb. 12.48 wiedergegebene Untersuchungsschema.

Oberflächlich ähnelt dieser Plan dem in Abb. 12.37 wiedergegebenen zweifaktoriellen Untersuchungsplan; dennoch besteht zwischen beiden Plänen ein gravierender Unterschied: Der zweifaktorielle Plan ohne Messwiederholungen untersucht für jede Faktorstufenkombination eine andere Stichprobe, während im **Messwiederholungsplan** dieselben Stichproben mehrfach untersucht werden.

Man muss allerdings bei wiederholten Messungen einer abhängigen Variablen mit **Transfer-Effekten** (Ermüdung, Lerneffekte, Motivationsverlust etc.) rechnen, die die eigentliche Treatment-Wirkung verzerren können. Würde man bei der experimentellen Prüfung der Wirksamkeit eines Anti-Rauch-Programms zur Messung der abhängigen Variablen „Anzahl täglich gerauchter Zigaretten" ein „Zigarettentagebuch" führen lassen (zur Tagebuchmethode ▶ Abschn. 10.3.3), könnte allein das Tagebuch zu einem veränderten Rauchverhalten führen – etwa durch das ständige Bewusstmachen des Zigarettenkonsums. In diesem Falle wäre dem Messwiederholungsplan der folgende **randomisierte Blockplan** vorzuziehen:

Wenn im Messwiederholungsplan z. B. 50 Raucher unter Experimentalbedingung und 50 Raucher unter Kontrollbedingung 10 Wochen lang pro Woche einmal beobachtet werden sollten, würde man für einen analogen Blockplan 2 × 50 Blöcke à 10 Personen benötigen. Die 10 Personen eines jeden Blocks sollten bzgl. untersuchungsrelevanter Störvariablen (z. B. Alter, Geschlecht, Anzahl täglich gerauchter Zigaretten, Dauer des Rauchens) nach Ähnlichkeit ausgewählt worden sein **(Matched Samples)**. Es wird für jede Person eines jeden Blocks per Zufall entschieden, in welcher Woche die abhängige Variable gemessen wird (Tagebuch führen) und welcher Bedingung (Experimental- oder Kontrollbedingung) der Block zugeordnet wird. Jede Person wird also nur einmal untersucht und nicht zehnmal wie im Messwiederholungsplan. In beiden Plänen erhält man 1 000 Messungen der abhängigen Variablen: Im Messwiederholungsplan 10 Messungen von 100 Personen und im Blockplan 1 Messung von 1 000 Personen. In beiden Plänen dauert die Untersuchung insgesamt 10 Wochen, wobei den Personen unter der Kontrollbedingung zum Untersuchungsbeginn lediglich mitgeteilt wird, dass sie an einer Studie teilnehmen. Für Personen der Experimentalgruppe endet die Studie nach der zufällig ausgewählten Woche, in der das Tagebuch geführt wurde.

Blockplan – Wenn bei wiederholter Untersuchung derselben Untersuchungspersonen Transfer-Effekte drohen, sollte ein randomisierter Blockplan („randomized block plan") eingesetzt werden. Die k-fache Messung einer Untersuchungsperson wird hierbei durch Einzelmessungen von k Untersuchungspersonen ersetzt, wobei die k Untersuchungspersonen eines Blocks nach Ähnlichkeit auszuwählen sind ("Matched Samples") und zufällig den k Messzeitpunkten zugeordnet werden. Die Blöcke werden zufällig der Experimental- bzw. Kontrollbedingung zugeordnet.

Für die statistische Auswertung dieses Blockplans oder eines Messwiederholungsplans wird üblicherweise eine

spezielle Variante der Varianzanalyse, die **Varianzanalyse mit Messwiederholungen** („repeated measures analysis of variance“, „repeated measures ANOVA“) eingesetzt. Neben dem oder den Treatmentfaktoren fungiert der Messzeitpunkt als Messwiederholungsfaktor. Man spricht deswegen auch von der Varianzanalyse mit Messwiederholungen auf einem Faktor oder von **gemischten Designs**, um zum Ausdruck zu bringen, dass das Design neben dem „within-subject factor“ (Messwiederholungsfaktor) auch mindestens einen „between-subject factor“ (experimentellen Behandlungsfaktor) enthält. Die Varianzanalyse mit Messwiederholungen setzt u. a. voraus, dass die zu verschiedenen Zeitpunkten erhobenen Messungen gleichförmig miteinander korrelieren, dass also z. B. die erste Messung mit der zweiten Messung genauso hoch korreliert wie mit der letzten Messung – eine Voraussetzung, die in vielen Messwiederholungsplänen verletzt ist (Ludwig, 1979). Wie man diese Voraussetzung überprüft und wie zu verfahren ist, wenn das Datenmaterial diesen Voraussetzungen nicht genügt, wird z. B. bei Bortz & Schuster (2010, Kap. 18) näher erläutert (zum Vergleich bzw. zur Indikation von Varianzanalysen mit bzw. ohne Messwiederholungen bzw. von Between- oder Within-Subject-Designs verweisen wir auf Keren, 1993).

Ferner wird vorausgesetzt, dass die individuellen Datensätze vollständig sind, dass also von allen Versuchspersonen zu allen Messzeitpunkten Messungen vorliegen. Möglichkeiten, mit unvollständigen Datensätzen („missing data“) umzugehen, werden bei Hedeker und Gibbons (1997) bzw. Davis (2002) erörtert (zur Behandlung von fehlenden Werten im Zuge der Datenaufbereitung ▶ Kap. 11).

Veränderungshypothesen, die wie in ◘ Abb. 12.48 mit zweifaktoriellen Messwiederholungsplänen mit einem Messwiederholungsfaktor überprüft werden, gelten als bestätigt, wenn der Haupteffekt „Experimental- vs. Kontrollgruppe“ signifikant ist (in diesem Falle unterscheiden sich die beiden Gruppen gleichförmig über alle Messungen hinweg) und/oder wenn die **Interaktion zwischen dem Gruppierungs-/Treatmentfaktor und dem Messwiederholungsfaktor** statistisch bedeutsam ist (was ein Beleg dafür wäre, dass sich die Experimentalgruppe im Verlauf der Zeit anders verändert als die Kontrollgruppe).

Messwiederholungspläne werden nicht nur für die Überprüfung von Veränderungshypothesen im engeren Sinne benötigt (ein Treatment verändert die abhängige Variable), sondern können generell eingesetzt werden, wenn von einer Stichprobe wiederholte Messungen erhoben werden. In ▶ Abschn. 12.5.3 erwähnten wir ein Beispiel, bei dem es um die Ablesbarkeit von Anzeigegeräten ging, die sich bzgl. der Faktoren A („Form“) und B („Art der Zahlendarstellung“) unterschieden. Es wurde

	A_1	A_2	A_3	…	A_p
Abfolge 1	S_1	S_1	S_1	…	S_1
Abfolge 2	S_2	S_2	S_2	…	S_2
Abfolge 3	S_3	S_3	S_3	…	S_3

◘ **Abb. 12.49** Zweifaktorieller Messwiederholungsplan zur Kontrolle von Sequenzeffekten

ein zweifaktorieller Plan vorgestellt, der die Unterschiedshypothese überprüft, die Ablesefehler seien von der Form des Anzeigegerätes sowie der Art der Zahlendarstellung abhängig. Dieser Plan benötigte 3×3 (bzw. allgemein $p \times q$) Stichproben.

Die gleiche Unterschiedshypothese ließe sich auch mit einem Messwiederholungsplan prüfen, in dem eine Stichprobe z. B. alle zur Stufe A_1 gehörenden Anzeigegeräte, eine weitere Stichprobe alle zur Stufe A_2 gehörenden Anzeigegeräte etc. beurteilt. Statt der zwei Stichproben in ◘ Abb. 12.48 (Experimental- und Kontrollgruppe) benötigt man also p Stichproben (eine Stichprobe pro Faktorstufe von A), die jeweils q Anzeigegeräte beurteilen. Werden Untersuchungspläne zur Prüfung von Unterschiedshypothesen mit Messwiederholungen realisiert, werden dadurch **Untersuchungsaufwand und Versuchspersonen eingespart**. Dieses Vorgehen ist jedoch nur sinnvoll, wenn die Messwiederholung das Ergebnis nicht verfälscht (z. B. durch Übungs-, Vertrautheits-, Ermüdungseffekte etc.).

▪ Kontrolle von Sequenzeffekten

Bei Untersuchungen, in denen von einer Stichprobe unter mehreren Untersuchungsbedingungen Messungen erhoben werden, kann die Abfolge der Untersuchungsbedingungen von ausschlaggebender Bedeutung sein. Zur Kontrolle derartiger Sequenzeffekte empfiehlt sich der in ◘ Abb. 12.49 wiedergegebene experimentelle Untersuchungsplan.

Mit diesem Plan wird der Einfluss von drei verschiedenen Abfolgen ermittelt. Jeder Abfolge wird eine Stichprobe zugewiesen, die die Untersuchungsbedingungen in der entsprechenden Reihenfolge durchläuft. (Man beachte, dass das in ◘ Abb. 12.49 wiedergegebene Datenschema nur eine Abfolge: $A_1, A_2 \ldots A_p$ enthält, d. h., die Untersuchungsergebnisse der einzelnen Stichproben müssen für dieses Datenschema jeweils „umsortiert“ werden.) Unterscheiden sich die Stichproben nicht bzw. ist der **Abfolgefaktor** in der Varianzanalyse nicht signifikant, ist die Reihenfolge der Untersuchungsbedingungen unerheblich. Eine Interaktion zwischen den Untersuchungsbedingun-

gen und den Abfolgen weist auf Positionseffekte hin, mit denen man z. B. rechnen muss, wenn die Untersuchungsteilnehmenden im Verlauf der Untersuchung ermüden, so dass z. B. auf die erste Untersuchungsbedingung unabhängig von der Art dieser Bedingung anders reagiert wird als auf die letzte.

Der Abfolgefaktor kann als „fixed factor“ oder als „random factor“ konzipiert werden (Bortz & Schuster, 2010, S. 212 ff.). Bei einem „fixed factor“ wählt man systematisch bestimmte Abfolgen aus und bei einem „random factor“ wählt man aus allen möglichen Abfolgen einige zufällig aus.

Sequenzeffekte – Durchläuft dieselbe Person nacheinander mehrere Untersuchungsbedingungen, können **Sequenzeffekte** („sequence effects“, „order effects“) auftreten. Um einen verfälschenden Einfluss von Sequenzeffekten zu vermeiden, durchlaufen Teilgruppen der Untersuchungspersonen die Untersuchungsbedingungen systematisch in verschiedener Abfolge. Die verschiedenen Abfolgen werden als zusätzliche unabhängige Variable (Abfolgefaktor) in das Design aufgenommen und im Zuge einer Varianzanalyse statistisch geprüft. Wird der Abfolgefaktor nicht signifikant (und war die Teststärke ausreichend), so spricht dies gegen einen Sequenzeffekt.

Kontrolle von Pretest-Effekten durch den Solomon-Viergruppenplan

Nehmen wir an, die Tauglichkeit einer Software zum Lernen von Grammatikregeln soll experimentell überprüft werden. Nachdem die Untersuchungsteilnehmenden einer Experimentalgruppe und einer Kontrollgruppe zufällig zugewiesen wurden, will man vor dem Training überprüfen, ob die beiden Stichproben im Durchschnitt annähernd gleich gute Grammatikkenntnisse aufweisen oder ob durch die Randomisierung zufällig zwei Stichproben entstanden sind, die sich in Bezug auf ihre Grammatikkenntnisse unterscheiden. Hierzu werden Pretests durchgeführt.

Nun kann man allerdings nicht ausschließen, dass bereits die im Pretest gestellten Grammatikfragen Lerneffekte auslösen, in dem sie z. B. vergessenes Wissen reaktivieren oder zum Nachdenken über grammatische Regeln anregen. Der Pretest selbst übt eine Treatmentwirkung aus und verändert die abhängige Variable, d. h., die Posttest-Ergebnisse können durch die Pretests verfälscht sein.

Zur Kontrolle derartiger **Pretest-Effekte** wurde ein spezielles Untersuchungsschema entwickelt, das in der Literatur unter der Bezeichnung Solomon-Viergruppenplan geführt wird (Abb. 12.50).

Der Plan erfordert **vier randomisierte Gruppen**. Die erste Gruppe ist eine „klassische“ Experimentalgruppe (mit Pretest, Treatment und Posttest) und die zweite Gruppe eine „klassische“ Kontrollgruppe (Pretest und Posttest ohne Treatment). Die dritte Gruppe realisiert ein One-Shot-Case-Design, bei dem nach appliziertem Treatment nur eine Posttest-Messung durchgeführt wird. Die vierte Gruppe schließlich wird nur einer „Posttest-Messung“ unterzogen.

Gruppe 1:	Pretest	–	Treatment	–	Posttest
Gruppe 2:	Pretest		–		Posttest
Gruppe 3:	–		Treatment	–	Posttest
Gruppe 4:	–		–		Posttest

Abb. 12.50 Solomon-Viergruppenplan

Dieser Plan eröffnet zahlreiche Kontrollmöglichkeiten. Das Posttest-Ergebnis in der ersten Gruppe (PT_1) enthält neben einem möglichen Effekt des Treatments (T) und des Pretests (P) auch Effekte zeitgebundener Störvariablen (Z; externe zeitliche Einflüsse, Reifungsprozesse, Testübung). Symbolisch schreiben wir:

$$PT_1 = f(T, P, Z).$$

Das Posttest-Ergebnis in der ersten Gruppe ist eine Funktion von Treatment-, Pretest- und Zeit-Effekten. Mit dieser Symbolik können wir die Posttest-Ergebnisse in den übrigen Gruppen wie folgt charakterisieren:

$$PT_2 = f(P, Z)$$
$$PT_3 = f(T, Z),$$
$$PT_4 = f(Z).$$

Eine Gegenüberstellung der Veränderungen in den Gruppen 1 und 2 (Pretest-Posttest-Differenzen) informiert damit über „reine“ Treatment-Effekte (Nettoeffekt; Tab. 12.23). Das Resultat dieses Vergleichs müsste dem Vergleich von PT_3 und PT_4 entsprechen, denn auch dieser Vergleich isoliert den „reinen“ Effekt des Treatments. Es wäre allerdings möglich, dass der Pretest in der Experimentalgruppe andere Wirkungen hat als in der Kontrollgruppe (Interaktion Pretest × Gruppen), was dazu führen würde, dass die Ergebnisse dieser Vergleiche nicht übereinstimmen.

Der Vergleich von PT_2 und PT_4 dient der Abschätzung von Pretest-Effekten. Beide Gruppen sind ohne Treatment und unterscheiden sich nur darin, dass Gruppe 2, aber nicht Gruppe 4 vorgetestet wurde. Will man erfahren, ob das Treatment in Kombination mit dem Vortest anders wirkt als ohne Vortest (Interaktion Pretest × Treatment), wären der Durchschnitt von PT_2 (Pretest- und Zeit-Effekte) und PT_3 (Treatment- und Zeit-Effekte) mit PT_1 (Treatment-, Pretest- und Zeit-Effekte) zu vergleichen.

Ausführliche Hinweise zur statistischen Auswertung dieses Planes findet man bei Braver und Braver (1988).

Der Solomon-Viergruppenplan lässt sich auch in komplexere mehrfaktorielle Pläne einbauen. Entscheidend ist, dass grundsätzlich ein weiterer Faktor einbezogen wird, der vorgetestete und nicht-vorgetestete Versuchspersonen des Experiments unterscheidet.

Solomon-Viergruppenplan – Der Solomon-Viergruppenplan („Solomon four group design") stellt eine Erweiterung des klassischen experimentellen Pretest-Posttest-Designs dar (zweifaktorieller Plan mit Treatmentfaktor und Messwiederholungsfaktor). Er dient dazu, die mögliche Wirkung von Pretest-Effekten zu überprüfen.

Quasi-experimentelle Untersuchungen mit Messwiederholungen

Veränderungshypothesen führen eine Senkung oder Steigerung der Werte der abhängigen Variablen kausal auf die Wirkung einer unabhängigen Variablen (z. B. Behandlungsmethode vs. Placebo) zurück. Kausale Hypothesen über Veränderungen lassen sich am eindeutigsten bzw. mit höchster interner Validität anhand von Daten prüfen, die aus **echten Experimenten** stammen: Hierbei wird zunächst durch Randomisierung aktiv sichergestellt, dass Experimental- und Kontrollgruppe vergleichbar sind und dies meist auch durch Erhebung der Pretest-Werte erfasst. Dann wird die Intervention durchgeführt und die Wirkung auf der/den abhängigen Variablen gemessen (Posttest-Messung).

Sehr viel weniger eindeutig können kausale Veränderungshypothesen geprüft werden, wenn Kontroll- und Experimentalgruppe bereits vor der eigentlichen Behandlung Unterschiede aufweisen (also nicht randomisiert gebildet wurden). Diese Situation ist insbesondere bei Feldexperimenten und in der Evaluationsforschung verbreitet, weil hier oft Interventionen an natürlich **vorgefundenen Gruppen** (z. B. Schulklassen, Krankhausstationen, Fachabteilungen) oder **nicht-randomisiert gebildeten Gruppen** (z. B. Zuordnung der Versuchspersonen zu den Gruppen nicht strikt zufallsgesteuert, sondern durch Praktikabilität bestimmt) hinsichtlich ihrer Veränderungswirkung untersucht werden. Eine Randomisierung würde die Bedingungen im Feld zu sehr stören, wäre also forschungsökonomisch und teilweise auch forschungsethisch nicht möglich. Derartige **Quasi-Experimente**, bei denen nicht-randomisiert gebildete Gruppen unterschiedlich behandelt werden, um die Veränderungswirkungen zu messen, sind in ihrer internen Validität eingeschränkt.

Beispiele für quasi-experimentelle Studien, in denen Hypothesen geprüft werden, die behaupten, dass eine Maßnahme in verschiedenen, real existierenden Populationen unterschiedlich verändernd wirkt: Informierte Menschen sind in ihren Einstellungen weniger leicht beeinflussbar als uninformierte Menschen. Ein neues Entspannungsprogramm wirkt bei jüngeren Erwerbstätigen stärker stressreduzierend als bei älteren. Eine bestimmte Anti-Mobbing-Maßnahme reduziert Mobbingverhalten bei Schülern stärker als bei Schülerinnen etc.

Alle Beispielen thematisieren eine abhängige Variable, die sich laut Hypothese bei den jeweils verglichenen Populationen unterschiedlich ändert. Ein bestimmtes „Treatment" hat in verschiedenen Populationen unterschiedliche Auswirkungen. Um das Ausmaß der Wirkung abschätzen zu können, müssen die Merkmalsausprägungen vor der Behandlung durch sog. Vortests **(Pretests, Prätests)** bekannt sein. Vortests haben in quasi-experimentellen Untersuchungen die Funktion, Unterschiede zwischen den Stichproben zu Beginn der Untersuchung festzustellen. Die stichprobenspezifischen „Startbedingungen" sind die Referenzdaten, auf die sich treatmentbedingte Veränderungen beziehen.

Veränderungsmessung durch Differenzwerte

Veränderung wird in quasi-experimentellen Untersuchungen durch Differenzen zwischen Durchschnittswerten angezeigt, die für eine Stichprobe zu zwei oder mehr Messzeitpunkten ermittelt wurden (z. B. Prä-Post-Differenzen). Anders als Differenzen **zwischen** Stichproben („between-subject factor"), bereiten Differenzen „innerhalb" von Stichproben („within-subject factor") bei der statistischen Auswertung und Interpretation einige Schwierigkeiten, auf die im Folgenden eingegangen wird.

Das **einfache Differenzmaß** innerhalb von Stichproben (Vorher- minus Nachher-Messung) als Indikator von Veränderung war in der Vergangenheit häufig heftiger Kritik ausgesetzt (Bereiter, 1963; Bohrnstedt, 1969; Cronbach & Furby, 1970; Linn & Slinde, 1977; O'Connor, 1972; Rennert, 1977). Das zentrale Argument betraf die **mangelnde Reliabilität dieser Differenzwerte**. Wenn schon die Reliabilität (zum Reliabilitätsbegriff ► Abschn. 10.4.1) vieler sozialwissenschaftlicher Messungen sehr zu wünschen übrig lässt, trifft dies – so die übliche Kritik – in noch stärkerem Maße auf Differenzen dieser Messungen zu. Allgemein gilt, dass in den Messfehler von Differenzwerten zweier Variablen X und Y sowohl der Messfehler von X als auch der Messfehler von Y eingehen. Bezogen auf die hier interessierende Pretest-Posttest-Situation besagt dieser Sachverhalt, dass ein Messinstrument, das eine Reliabilität von z. B. $r = .90$ aufweist (eine für sozialwissenschaftliche Messungen beachtliche Reliabilität), zu Messwertdifferenzen mit einer Reliabilität von .67 führt, wenn Pretest- und Posttestmessungen zu $r = .70$ miteinander korrelieren. (Zur rechnerischen Ermittlung der Reliabilität von Differenzwerten s. Guilford, 1954, S. 394, oder Rost, 2004, S. 276; Spezialfälle behandeln Williams

& Zimmermann, 1977.) Geht man davon aus, dass die Reliabilität der Messungen eher niedriger ist als im Beispiel, kommt man zu Reliabilitäten des Differenzmaßes, die in der Tat problematisch gering erscheinen.

Diese Auffassung gilt jedoch als überholt bzw. revisionsbedürftig (Collins, 1996; Gottman, 1995; Mellenberg, 1999; Rogosa, 1995; Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983, 1985; Zimmermann & Williams, 1982, zum Stichwort „Gain Scores"). Es wird argumentiert, dass die Reliabilität von Differenzmaßen nicht nur von der Reliabilität der Merkmalserfassung, sondern insgesamt von vier Einflussgrößen abhängt:

1. **Unterschiedlichkeit der wahren individuellen Veränderungen**

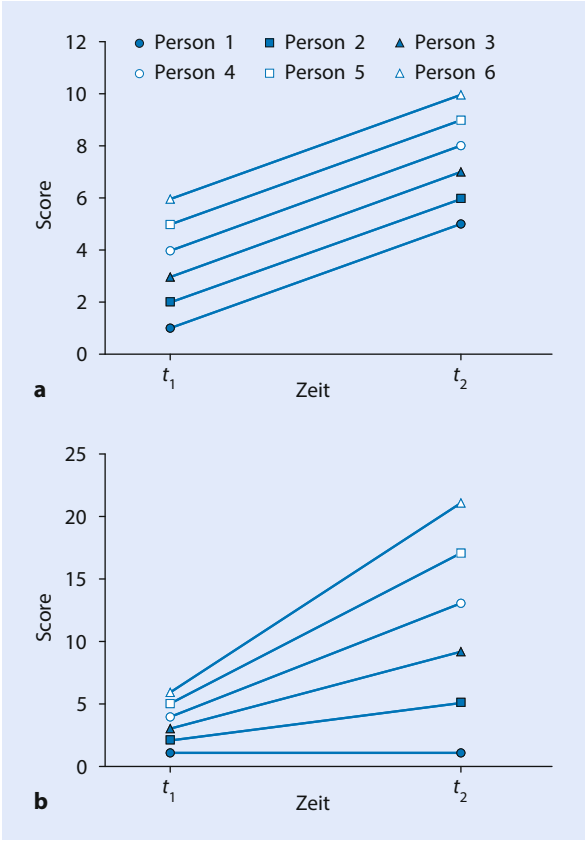
   Bei nur zwei Messpunkten t_1 und t_2 entspricht die gemessene Veränderung eines Individuums i der Differenz d_i der Messwerte für die Zeitpunkte t_1 und t_2. Je stärker sich die wahren, den d_i-Werten zugrunde liegenden Veränderungen in einer Stichprobe von Individuen unterscheiden, desto größer ist die Reliabilität der Differenzwerte. Die Streuung der d_i-Werte ist damit ein wichtiger Indikator für die Reliabilität von Differenzmaßen.

   Nach Rogosa et al. (1982) zeigt die Reliabilität der Differenzen an, wie verlässlich die untersuchten Individuen nach Maßgabe ihrer d_i-Werte in eine Rangreihe gebracht werden können. Dies gelingt natürlich umso besser, je größer die Streuung der d_i-Werte ist. Unterscheiden sich die d_i-Werte hingegen nur wenig, ist damit auch deren Reliabilität gering.

   Eine niedrige Reliabilität impliziert jedoch keineswegs zwangsläufig, dass die Veränderungsmessungen unpräzise sind. Wie Rogosa et al. (1982) zeigen, können nahezu identische wahre Veränderungen in einer Stichprobe sehr präzise gemessen werden, auch wenn die Reliabilität der Differenzwerte wegen ihrer geringen Streuung nahezu 0 ist.

   Um uns diesen scheinbar widersprüchlichen Sachverhalt zu veranschaulichen, betrachten wir zunächst ◘ Abb. 12.51a. Eine Stichprobe von 6 Personen wurde einmal vor (t_1) und ein zweites Mal nach einem Treatment (t_2) mit einem nahezu perfekt reliablen Messinstrument untersucht. Wir entnehmen der Abbildung für alle Personen identische Veränderungen, d. h., die Varianz der Veränderungs- bzw. Differenzwerte d_i ist Null ($s_D^2 = 0$).

   In ▶ Abschn. 10.4.1 wurde die Reliabilität eines Tests allgemein als Quotient aus wahrer und beobachteter Varianz definiert $\left(\text{Rel} = s_T^2/s_X^2 = s_T^2/\left(s_T^2 + s_E^2\right)\right)$. Akzeptieren wir nun s_D^2 als Schätzwert für die wahre Varianz der Veränderungswerte, ist festzustellen, dass die Reliabilität der Differenzwerte (r_D) Null ist: $r_D = s_D^2/\left(s_D^2 + s_E^2\right) = 0/\left(0 + s_E^2\right) = 0$. Dieses Ergebnis bedeutet jedoch keineswegs, dass die Veränderungen ungenau gemessen wurden, denn – so unsere Annahme – das bei Pre- und Posttest eingesetzte Messinstrument hat eine nahezu perfekte Reliabilität. Betrachten wir nun ◘ Abb. 12.51b. Auch hier ist die Rangreihe der 6 Personen im Pre- und im Posttest identisch; allerdings unterscheiden sich die Veränderungsraten d_i von Person zu Person. Hier ist also $s_D^2 > 0$, d. h., bei geringen Messfehleranteilen in den Differenzwerten ist die Reliabilität der Differenzwerte nahezu perfekt. Wir stellen also fest, dass Differenzwerte, die mit ein- und demselben Messinstrument gewonnen wurden, manchmal sehr reliabel und manchmal überhaupt nicht reliabel sein können. Die Reliabilität der Differenzwerte wird bei gleichbleibender Reliabilität des Messinstrumentes von der Streuung der Pretest- und der Posttestwerte (s_1 und s_2) sowie der Korrelation zwischen Pre- und Posttestmessungen bestimmt (r_{12}).



◘ **Abb. 12.51** Zwei Beispiele für Veränderungsmuster mit unterschiedlicher Reliabilität (r_D): **a** $r_D = 0$; **b** $r_D \approx 1$. (Mod. nach Collins, L. M. (1996). Is reliability obsolete? A commentary on „Are simple gain scores obsolete?". *Applied Psychological Measurement, 20*(3), 289–292. Copyright © 2013 by SAGE. Reprinted by permission of SAGE Publications)

Ist der Quotient $s_1/s_2 \approx 1$ und liegt r_{12} nahe bei 1, haben Differenzwerte eine geringe Reliabilität. Sie steigt mit größer werdendem Unterschied von s_1 und s_2 bei einem hohen r_{12}-Wert.

Diese Zusammenhänge legen die Schlussfolgerung nahe, dass das Reliabilitätskonzept der Klassischen Testtheorie bei der Erfassung der Genauigkeit von Differenzwerten offenbar versagt. Eine mögliche Alternative hierzu bieten Veränderungsmessungen im Rahmen der Probabilistischen Testtheorie (Fischer, 1995; Formann & Ponocny, 2002; ▶ Abschn. 10.4.4).

Werden mehr als zwei Messungen vorgenommen, so tritt an die Stelle der einfachen Differenz die Steigung einer an die zeitabhängigen Messungen angepassten Geraden (Steigung der Regressionsgeraden zur Vorhersage der individuellen Messungen aufgrund der Messzeitpunkte). Dieser Steigungsparameter charakterisiert die individuelle Wachstumsrate pro Zeiteinheit (je nach Untersuchungsanlage sind dies Stunden, Tage, Wochen, Monate, Jahre). Dass eine Gerade (anstelle einer nichtlinearen Funktion) zur Charakterisierung eines individuellen Veränderungsverlaufes meist ausreichend ist, wird bei Willet (1989) begründet.

Wie für die einfachen Differenzmaße gilt auch für die Steigungsmaße, dass deren Reliabilität mit zunehmender Streuung der Steigungsmaße steigt (ausführlicher hierzu Maxwell, 1998).

2. **Genauigkeit der Messungen**
   Über die Abhängigkeit der Veränderungsmessungen von der Genauigkeit der Messungen bzw. deren Reliabilität wurde eingangs bereits berichtet. Mit zunehmendem Messfehler bzw. mit abnehmender Reliabilität der Messungen sinkt die Reliabilität der Differenzmaße. Man beachte, dass niedrige Reliabilität der Messungen nicht zwangsläufig niedrige Reliabilität der Differenzen bzw. – bei mehr als zwei Messungen – der Steigungskoeffizienten bedeutet. Obwohl die Reliabilität der Messungen die Reliabilität der Differenzen beeinflusst, kann die Reliabilität der Differenzen beachtlich sein, wenn die wahren individuellen Veränderungsraten sehr heterogen sind.
3. **Verteilung der Messzeitpunkte**
   Die wohl wichtigste, weil untersuchungstechnisch einfach zu manipulierende Determinante der Reliabilität von Veränderungsmessungen ist die Anzahl der pro Person vorgenommenen Messungen bzw. die Art ihrer Verteilung über die Zeit. Bezogen auf die Verteilung der Messpunkte argumentiert Willett (1989), dass mehrere Messungen zu Beginn und am Ende der Untersuchungsperiode äquidistanten Messintervallen deutlich überlegen seien. Diesem statistisch begründeten Vorschlag steht allerdings entgegen, dass die individuelle Veränderungscharakteristik bei gleichförmig verteilten Messpunkten besser erkannt werden kann. Dennoch sollte – soweit die Untersuchungsanlage dies zulässt – darauf geachtet werden, dass die Messungen am Anfang und am Ende des Untersuchungszeitraumes häufiger wiederholt werden als im mittleren Bereich.
4. **Anzahl der Messzeitpunkte**
   Die Reliabilität der Veränderungsmaße lässt sich zudem drastisch verbessern, wenn die Anzahl der Messpunkte erhöht wird, wobei der Reliabilitätszugewinn am größten ist, wenn der Untersuchungsplan statt zwei Messzeitpunkten (z. B. Pre- und Posttest) drei Messzeitpunkte vorsieht. Willett (1989) berichtet, dass die Reliabilität allein durch das Hinzufügen eines dritten Messzeitpunktes um 250 % und mehr erhöht werden kann. Mit wachsender Anzahl der Messzeitpunkte wird der Einfluss eines fehlerhaften bzw. wenig reliablen Messinstrumentes auf die Reliabilität der Veränderungsmaße zunehmend kompensiert.

Schlussfolgerungen bezüglich Differenzwerten Für die Überprüfung von Veränderungshypothesen mit quasiexperimentellen Untersuchungen lässt sich hieraus zusammenfassend folgern, dass man in verstärktem Maße auf einfache Pretest-Posttest-Pläne bzw. Pläne mit nur zwei Messungen verzichten und stattdessen **Untersuchungspläne mit mehr als zwei Messzeitpunkten** vorsehen sollte. Wenn es zudem möglich ist, die Messzeitpunkte am Anfang und am Ende des Untersuchungszeitraumes stärker zu konzentrieren als im mittleren Bereich, erhält man verlässliche Schätzungen der wahren individuellen Veränderungsraten, auch wenn das eingesetzte Messinstrument weniger reliabel ist.

Falls aus untersuchungstechnischen Gründen Pläne mit mehr als zwei Messzeitpunkten nicht umsetzbar sind, ist gegen die Verwendung **einfacher Differenzmaße** als Veränderungsindikatoren nichts einzuwenden. Wird ein Messinstrument eingesetzt, dessen Reliabilität bekannt ist, kann diese zu einer verbesserten Schätzung der wahren individuellen Veränderungen genutzt werden. Einzelheiten hierzu findet man bei Rogosa et al. (1982).

Maxwell (1994) macht zudem darauf aufmerksam, dass sich die **Teststärke** eines Pretest-Posttest-Plans (also die Wahrscheinlichkeit, mit diesem Plan einen Treatmenteffekt nachzuweisen), beträchtlich erhöhen lässt, wenn ca. 25 % der gesamten Erhebungszeit auf den Pretest und ca. 75 % auf den Posttest entfallen. Praktisch bedeutet dies, dass der Aufwand zur Operationalisierung der abhängigen Variablen (z. B. Anzahl der Items einer Testskala oder Dauer der Verhaltensbeobachtung) im Pretest gegenüber dem Posttest reduziert werden kann.

Über den Effekt, den das Hinzufügen eines einzigen zusätzlichen Messpunktes in einem Pretest-Posttest-Design auf die Teststärke der Veränderungsprüfung hat, berichten Venter, Maxwell, & Bolig (2002).

Regression zur Mitte bei Veränderungsmessungen

Bei einer quasi-experimentellen Untersuchung zur Überprüfung von Veränderungshypothesen besteht die Gefahr, dass die Ergebnisse durch sog. Regressionseffekte verfälscht werden: **Extreme Pretest-Werte** haben die Tendenz, sich bei einer wiederholten Messung zur Mitte der Merkmalsverteilung hin zu verändern (Regression zur Mitte) bzw. – genauer – zur größten Dichte der Verteilung (zum Dichtebegriff ▶ Abschn. 12.4.1). Bei unimodalen symmetrischen Verteilungen (z. B. Normalverteilung) entspricht der Bereich mit der größten Dichte dem mittleren Merkmalsbereich. Diese Veränderung erfolgt unabhängig vom Treatment.

Das von dem britischen Naturforscher Francis Galton (1886) erstmals beschriebene Phänomen der **Regression zur Mitte** beruht auf der Beobachtung, dass die Kinder großer Eltern der Tendenz nach über eine kleinere Körpergröße verfügen als die Eltern. Wie ist dieses Phänomen zu erklären? Nehmen wir einmal an, ein Weitspringer absolviert 100 Trainingssprünge. Wenn die Bedingungen für alle Sprünge exakt identisch sind, wenn durch das Training keine Leistungsverbesserung erzielt wird und zudem die Messungen der Sprungweiten absolut fehlerfrei sind, müsste – eine konstante „wahre“ Weitsprungleistung vorausgesetzt – mit allen Sprüngen die gleiche Weite erzielt werden. Dies entspricht natürlich nicht der Realität. Manche Sprünge gelingen besonders gut, weil „alles stimmte“, und andere weniger, weil mehrere „Störfaktoren“ gleichzeitig wirksam waren. Kurz: Die Messungen des Merkmals „Weitsprungleistung“ sind nicht identisch, d. h., wiederholte Messungen desselben Merkmals führen zu unterschiedlichen Ergebnissen, auch wenn die Messungen der Sprungweiten sehr genau bzw. perfekt reliabel sind. Nimmt man an, dass mit den Trainingssprüngen keine merkbaren Leistungsverbesserungen einhergehen und dass Störfaktoren zufällig wirksam sind, werden sich die Weitsprungleistungen des Sportlers normal verteilen (zur Begründung dieser Behauptung s. Bortz & Schuster, 2010, S. 70ff.).

Regression zur Mitte – Als Regression zur Mitte („regression to the mean“, „regression toward the mean“) bezeichnet man bei Messwiederholungen den Effekt, dass extreme Vorher-Messungen bei Wiederholungs-Messungen zur Mitte der Verteilung (allgemein: zur höchsten Dichte der Verteilung) tendieren. Dies gilt vor allem bei geringer Retest-Reliabilität des Merkmals.

Wir beobachten nun einen besonders gelungenen Sprung, bei dem die Weite deutlich über dem individuellen Durchschnitt liegt. Wird nun der nächste Sprung vergleichbar weit oder gar noch weiter sein? Vermutlich eher nicht, denn die Wahrscheinlichkeit, dass sich die Sprungbedingungen erneut so günstig fügen, ist geringer als die Wahrscheinlichkeit für die am häufigsten anzutreffenden „durchschnittlichen“ Sprungbedingungen. Man wäre deshalb mit einer Wette gut beraten, die darauf setzt, dass auf eine hervorragende Sprungweite eine mäßigere folgt. (Betrachten wir die einzelnen Sprünge als stochastisch voneinander unabhängige Ereignisse, ist die Wahrscheinlichkeit jeder beliebigen Sprungweite natürlich unabhängig von der vorangegangenen Sprungweite. Der einfache Hintergrund dieser auf Regressionseffekte zugespitzten Argumentation lautet, dass bei normalverteilten Merkmalen mittlere Ausprägungen häufiger auftreten als extreme.)

Nun registrieren wir statt vieler Sprünge eines Springers jeweils einen Sprung vieler Springer. Auch diese Sprungleistungen mögen sich normal verteilen. Greifen wir nun einen Springer heraus, dessen Sprungleistung weit über dem Mittelwert der Stichprobe liegt, kann man vermuten, dass am Zustandekommen dieser Sprungleistung neben der „wahren“ Sprungstärke auch günstige Bedingungen beteiligt waren. Sofern die Sprungbedingungen von der „wahren“ Sprungleistung unabhängig sind, ist damit zu rechnen, dass diese bei einem zweiten Sprung nicht so günstig ausfallen werden wie beim ersten Sprung, d. h., der zweite Sprung wäre weniger weit. Dies hat zur Folge, dass die ersten Sprünge einer Stichprobe von Weitspringern nicht perfekt mit deren zweiten Sprüngen korrelieren. Diese Korrelation bezeichneten wir in ▶ Abschn. 10.4.1 als Retest-Reliabilität (Stabilität), die gering ausfallen kann, auch wenn die Weitenmessungen selbst perfekt reliabel sind.

Nun kann natürlich die überdurchschnittliche Sprungweite auch von einem sehr guten Springer erzielt worden sein, der mit diesem Sprung (wegen ungünstiger Bedingungen) unter seiner individuellen Norm bleibt. Dieser Springer würde sich bei einem zweiten Sprung vermutlich verbessern. Die Wahrscheinlichkeit, dass gute Sprungleistungen (im Vergleich zur Gruppennorm) unter günstigen Bedingungen erzielt werden, ist jedoch größer als die Wahrscheinlichkeit guter Sprungleistungen unter schlechten Bedingungen.

Die mangelnde Stabilität eines Merkmals hat zur Folge, dass wiederholte Messungen nicht perfekt miteinander korrelieren. Bei völlig instabilen Merkmalen korrelieren wiederholte Messungen mit den ersten Messungen zu Null, d. h., Personen, die bei der ersten Messung einheitlich einen bestimmten Wert erzielen, der deutlich vom Gesamtmittel aller Erstmessungen abweicht, haben bei der zweiten Messung beliebige Werte, deren Mittelwert allerdings weniger vom Stichprobenmittel aller Zweitmessungen abweicht. Dieser Sachverhalt wird „Re-

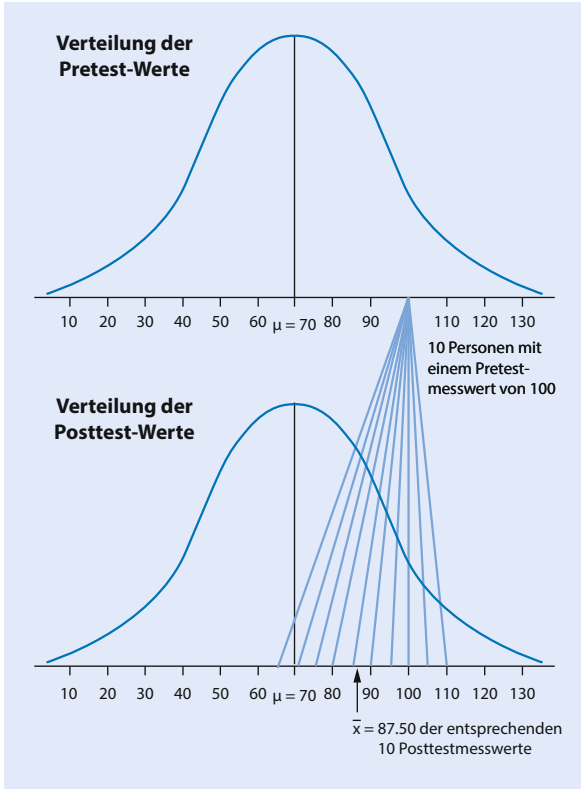


Abb. 12.52 Regressionseffekt bei Pretest-Posttest-Untersuchungen

gression zur Mitte“ genannt. Die Regression extremer Werte zur Mitte der Verteilung nimmt mit abnehmender Retest-Reliabilität des Merkmals zu.

In Abb. 12.52 sieht man den Regressionseffekt für eine Testskala mit mittlerer Retest-Reliabilität. Zehn Personen, die im Pretest einen Wert von 100 erzielten, haben im Posttest Werte zwischen 65 und 110 mit einem Mittelwert von $\bar{x}_2 = 87.50$. Dieser Mittelwert unterscheidet sich weniger von $\mu = 70$ als der Mittelwert der Pretest-Messungen ($\bar{x}_1 = 100$).

Eine weitere Veranschaulichung des Regressionseffekts haben Preacher, Rucker, MacCallum und Nicewander (2005, S. 187f.) vorgeschlagen. In einer Simulationsstudie wurden aus einer bivariat normal verteilten Population $2 \times n = 1\,000$ Messungen generiert, die zu $r = .80$ miteinander korrelieren. Diese Korrelation ist als Retest-Reliabilität zu interpretieren. Die $n = 1\,000$ Messwerte wurden auf der Basis der 1. Messung und auf der Basis der 2. Messung in das untere, mittlere oder obere Drittel der Messwertverteilung klassifiziert, so dass erkennbar wird, wie viele Werte ihre Drittelkategorie verändern bzw. beibehalten (Tab. 12.22).

Man erkennt, dass von den 333 Fällen, die aufgrund der 1. Messung in das untere Drittel fielen, nur 241 Fälle in der 2. Messung in dieser Kategorie verblieben, 76 „regredierten“ in das mittlere Drittel und 16 gar in das obere Drittel. Zusammengenommen ist festzustellen, dass nur ca. 74 % der Fälle ihren Extremgruppenstatus aufrechterhalten (241 von 333 bleiben im unteren und 249 von 333 im oberen Drittel). Würde man nun eine Veränderungshypothese mit einer Extremgruppe (z. B. dem oberen Drittel aufgrund der 1. Messung) durchführen, käme es zu einer Veränderung „hin zur Mitte“, für deren Erklärung allein der Regressionseffekt ausreichte.

Die Merkmalsverteilung in der gesamten Stichprobe wird durch den Regressionseffekt jedoch nicht verändert. Auch dies ist Tab. 12.22 zu entnehmen: Von den 334 Fällen des mittleren Drittels aufgrund der 1. Messung verändern sich $83 + 68 = 151$ Fälle bzw. ca. 45 % weg von der Mitte bzw. in die Extremgruppen. Dieser Austausch – hin zur Mitte und weg von der Mitte – fällt umso deutlicher aus, je geringer die Stabilität bzw. die Retest-Reliabilität ist.

Bei Pretest-Posttest-Messungen mit einem **Extremgruppendesign** würde man sich natürlich v. a. für die veränderten Posttest-Messungen interessieren. Aber der Regressionseffekt „kennt“ keine zeitliche Richtung. Extreme Posttest-Messungen sind mit Pretest-Messungen verbunden, die ebenfalls näher an deren Mittelwert liegen als die extremen Posttest-Messungen. Dies zeigen Nachtigall und Suhl (2002) am Beispiel der Körpergrößen von Psychologiestudentinnen. Regression zur Mitte findet man bei den Töchtern großer Mütter genauso wie bei den Müttern großer Töchter.

Eine formale Analyse der Regressionseffekte liefern Rogosa und Willett (1985, S. 217f.). Danach sind Regressionseffekte an die Voraussetzung geknüpft, dass die Erstmessungen mit den Veränderungsraten negativ korrelieren. Weitere Informationen zu Regressionsartefakten haben Campbell und Kenny (1999) zusammengestellt.

Praktische Implikationen von Regressionseffekten Welche Konsequenzen haben nun Regressionseffekte für quasi-experimentelle Untersuchungen zur Überprüfung von Veränderungshypothesen? Sie können konsequenzenlos sein oder aber zu völlig falschen Schlüssen führen. Beide Fälle seien an einem einfachen Beispiel verdeutlicht.

Es geht um die Frage, ob ein spezielles sportmedizinisches Programm zur Rehabilitation nach einem Bandscheibenvorfall nur für die Betroffenen oder auch für die Allgemeinbevölkerung geeignet ist, um Rückenschmerzen zu reduzieren. Der Einfachheit halber nehmen wir an, das Merkmal Rückenschmerzen sei sowohl bei den von einem Bandscheibenvorfall Betroffenen als auch bei den Nichtbetroffenen normal verteilt. Zudem gehen wir realistischerweise davon aus, dass Rückenschmerzen nicht stabil sind, da sowohl Häufigkeit als auch Intensität der

Tabelle 12.22 Veränderungen durch Regressionseffekte (Erläuterungen s. Text)

		Zweite Messung			
		Unteres Drittel	Mittleres Drittel	Oberes Drittel	Zeilensumme
Erste Messung	Unteres Drittel	241	76	16	333
	Mittleres Drittel	83	183	68	334
	Oberes Drittel	9	75	249	333
	Spaltensumme	333	334	333	1 000

Schmerzbelastung variieren bzw. ungenau erinnert und wiedergegeben werden. Wir ziehen nun aus der Population der Betroffenen und Nichtbetroffenen jeweils eine **Zufallsstichprobe** und stellen anhand eines Vortests fest, dass die von einem Bandscheibenvorfall Betroffenen im Durchschnitt generell stärker an Rückenschmerzen leiden als die Nichtbetroffenen. Nach Absolvierung des Rückentrainings wird die Schmerzbelastung im Posttest erneut gemessen. Hat das Rückentraining keine Wirkung, dürften sich weder der Durchschnittswert der Betroffenen noch der Mittelwert der Nichtbetroffenen bedeutsam geändert haben (wenn man von Störeffekten wie instrumenteller Reaktivität einmal absieht).

Regressionseffekte sind hier ausgeschlossen, da aus beiden Populationen repräsentative Zufallsstichproben gezogen wurden. Zwar werden innerhalb der Stichproben extreme Pretestwerte im Posttest zur Mitte tendieren; gleichzeitig verändern sich jedoch mittlere Werte zu den Extremen hin, d. h., insgesamt bleiben Pretest- und Posttestverteilung unverändert.

Nun wollen wir annehmen, dass aus den Populationen statt repräsentativer Stichproben **selektierte Stichproben** gezogen werden. Einen solchen Selektionseffekt handelt man sich häufig unwillentlich ein, weil die Ziehung einer echten Zufallsstichprobe pragmatisch nicht realisierbar ist (▶ Abschn. 9.3.2). So würde man in der Praxis etwa die Probandenanwerbung für das Rückentraining in einem Rehabilitationszentrum (Betroffene) sowie in einem Sportzentrum (Nichtbetroffene) durchführen. Angenommen man entschließt sich nun, beide Stichproben zu parallelisieren, um die Wirkung personengebundener Störvariablen zu neutralisieren (es handelt sich schließlich um ein quasi-experimentelles Design). Eine Parallelisierung anhand der Vortestergebnisse führt dazu, dass beide Untersuchungsgruppen so zusammengestellt werden, dass sie als Startbedingung im Durchschnitt dieselbe Belastung mit Rückenschmerzen aufweisen. Ist das Training wirkungslos, sollten die Posttest-Werte den Pretest-Werten entsprechen. Tatsächlich zeigen sich aber Veränderungen: Der durchschnittliche Schmerzwert der Betroffenengruppe steigt an, der der Nichtbetroffenengruppe fällt ab. Man würde also fälschlich schließen dass das Training den Betroffenen nicht nur nicht nutzt, sondern sogar schadet, dafür aber den Nichtbetroffenen hilft. Ein solcher Schluss wäre jedoch völlig unangebracht, da die registrierten Veränderungen ausschließlich durch Regressionseffekte erklärbar sind: Die Parallelisierung beider Gruppen war nämlich nur möglich, weil sich in der Betroffenengruppe überwiegend unterdurchschnittlich Schmerzbelastete befanden, während in der Nichtbetroffenengruppe besonders viele überdurchschnittlich Schmerzbelastete an der Untersuchung teilnahmen (andernfalls wären bei gegebener Mittelwertdifferenz in den Populationen ja keine Stichproben mit gleichem Mittelwert konstruierbar gewesen). Die Posttestwerte spiegeln allein die Regression beider Stichproben zum Mittelwert ihrer jeweiligen Referenzpopulationen wider.

Allgemein: Will man die differenzielle Wirkung eines Treatments an **Extremgruppen** überprüfen (z. B. ängstliche versus nicht ängstliche Personen etc.), muss mit Regressionseffekten gerechnet werden (auf weitere Probleme des Extremgruppenvergleichs wurde bereits in ▶ Abschn. 12.5.2 hingewiesen; Möglichkeiten zur Korrektur von Untersuchungsergebnissen in Bezug auf Regressionseffekte diskutieren Thistlethwaite & Campbell, 1960, sowie Vagt, 1976).

Schlussfolgerungen bezüglich Regression zur Mitte Für die quasi-experimentelle Überprüfung von Veränderungshypothesen lässt sich zusammenfassend feststellen, dass die einfachen Differenzen zwischen den Messungen verschiedener Messzeitpunkte sinnvolle, unverzerrte Schätzungen für „wahre" Veränderungen darstellen (Corder-Bolz, 1978; Kenny, 1975; Zielke, 1980).

Andere in der Literatur diskutierte Veränderungsmaße wie z. B. Regressionsresiduen (Du Bois, 1957; Lord, 1956, 1963; Malgady & Colon-Malgady, 1991; McNemar, 1958; Minsel & Langer, 1973), der „Change-Quotient" von Lacey und Lacey (1962) oder auch sog. „wahre" Differenzwerte (Lord, 1953, 1963; McNemar, 1958) sind unter Gesichtspunkten der Praktikabilität bzw. auch inhaltlich für die Erfassung von Veränderungen weniger

geeignet (zusammenfassend Rogosa et al., 1982). Zur Vermeidung von Regressionseffekten sollten die in quasi-experimentellen Untersuchungen eingesetzten Stichproben zufällig bzw. möglichst unverzerrt aus den zu vergleichenden Populationen ausgewählt werden; der Vergleich von Veränderungen in Extremgruppen ist äußerst problematisch.

Die Messungen sollten mindestens **intervallskaliert** sein, denn Differenzen sind bei einem niedrigeren Skalenniveau inhaltlich sinnlos (Beispiele hierfür gibt Stelzl, 1982, Abschn. 7.1). Abzuraten ist ferner von Messskalen, die in extremen Merkmalsbereichen begrenzt sind (z. B. Ratingskalen). Extrem hohe Messwerte können sich dann nicht mehr vergrößern **(Ceiling- oder Deckeneffekt)** und extrem niedrige Messwerte nicht mehr verringern **(Floor- oder Bodeneffekt)**. Für die Auswertung von Untersuchungen, bei denen Veränderungen mit nominalen Daten erfasst wurden, findet man bei Langeheine und van de Pol (1990) einschlägige Verfahren (zur Analyse kategorialer Daten s. Agresti, 2007).

Mit Nachdruck ist darauf hinzuweisen, dass Untersuchungen mit **drei oder mehr Messzeitpunkten** erheblich vorteilhafter sind als Untersuchungen mit nur zwei Messzeitpunkten. Die Präzision der Veränderungsmessung lässt sich zudem erhöhen, wenn sich die Messzeitpunkte am Anfang und am Ende des Untersuchungszeitraumes stärker konzentrieren als im mittleren Bereich.

12

■ Pläne für quasi-experimentelle Studien mit Messwiederholungen

Im Folgenden werden einige quasi-experimentelle Untersuchungspläne vorgestellt, die der Prüfung von Veränderungshypothesen zu Treatment-Effekten dienen und in der Praxis häufig eingesetzt werden bzw. die für die Praxis besonders wichtig erscheinen. Einen Vergleich verschiedener Auswertungsverfahren, auch unter dem Blickwinkel unvollständiger Daten, findet man bei Delucchi und Bostrom (1999).

Eingruppen-Pretest-Posttest-Pläne Bei einem Eingruppen-Pretest-Posttest-Plan wird idealerweise eine möglichst repräsentative Stichprobe der interessierenden Zielpopulation einmal vor und einmal nach dem Treatment untersucht. Die durchschnittliche Differenz auf der abhängigen Variablen gilt behelfsweise als Indikator für die Treatment-Wirkung, obwohl praktisch alle in ► Abschn. 3.2.2 genannten Störeinflüsse die Veränderung bzw. Nichtveränderung ebenfalls bewirkt haben könnten. Die interne Validität dieses Designs ist also gering. Sie lässt sich jedoch durch die vorsorgliche Erhebung zeitabhängiger Variablen verbessern, die die abhängige Variable ebenfalls beeinflussen können (Kontrollvariablen) und deren Einfluss nachträglich kontrolliert wird (Partialkorrelation; ► Abschn. 12.5.2).

Gelegentlich ist man auf den Einsatz dieses Planes angewiesen. Dies gilt v. a. für Fragestellungen, bei denen ein Treatment interessiert, von dem praktisch alle Personen betroffen sind, so dass auf die Bildung einer Kontrollgruppe verzichtet werden muss. Beispiele hierfür sind Untersuchungen zur Wirkung einer neuen Fernsehwerbung oder eines neuen Gesetzes. Auch ethische Gründe können den Einsatz einer Kontrollgruppe unmöglich machen. Die statistische Auswertung erfolgt bei einer kardinalskalierten abhängigen Variable über den ***t*-Test für abhängige Stichproben**, bei einer dichotom-nominalskalierten Variablen über den **McNemar-Chi-Quadrat-Test** auf Veränderungen (z. B. Anteil der Raucher und Nichtraucher vor und nach einer Anti-Rauch-Kampagne). Man beachte, dass beim Eingruppen-Plan das Treatment keine unabhängige Variable darstellt, da es nicht variiert wird (kein Vergleich des Treatments mit einem Alternativtreatment oder einer Kontrollgruppe). Stattdessen fungiert der Messzeitpunkt (Vorher- versus Nachher-Messung) als unabhängige Variable.

Eine noch geringere interne Validität als Eingruppen-Pretest-Posttest-Pläne, bei denen Messwerte einer Stichprobe vor und nach einem Treatment erhoben werden, haben Eingruppen-Pläne ohne konkret benanntes Treatment. Veränderungshypothesen, die behaupten, eine abhängige Variable verändere sich **im Laufe der Zeit** ohne eine konkret zu benennende Treatment-Wirkung (Beispiele: Das Konzentrationsvermögen von Kindern ist morgens höher als abends; Arbeitsausfälle durch Krankmeldungen treten am Anfang der Woche häufiger auf als in der Mitte der Woche; die Bereitschaft der Bevölkerung, aktiv etwas gegen die Zerstörung der Umwelt zu unternehmen, hat in den letzten Jahren zugenommen) sprechen mit der „Zeit" einen **globalen Variablenkomplex** an, dessen verändernde Wirkung auf viele unkontrollierte und zeitabhängige Merkmale zurückgeht. Insbesondere sind bei der Betrachtung von Veränderungen über längere Zeitspannen (z. B. Vorher- und Nachher-Messung im Abstand eines Jahres) die Effekte des Lebensalters, der Geburtskohorte der Untersuchten sowie der historischen Situation des Erhebungszeitpunktes konfundiert (auf die Prüfung entsprechender **Entwicklungshypothesen** gehen wir in einem späteren Abschnitt ein).

Veränderungshypothesen zu Interventionen können mit höherer interner Validität durch quasi-experimentelle Untersuchungen geprüft werden, die **mehrere Pretest- und Posttest-Messungen** umfassen. Wenn sich hierbei zeigt, dass sich das Niveau der Pretest-Messungen deutlich vom Niveau der Posttest-Messungen unterscheidet, ist dies ein guter Beleg für Treatmentwirkungen. Man beachte jedoch, dass die interne Validität derartiger Un-

tersuchungen, die sich meist über einen längeren Zeitraum erstrecken, besonders durch Testübung und experimentelle Mortalität gefährdet ist. Bei mehr als zwei Messungen erfolgt die statistische Hypothesenprüfung über die **einfaktorielle Varianzanalyse mit Messwiederholungen**, wobei der Messzeitpunkt den Messwiederholungsfaktor bzw. die unabhängige Variable darstellt. Kommt eine varianzanalytische Auswertung nicht in Betracht (z. B. wegen verletzter Voraussetzungen), kann auf verteilungsfreie Verfahren (z. B. H-Test), eine regressionsanalytische Auswertungstechnik von Swaminathan und Algina (1977) oder – bei vielen Messzeitpunkten – ggf. auf zeitreihenanalytische Technikenzurückgegriffen werden.

Eine Modifikation des einfachen Pretest-Posttest-Planes wurde von Johnson (1986) vorgeschlagen. Ein typischer Anwendungsfall dieses Planes könnte ein zu evaluierendes Bildungsprogramm sein, das der interessierten Öffentlichkeit in mehreren sich wiederholenden Workshops angeboten wird (z. B. Workshop über Steuerrecht). Es seien z. B. vier Termine mit identischem Unterrichtsangebot vorgesehen, auf die die Voranmeldungen zufällig verteilt werden. Die einem Termin zugeordneten Personen werden überdies zufällig in vier Gruppen eingeteilt: Eine Pre-Pretestgruppe, eine Pretestgruppe, eine Posttestgruppe und eine Post-Posttestgruppe. Die erste Gruppe wird z. B. zwei Wochen, die zweite Gruppe eine Woche vor dem Workshop-Termin, die dritte Gruppe eine Woche nach und die vierte Gruppe zwei Wochen nach dem Workshop-Termin hinsichtlich der abhängigen Variablen (Steuerkenntnisse) getestet. Zusammengefasst über die verschiedenen Workshop-Termine erhält man so 4 Gruppen (Pre-Pre, Pre, Post, Post-Post), die 4 Stufen einer unabhängigen Variablen für eine **einfaktorielle Varianzanalyse ohne Messwiederholungen** bilden. Nach Johnson (1986) ist die interne Validität dieses Planes der einer experimentellen Untersuchung mit einer Kontrollgruppe nahezu ebenbürtig. Dadurch dass jede Stichprobe nur einmal untersucht wird, werden v. a. Testübungseffekte bzw. instrumentelle Reaktivität vermieden.

Zweigruppen-Pretest-Posttest-Pläne Eine Verbesserung der internen Validität lässt sich in quasi-experimentellen Untersuchungen dadurch erzielen, dass neben der Experimentalgruppe eine Kontrollgruppe geprüft wird und somit neben der unabhängigen Variable Messzeitpunkt noch die unabhängige Variable Treatment berücksichtigt wird (zweifaktorieller Plan bzw. im einfachsten Fall: 2×2-Plan).

Beispiel: Es geht um die Evaluation von computergestütztem Unterricht in Mathematik. Da eine individuelle Randomisierung von Experimental- und Kontrollgruppe seitens der Schulleitung abgelehnt wird, ist man auf den Vergleich „natürlicher" Gruppen (hier: Schulklassen) angewiesen: Zwei Schulklassen werden für die Untersuchung ausgewählt und mit einem einheitlichen Instrument vorgetestet. Nach den Vortests erhält eine Klasse computergestützten Unterricht (Experimentalgruppe) und die Parallelklasse vom gleichen Lehrer Normalunterricht. Den Abschluss der Untersuchung bilden Posttests in beiden Klassen.

Tabelle 12.23 Schema zur Ermittlung eines Treatmenteffektes

	Pretest	Posttest	Differenz
Experimentalgruppe	E_1	E_2	$E = E_1 - E_2$
Kontrollgruppe	K_1	K_2	$K = K_1 - K_2$
			Nettoeffekt $= E - K$

Für die statistische Auswertung dieses Planes empfiehlt sich eine **zweifaktorielle Varianzanalyse mit Messwiederholungen auf dem Faktor Messzeitpunkt**. Um den „Nettoeffekt" des Treatments zu ermitteln, berechnet man nach Rossi und Freeman (1985, S. 238) die Differenz der Veränderung in der Experimental- und der Kontrollgruppe (Tab. 12.23).

Die Buchstaben E und K stehen hier für Durchschnittswerte in der Experimental- bzw. Kontrollgruppe. Ein statistisch signifikanter „Nettoeffekt" wird durch eine **signifikante Interaktion** zwischen dem Gruppenfaktor und dem Messwiederholungsfaktor nachgewiesen. Der signifikante Interaktionseffekt der zweifaktoriellen Varianzanalyse besagt, dass sich die Experimentalgruppe überzufällig anders verändert als die Kontrollgruppe, bei der ja von keinem Effekt ausgegangen wird.

Die interne Validität dieses Planes ist akzeptabel, solange sich die durchschnittlichen Vortestwerte aus Experimental- und Kontrollgruppe (und auch ihre Streuungen) nicht allzu stark unterscheiden. Bei großen Diskrepanzen besteht die Gefahr von Regressionseffekten, die sich darin äußern würden, dass sich eine hohe Pretest-Differenz im Posttest verkleinert. (Im Beispiel bestünde diese Gefahr, wenn man eine Schulklasse mit guten Mathematikkenntnissen und eine Schulklasse mit schlechten Mathematikkenntnissen vergleicht.)

Externe zeitliche Einflüsse, Reifungsprozesse und Testübung werden in diesem Plan durch die Berücksichtigung einer Kontrollgruppe kontrolliert. Falls derartige Effekte wirksam sind, würden sie beide Gruppen in gleicher Weise beeinflussen, es sei denn, eine der beiden Gruppen ist für diese Störeffekte „anfälliger" als die andere (Interaktion von Störeinflüssen mit dem Gruppierungsfaktor).

		Pretest (C_1)	Posttest (C_2)	(weitere Posttests)
A_1 (männlich)	B_1 (Experimentalgruppe)	S_1	S_1	(S_1)
	B_2 (Kontrollgruppe)	S_2	S_2	(S_2)
A_2 (weiblich)	B_1 (Experimentalgruppe)	S_3	S_3	(S_3)
	B_2 (Kontrollgruppe)	S_4	S_4	(S_4)

Abb. 12.53 Dreifaktorieller Pretest-Posttest-Plan

Experimentelle Mortalität kann zu einem Problem werden, wenn der zeitliche Abstand zwischen Pre- und Posttest groß ist und einige Untersuchungspersonen für den Posttest nicht mehr zur Verfügung stehen. Kommt es hierbei zu systematischen Selektionsfehlern, weil die Ausfälle in Experimental- oder Kontrollgruppe nicht zufällig sind, dann ist die interne Validität der Untersuchung erheblich gefährdet.

Zur Steigerung der internen Validität ist ferner zu erwägen, ob sich das einfache Pretest-Posttest-Design durch mehrere Pretests und/oder mehrere Posttests erweitern lässt. Nicht zuletzt ist in manchen Fällen, in denen eine individuelle Randomisierung nicht möglich ist, in größer angelegten Studien eine sog. **Cluster-Randomisierung** realisierbar, bei der natürliche Gruppen randomisiert den Untersuchungsbedingungen zugeordnet und entsprechend ausgewertet werden. Damit würde das quasi-experimentelle Design zum experimentellen Design aufgewertet und die interne Validität gesteigert.

Mehrfaktorielle Pretest-Posttest-Pläne Mehrfaktorielle (d. h. mehr als zweifaktorielle) Pretest-Posttest-Pläne überprüfen differenzielle Wirkungen eines Treatments (z. B. Kopfschmerztherapie) auf verschiedene Populationen (z. B. Patienten unterschiedlichen Geschlechts, unterschiedlicher Altersgruppen, unterschiedlicher Kopfschmerztypen etc.). Hierfür sind zunächst aus den jeweiligen Referenzpopulationen Stichproben zu ziehen. Jede Stichprobe wird (möglichst zufällig) in eine Kontrollgruppe und eine Experimentalgruppe aufgeteilt. Mit Pretests der abhängigen Variablen ermittelt man für alle Gruppen die Ausgangsbedingungen. Unterschiede im Pretest zwischen Experimental- und Kontrollgruppen, die aus derselben Population stammen, sind durch Parallelisierung (ggf. Matching) auszugleichen. Pretest-Unterschiede zwischen Stichproben verschiedener Populationen werden akzeptiert und – um Regressionseffekte zu vermeiden – nicht durch eine selektive Auswahl von Untersuchungseinheiten ausgeglichen. Nach Einführung des Treatments erhebt man eine Posttest-Messung oder – besser noch – mehrere Wiederholungsmessungen. Über alle Pretest- und Posttestwerte wird eine **mehrfaktorielle Varianzanalyse mit Messwiederholungen** gerechnet.

Wird z. B. in einem Pretest-Posttest-Design (Messwiederholungsfaktor C) eine Kopfschmerztherapie mit einer Kontrollbedingung verglichen (Treatmentfaktor B) und wird dabei das biologische Geschlecht als weiterer zweifach gestufter Faktor A berücksichtigt, so ergibt sich ein dreifaktorieller $2 \times 2 \times 2$-Plan, der durch weitere Posttestmessungen erweiterbar ist (Abb. 12.53). Ist eine **signifikante Interaktion zweiter Ordnung** ($A \times B \times C$) darauf zurückzuführen, dass sich die beiden Experimentalgruppen unterschiedlich und die beiden Kontrollgruppen nicht verändert haben, wird damit eine differenzielle Veränderungshypothese bestätigt, nach der z. B. nur Patientinnen von der Therapie profitieren. (Dieses Interaktionsmuster sollte durch Einzelvergleiche bestätigt werden. Zur Konstruktion von Kontrasten und Prüfung von Einzelvergleichshypothesen im Rahmen von Messwiederholungsanalysen s. z. B. Furr & Rosenthal, 2003; Bortz & Schuster, 2010, Kap. 18; Eid, Gollwitzer, & Schmitt, 2010, Kap. 14). Verändern sich auch die Mittelwerte der Kontrollgruppen, muss man damit rechnen, dass außer dem Treatment weitere Variablen wirksam sind. Populationsspezifische Treatmentwirkungen sind dann nicht mehr eindeutig, sondern nur in Verbindung mit den Veränderungen der Kontrollgruppen interpretierbar.

Nicht alle Fragestellungen lassen die Bildung von Experimental- und Kontrollgruppen innerhalb der zu vergleichenden Stichproben zu. Will man z. B. überprüfen, wie sich die Herabsetzung der Regelstudienzeit von zehn auf acht Semester auf die durchschnittliche Studienleistung in verschiedenen Fächern auswirkt, so kann man innerhalb der einzelnen Studierendenstichproben nicht zwischen Untersuchungsteilnehmern, die von der Maßnahme betroffen sind (Experimentalgruppe), und solchen, die sie nicht betrifft (Kontrollgruppe), unterscheiden. Die Untersuchung könnte deshalb nur die Leistungen von Stichproben vor dieser Maßnahme mit Leistungen danach

vergleichen. Führt die statistische Auswertung des Materials (**zweifaktorielle Varianzanalyse** ohne Messwiederholungen bzw. mit Messwiederholungen, wenn „matched samples“ untersucht werden) zu einer signifikanten Interaktion, ist dies allerdings nur ein schwacher Beleg für eine differenzielle Wirkung der Maßnahme, denn man kann nicht ausschließen, dass andere Ursachen als die Verkürzung der Studienzeit für die Leistungsveränderungen in den einzelnen Studienfächern verantwortlich sind.

Solomon-Viergruppenplan Der bereits beschriebene Solomon-Viergruppenplan kann auch quasi-experimentell, d. h. mit nicht-randomisierten Gruppen, eingesetzt werden. Man beachte allerdings, dass die Einbeziehung der Gruppen 3 und 4 in die Designanalyse Probleme bereiten kann, wenn man davon ausgehen muss, dass diese Gruppen zu den Gruppen 1 und 2 nicht äquivalent sind. (Da die Gruppen 3 und 4 nicht vorgetestet werden, muss im experimentellen Ansatz unterstellt werden, dass diese Gruppen zu den anderen äquivalent sind. Die Rechtfertigung hierfür liefert die Randomisierung, auf die bei quasi-experimentellem Vorgehen verzichtet werden muss.)

Regressions-Diskontinuitäts-Analyse Quasi-experimentelle Untersuchungen mit Experimentalgruppe und nicht-äquivalenter Kontrollgruppe sind nur bedingt aussagekräftig, weil man nicht weiß, ob die Posttestunterschiede zwischen den Gruppen bzgl. der abhängigen Variablen allein auf das Treatment oder auf andere Besonderheiten der verglichenen Gruppen zurückzuführen sind. Bei der Auswahl der Untersuchungspersonen wird man deshalb besonders darauf achten, dass Experimental- und Kontrollgruppe möglichst ähnlich sind.

Einen anderen Weg beschreitet die **Regressions-Diskontinuitäts-Analyse („regression discontinuity design/analysis“**; Thistlethwaite & Campbell, 1960). Hier werden Unterschiede zwischen Experimental- und Kontrollgruppe bewusst herbeigeführt: Personen, die einen bestimmten Wert (Cut-off-Point) einer kontinuierlichen **Assignment- oder Zuweisungsvariablen** unterschreiten, zählen zur Kontrollgruppe und Personen oberhalb dieses Wertes zur Experimentalgruppe (oder umgekehrt). Eine Treatmentwirkung liegt in diesem Untersuchungsplan vor, wenn die Regressionsgerade zur Beschreibung des Zusammenhangs zwischen der Zuweisungsvariablen und der abhängigen Variablen am Cut-off-Point diskontinuierlich verläuft und gleichzeitig die entsprechende Regression ohne Treatment einen kontinuierlichen Verlauf nimmt.

Ein kleines Beispiel soll diesen Versuchsplan verdeutlichen. Man habe festgestellt, dass der Gesundheitszustand vieler Kinder, deren Eltern über ein geringes Einkommen verfügen, zu wünschen übrig lässt und vermutet

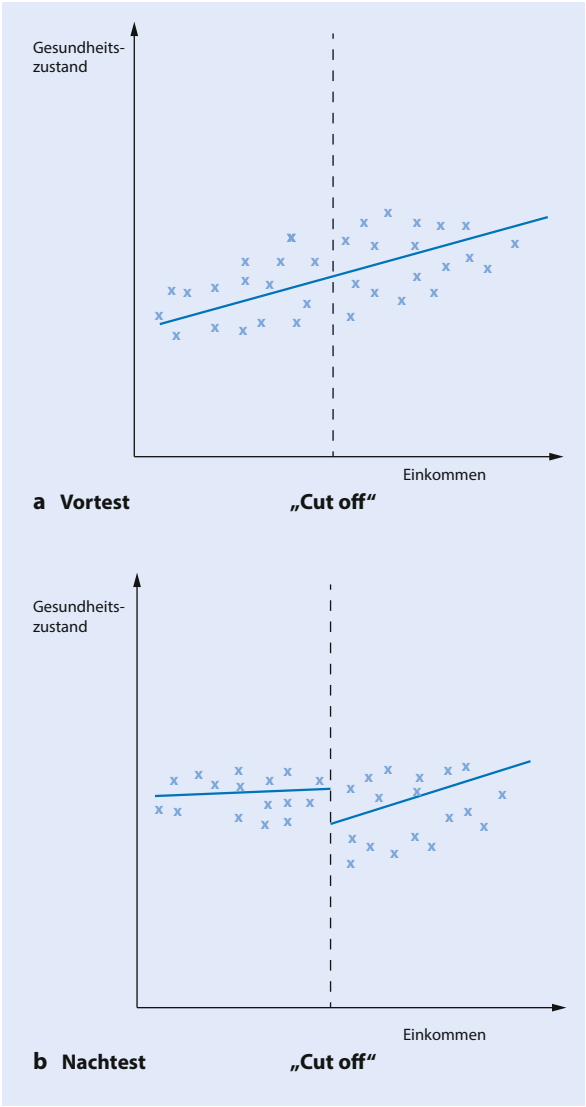


Abb. 12.54 Beispiel für eine Regressions-Diskontinuitäts-Analyse. **a** Vortest, **b** Nachtest

als Ursache hierfür eine schlechte bzw. unausgewogene Ernährung. Man plant eine Aufklärungsaktion „gesunde Ernährung“ und will diese anlässlich eines Ferienlageraufenthaltes mit ausgewählten Kindern evaluieren. Eine Regressions-Diskontinuitäts-Analyse könnte hier wie folgt aussehen: Eine Zufallsstichprobe von Kindern wird in einem Vortest bzgl. ihres Gesundheitszustandes untersucht. Zusätzlich wird das Einkommen der Eltern erfragt. Den Zusammenhang dieser beiden Variablen verdeutlicht Abb. 12.54a aufgrund der Vortestergebnisse.

Zur Bildung von Experimental- und Kontrollgruppe legt man ein Mindesteinkommen fest (Cut-off-Point) und entscheidet, dass alle Kinder der Stichprobe mit Eltern, deren Einkommen unterhalb des Cut-off-Points liegen,

zur Experimentalgruppe gehören. Die Kinder mit Eltern höheren Einkommens bilden die Kontrollgruppe (ohne Ferienlager). Vier Wochen nach Abschluss des Ferienlagers „gesunde Ernährung“ wird der Gesundheitszustand der Kinder erneut geprüft. Das Ergebnis zeigt Abb. 12.54b.

Offensichtlich hat das Treatment gewirkt. Die Vortestergebnisse zeigen einen tendenziell besseren Gesundheitszustand mit wachsendem Einkommen und die Regressionsgerade, die den Trend der „Punktewolke“ kennzeichnet, ist kontinuierlich. Im Nachtest hingegen sind es 2 Regressionsgleichungen, die den Trend in der Experimental- und Kontrollgruppe am besten beschreiben. Die Kinder mit Eltern geringeren Einkommens (links vom Cut-off-Point) befinden sich nach dem Ferienlageraufenthalt in einem besseren Gesundheitszustand als vor dem Ferienlager, zudem ist die Regressionsgerade diskontinuierlich.

Bei einer Regressions-Diskontinuitäts-Analyse ist darauf zu achten, dass zwischen der Zuweisungsvariablen und der abhängigen Variablen ein Zusammenhang besteht. Nach Mosteller (1990) ist dieser Plan einer experimentellen Untersuchung qualitativ gleichwertig (s. auch Rubin, 1977; Shadish, Galindo, Wong, Steiner, & Cook, 2011). Weitere Informationen findet man bei Trochim (1984), Braden und Bryant (1990) sowie Trochim und Cappelleri (1992) und Hinweise zur Auswertung bei Bierhoff und Rudinger (1996).

Korrelate von Veränderung

Abschließend seien Veränderungshypothesen erwähnt, mit denen behauptet wird, dass die Veränderung eines Merkmals mit einem anderen Merkmal (Drittvariable) korreliert. Als Beispiele lassen sich die Hypothesen nennen, dass der Lernfortschritt von Schülern mit ihrer Intelligenz zusammenhängt, dass Fortschritte in der Genesung Kranker von ihrer Bereitschaft, gesund werden zu wollen, abhängen oder dass Einstellungsänderungen mit zunehmendem Alter unwahrscheinlicher werden. In allen Beispielen geht es um den Zusammenhang zwischen der Veränderung einer Variablen und den Ausprägungen einer Drittvariablen. Bei der Überprüfung derartiger Hypothesen unterscheiden wir drei Fälle:

1. Die Differenzen stehen in keinem Zusammenhang zu den Eingangswerten, d. h., Stärke und Richtung der Veränderungen sind von den Vortest-Messungen unabhängig. In diesem Fall überprüft eine **bivariate Korrelation** zwischen den Differenzwerten und der Drittvariablen die Veränderungshypothese.
2. Die Veränderungen hängen von den Vortest-Ergebnissen ab (z. B. in der Weise, dass mit wachsender Vortest-Messung auch größere Veränderungen auftreten), und diese Abhängigkeit soll bei der Überprüfung der Veränderungshypothese mitberücksichtigt werden. Auch in diesem Falle empfiehlt sich die Berechnung einer **bivariaten Korrelation** zwischen den Differenzen und der Drittvariablen.
3. Es besteht eine Abhängigkeit zwischen den Vortest-Ergebnissen und den Veränderungen, aber diese Abhängigkeit soll unberücksichtigt bleiben. In dieser Situation bestätigt eine signifikante **Partialkorrelation** zwischen den Differenzwerten und der Drittvariablen unter Ausschaltung des Einflusses der Vortestwerte die Veränderungshypothese. Man kommt zu identischen Resultaten, wenn in diese Partialkorrelation statt der Differenzwerte die Posttestwerte eingesetzt werden (s. Helmreich, 1977, Abschn. 4.4; eine formale Analyse dieser Thematik findet man bei Rogosa & Willett, 1985).

Allgemeine Designempfehlungen

Wie bereits mehrfach angesprochen können Defizite beim Untersuchungsdesign oder bei der Operationalisierung der Variablen durch die spätere statistische Datenanalyse kaum kompensiert werden. Um Veränderungshypothesen mithilfe quasi-experimenteller Studien zu prüfen und dabei zu aussagekräftigen Resultaten zu kommen, ist nicht nur eine korrekte Auswertung durch den passenden Signifikanztest ausschlaggebend, sondern bereits bei der Untersuchungsplanung auf eine möglichst hohe interne Validität der Untersuchung zu achten. Zur Erhöhung der internen Validität quasi-experimenteller Untersuchungen, mit denen die Wirksamkeit eines Treatments überprüft werden soll, seien die folgenden Maßnahmen empfohlen (Cook & Shadish, 1994):

- **Einsatz mehrerer abhängiger Variablen oder Wirkkriterien:** Neben der bzw. den theoretisch mit dem Treatment verbundenen abhängigen Variablen (erwartete Wirkungen) sind auch solche Variablen vorzusehen, die mögliche alternative Erklärungen der Maßnahmenwirkung ausgrenzen helfen (z. B. mögliche „Nebenwirkungen“; zu multivariaten Versuchsplänen ▶ Abschn. 12.5.3).
- **Wiederholte Treatmentphasen:** Falls es die Untersuchungsumstände zulassen, empfiehlt es sich, das Treatment bei derselben Stichprobe nach einem angemessenen Zeitabstand erneut oder sogar mehrfach einzusetzen. Zeigen sich identische Treatmentwirkungen wiederholt, so ist dies ein guter Beleg dafür, dass die Untersuchung intern valide ist.
- **Wiederholte Pretest-Messungen:** Werden Experimental- **und** Kontrollgruppe zwei oder mehreren Pretest-Messungen unterzogen, erfährt man, ob bzw. wie sich die verglichenen Stichproben auch ohne Treatmentwirkungen verändern. Differenzielle Veränderungen in der Pretestphase haben dann die Funktion

einer **„Baseline“**, die die Interpretation gruppenspezifischer Veränderungen während oder nach der Treatmentphase, die ursächlich auf Treatmentwirkungen in der Experimentalgruppe zurückgeführt werden sollen, präzisieren hilft.

- **Mehr als zwei Vergleichsgruppen:** Mehrere Experimentalgruppen neben der Kontrollgruppe sind von großem Vorteil, wenn sich theoretisch begründen lässt, dass bestimmte Gruppen stärker und andere weniger stark auf das Treatment reagieren. Werden derartige Erwartungen empirisch bestätigt, ist dies ein guter Beleg für die interne Validität der Studie (Holland, 1986).
- **Abgestufte Treatment-Intensität:** Bei manchen Untersuchungen ist es möglich, dass verschiedene – evtl. auch ex post gebildete – Teilgruppen das Treatment mit unterschiedlicher Intensität oder „Dosis“ erhalten (Abb. 12.55). Hier wäre – ähnlich wie bei Teilgruppen, die auf ein konstantes Treatment unterschiedlich sensibel reagieren – ebenfalls mit abgestuften Treatmentwirkungen zu rechnen.
- **Parallelisierung:** Soweit möglich, sollten die zu vergleichenden Gruppen parallelisiert sein. Das Matching sollte auf stabilen Merkmalen beruhen, die zudem – zumindest theoretisch – mit der abhängigen Variablen zusammenhängen. Man achte hierbei jedoch auf mögliche Regressionseffekte.
- **Analyse der Gruppenselektion:** Wie bereits erwähnt, ist es von großem Vorteil, wenn der Selektionsprozess, der zur Bildung von Experimental- und Kontrollgruppe führte, genau reanalysiert werden kann. Wenn schon in quasi-experimentellen Untersuchungen mit nichtäquivalenten Vergleichsgruppen gearbeitet werden muss, sollte zumindest – so gut wie möglich – in Erfahrung gebracht werden, bzgl. welcher Merkmale Gruppenunterschiede bestehen, um diese ggf. im Nachhinein statistisch zu kontrollieren.
- **Konfundierte Merkmale bzw. Störvariablen:** Zu betonen ist erneut die Notwendigkeit, nach allen Merkmalen zu suchen, die neben dem Treatment ebenfalls auf die abhängige Variable Einfluss nehmen können. Diese Merkmale sind unschädlich, wenn sie – wie in randomisierten Experimenten – in Experimental- und Kontrollgruppe vergleichbar ausgeprägt sind. Sie können eine quasi-experimentelle Untersuchung jedoch völlig invalidieren, wenn ihr Beitrag zur Nichtäquivalenz erheblich bzw. ihre Beeinträchtigung der abhängigen Variablen nicht kontrollierbar ist.

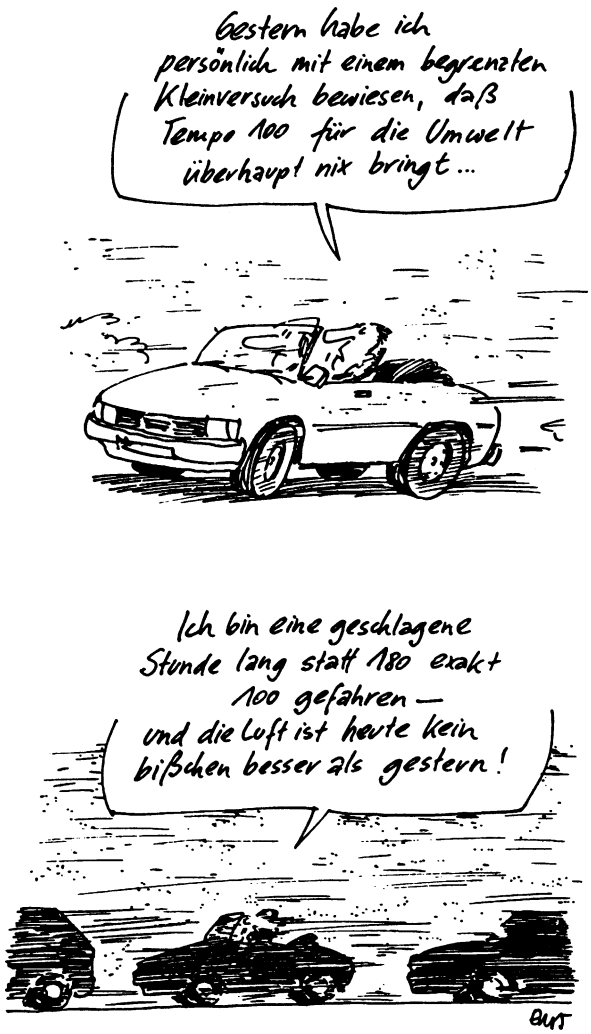


Abb. 12.55 Bei der experimentellen Prüfung von Veränderungshypothesen büßt man interne Validität ein, wenn das Treatment nicht angemessen dimensioniert wurde. (http://www.erich-rauschenbach.de. Mit freundlicher Genehmigung)

Fassen wir zusammen: Quasi-experimentelle Untersuchungen mit nichtäquivalenten Vergleichsgruppen sind hinsichtlich ihrer internen Validität experimentellen Untersuchungen mit randomisierten Vergleichsgruppen bei der Prüfung von Veränderungshpyothesen unterlegen. Dennoch sind sie für viele Fragestellungen zu Veränderungen unersetzbar. Eine Verbesserung der internen Validität dieser Untersuchungsart lässt sich „mechanisch“ oder „standardisiert“ kaum erzielen, denn die hier genannten Empfehlungen sind keineswegs durchgängig für jede Fragestellung praktikabel. Die Empfehlungen sollten jedoch ein Problembewusstsein fördern, durch eine kreative Designgestaltung auch quasi-experimentelle Untersuchungen so anzulegen, dass deren interne Validität bestmöglich gesichert ist, so dass dann bei der (meist varianzanalytischen) Datenauswertung ein statistisch signifikanter Veränderungseffekt möglichst zweifelsfrei auf

den kausalen Einfluss der untersuchten Intervention zurückgeführt werden kann.

Veränderungshypothesen für Entwicklungen: Alters-, Kohorten- und Epocheneffekte

Quasi-experimentelle Untersuchungen zur Überprüfung von Veränderungshypothesen (im Sinne von Hypothesen über Treatment-Wirkungen) führen – so zeigten die letzten Abschnitte – zu weniger eindeutigen Resultaten als experimentelle Untersuchungen. Gefährdungen der internen Validität sind zudem typisch für **nicht-experimentelle Studien**, die Veränderungen in Abhängigkeit vom Lebensalter (sog. **Entwicklungshypothesen**) untersuchen.

Neben dem Alter als unabhängige Variable berücksichtigt die entwicklungspsychologische Forschung auch die Wirkung zweier weiterer unabhängiger Variablen: **Zeiteffekte** (oder **epochale Effekte**) sowie **Kohorten- bzw. Jahrgangs-/Generationseffekte** (Baltes, 1967; Rudinger, 1981). Die folgenden drei Hypothesen sollen die Bedeutung der drei unabhängigen Variablen Alter, Epoche und Kohorte im Kontext der Prüfung von Veränderungshypothesen veranschaulichen:

- Die Gedächtnisleistung des Menschen lässt mit zunehmendem Alter nach. Hier werden Veränderungen des Gedächtnisses auf die unabhängige Variable **Alter** zurückgeführt. Nach Schaie (1965, zit. nach Hoppe, Schmid-Schönbein, & Seiler, 1977, S.141) sind mit Alterseffekten im engeren Sinne Verhaltensänderungen gemeint, die auf neurophysiologische Reifungs- bzw. Abbauprozesse der Individuen zurückgehen und prinzipiell für Menschen verschiedener Kohorten und in verschiedenen Epochen gelten sollten.
- Die Studierenden der frühen 1970er-Jahre waren politisch aktiver als die Studierenden der frühen 1990er-Jahre. Diese Hypothese behauptet unterschiedliche studentische Aktivitäten in verschiedenen historischen Zeitabschnitten oder **Epochen**. Allgemein betreffen epochale Effekte Verhaltensbesonderheiten, die für eine Population in einem begrenzten Zeitabschnitt typisch sind. Stichworte wie „Mode“, „Zeitgeist“, „gesellschaftlicher Wandel“ etc. sind für epochale Besonderheiten typisch. Sie sind Ausdruck kultureller, wissenschaftlicher, ökonomischer und ökologischer Veränderungen.
- Menschen einer bestimmten **Kohorte** bzw. **Generation** (ähnliche Geburtsjahrgänge) haben als Gleichaltrige dieselben Epochen durchgemacht und verfügen damit über homogenere biografische Erfahrungen als Menschen anderer Kohorten. Hypothesen können sich nun auf Merkmale oder Verhaltensweisen der Vertreter einer bestimmten Kohorte beziehen (z. B. politische und technische Einstellungen der „Kriegs-Generation“, der „Nachkriegs-Generation“ oder der „Internet-Generation“).

Beobachten wir zu einem bestimmten Zeitpunkt das Verhalten eines Menschen, wird dieses – neben weiteren Determinanten – immer auch von Alters-, Epochen- und Kohorteneffekten abhängen. Die drei Variablen sind untrennbar miteinander verknüpft bzw. konfundiert. Fragestellungen und Methoden, die auf die zumindest partielle Isolierung dieser drei unabhängigen Variablen abzielen, sind damit naheliegend und z. B. Gegenstand eines großen Teiles der entwicklungspsychologischen Grundlagenforschung.

Wir betrachten im Folgenden einfaktorielle Pläne, bei denen jeweils eine der drei angesprochenen veränderungsrelevanten unabhängigen Variablen – Alter, Kohorte, Epoche – variiert wird, sowie zweifaktorielle Pläne, bei denen zwei dieser drei unabhängigen Variablen systematisch variiert werden. Die Pläne können zu höherfaktoriellen Plänen erweitert werden, indem **zusätzliche unabhängige Variablen** berücksichtigt werden, die persönliche Eigenschaften oder Interventionen widerspiegeln (z. B. Längsschnittstudien, in denen die Entwicklungsverläufe von Menschen mit/ohne chronischer Erkrankung, mit/ohne Missbrauchserfahrung in der Kindheit, mit/ohne Scheidung der Eltern, mit/ohne exzessiver Computerspielnutzung in der Kindheit etc. untersucht werden). Weiterhin lassen sich die betrachteten Pläne durch die Berücksichtigung von **Kontrollvariablen** ausbauen, was dann anstelle der üblichen varianzanalytischen eine kovarianzanalytische Auswertung erfordert. Sofern Alters-, Epochen- und/oder Kohorteneffekte für sich genommen untersucht werden, handelt es sich um nicht-experimentelle Studien, wenn zusätzlich auch Interventionen gesetzt und mit Kontrollgruppen verglichen werden, handelt es sich um (quasi-)experimentelle Studien (je nachdem, ob die Gruppenzuordnung randomisiert erfolgt oder nicht).

Methodische Probleme bei einfaktoriellen Plänen

Für die Untersuchung der Bedeutung der drei unabhängigen Variablen Alter (A), Generation (G) und Epoche (E) sowie deren Kombinationen für eine abhängige Variable wäre zweifellos ein vollständiger dreifaktorieller Untersuchungsplan mit den Faktoren A, G und E ideal. Dieser Plan ist jedoch leider nicht realisierbar, denn die hierfür erforderliche vollständige Kombination der Stufen aller drei Faktoren ist aufgrund der Konfundierung der Variablen nicht möglich (z. B. gehören 20-Jährige und 40-Jährige zu einem bestimmten Zeitpunkt zwei verschiedenen Generationen bzw. Kohorten an, d. h., die Stufen des Generationsfaktors lassen sich für einen bestimmten Zeitpunkt nicht mit verschiedenen Altersstufen kombinieren).

Es soll deshalb überprüft werden, welche Möglichkeiten bestehen, die Wirkung der drei unabhängigen Variablen (A, G und E) einzeln zu überprüfen (Haupteffekte ohne Interaktionen). Eine Untersuchungsvariante, die dies zumindest theoretisch gestattet, besteht darin, einen Faktor systematisch zu variieren und die beiden übrigen konstant zu halten (einfaktorieller Untersuchungsplan analysierbar mit **einfaktorieller Varianzanalyse mit oder ohne Messwiederholung**). Veränderungen der abhängigen Variablen wären dann auf den variierten Faktor zurückzuführen. Die folgenden Ausführungen prüfen, ob sich dieser Ansatz mit den unabhängigen Variablen A, G und E realisieren lässt.

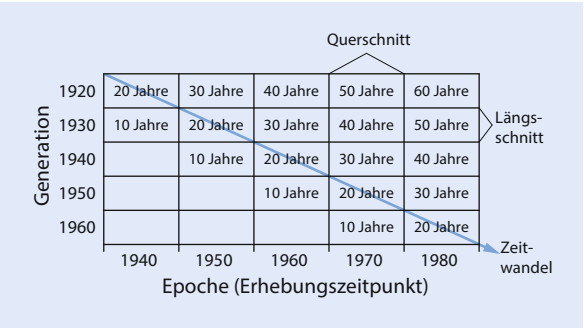


Abb. 12.56 Querschnittstudie, Längsschnittstudie und Zeitwandelstudie

Alterseffekte Um Alterseffekte zu isolieren, müssen die unabhängigen Variablen Generation und Epoche konstant gehalten werden. Dies ist jedoch nicht möglich. Entweder man untersucht Menschen verschiedenen Alters zu einem bestimmten Zeitpunkt (E konstant). Diese gehören dann jedoch verschiedenen Generationen an, d. h., der Faktor G kann nicht konstant gehalten werden (Untersuchungstyp 1). Oder man verfolgt Menschen einer Generation über mehrere Altersstufen hinweg (G konstant). Dies jedoch bedeutet, dass die Untersuchungen zu verschiedenen Zeitpunkten stattfinden, d. h., der Faktor E kann nicht konstant gehalten werden (Untersuchungstyp 2).

Der erste Untersuchungstyp entspricht der klassischen **Querschnittuntersuchung** („cross-sectional study“, Abb. 12.56). Sie vergleicht zu einem Zeitpunkt Stichproben verschiedenen Alters (d. h. gleichzeitig Personen aus unterschiedlichen Geburtskohorten bzw. Generationen). Um nun Unterschiede zwischen den Altersgruppen auf die unabhängige Variable „Alter“ zurückführen zu können, darf es keine Kohorteneffekte geben. Andernfalls könnte die Querschnittuntersuchung auch zum Nachweis von Kohorteneffekten eingesetzt werden, was dann allerdings voraussetzen würde, dass Alterseffekte zu vernachlässigen sind. Kurz: Bei **Querschnittuntersuchungen sind Alters- und Kohorteneffekte konfundiert**.

Eine Querschnittstudie, in der verschiedene Altersgruppen miteinander verglichen werden, wird mit einer **einfaktoriellen Varianzanalyse ohne Messwiederholung** (UV: Alter; AV: interessierende abhängige Variable) ausgewertet. Interessieren mehrere abhängige Variablen, kann eine **einfaktorielle multivariate Varianzanalyse** zum Einsatz kommen. Soll geprüft werden, inwiefern der Effekt des Alters auf die interessierende(n) AV durch weitere Variablen (Kontrollvariablen) beeinflusst wird, können diese als zusätzlicher Faktor (z. B. UV: Geschlecht) einbezogen oder – sofern es sich um kardinalskalierte Variablen handelt – im Zuge einer **Kovarianzanalyse** aus den Werten der abhängige Variablen herauspartialisiert werden.

Die zweite Untersuchungsart, die häufig zur Überprüfung von Alterseffekten eingesetzt wird, heißt **Längsschnittuntersuchung** oder **Longitudinalstudie** („longitudinal study“; Abb. 12.56). Hier wird die Variation des Alters dadurch erreicht, dass man eine Kohortenstichprobe zu verschiedenen Zeitpunkten (d. h. mit unterschiedlichem Alter) untersucht. Bei Längsschnittuntersuchungen wird eine Stichprobe wiederholt untersucht, d. h., hier sind Auswertungsmodelle, die die Abhängigkeit der Messungen berücksichtigen, einschlägig, etwa **Varianzanalysen mit Messwiederholungen**. Über die Auswertung nominaler Daten im Rahmen von Längsschnittuntersuchungen berichtet Plewis (1981).

Die Analyse der Veränderungen einer Stichprobe aus einer Generation bzw. Kohorte führt jedoch nur dann zu brauchbaren Angaben über den Alterseinfluss, wenn epochale Effekte zu vernachlässigen sind. Umgekehrt kann die Längsschnittuntersuchung unter der Annahme, Altersseffekte seien zu vernachlässigen, zur Überprüfung der unabhängigen Variablen „Epoche“ herangezogen werden. Da man in einer konkreten Untersuchung weder epochale noch altersbedingte Effekte völlig ausschließen kann, muss man damit rechnen, dass **in Längsschnittuntersuchungen Alters- und Epocheneffekte konfundiert** sind.

Wegen ihrer Bedeutung für die entwicklungspsychologische Forschung seien im Folgenden weitere Schwächen der Querschnittanalyse und der Längsschnittanalyse aufgezeigt (ausführlicher hierzu Hoppe et al., 1977).

- **Probleme bei Querschnittuntersuchungen:**
  - **Selektive Populationsveränderung:** Mit fortschreitendem Alter verändern sich die Stichproben systematisch in Bezug auf einige Merkmale. Nehmen wir an, wir wollen das menschliche Körpergewicht in Abhängigkeit vom Alter untersuchen. Dabei müsste man davon ausgehen, dass die Wahrscheinlichkeit, an Übergewicht zu sterben, nicht konstant ist, sondern mit zunehmendem

Alter steigt. In der Population alter Menschen wären dann prozentual weniger Übergewichtige anzutreffen als in der Population jüngerer Menschen. Hieraus zu folgern, der Mensch verliert im Verlauf seines Lebens an Gewicht, wäre sicherlich falsch.

- **Vergleichbarkeit der Messinstrumente**: Die Validität eines Messinstrumentes kann vom Alter der untersuchten Personen abhängen. Testaufgaben, die bei jüngeren Menschen kreative Denkleistungen erfordern, können von älteren Menschen durch Erfahrung und Routine gelöst werden (Eckensberger, 1973; Gulliksen, 1968; Vagt, 1977)

- **Probleme bei Längsschnittuntersuchungen:**
  - **Ausfälle von Untersuchungseinheiten**: Wird eine Stichprobe über einen langen Untersuchungszeitraum hinweg beobachtet, muss man damit rechnen, dass sich die Stichprobe durch Ausfall von Untersuchungsteilnehmenden im Verlauf der Zeit systematisch verändert (Drop-outs).
  - **Vergleichbarkeit der Messinstrumente**: Dieser schon auf die Querschnittuntersuchung bezogene Kritikpunkt trifft auch auf Längsschnittuntersuchungen zu. Mit zunehmendem Alter kann sich die Bedeutung eines Messinstrumentes verändern.
  - **Generationsspezifische Aussagen**: Die Resultate einer Längsschnittuntersuchung gelten nur für die untersuchte Generation/Kohorte und sind auf andere Kohorten nicht ohne weiteres übertragbar.
  - **Testübung**: Die häufige Untersuchung einer Stichprobe birgt die Gefahr, dass die Ergebnisse durch Erinnerungs-, Übungs- oder Gewöhnungseffekte verfälscht sind.
  - **Untersuchungsaufwand**: Längsschnittuntersuchungen erfordern einen erheblichen Zeitaufwand bzw. langfristiges Engagement (von Forschenden wie Untersuchungsteilnehmenden)

Kohorteneffekte Der erste Plan, um Kohorteneffekte zu prüfen, variiert die Generationen und hält das Alter konstant. Hierbei muss zwangsläufig auch eine Veränderung der Epochen, in denen untersucht wird, in Kauf genommen werden. Bezogen auf das Schema in ◘ Abb. 12.56 werden z. B. 10-Jährige des Jahrganges 1930 im Jahre 1940 untersucht, 10-Jährige des Jahrganges 1940 im Jahre 1950 etc. Baltes (1967) bezeichnet dieses Vorgehen als **Zeitwandelmethode**. Bei dieser querschnittlichen Untersuchungsvariante sind die Generation und die Epoche konfundiert.

Der zweite Plan variiert die Generationen und hält die Epoche (Erhebungszeitpunkt) konstant, d. h., er vergleicht z. B. im Jahre 1980 Personen der Jahrgänge 1930, 1940, 1950 etc. Diese Untersuchung ist nur möglich, wenn man auch eine Variation des Alters zulässt. Damit entspricht dieser Untersuchungstyp der bereits behandelten **Querschnittuntersuchung**. Beide Untersuchungsvarianten werden üblicherweise mit **einfaktoriellen Varianzanalysen ohne Messwiederholungen** ausgewertet.

Epochale Effekte Der erste Plan zur Prüfung von Epochen-Effekten variiert die Epoche und hält das Alter konstant, d. h., er untersucht z. B. die 10-Jährigen im Jahr 1940, die 10-Jährigen im Jahr 1950 etc.; damit variieren gleichzeitig auch die Generationen. Dieser Plan entspricht also dem ersten Plan zur Überprüfung von Generationseffekten, der als **Zeitwandelmethode** bezeichnet wurde. Der zweite Plan variiert die Epochen und hält die Generationen konstant. Damit muss zwangsläufig auch das Alter variiert werden, so dass die bereits behandelte **Längsschnittuntersuchung** resultiert.

Zusammenfassend führt also keiner der 6 Pläne (unter denen sich nur drei tatsächlich verschiedene Pläne befinden: Querschnitt-, Längsschnitt- und Zeitwandelstudie) zu eindeutigen Resultaten. Die Problematik nichtexperimenteller Pläne, dass die vorgefundene (d. h. nicht experimentell variierbare) unabhängige Variable von anderen Variablen überlagert ist, die die abhängige Variable möglicherweise ebenfalls beeinflussen, zeigt sich hier besonders drastisch. Es ist untersuchungstechnisch unmöglich, die Bedeutung einer der drei unabhängigen Variablen Alter, Generation/Kohorte und Epoche isoliert zu erfassen.

Mit den drei „klassischen" entwicklungspsychologischen Untersuchungsansätzen – *Querschnitt, Längsschnitt und Zeitwandel* – ist es nicht möglich, Effekte des Alters, der Generation/Kohorte und der Epoche isoliert zu erfassen, diese drei Variablen sind untrennbar miteinander verknüpft (konfundiert) Auch eine statistische Isolierung der Effekte ist nicht möglich, vielmehr müssen die statistischen Ergebnisse einfaktorieller varianzanalytischer Pläne (z. B. signifikanter Haupteffekt für Alter) hinsichtlich des möglichen Einflusses der anderen beiden Variablen (z. B. Generation und Epoche) relativiert werden.

Methodische Probleme bei zweifaktoriellen Plänen

Wenn man in einer entwicklungspsychologischen Untersuchung nicht nur eine, sondern zwei der einschlägigen unabhängigen Variablen Alter, Generation und Epoche systematisch variiert, resultieren zweifaktorielle Pläne (sequenzielle Untersuchungspläne nach Schaie, 1977, 1994). Es sind dann drei verschiedene Untersuchungstypen denkbar, für die ◘ Abb. 12.57 jeweils ein Beispiel gibt.

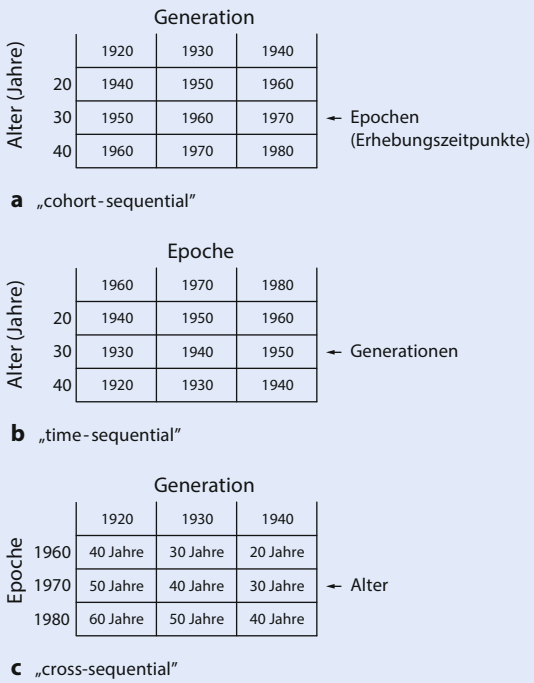


Abb. 12.57 Sequenzmodelle. **a** Längsschnitt- und Zeitwandeluntersuchungen, **b** Querschnitt- und Zeitwandeluntersuchungen, **c** Längs- und Querschnittuntersuchungen

In diesen – untersuchungspraktisch sehr aufwändigen – Plänen sind Replikationen von Längsschnitt- und Zeitwandeluntersuchungen (Abb. 12.57a), von Querschnitt- und Zeitwandeluntersuchungen (Abb. 12.57b) und von Längs- und Querschnittuntersuchungen (Abb. 12.57c) kombiniert. Im Einzelnen ergeben sich die im Folgenden aufgeführten Untersuchungspläne:

Cohort-Sequential-Design Betrachten wir zunächst eine Untersuchung, in der die **Generation** und das **Alter** der Untersuchungsteilnehmer systematisch variiert werden (Abb. 12.57a). Da jede Generationsstichprobe (1920, 1930, 1940) wiederholt (im Alter von 20, 30 und 40 Jahren) untersucht wird, handelt es sich um die Kombination von 3 Längsschnittstudien. Dies entspricht der Kombination von 3 Zeitwandelstudien: Untersucht werden 20-Jährige aus den Generationen 1920, 1930 und 1940, 30-Jährige aus diesen Generationen und auch 40-Jährige. Damit ist der Haupteffekt „Generationen" in Bezug auf das Alter und der Haupteffekt „Alter" in Bezug auf die Generationen ausbalanciert. Beide Effekte sind jedoch mit epochalen Effekten konfundiert, so dass die Haupteffekte „Generationen" und „Alter" nur bei zu vernachlässigenden epochalen Effekten isoliert interpretierbar sind.

Der Vorteil zweifaktorieller Pläne gegenüber einfaktoriellen Plänen besteht im Allgemeinen darin, dass neben Haupteffekten auch **Interaktionen** geprüft werden können. Was aber – so wollen wir fragen – bedeutet der Interaktionseffekt im Kontext sequenzieller Pläne?

Nehmen wir einmal an, als abhängige Variable wird das Konstrukt „Selbstwert" untersucht. Im „Cohort-Sequential-Ansatz" könnte ein **signifikanter Interaktionseffekt in der zweifaktoriellen Varianzanalyse** bedeuten, dass sich die Selbstwert-Scores mit zunehmendem Alter (von 20–40 Jahren) nur unbedeutend verändern, wenn Personen aus der Generation 1920 untersucht werden, dass aber deutliche Selbstwertänderungen (z. B. höheres Selbstwertgefühl mit 40 Jahren als mit 20 Jahren) bei Personen aus der Generation 1930 registriert werden. Diesen Befund als reinen Interaktionseffekt zu interpretieren, wäre insoweit problematisch, als die Untersuchungszeiträume (Epochen) für die hier verglichenen Generationen divergieren: Die Generation 1920 wird in den Jahren 1940–1960 untersucht und die Generation 1930 in den Jahren 1950–1970. Eine Interpretation der Interaktion Generation × Alter wäre also nur zulässig, wenn die Untersuchungszeiträume bzw. epochalen Effekte ohne Bedeutung sind.

Daten aus Kohorten-Sequenz-Längsschnittstudien sind mit mindestens **zweifaktoriellen Varianzanalysen mit Messwiederholungen auf dem Altersfaktor** auszuwerten (UV1: Generation, UV2: Alter, UV3, UV4. . .: ggf. weitere interessierende unabhängige Variablen wie z. B. Geschlecht). Zudem ist eine Datenanalyse über Strukturgleichungsmodelle (▶ Kap. 17) – speziell **latente Wachstumsmodelle** („latent growth models") möglich (für eine entsprechende Studie zu Veränderungen von Familienkonflikten mit zunehmendem Alter jugendlicher Kinder s. z. B. Baer & Schmitz, 2000; zur Veränderung des Selbstwertgefühls mit dem Lebensalter s. z. B. Orth, Trzesniewski, & Robins, 2010).

Time-Sequential-Design Werden **Epochen** und **Alter** systematisch variiert (Abb. 12.57b), resultiert ein Plan mit mehreren Querschnittuntersuchungen (20-, 30- und 40-Jährige werden 1960, 1970 und 1980 untersucht) bzw. mehrere Zeitwandelstudien (20-, 30- und 40-Jährige werden jeweils 1960, 1970 und 1980 untersucht). Wie man Abb. 12.57b entnehmen kann, sind die Haupteffekte „Epoche" und „Alter" mit Generationseffekten konfundiert. Dies gilt auch für die Interaktion Epoche × Alter, die nur interpretiert werden kann, wenn Generationseffekte zu vernachlässigen sind. Die Auswertung erfolgt mit mindestens **zweifaktorieller Varianzanalyse ohne Messwiederholung** (UV1: Alter, UV2: Messzeitpunkte).

Beispiel: Eine Studie untersuchte Veränderungen der biografischen Zeitperspektive (d. h. Bewertung von Ver-

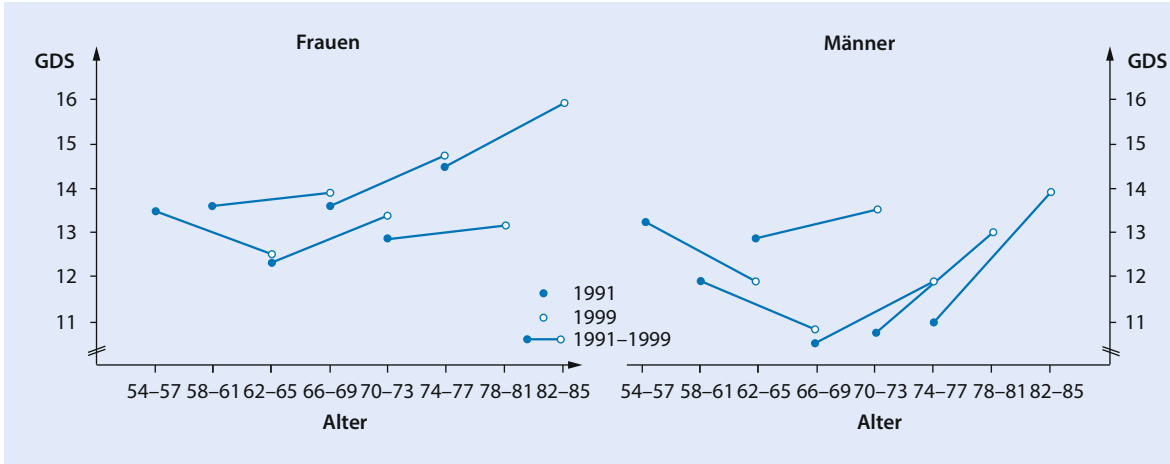


Abb. 12.58 Cross-sequenzielle Muster der Mittelwerte der „geriatric depression scale" *(GDS)*. (Aus Rothermund, K. & Brandstädter, J. (2003). Depression in later life: Cross-sequential patterns and possible determinants. *Psychology and Aging, 18*(1), 80–90. Copyright © 2013 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

gangenheit und Zukunft) über die Lebensspanne: Dazu wurden in einem US-Bundesstaat zwei bevölkerungsrepräsentative Querschnittstudien (Datenerhebung 1972 und 1973: zwei „Epochen") durchgeführt, wobei die Befragten jeweils 6 Altersgruppen entstammten (Hultsch & Bortner, 1974). Die Daten wurden mit einer dreifaktoriellen **2 × 2 × 6-Varianzanalyse** (Messzeitpunkt, Geschlecht, Alter) ausgewertet. Es zeigte sich, dass nur der Alterseffekt signifikant wurde, nicht jedoch der Messzeitpunkt allein oder in Interaktion mit den anderen UV. Die Autoren schlussfolgern, dass zwischen 1972 und 1973 dementsprechend keine kulturellen bzw. „epochalen" Änderungen der biografischen Zeitbewertung auftraten, sondern primär der chronologische Alterseffekt wirkte: Dementsprechend bewerten Jüngere ihre Vergangenheit in Relation zur Gegenwart eher negativ und die Zukunft positiv. Dieser Zukunftsoptimismus schwächt sich mit zunehmendem Alter ab, bis am Ende von Hochbetagten die Vergangenheit positiver als die Zukunft bewertet wird.

Cross-Sequential-Design Der dritte Plan variiert den **Epochen-** und den **Generationsfaktor** (Abb. 12.57c), was zu replizierten Längsschnittuntersuchungen (jede der drei Generationen 1920, 1930 und 1940 wird wiederholt in den Jahren 1960, 1970 und 1980 untersucht) und replizierten Querschnittuntersuchungen führt (in den Jahren 1960, 1970 und 1980 werden jeweils die Generationen 1920, 1930 und 1940 untersucht). Eindeutige Interpretationen der Haupteffekte „Generation" und „Epoche" bzw. der Interaktion „Generation × Epoche" setzen hier voraus, dass Alterseffekte zu vernachlässigen sind. Die statistische Hypothesenprüfung von Haupt- und Interaktionseffekten erfolgt mit einer mindestens **zweifaktoriellen Varianzanalyse mit Messwiederholungen** auf dem Faktor Epoche bzw. Messzeitpunkt.

Beispiel: Eine Studie untersuchte Depression im höheren Lebensalter (Rothermund & Brandstädter, 2003). Dazu wurden 6 Kohorten jeweils als Längsschnitt 2-mal (1991 und 1999: Epoche) untersucht. Als dritte unabhängige Variable wurde das Geschlecht berücksichtigt (Abb. 12.58). Die statistische Hypothesenprüfung erfolgte über eine **dreifaktorielle multivariate Varianzanalyse mit Messwiederholung auf dem Faktor Messzeitpunkt**. Es zeigte sich ein signifikanter Haupteffekt für Geschlecht (Frauen waren stärker von Depression betroffen als Männer) sowie ein signifikanter Interaktionseffekt für Messzeitpunkt und Kohorte (in den acht Jahren zwischen erster und zweiter Messung war die Depression v. a. in den älteren Generationen gestiegen, in den jüngeren Generationen dagegen nicht).

Resümee Ein Vergleich dieser Pläne mit einem lateinischen Quadrat (▶ Abschn. 12.5.3) zeigt ihre Eigenschaften auf einer formaleren Basis. (Dieser Vergleich bietet sich an, weil für die Beispiele in Abb. 12.57 quadratische Pläne mit gleicher Stufenzahl der jeweils variierten Faktoren ausgewählt wurden, was natürlich nicht zwingend ist. Die Argumente gelten jedoch analog für nicht-quadratische Sequenzpläne.) Lateinische Quadrate sind vollständig in Bezug auf die drei Haupteffekte, aber nur partiell hinsichtlich der Interaktionen ausbalanciert. Hieraus wurde gefolgert, dass die Haupteffekte nur interpretierbar sind, wenn man die Interaktionen vernachlässigen kann. Da man jedoch meistens nicht weiß, ob bzw. welche Faktoren miteinander interagieren, sind Ergebnisse, die man mit einem lateinischen Quadrat findet, nur bedingt verwertbar.

Die Pläne in ◘ Abb. 12.57 müssen nicht nur auf eine Ausbalancierung in Bezug auf die Interaktion verzichten (diese läge vor, wenn jede Faktorstufe eines Faktors mit allen Faktorstufenkombinationen der beiden übrigen Faktoren aufträte, wenn also z. B. 20-Jährige aus allen Generationen zu allen Erhebungszeitpunkten untersuchbar wären), sondern zusätzlich auf eine Ausbalancierung in Bezug auf die Haupteffekte (bei der Kombination zweier Faktoren kann der dritte Faktor nicht konstant gehalten werden bzw. den Kombinationen zweier Faktoren werden Stichproben aus unterschiedlichen Populationen zugeordnet). Damit sind die beiden jeweils als Haupteffekte variierten unabhängigen Variablen nur interpretierbar, wenn die dritte unabhängige Variable, deren Interaktion mit den beiden Haupteffekten sowie die Interaktion zweiter Ordnung (Triple-Interaktion) zu vernachlässigen sind. Die interne Validität dieser Pläne liegt also unter der eines lateinischen Quadrates.

Eine formalstatistische Analyse der Sequenzmodelle, die ebenfalls darauf hinausläuft, dass Alters-, Generations- und epochale Effekte nicht voneinander unabhängig bestimmbar sind, findet man bei Adam (1978), und weitere Hinweise bei Erdfelder, Rietz, & Rudinger (1996). Bei Schaie (1994) ist nachzulesen, wie man empirisch überprüfen kann, ob ein bestimmter Effekt – Alter, Generation oder Epoche – unwahrscheinlich ist, so dass die beiden übrigen Effekte bzw. ihre Interaktion interpretierbar werden. Wie man durch Kombination mehrerer aufeinander folgender Generationen und relativ kurze Untersuchungszeiträume Interaktionen zwischen Alter und Generation prüfen kann, um so die interne Validität des CrossSequentialDesigns zu erhöhen, wird bei Miyazaki und Raudenbusch (2000) gezeigt.

Zweifaktorielle entwicklungspsychologische Pläne (z. B. Alter und Generation) sind eindeutig zu interpretieren, wenn der jeweils dritte Faktor (im Beispiel: die Untersuchungsepoche bzw. Untersuchungszeitpunkt) keinen Einfluss auf die abhängige Variable ausübt.

Wir haben es bei Alter, Generation und Epoche mit drei zentralen, veränderungsrelevanten unabhängigen Variablen zu tun, deren wechselseitige Konfundierung untersuchungstechnisch und rechnerisch-statistisch nicht zu beseitigen ist. Hypothesen, die sich z. B. auf epochale Effekte beziehen, können immer nur für bestimmte Kombinationen von Altersgruppen und Generationen überprüft werden. Man kann z. B. fragen, ob die 20-Jährigen des Jahres 1990 politisch aktiver waren als die 20-Jährigen des Jahres 1950. Eine Antwort auf diese Frage muss jedoch immer in Rechnung stellen, dass mögliche Unterschiede nicht nur epochal, sondern auch kohortenbedingt sein können (weitere Überlegungen zu dieser Thematik findet man bei Mayer & Huinink, 1990).

Veränderungshypothesen für Zeitreihen

Ausprägungen einer Variablen, die in gleichen Zeitabständen wiederholt gemessen werden, bilden eine **Zeitreihe** („time series“). Für eine Zeitreihe ist es unerheblich, ob die wiederholten Messungen einer Variable von einer **einzelnen Person** stammen (z. B. tägliche Trainingsergebnisse einer Spitzensportlerin; ▶ Abschn. 12.5.5 zu hypothesenprüfenden Einzelfalluntersuchungen) oder ob jeweils Messwerte einer **ganzen Stichprobe** vorliegen, so dass pro Messzeitpunkt ein Durchschnittswert zu betrachten ist (z. B. bei einer Schülerstichprobe wird mit wöchentlichen Vokabeltests die durchschnittliche Anzahl der gelernten Vokabeln der unterrichteten Fremdsprache ermittelt). Auch wenn regelmäßig anhand **unterschiedlicher Stichproben** derselbe Index ermittelt wird, bildet dies eine Zeitreihe (z. B. jährlich berechnete Lebenserwartung von Frauen und Männern in Deutschland über die vergangenen 50 Jahre hinweg). Formal unterscheiden wir im Kontext von Zeitreihenanalysen drei verschiedene Hypothesenarten:

- **Vorhersagemodelle**: Die in einer Zeitreihe entdeckten Regelmäßigkeiten setzen sich laut Veränderungshypothese auch zukünftig fort.
- **Interventionsmodelle**: Ein „Treatment“ oder „Ereignis“ verändert laut Hypothese eine Zeitreihe in einer bestimmten Weise.
- **Transferfunktionsmodelle**: Die Werte einer Zeitreihe werden laut Hypothese als Funktion der beobachteten Veränderungen einer oder mehrerer anderer Zeitreihen modelliert, wobei diese Wirkungen auch verzögert und über die Zeit verteilt auftreten können.

Zeitreihe – Eine Zeitreihe („time series“) ist ein Datensatz, der dadurch entsteht, dass dieselbe Variable bei derselben Person bzw. Untersuchungseinheit (Einzelfalluntersuchung) oder bei denselben Personen bzw. Untersuchungseinheiten (Stichprobenuntersuchung oder Vollerhebung) immer wieder im gleichen Abstand erhoben wird.

In diesem Abschnitt beschäftigen wir uns mit **langen Zeitreihen**, die aus mindestens $n = 50$ Messpunkten bestehen. Von besonderer Bedeutung für die Analyse langer Zeitreihen ist ein von den britischen Statistikern George Box und Gwilym Jenkins entwickeltes Verfahren, das unter der Bezeichnung **Box-Jenkins-Methode** (Box & Jenkins, 1976) bekannt wurde. Die Box-Jenkins-Methode spielt in der Ökonometrie eine wichtige Rolle und kommt darüber hinaus in den Sozial- und Humanwissenschaften vielfältig zur Anwendung (Deutsch & Alt, 1977; Glass, Tiao & Maguire, 1971; Gudat & Revenstorff, 1976; Hennigan et al., 1979; Hibbs, 1977; Kette, 1990; Meier, 1988; Metzler & Nickel, 1986; Pawlik & Buse, 1994).

Es kann nicht Aufgabe dieses Textes sein, die mathematisch aufwändige Box-Jenkins-Methode im Detail darzustellen. Einführungen findet man z. B. bei Glass, Willson, & Gottman (1975); Gudat und Revenstorff (1976); Hamilton (1994); McDowall et al. (1980); McCain und McCleary (1979); Nelson (1973); Rottleuthner-Lutter (1985); Schlittgen und Streitberg (1994); Shumway und Stoffer (2000); Schmitz (1989) sowie Thome (2005). Ziel dieses Abschnittes ist es, eine erste Orientierung über den Aufbau dieser Methode zu geben, den wichtigen Schritt der Modellidentifikation zu erläutern und Hinweise zur Überprüfung der eingangs erwähnten Typen von Veränderungshypothesen (Vorhersage-, Interventions-, Transferfunktionsmodelle) zu geben. Die Box-Jenkins-Methode ist Bestandteil gängiger Statistik-Programmpakete, (▶ Abschn. 12.2.3), zudem existieren spezielle Programme (z. B. Autobox: http://www.autobox.com; für Software-Hinweise s. Thome, 2005, S. 354ff.).

Verhaltensänderungen können in unterschiedlichen Zeitphasen unterschiedlich stark ausfallen. Fragen wie „Ist der Veränderungsprozess gleichförmig oder gibt es Phasen mit Veränderungssprüngen oder beschleunigter Veränderung?“ sind wichtig für viele entwicklungspsychologische oder lerntheoretische Themen. Mit der Analyse derartiger „Sprungstellen“ bzw. Zeitpunkte, zu denen deutliche Veränderungen der Verlaufscharakteristik einer Zeitreihe auftreten, befasst sich u. a. eine Arbeit von Cudeck und Klebe (2002).

Die Idee des Box-Jenkins-Verfahrens: Wenn schon die Mathematik des Box-Jenkins-Verfahrens hier nicht dargestellt werden kann, soll zumindest eine ungefähre Vorstellung davon vermittelt werden, wie eine Zeitreihenanalyse nach Box-Jenkins funktioniert.

Die Systematik einer Zeitreihe zu erfassen bzw. in einem mathematischen Modell abzubilden, ist die zentrale Aufgabe der Zeitreihenanalyse nach Box-Jenkins. Erst wenn man die Eigenheiten einer Zeitreihe erkannt und formal abgebildet hat, ist es möglich, genaue Vorhersagen zu machen oder Interventionswirkungen zu prüfen. Man spricht von **univariaten Modellen**, wenn per Zeitreihenanalyse zunächst nur die Veränderungen einer Variablen über die Messzeitpunkte hinweg betrachtet werden.

Ob ein gefundenes statistisches Modell geeignet ist, die Systematik der Zeitreihe optimal zu erfassen, erkennt man, wenn man die Ausprägungen des geprüften Merkmals für jeden Zeitpunkt der Zeitreihe auf der Basis der zeitlich vorangegangenen Messwerte vorhersagt. Ergeben sich hierbei **Residuen** (d. h. Abweichungen zwischen den vom Modell vorhergesagten und den tatsächlichen empirischen Messwerten), die zufällig verteilt sind (im Kontext von Box-Jenkins-Analysen spricht man von **„White Noise“** oder „Weißem Rauschen“), dann ist dies ein Indikator für ein richtig gewähltes Modell, mit dem man Vorhersagen machen und Interventionen prüfen kann.

Entscheidend für die Identifikation der univariaten Modelle ist die **Autokorrelationsfunktion**. Im Kontext der Zeitreihenanalyse versteht man unter Autokorrelationen die Korrelationen der Zeitreihe mit sich selbst, wenn diese um einen oder mehrere Messzeitpunkte verschoben ist.

Die Charakterisierung der Systematik univariater Abhängigkeitsbeziehungen hat im Rahmen einer Box-Jenkins-Analyse drei Modellbestandteile:

1. Angaben zur sog. Autokorrelationsstruktur (**AR-Parameter**; „autoregressive parameter“). Hierbei wird die Merkmalsausprägung zu einem Zeitpunkt t_i als Funktion zeitlich vorangegangener Messwerte angesehen.
2. Angaben zu Moving-Average-Prozessen (Gleitmittelprozesse; **MA-Parameter**, „ moving average parameter“). Hierbei wird die Merkmalsausprägung zu einem Zeitpunkt t_i als Funktion von vorangehenden Zufallskomponenten („random shocks“) angesehen. AR- und MA-Prozesse können jeweils alleine oder gemeinsam (ARMA-Prozess) auftreten und werden jeweils mittels der Autokorrelation empirisch identifiziert.
3. Bestimmung von Trends in der Zeitreihe (**I-Parameter**, „integrated [difference] parameter“). Hierbei werden die für die Zeitreihe charakteristischen stochastischen Trends analysiert.

Mit diesen drei Modellparametern (die wir weiter unten noch präzisieren werden) kann die Systematik einer Zeitreihe vollständig erfasst werden (wobei hier Zyklen in Form saisonaler Schwankungen unberücksichtigt bleiben: sog. SARIMA-Modelle).

Es resultiert ein **ARIMA-Modell** („auto regressive integrated moving average model“), das die Ergebnisse der Zeitreihenanalyse nach Box-Jenkins zusammenfasst. Die Bestimmung eines ARIMA-Modells erfolgt in mehreren Schritten, die in ◘ Abb. 12.59 für univariate Modelle schematisch dargestellt sind (die Bestimmung von Transferfunktionsmodellen ist komplexer). Bei der Box-Jenkins-Methode ist der **„White-Noise“-Charakter der Residuen** das wichtigste Kriterium der Bewertung eines ARIMA-Modells und damit der Prüfung von Veränderungshypothesen für Zeitreihen. Es ist aber auch eine **Signifikanzaussage** möglich, wobei ein statistisch signifikantes Ergebnis zur Annahme der Forschungshypothese veranlasst, die z. B. bei Transferfunktionsmodellen postuliert, dass eine Zeitreihe einen systematischen Einfluss auf eine andere Zeitreihe hat (für ein ▶ Studienbeispiel „Zeitreihenanalyse: Der Einfluss des Alkoholkonsums auf die Suizidrate in Russland“).

Die Identifikation des ARIMA-Modells, das der Zeitreihe vermutlich zugrunde liegt, erfordert die Bestimmung von drei Kennwerten: p charakterisiert den autoregressiven Anteil (AR-Parameter) der Zeitreihe, d

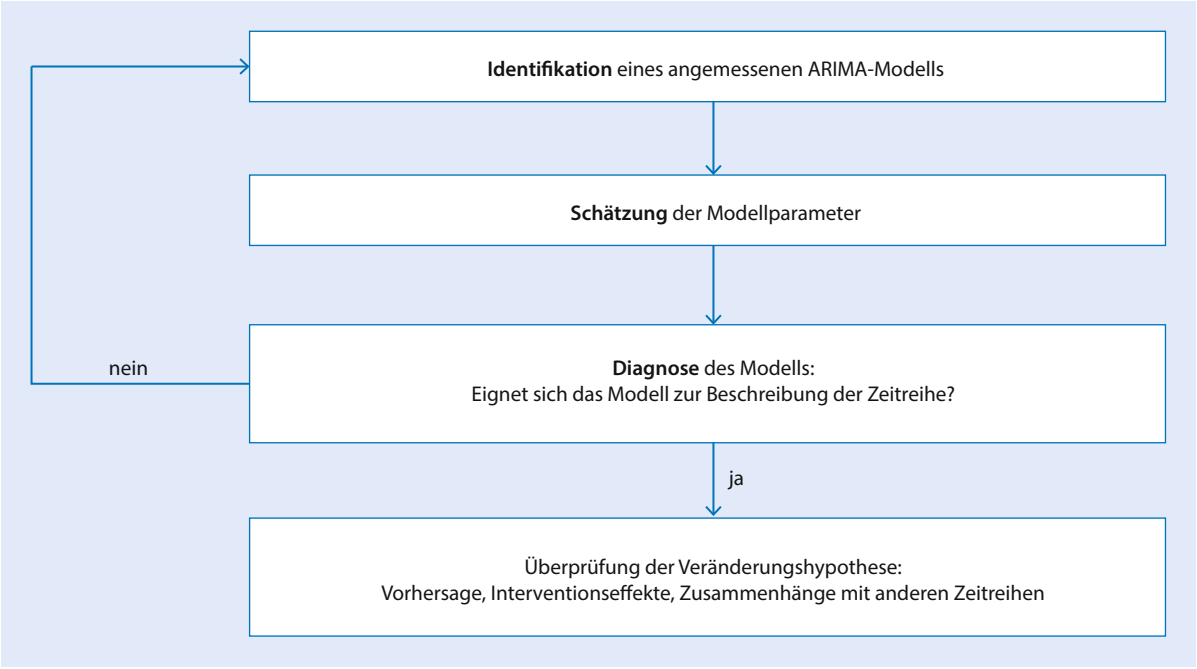


Abb. 12.59 Überprüfung von Veränderungshypothesen nach der Box-Jenkins-Methode

charakterisiert mögliche stochastische Trends (I-Parameter) in der Zeitreihe und q beschreibt den in einer Zeitreihe evtl. vorhandenen Gleitmittelprozess (MA-Parameter). Auf die Bedeutung dieser Kennwerte geht der folgende, an Rottleuthner-Lutter (1985) orientierte Abschnitt einführend ein.

Bedeutung von ARIMA(p, d, q)-Modellen

Im Folgenden werden die Parameter eines ARIMA-Modells (p, d und q) präzisiert. Wir beginnen mit dem d-Parameter.

d-Parameter Zur Illustration des d-Parameters nehmen wir an, eine Zeitreihe habe einen linearen Trend und bestehe aus den Messungen 1, 2, 3, 4, 5 . . . n. Bilden wir die Differenz zwischen einer Messung x_t und der zeitlich vorausgegangenen Messung x_{t-1}, resultiert eine **trendfreie bzw. stationäre Zeitreihe**:

$$\begin{aligned} 2-1&=1\\ 3-2&=1\\ 4-3&=1\\ 5-4&=1\\ &\vdots\\ n-(n-1)&=1 \end{aligned}$$

Allgemein ergibt sich bei einer Zeitreihe mit einem linearen Trend vom Typ $x_t = a + b \times t$ ($t = 1, 2, \ldots n$; $n =$ Anzahl der Zeitpunkte) durch **sukzessive Differenzenbildung** eine Konstante, die dem Steigungsparameter (b) der Regression entspricht (Thome, 2005, Abschn. 2.3.9). Eine Zeitreihe, die durch (einmalige) Differenzenbildung trendfrei bzw. stationär wird, kennzeichnet man mit dem Parameter $\boldsymbol{d = 1}$ allgemein als **ARIMA ($\boldsymbol{p}$, 1, $\boldsymbol{q}$)**. Für das hier gewählte Beispiel haben p und q jeweils den Wert Null, es resultiert für eine Zeitreihe mit ausschließlich linearem Trend ARIMA (0,1,0), dargestellt auch in Abb. 12.60.

Gelegentlich reicht – wie das folgende Beispiel zeigt – eine einfache Differenzenbildung nicht aus, um eine Trendbereinigung zu erzielen. Für die Zeitreihe 1, 4, 9, 16, 25 . . . resultiert folgende Differenzenbildung:

$$\begin{aligned} 4-1&=3\\ 9-4&=5\\ 16-9&=7\\ 25-16&=9\\ &\vdots\\ n^2-(n-1)^2& \end{aligned}$$

Die Differenzierung führt zu einer Zeitreihe mit einem (linearen) Trend. Erst eine zweite Differenzenbildung macht

Tabelle 12.24 Datenschema für die Berechnung von Autokorrelationen einer Zeitreihe

r_1	r_2	r_3	r_4	...
$x_1\ x_2$	$x_1\ x_3$	$x_1\ x_4$	$x_1\ x_5$	
$x_2\ x_3$	$x_2\ x_4$	$x_2\ x_5$	$x_2\ x_6$	
$x_3\ x_4$	$x_3\ x_5$	$x_3\ x_6$	$x_3\ x_7$	
⋮	⋮	⋮	⋮	
$x_{n-1}\ x_n$	$x_{n-2}\ x_n$	$x_{n-3}\ x_n$	$x_{n-4}\ x_n$	

r_1 Autokorrelation 1. Ordnung, r_2 Autokorrelation 2. Ordnung, r_3 Autokorrelation 3. Ordnung etc.

diese Reihe trendfrei:

$$5-3=2$$
$$7-5=2$$
$$9-7=2$$
$$11-9=2$$
$$\vdots$$

Allgemein ergibt sich durch die erste Differenzenbildung aus einem Polynom zweiten Grades ($x_t = a+b_1 \cdot t+b_2 \cdot t^2$) ein Polynom ersten Grades und durch die zweite Differenzenbildung ein Polynom nullten Grades mit einer Konstanten, die dem doppelten Betrag des Steigungskoeffizienten b_2 der Regression entspricht.

Für Zeitreihen, die erst nach zweimaliger Differenzierung trendfrei werden, ist das Modell mit dem Parameter $\boldsymbol{d = 2}$ charakteristisch: **ARIMA ($\boldsymbol{p}$, 2, $\boldsymbol{q}$)**. Für das hier gewählte Beispiel haben p und q jeweils den Wert Null: ARIMA (0, 2, 0).

***p*-Parameter** Korrelationen, die man durch zeitliche Versetzungen der Messwerte einer Zeitreihe errechnet, heißen **Autokorrelationen**. Je nachdem, um wie viele Zeitintervalle **(Lags)** die Zeitreihen versetzt sind, unterscheidet man Autokorrelationen 1. Ordnung (1 Lag), Autokorrelationen 2. Ordnung (2 Lags), 3. Ordnung (3 Lags) etc. (Tab. 12.24).

Für die Identifizierung des ARIMA-Modells einer Zeitreihe ist es wichtig, ihren **autoregressiven Anteil** bzw. ihre Autokorrelationsstruktur zu kennen. Besteht nur eine Abhängigkeit zwischen benachbarten Messungen, gilt das folgende Regressionsmodell:

$$x_t = \varphi_1 \cdot x_{t-1} + a. \qquad (12.40)$$

$\boldsymbol{\varphi_1}$ (griech. phi) beschreibt die **Enge des Zusammenhanges zwischen direkt benachbarten (1 Lag) Beobachtungen**. Die Ausprägung einer Messung zum Zeitpunkt t hängt nur von der unmittelbar vorangehenden Messung x_{t-1} sowie einer Zufallskomponente a ab. (Diese um Null normalverteilte Zufallskomponente a heißt im Rahmen der Box-Jenkins-Modelle „**Random-Shock**".) Eine Zeitreihe mit diesen Eigenschaften wird mit dem Parameter $p = 1$ durch ein **ARIMA(1, *d*, 0)–Modell** beschrieben (warum hier $q = 0$ gilt, wird unten erläutert).

Besteht nicht nur zwischen benachbarten, sondern auch zwischen Messungen mit 2 Lags eine Abhängigkeit, lautet das autoregressive Modell:

$$x_t = \varphi_1 \cdot x_{t-1} + \varphi_2 \cdot x_{t-2} + a. \qquad (12.41)$$

Die Tatsache, dass zwei autoregressive Komponenten substanziell sind ($p = 2$) und gleichzeitig keine Gleitmittelkomponenten ($q = 0$, s. unten), wird durch ARIMA (2, d, 0) zum Ausdruck gebracht. ARIMA-Modelle mit mehr als zwei autoregressiven Komponenten ($p > 2$) kommen in der Praxis selten vor.

***q*-Parameter** Die zu einem Zeitpunkt t erhobene Messung x_t kann nicht nur von den vorangegangenen Messungen x_{t-1}, x_{t-2} etc. abhängen, sondern auch – zusätzlich oder ausschließlich – von den **zeitlich vorangehenden Zufallskomponenten**, Gleitmittelkomponenten bzw. **Vorhersagefehlern** (Random Shocks) a_{t-1}, a_{t-2} etc. Regressionsmodelle dieser Art bezeichnet man als Gleitmittelmodelle (Moving Average bzw. **MA-Modelle**). Dieses lautet im einfachsten Fall:

$$x_t = a_t - \theta_1 \cdot a_{t-1}. \qquad (12.42)$$

Formal wird dieser Sachverhalt mit dem Parameter $q = 1$ durch **ARIMA ($\boldsymbol{p}$, $\boldsymbol{d}$, 1)** zum Ausdruck gebracht. Hängt eine Messung nicht nur von der jeweils letzten Zufallskomponente a_{t-1}, sondern auch von der vorletzten Zufallskomponente a_{t-2} ab, resultiert mit $q = 2$ somit allgemein **ARIMA ($\boldsymbol{p}$, $\boldsymbol{d}$, 2)** bzw.

$$x_t = a_t - \theta_1 \cdot a_{t-1} - \theta_2 \cdot a_{t-2}. \qquad (12.43)$$

MA-Modelle mit $q > 2$ findet man in der Praxis selten. θ_1 und θ_2 (griech. theta) sind Gewichte, deren Berechnung wir hier nicht näher erläutern.

Die Systematik einer Zeitreihe ist in vielen Fällen durch ein *ARIMA*(p, d, q)*-Modell* („auto regressive integrated moving average model") vollständig beschreibbar: p gibt die Anzahl der autoregressiven Anteile an, d entspricht der Anzahl der Differenzierungen, die erforderlich sind, um die Zeitreihe trendfrei zu machen, und q informiert über die Anzahl der Gleitmittelkomponenten.

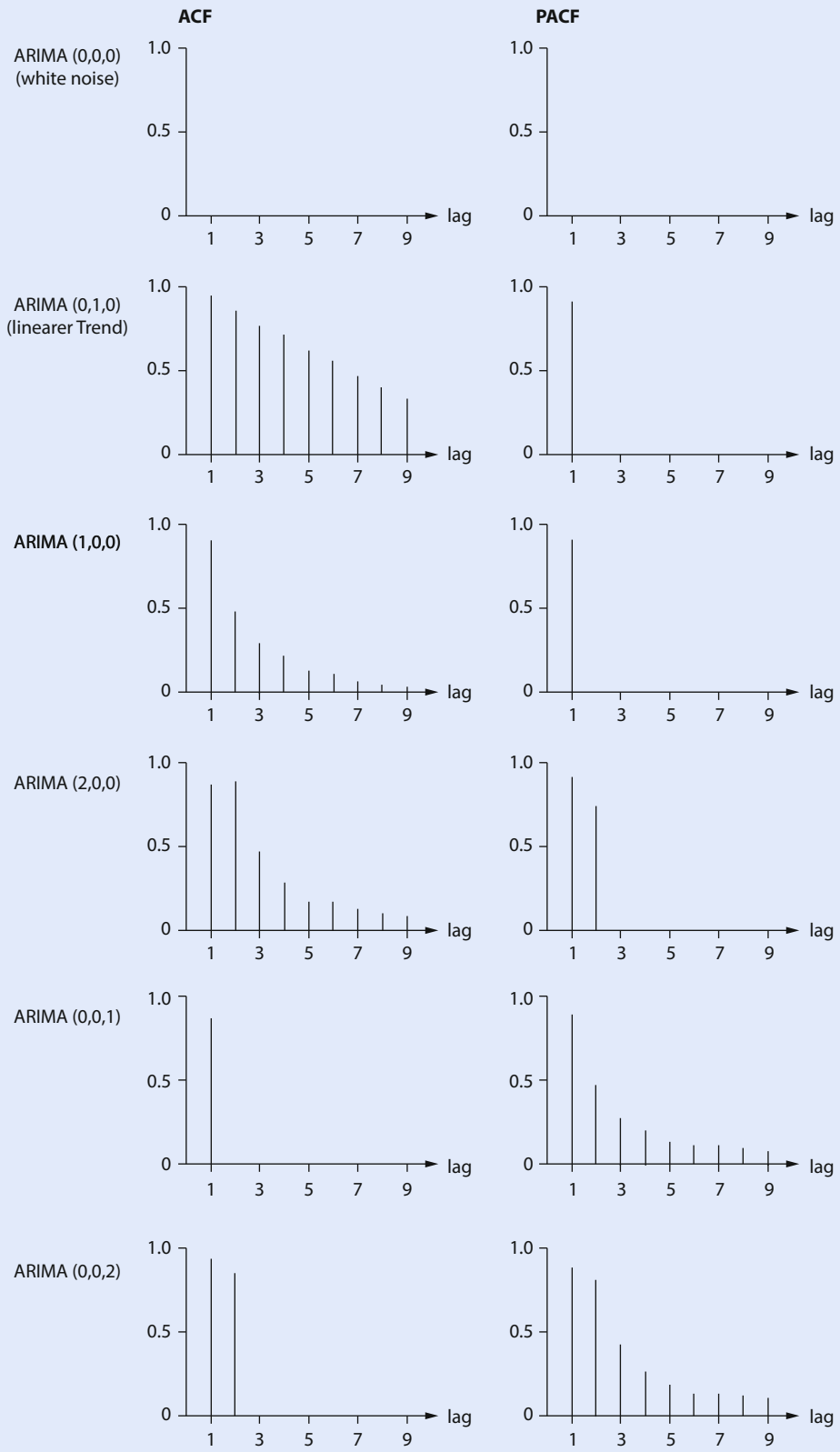


Abb. 12.60 Erwartete Autokorrelationen *(ACF)* und Partialautokorrelationen *(PACF)* für einige ARIMA (*p*, *d*, *q*)-Modelle (betrachtet werden hier nur AR- bzw. MA-Modelle, keine Mischmodelle)

Tabelle 12.25 Zusammenfassung der Identifikationshilfen für die wichtigsten ARIMA-Modelle

	White Noise	AR-Prozess	MA-Prozess	Mischprozess
Autokorrelationen	Für alle Lags: Null	Für alle Lags, beginnend mit Lag 1: rascher exponentieller Abfall bzw. gedämpfte Sinusschwingung	Die ersten q Lags: ungleich Null. Für Lag k, $k > q$: Null	Für Lag k, $k > q - p$: gedämpfte exponentielle und/oder Sinusschwingung
Partialautokorrelationen	Für alle Lags: Null	Die ersten p Lags: ungleich Null. Für Lag k, $k > p$: Null	Für alle Lags, beginnend mit Lag 1: rascher exponentieller Abfall bzw. gedämpfte Sinusschwingung	Für Lag k, $k > p - q$: gedämpfte exponentielle und/oder Sinusschwingung

Ein ARIMA(1, 1, 1)-Modell hat demnach eine autoregressive Komponente 1. Ordnung und eine Gleitmittelkomponente 1. Ordnung und zeigt zudem einen linearen Trend, der durch einfache Differenzierung zu beseitigen ist. Wie das ARIMA-Modell einer empirisch gefundenen Zeitreihe identifiziert wird, erläutert der folgende Abschnitt.

Identifikation eines ARIMA(p, d, q)-Modells

Jede durch bestimmte φ- und θ-Gewichte beschreibbare Zeitreihe zeigt für sie typische Autokorrelationen, die in einem **Autokorrelogramm** (Autokorrelationsfunktion, AKF, bzw. „autocorrelational function", **ACF**) darstellbar sind (Abb. 12.60). Folgt das Autokorrelogramm z. B. einem exponentiell abschwingenden Verlauf, wird dies als Hinweis gewertet, dass die empirische Zeitreihe eine Realisation eines autoregressiven Prozesses ist. Bei einem **AR-Modell** – z. B. ARIMA (1, 0, 0) – geht die ACF immer allmählich gegen Null, anders als beim **MA-Modell**, z. B. ARIMA (0, 0, 1). Man versucht, aus dem Muster der ACF sowohl die AR- als auch die MA-Parameter zu erschließen, was nur geht, wenn die Reihe stationär ist bzw. durch Differenzenbildung stationär gemacht worden ist.

Eine zusätzliche Identifikationshilfe stellt das sog. **Partialautokorrelogramm** („partial autocorrelational function", **PACF**) dar. In ihm werden die Partialautokorrelationen zwischen verschiedenen Messpunkten abgetragen, wobei die zwischen den Zeitpunkten liegenden Messungen herauspartialisiert werden. Die Partialautokorrelation 1. Ordnung heißt $r_{13.2}$ und gibt den Zusammenhang zwischen x_t und x_{t-2} (Lag 2) wieder, wobei der Einfluss der dazwischen liegenden Messungen x_{t-1} neutralisiert wird. Die Partialautokorrelation 2. Ordnung erfasst den Zusammenhang der Messungen x_t und x_{t-3} (Lag 3) unter Ausschaltung des Einflusses von x_{t-1} und x_{t-2}. Entsprechendes gilt für Partialautokorrelationen höherer Ordnung. (Definitionsgemäß ist die Partialautokorrelation 0. Ordnung mit Lag 1 die einfache Autokorrelation zwischen den Messungen x_t und x_{t-1}.)

In Abb. 12.60 finden sich die typischen Muster des Autokorrelogramms (ACF) und Partialautokorrelogramms (PACF) für die in der Praxis am häufigsten vorkommenden Zeitreihen. Die Abbildungen gehen von positiven Autokorrelationen 1. Ordnung aus. Bei negativen Autokorrelationen 1. Ordnung folgen abwechselnd negative und positive Autokorrelationen aufeinander. An den Verlaufsmustern ändert sich jedoch nichts, wenn man nur die Beträge betrachtet.

Identifikationshilfen Gelegentlich hat eine Zeitreihe sowohl AR-Anteile als auch MA-Anteile, was die Identifikation erschwert. Deshalb empfiehlt es sich, bei der Identifikation des ARIMA(p, d, q)-Modells einer Zeitreihe in folgenden Schritten vorzugehen (McCain & McCleary, 1979, S. 249f.):

1. Wenn die ACF nicht schnell absinkt, ist die Zeitreihe trendbehaftet. Sie muss (ggf. wiederholt) differenziert werden, bis die ACF schnell absinkt. Die Anzahl der hierfür erforderlichen Differenzierungen entspricht dem Parameter d.
2. Für eine trendfreie (oder trendbereinigte) Zeitreihe sind als nächstes die ACF und die PACF zu prüfen. Fällt die ACF exponentiell ab, ist dies als Hinweis auf ein AR-Modell zu werten. Eine exponentiell abfallende PACF deutet auf einen MA-Prozess hin.
3. Ist es möglich, die Zeitreihe entweder als AR-Prozess oder als MA-Prozess zu identifizieren, gibt die Anzahl der signifikanten Spikes (Spitzen) in der PACF den p-Wert des AR-Prozesses bzw. die Anzahl der signifikanten Spikes der ACF den q-Wert des MA-Prozesses an. Hierbei sollte man mit möglichst niedrigen Werten als Anfangsschätzungen für p und q beginnen, denn zu kleine Werte werden in der anschließenden Modelldiagnose erkannt, zu große Werte hingegen nicht. So ist in Abb. 12.60 für das ARIMA(1, 0, 0)-Modell erkennbar, dass die ACF exponentiell absinkt (Indikator für ein AR-Modell) und dass die PACF einen Spike aufweist (Indikator für $p = 1$).
4. Wenn sowohl die ACF als auch die PACF exponentiell fallen, hat die Zeitreihe AR- und MA-Anteile. Für p und q sollte dann probeweise zunächst der Wert 1

angenommen werden. Ist dieses ARIMA-Modell unangemessen, sind p und q abwechselnd (ggf. auch gemeinsam) auf 2 zu erhöhen.

5. Wenn sich kein angemessenes ARIMA-Modell identifizieren lässt, besteht schließlich die Möglichkeit, die Zeitreihe zu transformieren (näheres hierzu bei McCain & McCleary, 1979, S. 250)

In Tab. 12.25 werden die wichtigsten Identifikationshilfen nochmals zusammengefasst.

Saisonale Modelle

Gelegentlich zeigen Zeitreihen saisonale Schwankungen bzw. periodisch wiederkehrende Regelmäßigkeiten. Erhebt man z. B. monatliche Messungen über viele Jahre hinweg, können die Jahresverläufe einander stark ähneln. Dies hat zur Folge, dass sich in der ACF (und/oder ggf. in der PACF) nicht nur die bereits beschriebenen anfänglichen Spikes, sondern zusätzlich hohe Korrelationen für Lag 12, Lag 24, Lag 36 etc. zeigen. Das ARIMA(p, d, q)-Modell ist dann zu einem saisonalen Modell zu erweitern, zu einem **SARIMA(p, d, q) – (P, D, Q)-Modell**.

Die Werte P, D und Q charakterisieren hierbei das saisonale ARIMA-Modell. Es wird genauso identifiziert wie das ARIMA-Modell der regulären Zeitreihe. Wählen wir als Beispiel eine Zeitreihe mit Jahresschwankungen, gehen wir wie folgt vor:

Folgen die Jahresdurchschnitte einem (steigenden oder fallenden) Trend, zeigen sich allmählich abfallende Autokorrelationen für die Lags 12, 24, 36 etc. Saisonale Trends werden durch eine jahresweise vorgenommene Differenzierung beseitigt.

Eine autoregressive saisonale Komponente führt zu einem exponentiellen Abfall der Autokorrelationen für die Lags 12, 24, 36 etc. Für P wird 1 gesetzt, wenn die PACF nur bei Lag 12 einen Spike zeigt. Ist die Partialautokorrelation für Lag 24 ebenfalls hoch, nimmt man für P den Wert 2 an.

Für saisonal beeinflusste Gleitmittelprozesse erwarten wir einen exponentiellen Abfall der PACF für die Lags 12, 24, 36 etc. In Abhängigkeit davon, ob die ACF nur bei Lag 12 oder zusätzlich auch bei Lag 24 einen Spike hat, ist $Q = 1$ oder $Q = 2$ zu setzen.

Modelldiagnose

Nach einer (ggf. vorläufigen) Identifikation des ARIMA-Modells werden die Parameter φ und θ geschätzt (zur Technik vgl. die eingangs erwähnte Spezialliteratur). Es schließt sich eine Diagnostik an, die überprüft, ob das ARIMA-Modell die Zeitreihe hinreichend genau beschreibt oder ob die Abweichungen der empirischen Zeitreihe von der für das ARIMA-Modell vorhergesagten Zeitreihe substanziell sind. Hierfür werden jeweils eine **residuale ACF und PACF** errechnet, die bei guter Anpassung des Modells nur noch White Noise, d. h. nichtsignifikante Autokorrelationen und Partialautokorrelationen aufweisen dürfen.

In ►Die Belegschaft eines Betriebes im Verlauf von 20 Jahren. Identifikation eines ARIMA-Modells wird die

Die Belegschaft eines Betriebes im Verlauf von 20 Jahren. Identifikation eines ARIMA-Modells

Die in der folgenden Abbildung wiedergegebene Zeitreihe stellt – nach Monaten aufgeschlüsselt – die Anzahl der Werktätigen eines Betriebes in den Jahren 1945 bis 1966 dar (Abb. 12.61). Der Grafik ist – bei starken monatlichen Schwankungen – eine ständige Zunahme der Zahl der Betriebsangehörigen zu entnehmen. Jeweils im August werden die meisten Werktätigen gezählt, dies drückt sich in den jährlichen Spitzen im Liniendiagram der Zahl der Angestellten im Zeitverlauf aus. Die beiden folgenden Diagramme verdeutlichen den Verlauf der Autokorrelation (ACF) und der Partialautokorrelation (PACF; Abb. 12.62).

Die Klammern geben die Signifikanzgrenzen der jeweiligen Korrelation wieder. Die Autokorrelationen bleiben hier über mehrere Lags hinweg signifikant und weisen damit – wie schon vermutet – die Zeitreihe als trendbehaftet aus. Sie wird deshalb zunächst differenziert. Das Ergebnis dieser Differenzierung zeigt die folgende Grafik (Abb. 12.63):

Es zeigen sich nun deutliche Spikes in der ACF für die Lags 12 und 24. Da die Korrelationen für diese beiden Lags nur geringfügig verschieden sind, hat die Zeitreihe auch einen saisonalen Trend. Eine erneute Differenzierung für Lag 12 ist erforderlich. Es resultieren die in Abb. 12.64 wiedergegebenen ACF und PACF.

Die PACF sinkt nach Lag 1 und nach Lag 12 relativ rasch ab, und die ACF hat bei Lag 1 und bei Lag 12 jeweils einen Spike. Man kann deshalb vermuten, dass für diese Zeitreihe das ARIMA (0, 1, 1)(0, 1, 1)$_{12}$-Modell angemessen ist. Diese Vermutung bestätigt Abb. 12.65. Die Residuen dieses Modells sind statistisch nicht mehr signifikant und stellen „White Noise“ dar.

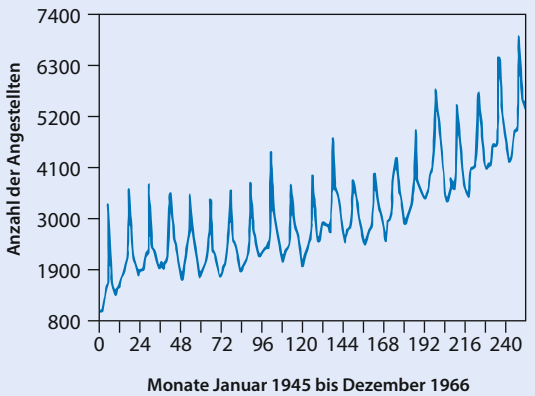


Abb. 12.61 Zeitreihe

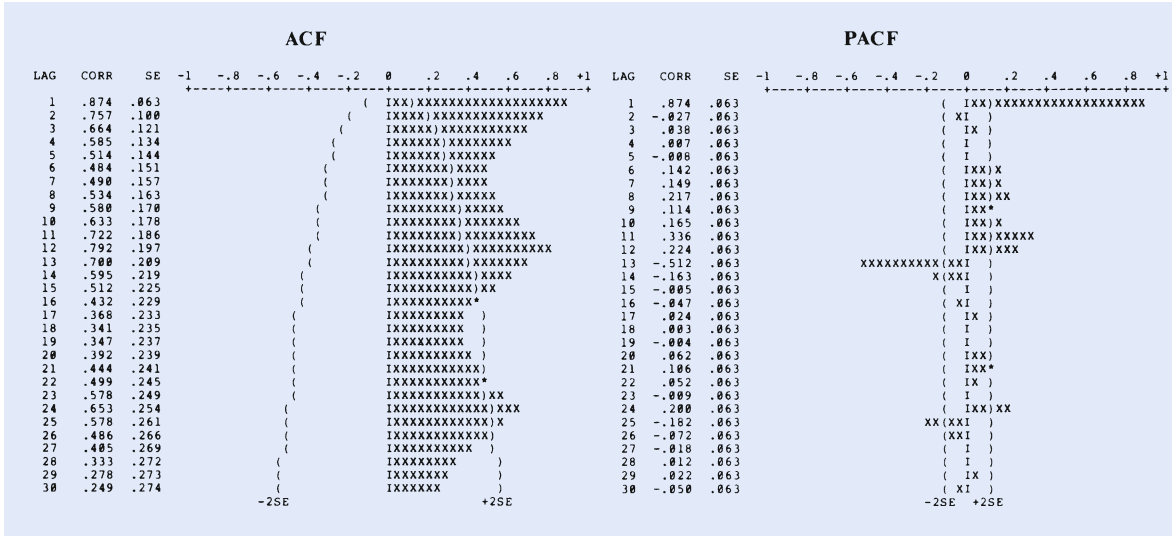


Abb. 12.62 Erwartete Autokorrelationen *(ACF)* und Partialautokorrelationen *(PACF)* der Zeitreihe

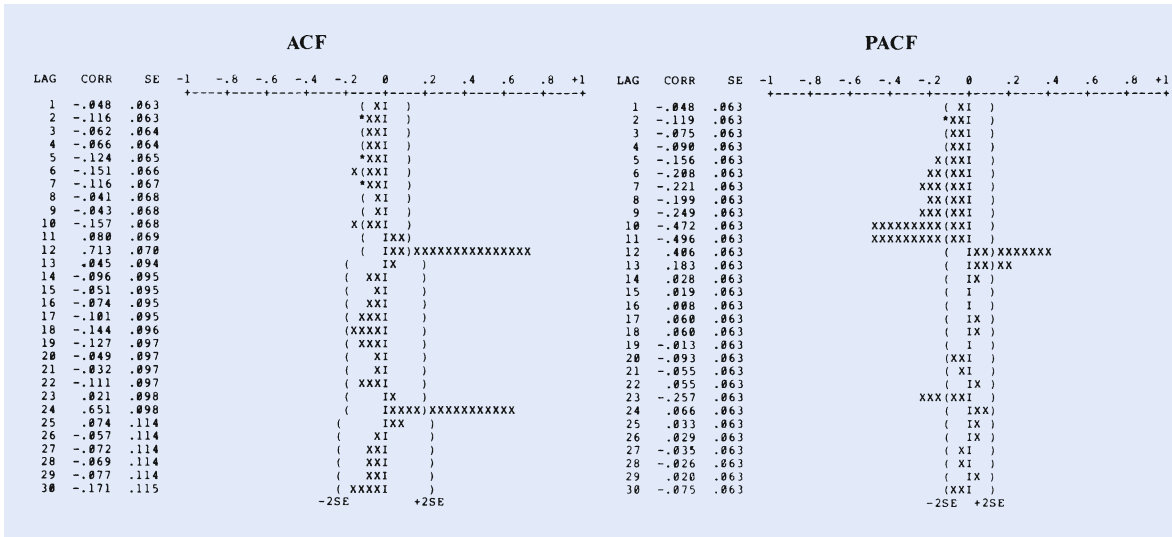


Abb. 12.63 Erwartete Autokorrelationen *(ACF)* und Partialautokorrelationen *(PACF)* der differenzierten Zeitreihe

Analyse einer Zeitreihe an einem Beispiel gezeigt (nach McCleary & Hay, 1980, S. 104ff.)

Vorhersagen, Interventionen und Zusammenhänge

Die Beschreibung einer Zeitreihe durch ein ARIMA-Modell ist für sich genommen belanglos (deshalb wurde in ▶ Die Belegschaft eines Betriebes im Verlauf von 20 Jahren. Identifikation eines ARIMA-Modells auf die Wiedergabe der Parameter φ und θ verzichtet, da diese in den meisten Fällen ohnehin nichtssagend sind). Liegt das ARIMA-Modell einer Zeitreihe hingegen fest, kann dieses zur Hypothesenprüfung im Sinne der Vorhersage zukünftiger Entwicklungen (Forecasting), zur Überprüfung der verändernden Wirkung einer Intervention (Treatment) oder zur Ermittlung des Einflusses einer anderen Zeitreihe auf die untersuchte Zeitreihe herangezogen werden (Transferfunktionsmodell).

Vorhersagemodelle Der erste Anwendungsfall, die Vorhersage, ist unkompliziert. Man nutzt die im ARIMA-Modell zusammengefassten Informationen bzgl. der vergangenen Entwicklungen für die Prognose eines oder mehrerer zukünftiger Messpunkte. Die Vorhersagen sind umso genauer, je länger und stabiler die Zeitreihe ist. Weit in die Zukunft reichende Vorhersagen sind natürlich weniger präzise als die Vorhersage des sich unmittelbar an

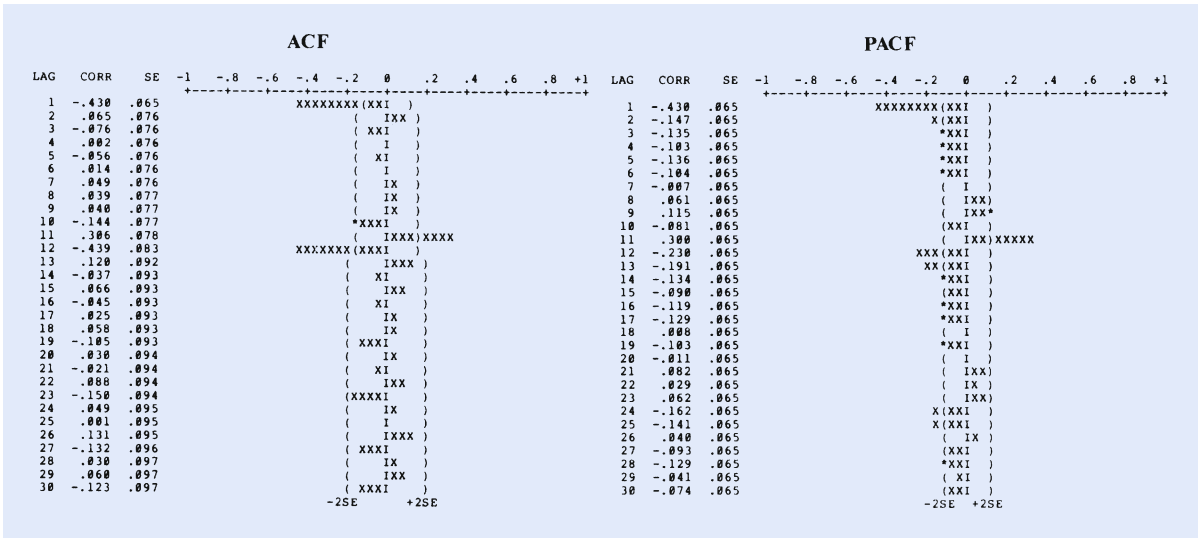


Abb. 12.64 Erwartete Autokorrelationen *(ACF)* und Partialautokorrelationen *(PACF)* der zusätzlich saisonal differenzierten Zeitreihe

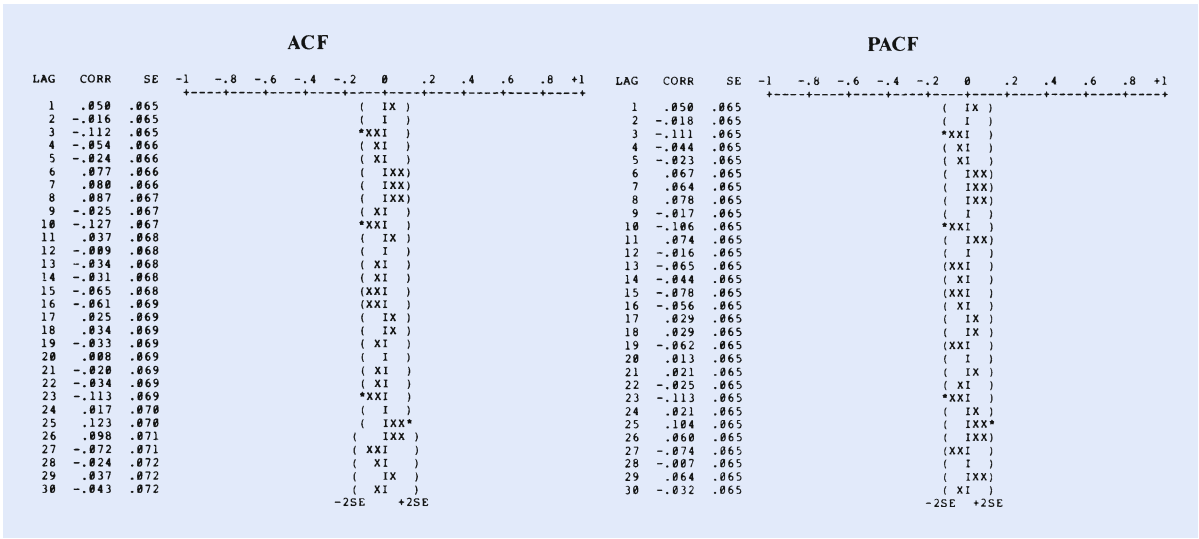


Abb. 12.65 Diagnose: Erwartete Autokorrelationen *(ACF)* und Partialautokorrelationen *(PACF)* der Modellresiduen

die Zeitreihe anschließenden nächsten Messpunktes. Bei Vorhersagemodellen ist nur eine einzige mehrfach gemessene Variable beteiligt **(univariates Modell)**.

Interventionsmodelle Der zweite Anwendungsfall betrifft die Überprüfung von Veränderungen, die eine Intervention, ein Ereignis oder ein Treatment bei einer Zeitreihe bewirken. Neben der **Zeitreihe als abhängiger Variable** (**Outputvariable** bzw. Outputreihe) kommt hier also noch die **Intervention als binäre unabhängige Variable** mit den Ausprägungen Eins und Null (**Inputvariable**) ins Spiel. Für jeden Messzeitpunkt ist anzugeben, ob die Intervention wirkt (1) oder nicht einwirkt (0). Hierbei unterschieden wir drei Arten von Interventionen:

- **Einmalige Intervention** (Puls-Input). Beispiel: Wie wirkt sich der zum Messzeitpunkt 6 aufgetretene einmalige Aufruf des Kinderschutzbundes, dass Familien wöchentlich einen fernsehfreien Tag einlegen sollten, auf das Fernsehverhalten aus, das in einer Zeitreihe regelmäßig gemessen wird? (0000010000 . . .)
- **Wiederholte Interventionen**. Beispiel: Wie wirken sich wiederholte Umweltschutzappelle (z. B. zu Messzeitpunkt 3, 8 und 11) auf das Stromsparverhalten aus? (00100001001 . . . ; diese Kodierung setzt vor-

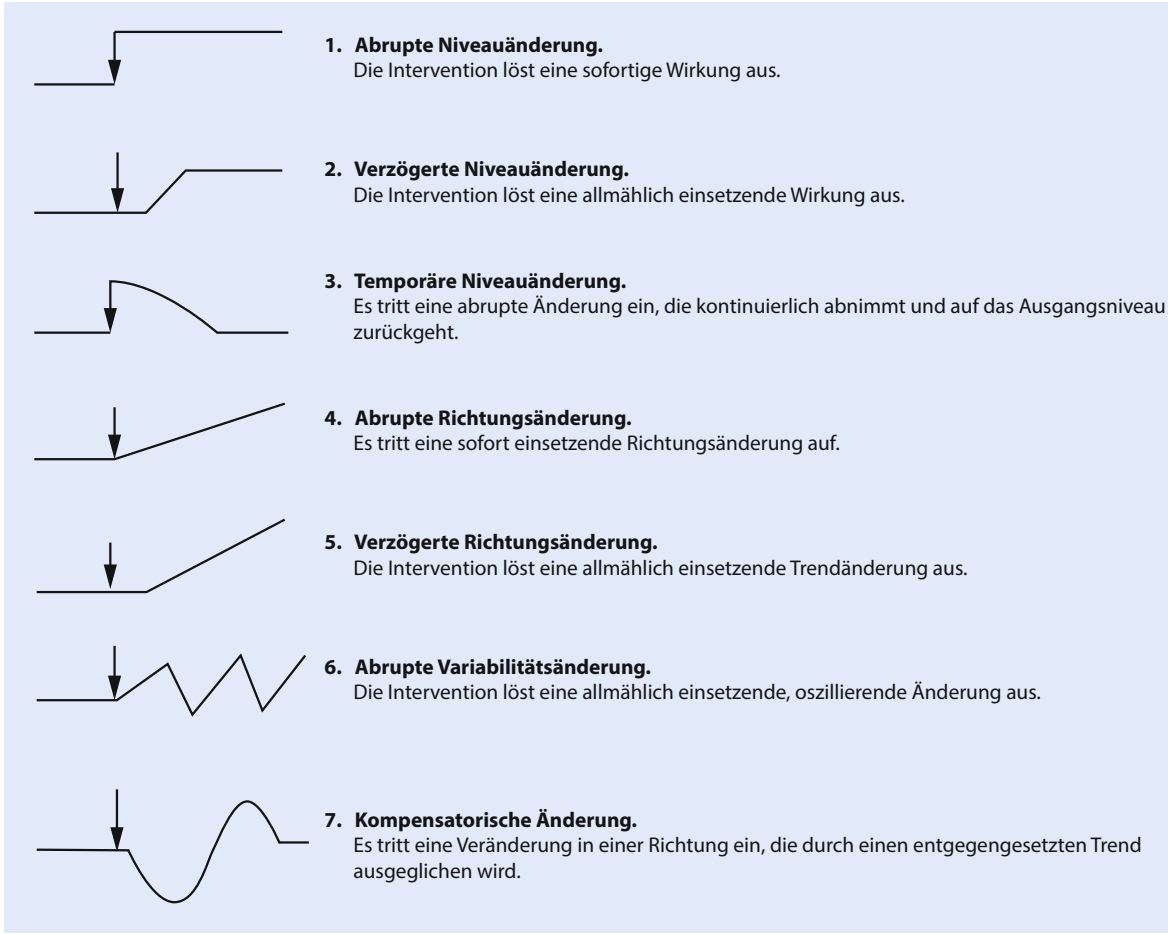


Abb. 12.66 Interventionseffekte in Zeitreihenanalysen

aus, dass die Wirkungen gleich bleiben. Eine andere Variante wäre, für jede Intervention eine eigene Interventionsvariable zu kreieren.)

- **Dauerhaft wirkende Interventionen** (Stufen-Input). Beispiel: Wie wirkt sich die Verabschiedung eines neuen Scheidungsgesetzes, das zu Messzeitpunkt 6 in Kraft getreten ist und seitdem gilt, auf die Anzahl der Scheidungen aus? (0000011111 . . .)

Bei einem Interventionsmodell gilt es zu überprüfen, wie die jeweilige Intervention die Zeitreihe ändert. Das ARIMA-Modell der Zeitreihe sollte anhand jener Daten ermittelt werden, die vor dem Zeitpunkt der Intervention liegen. Ist die Periode vor der Intervention zu kurz, um eine eindeutige Identifikation zu erlauben, schlägt Jenkins (1979, S. 72; zit. nach Rottleuthner-Lutter, 1985) vor, für die Bestimmung des ARIMA-Modells die gesamte Zeitreihe einschließlich der auf die Intervention folgenden Messzeitpunkte zu verwenden.

Die Art der Wirkung, die eine Intervention auslöst, kann beliebig sein oder vorher hypothetisch festgelegt und entsprechend überprüft werden. In Abb. 12.66 werden einige Beispiele für mögliche Interventionseffekte gezeigt (Glass et al., 1975, zit. nach Petermann, 1978, S. 95).

Für die Berechnung von Interventionsmodellen werden Transferfunktionen benötigt. Dabei ist die Transferfunktion der Filter, mit dem die 0/1-kodierte Intervention als Inputvariable auf den Output einwirkt (Thome, 2005, S. 176).

Transferfunktionsmodelle Beim dritten Anwendungsfall der Prüfung von Veränderungshypothesen mittels Box-Jenkins-Zeitreihenanalyse – den Transferfunktionsmodellen – wird nicht nur eine binär kodierte Inputvariable, sondern eine beliebige **andere Zeitreihe als Input** berücksichtigt, deren Effekte auf die interessierende abhängige Zeitreihe zu prüfen sind.

Beispiele: Wie beeinflusst die Anzahl berufstätiger Frauen die Geburtenrate? Verändert die selbst eingeschätzte Befindlichkeit eines Therapeuten das Befinden seines Patienten? Gehen Schwankungen der Arbeitslosenquote mit Veränderungen der Kriminalitätsrate bei Eigentumsdelikten einher? Hierbei wird explizit zwischen einer abhängigen Zeitreihe **(Outputreihe)** und einer unabhängigen Zeitreihe **(Inputreihe)** unterschieden, d. h. es wird theoretisch bzw. inhaltlich eine bestimmte **Kausalrichtung** vorausgesetzt (z. B. wird angenommen, dass durch eine erhöhte Arbeitslosenquote die Neigung zu Eigentumsdelikten steigt, nicht aber umgekehrt, dass durch mehr Wohnungseinbrüche die Arbeitslosenquote steigt). Es wird ein beide univariate Zeitreihen umfassendes bivariates Transferfunktions-ARIMA-Modell erstellt, dessen Interpretation z. B. Fragen folgender Art beantwortet:

- Mit welchem zeitlichen Verzug wirkt sich eine unabhängige Zeitreihe auf die abhängige Zeitreihe aus?
- Welche Veränderungen der unabhängigen Zeitreihe lösen welche Veränderungen der abhängigen Zeitreihe aus?
- Lassen sich Vorhersagen der Entwicklung einer Zeitreihe verbessern, wenn die gemeinsamen Regelmäßigkeiten von weiteren Zeitreihen berücksichtigt werden?

Das ▶ Studienbeispiel „Zeitreihenanalyse: Der Einfluss des Alkoholkonsums auf die Suizidrate in Russland“ illustriert die Analyse gemäß Transferfunktionsmodell. Weitere Beispiele und Einzelheiten hierzu findet man z. B. bei McCleary und Hay (1980, Kap. 5). Zur Interpretation wechselseitiger Zusammenhänge von Zeitreihen sei auf Boker, Xu, Rotondo, & King (2002) verwiesen. Für die formale Struktur von Transferfunktionsmodellen s. z. B. Thome (2005).

Mit Transferfunktionsmodellen können nicht nur punktuelle bzw. **statische Modelle**, sondern auch verzögerte und über die Zeit sich ausbreitende Effekte modelliert werden (**dynamische Modelle**, zu deren empirischer Identifikation man Kreuzkorrelationsfunktionen heranzieht). Außerdem ist auf zwei Auswertungsprobleme hinzuweisen, die in der Praxis bei Transferfunktionsmodellen häufig auftreten:

1. Für die korrekte Modellidentifikation und Parameterschätzung bei Zeitreihenanalysen ist es generell notwendig, **Ausreißerwerte** („outlier“) zu erkennen und korrekt zu behandeln (Thome, 2005).
2. Bei bestimmten Zeitreihen (Stichwort: kointegrierte Prozesse) führt eine **vorgängige Trendbereinigung**

Studienbeispiel

Zeitreihenanalyse: Der Einfluss des Alkoholkonsums auf die Suizidrate in Russland

Auf der Basis früherer Studien, die einen Zusammenhang zwischen Alkoholkonsum (insbesondere Binge Drinking) und Suizidalität belegen, wurde die Hypothese aufgestellt, dass der Alkoholkonsum in Russland über die letzten Dekaden hinweg die Suizidraten beeinflusst hat. Dazu wurden für Russland über 35 Jahre hinweg von 1970 bis 2005 der jährliche Pro-Kopf-Verbrauch an reinem Alkohol in Liter (Wodka, Bier und Wein) einerseits und die Suizidrate (Anzahl der Suizide pro 1 Mio. Einwohner) andererseits erfasst ($n = 35$ Jahre) und einer Zeitreihenanalyse nach dem Box-Jenkins-Verfahren unterzogen (Razvodosky, 2009). Die Hypothese lautete, dass sich mit der Veränderung des Alkoholkonsums (unabhängige Zeitreihe) einhergehend die Suizidraten (abhängige Zeitreihe) verändern.

Wie bei jeder inferenzstatistischen Auswertung ist auch bei einer Zeitreihenanalyse vor dem eigentlichen Signifikanztest zur Überprüfung der Veränderungshypothese eine deskriptivstatistische – insbesondere eine grafische – Inspektion der Daten sinnvoll. Es zeigt sich grafisch, dass die beiden Zeitreihen für Wodkakonsum und Suizid bei Männern in Russland sehr ähnliche Muster aufweisen (◘ Abb. 12.67). Dass Mitte der 1980er-Jahre eine nationale Anti-Alkoholkampagne stattfand, spiegelt sich deutlich im Liniendiagramm wider, wobei eine kompensatorische Änderung zu konstatieren ist: Der Alkoholkonsum geht zunächst deutlich zurück, steigt nach der Kampagne aber wieder an. Ziel der Studie war es jedoch nicht, die Interventionswirkung der Kampagne zu prüfen, sondern im Sinne eines Transferfunktionsmodells den Effekt der Inputzeitreihe (Alkoholkonsum) auf die Outzeitreihe (Suizidrate; ◘ Abb. 12.67).

Es wurde jeweils ein **Transferfunktionsmodell** bzw. ein „bivariates“ (= 2 Zeitreihen: Alkohol und Suizid) ARIMA(1, 1, 0)-Modell ($p = 1$: Autokorrelation 1. Ordnung, $d = 1$: trendfrei nach erster Differenzbildung, $q = 0$: keine Gleitmittelkomponente) für Männer und Frauen für den Einfluss von **Alkohol insgesamt** (Wodka, Bier, Wein) sowie von **Wodka allein** auf die Suizidraten berechnet. Die entsprechenden Kreuzkorrelationsanalysen sind hier nicht dargestellt. Alle vier Transferfunktionsmodelle sind auf dem 5 %-Niveau signifikant (◘ Abb. 12.68). Das Ergebnis ist nicht nur hypothesenkonform statistisch signifikant, so dass die Veränderungshypothese zum Effekt der Alkoholkonsumzeitreihe auf die Suizidratenzeitreihe angenommen werden kann, sondern zeigt auch eine nennenswerte Effektgröße: Eine Zunahme des Pro-Kopf-Alkoholkonsums um 1 Liter ging bei Männern mit einer Steigerung der Suizidrate um 4 %, bei Frauen um 2.8 % einher. Betrachtet man separat den Wodkakonsum, so betrug die entsprechende Steigerung der Suizidrate 9.3 % bei den Männern und 5.8 % bei den Frauen. In absoluten Zahlen lag die Suizidrate am Ende der Zeitreihe (2005) für Männer bei 599 und für Frauen bei 94 pro 1 Mio. Einwohner. Aufgrund von wissenschaftlichen Kenntnissen über Alkoholwirkungen kann die Kausalrichtung von Alkohol auf Suizid (und nicht umgekehrt) plausibilisiert werden, auch wenn die Studie nicht-experimentell angelegt ist. Welche weiteren Faktoren auf die Suizidrate (und ggf. gleichzeitig auch auf den Alkoholkonsum) einwirken, muss unter Rückgriff auf andere Studien diskutiert werden.

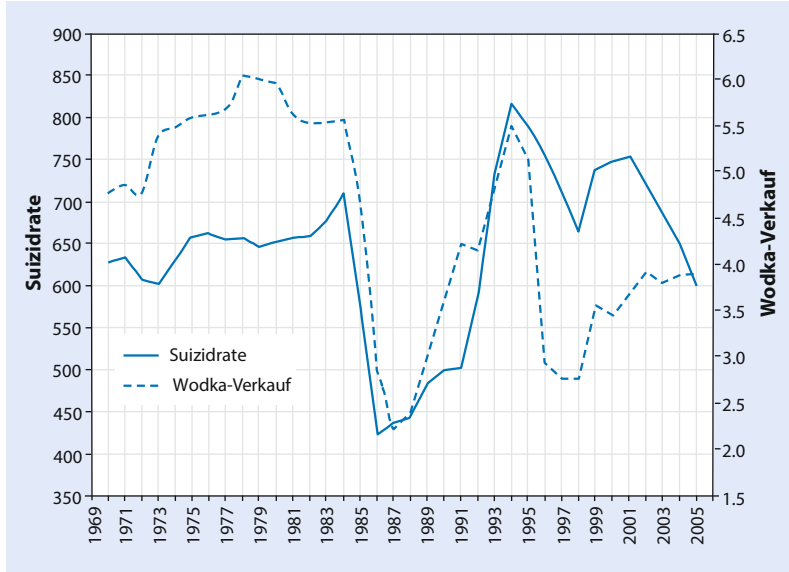


Abb. 12.67 Trends in der Höhe des Pro-Kopf-Wodkaverkaufs und der Suizidrate von Männern in Russland von 1970–2005. (Aus Razvodosky, 2009. Used by permission from Crisis 2009, Vol. 30 (4), S. 186–191. Copyright © 2009 Hogrefe Publishing; DOI 10.1027/0227-5910.30.4.186)

	Alkoholkonsum insgesamt			Wodka-Konsum		
	Model	Estimate	*p*	Model	Estimate	*p*
Suizidrate Männer	1,1,0	.040	.006	1,1,0	.093	<.0001
Suizidrate Frauen	1,1,0	.028	.005	1,1,0	.058	<.0001

Abb. 12.68 Auf der Basis eines Transferfunktionsmodells ARIMA (1, 1, 0) geschätzte Effekte des getränkespezifischen Alkoholverkaufs auf die Suizidrate. (Mod. nach Razvodosky, 2009. Used by permission from Crisis 2009, Vol. 30 (4), S. 186–191. Copyright © 2009 Hogrefe Publishing; DOI 10.1027/0227-5910.30.4.186)

(Differenzenbildung) zu einer Fehlspezifikation des Transferfunktionsmodells und unter Umständen zu dem Fehlschluss, zwischen beiden Zeitreihen bestünde kein Effekt (scheinbare Nichtkausalität; Thome, 2005, S. 279).

Lernquiz

1. Welches sind die typischen Anwendungsfälle der inferenzstatistischen Überprüfung von Veränderungshypothesen in den Sozialwissenschaften?
2. Welche drei Hypothesenarten können mit der Box-Jenkins-Zeitreihenanalyse statistisch geprüft werden?
3. Was sind Sequenzeffekte und wie kann man sie kontrollieren?
4. Skizzieren Sie die Vorgehensweise bei der Regressions-Diskontinuitäts-Analyse!
5. Diskutieren Sie Vor- und Nachteile von Querschnitt- und Längsschnittstudien und geben Sie an, welches statistische Auswertungsverfahren typisch ist!
6. Was versteht man im Kontext der Zeitreihenanalyse unter einer Autokorrelation?
7. Im Rahmen einer Längsschnittstudie zur Veränderung des Selbstwertgefühls über die Lebensspanne wird eine Stichprobe 20-Jähriger untersucht. Vier Folgeerhebungen finden jeweils im Abstand von 10 Jahren statt. Es liegen also Selbstwertdaten der Stichprobe im Alter von 20, 30, 40, 50 und 60 Jahren vor.
   a) Welche Hypothese wird mit einer einfaktoriellen Varianzanalyse geprüft?
   b) Lässt ein signifikanter Haupteffekt der Varianzanalyse den Schluss zu, dass sich der

Selbstwert durch das Lebensalter verändert? Begründung!

c) Mit welchen Verfahren ließe sich die Hypothese testen, dass der Selbstwert im Laufe des Lebens linear ansteigt?

d) Wie ließe sich die Hypothese prüfen, dass sich v. a. der erste und der letzte Messzeitpunkt unterscheiden?

8. Die monatlichen Umsatzzahlen des Online-Versandhauses Amazon.de liegen für März 2005, Februar 2006, Mai 2007, Februar 2008, Mai 2009, Juni 2010, Mai 2011, Juni 2012 und Juni 2013 vor. Handelt es sich um eine Zeitreihe? Begründung!
9. Auf welche drei Einflussfaktoren wird die Ausprägung eines Messwertes in einer Zeitreihe im Rahmen einer ARIMA-Analyse zurückgeführt?
10. Wozu dient der Solomon-Viergruppenplan?

Übungsaufgaben

1. Folgende Studie prüft die Veränderungshypothese, dass sich das **Selbstwertgefühl von britischen Kindern** im Alter von 11–12 Jahren durch sportliche Aktivität im Grünen verbessert (Open Access Fachartikel): Reed, K., Wood, C., Barton, J., Pretty, J.N., Cohen, D. et al. (2013). A Repeated Measures Experiment of Green Exercise to Improve Self-Esteem in UK School Children. PLoS ONE 8(7): e69176. doi:10.1371/journal.pone.0069176
   Sowohl körperliche Bewegung als auch der Aufenthalt in der freien Natur wirken sich positiv auf das Selbstwertgefühl aus, die Kombination beider Faktoren in Form sportlicher Aktivität im Grünen („green exercise“) gilt als besonders selbstwertförderlich. Ziel der Studie war es, diesen Effekt bei Schulkindern nachzuweisen. In einem randomisierten Prä-Post-Kontrollgruppen-Design absolvierten $n = 75$ Kinder der sechsten Schulklasse im Rahmen des Sportunterrichts einen Dauerlauf über rund 2.40 Kilometer in einem Park (Experimentalgruppe) sowie eine Woche später in der Stadt (Kontrollgruppe), wobei jeweils vor und nach dem Lauf der Selbstwert mit einer etablierten Selbstwert-Skala (Rosenberg Scale) erfasst wurde. Die Reihenfolge der Experimental- und Kontrollbedingung wurde zufällig variiert und hatte keinen Effekt. Eine zweifaktorielle **Varianzanalyse mit Messwiederholungen** auf beiden Faktoren wurde gerechnet: Wie viele Hypothesen wurden geprüft und welche Ergebnisse hatten sie? War die Teststärke ausreichend, um über alle Hypothesen eindeutig entscheiden zu können? Neben dem Selbstwertgefühl wurde u. a. auch der Spaß als abhängige Variable untersucht, war er beim Dauerlauf im Grünen größer als in der Stadt?
2. Folgende Studie prüft die Veränderungshypothese, dass sich die **Schlafqualität von Krankenhaus-Pflegekräften in Schweden** durch den Berufsalltag verschlechtert (Open Access Fachartikel):
   Hasson, D. & Gustavsson, P. (2010). Declining Sleep Quality among Nurses: A Population-Based Four-Year Longitudinal Study on the Transition from Nursing Education to Working Life. *PLoS ONE* 5(12): e14265. doi:10.1371/journal.pone.0014265
   Die Schlafqualität der Pflegekräfte ist zum einen relevant für deren Gesundheit und Wohlbefinden, zum anderen für die Sicherheit der Patienten, da übermüdetes Personal eher Fehler macht. Eine Längsschnittstudie über vier Jahre hinweg erhob jährlich die Schlafqualität der Pflegekräfte ($n = 988$ Befragte zu allen vier Messzeitpunkten), die sich bei der ersten Messung (Baseline) noch im Studium befanden und bei den drei Follow-Up-Messungen im Berufsleben standen. Die Schlafqualität wurde mit einem validierten Einzel-Indikator gemessen (fünfstufige Ratingskala von 5 = gute, 4 = eher gute, 3 = weder gute noch schlechte, 2 = eher schlechte, 1 = schlechte Schlafqualität). Die Datenanalyse erfolgte mit einer **Varianzanalyse (ANOVA) mit Messwiederholungen**: Handelte es sich um eine uni- oder multivariate, um eine ein- oder mehrfaktorielle Varianzanalyse? Die Schlafqualität im Sample sank von $M_{t1} = 4.26$ bei der Baseline-Messung über $M_{t2} = 4.10$ und $M_{t3} = 4.05$ auf $M_{t4} = 4.02$ bei der dritten Follow-up-Messung. War diese Veränderung statistisch signifikant? Welche Effektgröße hatte sie? Ergänzend wurden **Kovarianzanalysen (ANCOVA) mit Messwiederholungen** gerechnet? Welche Kovariaten wurden dabei betrachtet und welchen Einfluss hatten sie auf die Schlafqualität der schwedischen Pflegekräfte?
3. Folgende Studie prüft die Veränderungshypothese, dass durch die im März 2006 in Kraft getretene Anti-Rauch-Gesetzgebung die **Prävalenz des Rauchens in Schottland** zurückgegangen ist (Open Access Fachartikel):
   Mackay, D.F., Haw, S, Pell J.P. (2011). Impact of Scottish Smoke-Free Legislation on Smoking Quit Attempts and Prevalence. *PLoS ONE* 6(11): e26188. doi:10.1371/journal.pone.0026188
   Im Rahmen der Studie (Sekundäranalyse) wurde unter anderem die vierteljährlich in einer repräsen-

tativen Bevölkerungsstudie erhobene Prävalenz des Rauchens von Januar 1999 bis September 2010 ausgewertet. Wie viele Messzeitpunkte hatte diese Zeitreihe? Die Auswertung dieser Zeitreihe erfolgte über eine **Zeitreihenanalyse nach der Box-Jenkins-Methode (ARIMA)**: Wurde anhand der Daten ein AR- oder ein MA-Modell identifiziert? Die Einführung des neuen schottischen Anti-Rauch-Gesetzes im März 2006 wurde zudem im Rahmen der Zeitreihenanalyse mit einem **Interventionsmodell** getestet. Gehen die Autoren dabei von einem Puls- oder einem Stufen-Input aus? In welcher Weise und wie nachhaltig beeinflusste die Gesetzesänderung die Prävalenz des Rauchens?

Weiterführende Literatur

- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Hoboken: Wiley.
- Newsome, J. T., Jones, R. N. & Hofer, S. M. (Eds.). (2011). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences*. New York: Routledge Chapman & Hall.
- Senn, S. S. (2002). *Cross-over trials in clinical research* (2nd ed.). Hoboken: Wiley.
- Thome, H. (2005). *Zeitreihenanalyse. Eine Einführung für Sozialwissenschaftler und Historiker*. München: Oldenbourg.

Literatur

Adam, J. (1978). Sequential strategies and the separation of age, cohort, and time of measurement contributions to developmental data. *Psychological Bulletin, 85*(6), 1309–1316.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken: Wiley

Baer, J. & Schmitz, M. F. (2000). Latent growth curve modeling with *a* cohort sequential design. *Social Work Research, 24*(4), 243–248.

Baltes, P. B. (1967). *Längsschnitt- und Querschnittsequenzen zur Erfassung von Alters- und Generationseffekten.* Saarbrücken: Universität des Saarlandes.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measurement of change* (pp. 3–20). Madison: University of Wisconsin Press.

Bierhoff, H. W. & Rudinger, G. (1996). Quasi-experimentelle Untersuchungsmethoden. In E. Erdfelder, R. Mausfeld & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 47–58). Weinheim: Psychologie Verlags Union.

Blossfeld, H. P., Hamerle, A., & Mayer, K. U. (1986). *Ereignisanalyse. Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften*. Frankfurt/Main: Campus.

Bohrnstedt, G. W. (1969). Observations on the measurement of change. In E. F. Borgetta (Ed.), *Sociological methodology*. San Francisco: Bass.

Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002). Windowed crosscorrelation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods, 7*(3), 338–355.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. ed.). Berlin Heidelberg: Springer

Box, G. E. P. & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Braden, J. P. & Bryant, T. J. (1990). Regression discontinuity designs: Applications for school psychologists. *School Psychology Review, 19*(2), 232–239.

Braver, M. C. & Braver, S. L. (1988). Statistical treatment of the solomon four-groups design: A meta-analytic approach. *Psychological Bulletin, 104*(1), 150–154.

Campbell, D. T. & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guildford.

Collins, L. M. (1996). Is reliability obsolete? A Commentary on „Are simple gain scores obsolete?“. *Applied Psychological Measurement, 20*(3), 289–292.

Cook, T. D. & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*, 548–580.

Corder-Bolz, C. R. (1978). The evaluation of change: New evidence. *Educational and Psychological Measurement, 38*(4), 959–976.

Cronbach, L. J. & Furby, L. (1970). How should we measure „change“ – or should we? *Psychological Bulletin, 74*(1), 68–80.

Cudeck, R. & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods, 7*(1), 41–63.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York: Springer.

Delucchi, K. & Bostrom, A. (1999). Small sample longitudinal clinical trials with missing data: A comparison of analytic methods. *Psychological Methods, 4*(2), 158–172.

Deutsch, S. J. & Alt, F. B. (1977). The effect of massachusetts gun control law on gun-related crimes in the city of boston. *Evaluation Quarterly, 1*(4), 543–568.

Du Bois, P. H. (1957). *Multivariate correlational analysis*. New York: Harper.

Eckensberger, L. H. (1973). Methodological issues of cross-cultural research in development psychology. In J. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology – methodological issues*. (pp. 43–64). New York: Academic Press.

Eid, M., Gollwitzer, M., & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (3. Aufl.). Weinheim: Beltz.

Erdfelder, E., Rietz, C., & Rudinger, G. (1996). Methoden der Entwicklungspsychologie. In E. Erdfelder, R. Mausfeld & T. Meiser (Hrsg.), *Handbuch Quantitative Methoden* (S. 539–550). Weinheim: Beltz.

Fischer, G. H. (1995). Linear logistic models for change. In G. H. Fischer & J. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications* (pp. 157–180). New York: Springer.

Formann, A. K. & Ponocny, I. (2002). Latent change classes in dichotomous data. *Psychometrika, 52*, 263–267.

Furr, R. M. & Rosenthal, R. (2003). Repeated measures contrasts for “multiple–*p*attern“ hypothesis. *Psychological Methods, 8*(3), 275–293.

Galton, F. (1886). Family likeness in stature. *Procedings of the Royal Society, 40*, 49–53.
Glass, G. V., Tiao, G. O., & Maguire, T. O. (1971). Analysis of data on the 1900 revision of German divorce laws as *a*time-series quasi-experiment. *Law and Society Review, 5*(4), 539–562.
Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder: University Press.
Gottman, J. M. (Ed.). (1995). *The analysis of change*. Mahwah: Erlbaum.
Gudat, U. & Revenstorff, D. (1976). Interventionseffekte in klinischen Zeitreihen. *Archiv für Psychologie, 128*, 16–44.
Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw Hill.
Gulliksen, H. (1968). Methods of determining equivalence of measures. *Psychological Bulletin, 70*(6), 534–544.
Hamilton, J. (1994). *Time series analysis*. Princeton: Princeton University Press.
Hedeker, D. & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*(1), 64–78.
Helmreich, R. (1977). *Strategien zur Auswertung von Längsschnittdaten*. Stuttgart: Klett.
Hennigan, K. M., Del Rosario, M. L., Heath, L., Cock, T. D., Calder, B. J., & Wharton, J. D. (1979). *How the introduction of television affected the level of violent and instrumental crime in the united states*. Washington: National Science Foundation.
Hibbs, D. (1977). On analyzing the effects of policy interventions: Box–Jenkins and Box–Tiao vs. structural equation models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 137–179). San Francisco: Bass.
Holland, P. W. (1986). Statistical and causal inference (with discussion). *Journal of the American Statistical Association, 81*(396), 945–970.
Hoppe, S., Schmid-Schönbein, C., & Seiler, T. B. (1977). *Entwicklungssequenzen*. Bern: Huber.
Hultsch, D. F. & Bortner, R. W. (1974). Personal Time Perspective in Adulthood: A Time–Sequential Study. *Developmental Psychology, 10*(6), 835–837.
Jenkins, G. M. (1979). *Practical experiences with modelling and forcasting time series*. Jersey: Channel Islands.
Johnson, C. W. (1986). A more rigorous quasi-experimental alternative to the one-group pretest-posttest design. *Educational and Psychological Measurement, 46*(3), 585–591.
Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin, 82*(3), 345–362.
Keren, G. (1993). Between or within – subjects design: A methodological dilemma. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological Issues* (pp. 257–272). Hillsdale: Lawrence Erlbaum Associates.
Kette, G. (1990). Determinanten der Geschworenenentscheidung: Ein Anwendungsbeispiel für die Zeitreihenanalyse der Rechtspsychologie. *Archiv für Psychologie, 142*, 59–81.
Klemmert, H. (2004). *Äquivalenz- und Effekttests in der Psychologischen Forschung*. Frankfurt/Main: Lang.
Lacey, J. I. & Lacey, B. C. (1962). The law of initial value in the longitudinal study of autonomic constitution. *Annals of the New York Academy of Sciences, 98*(4), 1257–1290.
Langeheine, R. & Van de Pol, F. (1990). Veränderungsmessung bei kategorialen Daten. *Zeitschrift für Sozialpsychologie, 21*(2), 88–100.
Linn, R. L. & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research, 47*(1), 121–150.
Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*(12), 750–751.
Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421–437.
Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. (pp. 21–38) Madison: University of Wisconsin Press.
Ludwig, D. A. (1979). Statistical considerations for the univariate analysis of repeated-measures experiments. *Perceptual Motor Skills, 49*(3), 899–905.
Malgady, R. G. & Colon-Malgady, G. (1991). Comparing the reliability of difference scores and residuals in analysis of covariance. *Educational and Psychological Measurement, 51*(4), 803–807.
Maxwell, S. E. (1994). Optimal allocation of assessment time in randomized pretest-posttest designs. *Psychological Bulletin, 115*(1), 142–152.
Maxwell, S. E. (1998). Longitudinal designs in randomized group comparison. When will intermediate observation increase statistical power? *Psychological Methods, 3*(3), 275–290.
Mayer, K.-U., & Huinink, J. (1990). Alters-, Perioden- und Kohorteneffekte in der Analyse von Lebensverläufen oder: Lexis ade? In K.-U. Mayer (Hrsg.), *Lebensverläufe und sozialer Wandler, Sonderheft 31 der Kölner Zeitschrift für Soziologie und Sozialpsychologie* (S. 442–459). Opladen: Westdeutscher Verlag.
McCain, L. J. & McCleary, R. (1979). The statistical analysis of the simple interrupted times series quasi-experiment. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experiment: Design and analysis issues for field settings* (pp. 207–233). Chicago: Rand-McNally.
McCleary, R. & Hay, J. R. A. (1980). *Applied time series analysis for the social sciences*. Thousand Oaks: Sage.
McDowall, D., McCleary, R., Meidinger, E. E., & Hay Jr., R. A. (1980). *Interrupted time series*. Thousand Oaks: Sage.
McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement, 18*, 47–55.
Meier, F. (Hrsg.). (1988). *Prozeßforschung in den Sozialwissenschaften. Anwendungen zeitreihenanalytischer Methoden*. Stuttgart: Fischer.
Mellenberg, G. J. (1999). A note on simple gain score precisions. *Applied Psychological Measurement, 23*(1), 87–89.
Metzler, P. & Nickel, B. (1986). *Zeitreihen- und Verlaufsanalyse*. Leipzig: Hirzel.
Minsel, W. R. & Langer, I. (1973). Methodisches Vorgehen zum Erfassen von psychotherapeutisch bedingten Veränderungen. In G. Reinert (Hrsg.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970* (S. 646–648). Göttingen: Hogrefe.
Mittring, G. & Hussy, W. (2004). Die Ermittlung der kleinsten hinreichend großen Stichprobe bei wissenschaftlichen Experimenten mit Randomisierung. Kölner Psychologische Studien. *Beiträge zur natur-, kultur-, sozialwissenschaftlichen Psychologie, IX*(1), 1–57.
Miyazaki, Y. & Raudenbusch, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods, 5*(1), 44–63.
Mosteller, F. (1990). Improving research methodology: An overview. In L. Sechrest, E. Perrin & J. Bunker (Eds.), *Research methodology. Strengthening causal interpretations of nonexperimental data* (pp. 221–230). Rockville: AHCPR, PHS.
Nachtigall, C. & Suhl, U. (2002). *Der Regressionseffekt. Mythos und Wirklichkeit*. Jena: Friedrich-Schiller-Universität Jena.
Nelson, C. R. (1973). *Applied time series analysis for managerial forecasting*. San Francisco: Holden-Day.
O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42*(1), 73–98.
Orth, U., Trzesniewski, K. H., & Robins, R. W. (2010). Self-esteem development from young adulthood to old age: A cohort-sequential longitudinal study. *Journal of Personality and Social Psychology, 98*(4), 645–658.

Pawlik, K. & Buse, L. (1994). „Psychometeorologie“: Zeitreihenanalytische Ergebnisse zum Einfluß des Wetters auf die Psyche aus methodenkritischer Sicht. *Psychologische Rundschau, 45*, 63–78.

Petermann, F. (1978). *Veränderungsmessung*. Stuttgart: Kohlhammer.

Plewis, I. (1981). A comparison of approaches to the analysis of longitudinal categoric data. *British Journal of Mathematical and Statistical Psychology, 34*(1), 118–123.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178–192.

Razvodosky, Y. E. (2009). Beverage-specific alcohol sale and suicide in russia. *Crisis: The Journal of Crisis Intervention and Suicide Prevention, 30*(4), 186–191.

Rennert, M. (1977). Einige Anmerkungen zur Verwendung von Differenzwerten bei der Veränderungsmessung. *Psychologische Beiträge, 19*, 100–109.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using Significance test to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553–565.

Rogosa, D. R. (1995). Myths and methods: „Myths about longitudinal research“ plus supplemental questions. In J. M. Gottman (Ed.), *The Analysis of Change* (pp. 3–66). Mahwah: Erlbaum.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726–748.

Rogosa, D. R. & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20*(4), 335–343.

Rogosa, D. R. & Willett, J. B. (1985). Understanding correlates of change by modelling individual differences in growth. *Psychometrika, 50*(2), 203–228.

Rossi, P. H. & Freeman, H. E. (1985). *Evaluation: A systematic approach* (3rd ed.). Thousand Oaks: Sage.

Rost, J. (2004). *Lehrbuch Testtheorie Testkonstruktion* (2. Aufl.). Bern: Huber.

Rothermund, K. & Brandstädter, J. (2003). Depression in later life: Cross-sequential patterns and possible determinants. *Psychology and Aging, 18*(1), 80–90.

Rottleuthner-Lutter, M. (1985). *Evaluation mit Hilfe der Box-Jenkins-Methode.* Technische Universität Berlin, Berlin.

Rubin, D. B. (1977). Assignement to treatment groups on the basis of *a* covariate. *Journal of Educational Statistics, 2*(1), 1–26.

Rudinger, G. (1981). Tendenzen und Entwicklungen entwicklungspsychologischer Versuchsplanung – Sequenzanalysen. *Psychologische Rundschau, 32*, 118–136.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64*(2), 92–107.

Schaie, K. W. (1977). Quasi-experimental designs in the psychology of aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 39–58). New York: Van Nostrand Reinhold.

Schaie, K. W. (1994). Developmental designs revisited. In S. H. Cohen & H. W. Reese (Eds.), *Life-Span Developmental Psychology. Methodological Contributions* (pp. 45–64). Hillsdale: Erlbaum.

Schlittgen, R. & Streitberg, B. H. (1994). *Zeitreihenanalyse* (5. ed.). München: Oldenbourg.

Schmitz, B. (1989). *Einführung in die Zeitreihenanalyse. Modelle, Softwarebeschreibung, Anwendung*. Bern: Huber.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods, 16*(2), 179–191.

Shumway, R. H. & Stoffer, D. S. (2000). *Time series analysis and its appliccations*. New York: Springer.

Singer, J. D. & Willett, J. B. (1991). Modeling the days of our lifes: Using survival analysis when designing and analysing longitudinal studies of duration and the time of events. *Psychological Bulletin, 110*(2), 268–290.

Stelzl, I. (1982). *Fehler und Fallen der Statistik*. Bern: Huber.

Swaminathan, H. & Algina, J. (1977). Analysis of quasi-experimental time-series designs. *Multivariate Behavioral Research, 12*(1), 111–131.

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis. An alternative to the ex-post-facto-experiment. *Journal of Educational Psychology, 51*(6), 309–317.

Thome, H. (2005). *Zeitreihenanalyse. Eine Einführung für Sozialwissenschaftler und Historiker*. München: Oldenbourg.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression discontinuity approach*. Thousand Oaks: Sage.

Trochim, W. M. K. & Cappelleri, J. C. (1992). Cutoff assignement strategies for enhancing randomized clinical trials. *Controlled Clinical Trials, 13*(3), 190–212.

Vagt, G. (1976). Korrektur von Regressioneffekten in Behandlungsexperimenten. *Zeitschrift für experimentelle und angewandte Psychologie, 23*, 284–296.

Vagt, G. (1977). Meßinstrumente verändern sich im Laufe der Zeit. *Psychologie und Praxis, 21*, 117–122.

Venter, A., Maxwell, S. E., & Bolig, E. (2002). Power in randomized group comparison: the value of adding a single intermediate time point to a traditional pretest-posttest design. *Psychological Methods, 7*(2), 194–209.

Wellek, S. (1994). *Methoden zum Nachweis von Äquivalenz*. Stuttgart: Fischer.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49*(3), 587–602.

Williams, R. H. & Zimmermann, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement, 37*(3), 679–689.

Zielke, M. (1980). Darstellung und Vergleich von Verfahren zur individuellen Veränderungsmessung. *Psychological Beiträge, 22*, 592–609.

Zimmermann, D. W. & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*(2), 149–154.

12.5.5 Signifikanztests für Einzelfallhypothesen

Lernziele

- Verschiedene Verfahren zur statistischen Überprüfung von Einzelfallhypothesen kennen.
- Geeignete Untersuchungspläne und Datenanalyseverfahren zur Überprüfung von Einzelfallhypothesen auswählen können.
- Methodische Probleme von Längsschnittstudien mit Einzelfällen bei der Untersuchungsplanung berücksichtigen und vermeiden können.
- Prinzipien der Einzelfalldiagnostik verstehen.
- Ergebnisse der Überprüfung von Einzelfallhypothesen richtig interpretieren können.

Bei quantitativen Studien handelt es sich meist um **Gruppenstudien** (seien es Stichprobenuntersuchungen oder Vollerhebungen): Effekte werden in Gruppenstudien anhand von Aggregatwerten (z. B. Gruppenmittelwerten) untersucht. Es ist jedoch auch möglich und sinnvoll, quantitative hypothesenprüfende Studien mit jeweils einer einzelnen Untersuchungsperson (bzw. einem einzelnen Untersuchungsobjekt: eine Gruppe, eine Organisation) durchzuführen. Wenn bei derartigen **Einzelfalluntersuchungen** („single-subject research“, „single case research“, „single participant research“) Ursache-Wirkungs-Relationen untersucht werden sollen, fungiert die Untersuchungsperson gleichzeitig als ihre eigene Kontrollgruppe bzw. Kontrollperson: Es werden systematisch unter unterschiedlichen Bedingungen eine Reihe von Messwerten derselben Untersuchungsperson erhoben. Dabei können in einer Einzelfalluntersuchung die Ausprägungen der unabhängigen Variablen (Bedingungs-/Ursachenfaktoren), die einen Einfluss auf die abhängige Variable (z. B. Angstsymptome) haben, entweder von den Forschenden in einem experimentellen Design gesetzt (z. B. Phase mit Behandlung und Phase ohne Behandlung) oder natürlich vorgefunden werden (z. B. Phase mit besonderer beruflicher Belastung und Phase ohne besondere berufliche Belastung). Von der hypothesenprüfend angelegten quantitativen Einzelfallstudie, deren Messwerte grafisch und/oder statistisch ausgewertet werden, ist die **qualitative Fallstudie** („case study“) abzugrenzen.

Typische Anwendungsfelder von hypothesenprüfenden Einzelfalluntersuchungen sind die pädagogische, sonderpädagogische, klinische oder sportwissenschaftliche Forschung (Petermann, 1996). So möchte man etwa prüfen, welche Trainingspläne bei einer Profisportlerin zu den besten Leistungssteigerungen führen oder wie ein körperlich behindertes Kind von einer Reittherapie optimal profitiert. Da hier teilweise hochindividualisierte Trainings- oder Therapiepläne zum Einsatz kommen, sind Einzelfallstudien oft sinnvoller als Gruppenstudien (Barker, Richards, Jones & Monro, 2011; Kazdin, 2010; Morgan & Morgan, 2008).

Hypothesenprüfende Einzelfalluntersuchung – Bei der hypothesenprüfenden Einzelfalluntersuchung („single-subject research“, „single case research“, „single participant research“) geht es darum, **Einzelfallhypothesen** – also Annahmen über Merkmale oder Verhaltensweisen einer einzelnen Person bzw. eines einzelnen Objektes – anhand eines systematisch erhobenen quantitativen Datensatzes zu überprüfen. Für hypothesenprüfende Einzelfallanalysen werden Verhaltensstichproben derselben Person in verschiedenen Situationen, zu unterschiedlichen Zeitpunkten und/oder unter variierenden Aufgabenstellungen gezogen. Zur Auswertung eignen sich v. a. visuelle Verfahren sowie nonparametrische (verteilungsfreie) Signifikanztests.

Bevor wir uns verschiedenen Untersuchungstechniken zuwenden, ist es angebracht, sich zunächst ein wenig mit dem Grundprinzip hypothesenprüfender Einzelfalluntersuchungen auseinanderzusetzen. Als Beispiel wählen wir die Hypothese: „Frau M. reagiert auf berufliche Misserfolge mit Magenschmerzen“. Aus der Menge aller das Verhalten und Erleben von Frau M. beeinflussenden Merkmale wird ein (hier dichotomes) Merkmal als unabhängige Variable (beruflicher Misserfolg versus kein beruflicher Misserfolg) herausgegriffen, das der Hypothese nach die abhängige Variable „Magenschmerzen“ (mit den Stufen vorhanden versus nicht vorhanden bzw. mit unterschiedlichen Intensitätsstufen) beeinflusst. Zweifellos reicht es nicht aus, die Hypothese als bestätigt anzusehen, wenn auf ein einmaliges berufliches Versagen mit Magenschmerzen reagiert wird. Denn diese Koinzidenz könnte zufällig sein und bei vergleichbaren Anlässen nicht wiederholt auftreten. Für die Überprüfung der Hypothese müssen mehrere Phasen mit bzw. ohne berufliche Misserfolge hinsichtlich ihrer Auswirkungen auf die abhängige Variable überprüft werden.

Damit stellt sich das für die statistische Hypothesenprüfung wichtige Problem der „Stichprobe“ bzw. der Art von Generalisierung, die für das Ergebnis einer Einzelfalluntersuchung angestrebt wird. Solange der Geltungsbereich der Hypothese nicht durch die Nennung von (z. B. zeitlichen) Rahmenbedingungen eingeschränkt ist, besteht die Population, für die die Untersuchung aussagekräftig sein soll, aus allen erfolglosen bzw. nicht erfolglosen Phasen des Berufslebens von Frau M. Übertragen wir die Erfordernisse einer Zufallsstichprobe (► Abschn. 9.3.2) konsequent auf Einzelfalluntersuchungen, müsste zum einen die Liste aller „Elemente“ dieser Population bekannt sein, und zum anderen wäre die zu untersuchende Stichprobe (hier die zu untersuchenden Berufsphasen) nach einem statistischen Zufallsverfahren aus den Elementen der Population zusammenzustellen.

Die meisten Einzelfalluntersuchungen dürften weder das eine noch das andere Kriterium erfüllen: Die vollständige Liste aller Elemente der Population ist in der Regel unbekannt, und die Auswahl erfolgt üblicherweise willkürlich, indem z. B. mehrere in einem begrenzten Zeitabschnitt zeitlich aufeinander folgende Phasen untersucht werden. Für derartige Stichproben lassen sich bestenfalls fiktive Referenzpopulationen konstruieren (▶ Abschn. 9.1), die mit der eigentlich interessierenden Population (im Beispiel das gesamte Berufsleben von Frau M.) nur wenig zu tun haben (zum Problem der interindividuellen Generalisierbarkeit von Einzelfallergebnissen s. Westmeyer, 1979).

Will man Einzelfallhypothesen überprüfen, ist man darauf angewiesen, eine **Stichprobe vergleichbarer Verhaltensausschnitte** zu untersuchen, die über einen mehr oder weniger langen Zeitabschnitt verteilt sind. Dieses Faktum ist entscheidend, wenn man die interne Validität von Einzelfalluntersuchungen einschätzen will. Ob nun, wie im obigen Beispiel, die Bedeutung von nicht geplant auftretenden Ereignissen oder von willkürlich gesetzten Interventionen interessiert – der statistische Nachweis einer Veränderung des untersuchten Merkmals ist kein sicherer Beleg für die Wirksamkeit der Ereignisse oder der Interventionen. Wie bei allen Untersuchungen zur Überprüfung von Veränderungen, die sich über die Zeit erstrecken, muss damit gerechnet werden, dass andere, mit der Zeit kovariierende Merkmale die Veränderung verursachen. Derartige Störbedingungen lassen sich allerdings in Einzelfalluntersuchungen leichter kontrollieren als in Untersuchungen mit Stichproben. Um im genannten Untersuchungsbeispiel sicherzugehen, dass ein eindeutiger Rückschluss von beruflichem Misserfolg auf die von Frau M. berichteten Magenschmerzen möglich ist, müssen andere Einflussfaktoren (z. B. Mahlzeiten, Medikamente, familiäre Belastungen etc.) auf die abhängige Variable Magenschmerzen in geeigneter Weise kontrolliert werden.

Bei der Untersuchungsplanung ist auch bei Einzelfalluntersuchungen darauf zu achten, dass beobachtete Veränderungen in den abhängigen Variablen möglichst eindeutig auf die interessierende(n) unabhängige(n) Variable(n) zurückgeführt werden können und nicht auf andere zwischenzeitlich wirksame Einflussfaktoren zurückgehen (Frage der *internen Validität*).

Wiederholte Messungen können nicht nur die interne Validität von Einzelfalluntersuchungen beeinträchtigen, sondern auch deren statistische Auswertung erschweren. Viele „klassische“ Auswertungsroutinen basieren auf der **Annahme der Unabhängigkeit der Messungen** (bzw. der Messfehler), die bei wiederholten Messungen einer Person – zumal, wenn diese in kurzen Abständen erfolgen – meistens nicht gegeben ist (Gottman, 1973). Dies wurde bei vielen Auswertungen von Einzelfalldaten und auch bei der Entwicklung spezieller statistischer Verfahren zur Einzelfallanalyse übersehen, was häufig zur Folge hatte, dass die Bedeutung der überprüften Interventionseffekte **überschätzt** wurde (Nicolich & Weinstein, 1977, zit. nach Levin, Marascuilo, & Hubert, 1978; Revenstorf & Keeser, 1979).

Bei der statistischen Auswertung von Daten aus Einzelfalluntersuchungen ist zu beachten, dass es sich nicht um unabhängige Messungen, sondern um *abhängige Messungen* handelt.

Diese Überlegungen sind bei der Auswertung und der Interpretation zu beachten. Dessen ungeachtet wird empfohlen, auch Einzelfalluntersuchungen hypothesenprüfend anzulegen und die Ergebnisse entsprechenden Signifikanztests zu unterziehen (Barlow & Hersen, 1984; Franklin, Allison, & Gorman, 1996; Kratochwill, 1978; Levin & Wampold, 1999; Petermann, 1981, 1982).

Deshalb seien im Folgenden einige Untersuchungsstrategien zur Überprüfung von Einzelfallhypothesen zu individuellen Veränderungen dargelegt. Anschließend gehen wir auf Signifikanztests für die Einzelfalldiagnostik ein, in deren Rahmen Testergebnisse einer Person zufallskritisch zu bewerten und mit anderen Testergebnissen derselben Person zu vergleichen sind.

Signifikanztests für individuelle Veränderungen

Wie bei jeder hypothesenprüfenden Studie ist auch die **Aussagekraft einer quantitativen Einzelfalluntersuchung** nicht nur daran gebunden, dass der passende statistische Signifikanztest ausgewählt und korrekt angewendet wird und zu einem signifikanten Ergebnis führt. Bei der Ergebnisinterpretation ist immer auch die Qualität des Untersuchungsdesigns zu berücksichtigen: Ein identisches statistisches Ergebnis kann im einen Fall ein sehr überzeugender Test der Einzelfallhypothese sein, während dies in einem anderen Fall nicht zutrifft: Hier kann auch ein statistisch signifikantes Ergebnis nicht überzeugen, weil das Untersuchungsdesign zu geringe interne und/oder externe Validität aufweist.

- **Externe Validität von Einzelfalluntersuchungen**: Quantitative Einzelfallstudien sind zur statistischen Hypothesenprüfung geeignet, wenn es gelingt, für das interessierende Phänomen mögliche **repräsentative** bzw. **typische Verhaltensausschnitte** zu finden. Hierbei wird es sich in der Regel um eine mehrere Messzeitpunkte umfassende Zeitreihe handeln, deren

Systematik man in Abhängigkeit von bestimmten Interventionen oder Ereignissen mit den Vorhersagen der zu prüfenden Hypothesen vergleicht. Die externe Validität derartiger Untersuchungen hängt v. a. davon ab, wie gut mit der Zeitreihe „typisches" Verhalten repräsentiert wird.

- **Interne Validität von Einzelfalluntersuchungen**: Die interne Validität ist wegen der vielen durchzuführenden Messungen v. a. durch **Testübungseffekte** gefährdet. Um Treatmenteffekte zu isolieren, wäre prinzipiell auch hier der Einsatz einer „Kontrollperson" denkbar, die ebenfalls – ohne Anwendung des Treatments – wiederholt untersucht wird. Vergleiche der Zeitreihen von Experimental- und Kontrollperson(en) dürften jedoch nur aussagekräftig sein, wenn die „Äquivalenz" der beiden verglichenen Personen sichergestellt ist. Zeitbedingte Gefährdungen der internen Validität (**externe zeitliche Einflüsse, Reifungsprozesse** ▶ Abschn. 3.2.2) können in Einzelfallanalysen besser kontrolliert werden als in Gruppenstudien. Dies setzt allerdings voraus, dass die geprüfte oder behandelte Person bereit ist, über ihre persönliche Wahrnehmung des Untersuchungsgeschehens offen Auskunft zu erteilen bzw. dass im Untersuchungszeitraum stattfindende – aber nicht auf Einflüsse der unabhängigen Variablen bezogene – Veränderungen der Person möglichst umfassend registriert werden.

Wir behandeln nun **Untersuchungspläne**, mit denen Hypothesen über die Wirksamkeit eines Treatments bzw. einer Intervention (unabhängige Variable, meist nominalskaliert: Behandlung oder keine Behandlung) hinsichtlich einer oder mehrerer abhängiger Variablen (z. B. Depressivität, Lernerfolg etc.) überprüft werden können. Alle Untersuchungspläne beinhalten die wiederholte Erhebung von Messungen an einem Einzelfall, wobei wir zwischen Erhebungsphasen **ohne Intervention (*A*-Phasen)** und Erhebungsphasen **mit Intervention (*B*-Phasen)** unterscheiden. Bei Untersuchungen mit willkürlich manipulierbaren Interventionen kann der Wechsel von *A*- und *B*-Phasen von den Forschenden gesteuert werden (z. B. in einer Verhaltenstherapie oder einem Trainingsprogramm, in denen das erwünschte Verhalten phasenweise belohnt und phasenweise nicht belohnt wird). Geht es um die Wirkung von Ereignissen, deren zeitliche Abfolge nicht vorhersagbar ist (wie im genannten Beispiel von Magenschmerzen bei beruflichen Misserfolgen) wechseln *A*- und *B*-Phasen nach Maßgabe des Auftretens des jeweils untersuchten Ereignisses im natürlichen Lebensumfeld der Untersuchungsperson.

Um Auffälligkeiten und Veränderungen der abhängigen Variablen während einer *B*-Phase feststellen zu können, muss das **„Normalverhalten"** bzw. die **Baseline** der abhängigen Variablen bekannt sein. Man bestimmt die Baseline durch mehrere Messungen vor dem ersten Einsetzen einer Intervention, wobei die Anzahl der Messungen während dieser ersten *A*-Phasen genügend groß sein sollte, um mehr oder weniger regelmäßige Schwankungen im Normalverhalten identifizieren zu können. Mögliche Interventionseffekte während der *B*-Phase sind dann einfacher vom Normalverhalten zu unterscheiden (zur grafischen Aufbereitung individueller Zeitreihen s. Parsonson & Baer, 1978).

In *Einzelfalluntersuchungsplänen* werden Erhebungsphasen ohne Intervention als *A-Phasen* und Erhebungsphasen mit Intervention als *B-Phasen* bezeichnet. Man unterscheidet unterschiedliche Designs von Einzelfallstudien (*Einzelfallstudienpläne*) nach Anzahl und Abfolge der *A*- und *B*-Phasen. Der *A-B-A-B-Plan* kommt in der Forschungspraxis am häufigsten zum Einsatz.

Einzelfalluntersuchungspläne unterscheiden sich in erster Linie darin, wie häufig und auf welche Art sich *A*- und *B*-Phasen abwechseln. Soweit die Grenzen der Belastbarkeit des Einzelfalles nicht überschritten werden, sind hierbei relativ beliebige Kombinationen denkbar. Die sechs in der Literatur am häufigsten erwähnten Pläne seien im Folgenden kurz vorgestellt (Barlow & Hersen, 1973; Fichter, 1979; Kratochwill & Levin, 1992).

- ***A*-*B*-Plan**
  Untersuchungen nach diesem Plan bestehen nur aus einer *A*-Phase mit einer darauffolgenden *B*-Phase. Zur Feststellung der Baseline werden in der *A*-Phase unter kontrollierten Bedingungen mehrere Messungen erhoben, die anschließend mit den in der *B*-Phase anfallenden Messungen verglichen werden. Dieser Plan ist zur statistischen Überprüfung einer individuellen Veränderungshypothese wenig geeignet.
- ***A*-*B*-*A*-Plan**
  Auch dieser Plan beginnt mit einer Baseline-Phase. An die Interventions-Phase schließt sich eine weitere Baseline-Phase an, die eindeutigere Aussagen über die Wirksamkeit der Intervention zulässt als der einfache *A*-*B*-Plan. Gleicht sich die abhängige Variable in der zweiten *A*-Phase wieder der Baseline an, ist dies – soweit hierfür keine Zufallsschwankungen verantwortlich sind – ein deutlicher Beleg für eine kurzzeitige Wirksamkeit der Intervention.
- ***B*-*A*-*B*-Plan**
  In vielen klinischen Einzelfallstudien erweist es sich als ungünstig (bzw. ethisch bedenklich), wenn die Untersuchung wie im *A*-*B*-*A*-Plan mit einer Baseline-Phase (d. h. ohne Behandlung) endet. Dies wird im *B*-*A*-*B*-Plan vermieden. Die Aussagekraft dieses Planes

ist jedoch durch die zwischen zwei B-Phasen eingeschobene A-Phase erheblich eingeschränkt, wenn man damit rechnen muss, dass das Normalverhalten vor der erstmaligen Wirkung einer Intervention anders geartet ist als nach einer Intervention.

- ***A-B-A-B*-Plan**
  Dieser Plan verbindet die Vorteile des A-B-A-Planes und des B-A-B-Planes und kommt deshalb in der Einzelfallforschung **am häufigsten zur Anwendung**. Nach der Etablierung einer stabilen Baseline wird – wie im A-B-A-Plan – untersucht, ob ein möglicher Interventionseffekt nach Absetzen der Intervention verschwindet und nach erneuter Intervention wieder auftritt, was zusammengenommen die Wirksamkeit der Intervention besser belegt als alle bisher besprochenen Pläne. Die Untersuchung endet hier mit einer Behandlungsphase.
- ***A-BC-B-BC*-Plan**
  Auch dieser Plan besteht (in seiner einfachsten Form) aus 4 Phasen. Dennoch führt er zu anderen Aussagen als der A-B-A-B-Plan. Er erfordert zwei Interventionen B und C (z. B. eine medikamentöse und eine psychotherapeutische Behandlung), die in kombinierter Form und auch einzeln eingesetzt werden. Die A-Phase dient wiederum der Festlegung einer stabilen Baseline. Es folgt eine BC-Phase, in der beide Interventionen gleichzeitig eingesetzt werden. Die anschließende B-Phase liefert darüber Aufschluss, welcher Anteil der Kombinations-(Interaktions-)Wirkung von BC auf B zurückzuführen ist. Um die Wirkung beider Interventionen isoliert erfassen zu können, müsste der Plan um eine C-Phase (und ggf. um eine weitere BC-Phase) erweitert werden.
- **Multiple-Baseline-Plan**
  Dieser von Baer et al. (1968) beschriebene Plan findet v. a. in der verhaltenstherapeutischen Einzelfallanalyse Beachtung. Er überprüft die Auswirkungen einer Behandlung auf mehrere Variablen (z. B. phobische Reaktionen auf verschiedene Auslöser). Nachdem die Baselines für alle Variablen feststehen, beginnt zunächst die auf eine Variable ausgerichtete Behandlung. Danach wird die zweite Variable mit in die Behandlung einbezogen etc. Zusammengenommen besteht dieser Plan also aus mehreren zeitversetzten A-B-Plänen (Einzelheiten zu diesem Plan berichtet Kazdin, 1976, 1978, 1982).

Die inferenzstatistische Auswertung dieser (und ähnlicher) Pläne bereitet wegen der bereits erwähnten seriellen Abhängigkeit der Messungen erhebliche Schwierigkeiten (Kratochwill et al., 1974). Da „klassische" Routineauswertungen (wie z. B. der t-Test oder die einfaktorielle Varianzanalyse) für Einzelfalldaten ungeeignet sind, wollen wir uns dem Problem der statistischen Hypothesenüberprüfung in Einzelfalluntersuchungen im Folgenden etwas ausführlicher zuwenden. Wir unterscheiden hierbei zwischen kardinalskalierten abhängigen Variablen (psychometrische Testwerte, physiologische Messungen, Ratingskalen, Häufigkeiten eines Merkmals etc.) und nominalskalierten abhängigen Variablen (Tab. 12.26).

Kardinalskalierte abhängige Variable

Zeitreihenanalyse nach Box-Jenkins Die Zeitreihenanalyse (▶ Abschn. 12.5.4) ist auf Einzelfalluntersuchungen anwendbar, sofern die Einzelfallzeitreihe **mehr als 50 Messungen** umfasst. Als vergleichsweise voraussetzungsarmes Verfahren vermittelt die Zeitreihenanalyse nach dem Box-Jenkins-Modell Einblicke in die seriellen Abhängigkeitsstrukturen und periodischen Regelmäßigkeiten der Daten, sie überprüft direkte oder zeitlich versetzte Wirkungen von Interventionen auf die abhängige Variable und gestattet die Überprüfung spezieller Trendhypothesen. Daten aus Multiple-Baseline-Plänen, in denen mehrere Zeitreihen gleichzeitig anfallen, lassen sich mit multiplen Transferfunktionsmodellen erschöpfend auswerten. Es sei jedoch nicht verschwiegen, dass das erfolgreiche Arbeiten mit der Box-Jenkins-Zeitreihenanalyse erhebliche Vorkenntnisse und viel Routine voraussetzt.

Randomisierungstest Randomisierungstests sind auf Einzelfallzeitreihen mit weniger als 50 Messungen anwendbar. Die Signifikanztestgruppe der Randomisierungstests zählt zu den nonparametrischen (verteilungsfreien) Verfahren, die auch für kleine Stichproben geeignet sind und keine Anforderungen an die Verteilung der betrachteten Variablen in der Population stellen, also z. B. keine Normalverteilung, keine Varianzhomogenität und keine unabhängigen Messungen verlangen. Im Vergleich zur Zeitreihenanalyse sind Randomisierungstests wesentlich einfacher anwendbar, sie stellen allerdings auch eine weniger erschöpfende Auswertungsstrategie dar (Levin et al., 1978). Bei Randomisierungstests wird die vorliegende Stichprobe als Ausgangsmaterial verwendet und aus ihr werden (ohne Zurücklegen) alle möglichen Messwertkombinationen konstruiert, wobei jede Kombination nur einmal auftritt. Das empirische Stichprobenergebnis wird im Lichte aller möglichen anderen aus den Daten generierbaren Ergebniskombinationen bewertet, indem man die Wahrscheinlichkeit bestimmt, mit der das gefundene (oder ein extremeres) Ergebnis auftritt. Getestet wird meist gegen die inhaltliche Nullhypothese, dass das gefundene Ergebnis keinen systematischen Effekt darstellt, sondern ein reines Zufallsprodukt ist. Wenn das gefundene (oder ein extremeres) Ergebnis im Kontext aller möglichen Kombinationen sehr unwahrscheinlich ist ($p < 5\,\%$), dann gilt das Ergebnis als statistisch si-

Tabelle 12.26 Signifikanztests für Einzelfallhypothesen

Signifikanztests für Einzelfallhypothesen	Hypothesen
Kardinalskalierte abhängige Variable	
Zeitreihenanalyse nach Box-Jenkins (für einzelne Messwerte)	Alternativhypothese, dass es in Folge der Behandlung unmittelbar oder zeitlich versetzt zu Wirkungen auf die abhängige Variable kommt (inkl. spezieller Trendhypothesen), für Einzelfallzeitreihen ab 50 Messungen
Randomisierungstest (für Phasenmittelwerte)	Alternativhypothese, dass Unterschiede zwischen zusammengefassten A- und B-Phasen nicht zufällig, sondern systematisch ausfallen (z. B. H_1: $A < B$), ab 8 Phasenmittelwerten
Trendtest (für Phasenmittelwerte)	Alternativhypothese, dass die Reihenfolge der einzelnen A- und B-Phasen einem monotonen Trend folgt und diesem nicht zuwiderläuft (z. B. H_1: $A_1 < A_2 < B_1 < B_2$), ab 4 Phasenmittelwerten
Permutationstest (für Phasenmittelwerte): – Exakter Permutationstest (weniger als 8 Phasenmittelwerte) – Asymptotischer Permutationstest (ab 8 Phasenmittelwerte)	Alternativhypothese, dass die Reihenfolge der einzelnen A- und B-Phasen einem linearen (oder einem anderen polynomialen) Trend folgt und diesem nicht zuwiderläuft (z. B. H_1: $A_1 < A_2 < B_1 < B_2$, wobei sich die Mittelwerte um jeweils mindestens den Wert 10 erhöhen)
Nominalskalierte abhängige Variable	
Dichotome abhängige Variable	
Iterationshäufigkeitstest: – Exakter Iterationshäufigkeitstest – Asymptotischer Iterationshäufigkeitstest (ab n_1 und $n_2 > 30$)	Alternativhypothese, dass der Wechsel zwischen dem Auftreten von zwei Ausprägungen eines Merkmals nicht zufällig, sondern systematisch erfolgt
Rangsummentest (entspricht dem U-Test): – Exakter Rangsummentest – Asymptotischer Rangsummentest (ab n_1 oder $n_2 > 25$)	Alternativhypothese, dass im Sinne eines monotonen Trends eine von zwei Merkmalsausprägungen im Verlauf der Zeit immer häufiger auftritt
Polytome abhängige Variable	
Multipler Iterationshäufigkeitstest: – Exakter multipler Iterationshäufigkeitstest – Asymptotischer multipler Iterationshäufigkeitstest (ab $N > 13$)	Alternativhypothese, dass der Wechsel zwischen dem Auftreten der k verschiedenen Merkmalsausprägungen nicht zufällig, sondern systematisch erfolgt
Trendtests (z. B. von Jonckheere)	Alternativhypothese, dass der Wechsel zwischen dem Auftreten der k verschiedenen Merkmalsausprägungen einem monoton steigenden oder fallenden Trend entspricht (und nicht zufällig oder trendkonträr verläuft)

gnifikant. Randomisierungstests arbeiten im Unterschied zu klassischen parametrischen Signifikanztests nicht mit theoretischen Prüfverteilungen (z. B. t-, X^2-, F-Verteilung), sondern mit empirisch jeweils aus den Stichprobendaten generierten Verteilungen. Da man dafür aus der Ursprungsstichprobe alle möglichen Ergebnisse konstruiert, spricht man auch von einem Resampling-Verfahren (► Abschn. 12.2.1).

Die serielle Abhängigkeit von Einzelfalldaten verbietet es, diese wie Realisierungen von unabhängigen Zufallsvariablen zu behandeln. Die Abhängigkeit der einzelnen Messungen ist jedoch für praktische Zwecke zu vernachlässigen, wenn für mehrere Messungen der Zeitreihe jeweils zusammenfassende Statistiken, wie z. B. Mittelwerte, berechnet werden. Bei der Analyse von Interventionseffekten bietet es sich an, die Einzelmessungen verschiedener A- und B-Phasen zu Mittelwerten zusammenzufassen. Wenn z. B. in einem A-B-A-Plan pro Phase 15 Messungen vorliegen, stehen für die Hypothesenüberprüfung statt der 45 abhängigen Einzelmessungen drei weitgehend unabhängige **Phasenmittelwerte** zur Verfügung. Bei mäßiger Autokorrelation 1. Ordnung kann man davon ausgehen, dass Phasenmittelwerte, die mindestens auf **jeweils 10 Einzelmessungen beruhen, praktisch voneinander unabhängig** sind (Levin et al., 1978, S. 179, Tab. 3.1; bei der Auslegung dieser Tabelle folgen wir ei-

ner Einschätzung von Glass, Willson, & Gottman, 1975, nach der für die meisten sozialwissenschaftlichen Zeitreihen Autokorrelationen 1. Ordnung im Bereich $r \leq .50$ typisch sind).

Um das Problem der seriellen Abhängigkeit von Einzelmessungen in Einzelfalluntersuchungen zu umgehen, kann man Einzelmessungen z. B. zu *Phasenmittelwerten* zusammenfassen, die bei mindestens 10 Messwerten pro Phase in der Regel nahezu unabhängig sind.

Beispiel: Angenommen, eine Logopädin behandelt ein Kind mit schweren Sprachstörungen und möchte die Bedeutung kleiner Belohnungen für die Therapie dieses Kindes mit einem *A*-*B*-*A*-*B*-Plan überprüfen. Sie stellt 15 Blöcke von jeweils 10 schwierig auszusprechenden Wörtern zusammen und bittet das Kind, diese Wörter in der ersten *A*-Phase nachzusprechen. Für jeden Block wird die Anzahl richtig wiederholter Wörter notiert und die durchschnittliche Zahl korrekt ausgesprochener Wörter über alle Blöcke errechnet (ein Block entspricht damit einer Messung, d. h. pro Phase werden 15 Messwerte erhoben und zu einem Phasenmittelwert verrechnet).

Für die erste *A*-Phase möge sich ein Durchschnittswert von 2 ergeben haben. In der ersten *B*-Phase erhält das Kind nach jedem Block in Abhängigkeit von der Anzahl der richtig gesprochenen Wörter Belohnungen. Es resultiert ein Durchschnittswert von 7 richtigen Wörtern. In den beiden folgenden Phasen erreicht das Kind im Durchschnitt 3 richtige Wörter für die zweite *A*-Phase und 8 richtige Wörter für die zweite *B*-Phase. Damit führt der *A*-*B*-*A*-*B*-Plan insgesamt zu den **Phasenmittelwerten A_1: 2, B_1: 7, A_2: 3 und B_2: 8**.

Unter der Annahme, die Belohnungen seien wirkungslos (Nullhypothese), sind die Unterschiede zwischen den Phasen auf Zufälligkeiten zurückzuführen (von Wiederholungs- bzw. Übungseffekten wollen wir hier absehen; diese wären durch die Verwendung verschiedener Wörter in den einzelnen Phasen auszuschalten). Jeder dieser 4 Mittelwerte hätte bei Gültigkeit der H_0 in jeder Phase auftreten können (Tab. 12.27 stellt die möglichen Ergebnisse gemäß Nullhypothese dar). Fassen wir jeweils zwei gleiche Phasen zusammen, resultieren bei Gültigkeit der H_0 gemäß Binomialkoeffizient die folgenden $\binom{4}{2} = \frac{4 \cdot 3}{2 \cdot 1} = 6$ gleichwahrscheinlichen Kombinationen.

Bei Gültigkeit der Nullhypothese tritt jede dieser 6 Kombinationen mit gleicher Wahrscheinlichkeit ($p = 1/6$) auf. (Wir beziehen uns nur auf die beobachteten Mittelwerte. Über mögliche andere Mittelwerte, die in der Untersuchung auch hätten auftreten können, werden keine Aussagen gemacht.)

Tabelle 12.27 Zusammenfassung von Einzelwerten zu Phasenmittelwerten unter Annahme der Gültigkeit der Nullhypothese (d. h. Zuordnung aller Phasenmittelwert-Kombinationen gleichermaßen zu *A*- und *B*-Phasen)

	A-Phasen	*B*-Phasen
1. Kombination	$2 + 3 = 5$	$7 + 8 = 15$
2. Kombination	$2 + 7 = 9$	$3 + 8 = 11$
3. Kombination	$2 + 8 = 10$	$3 + 7 = 10$
4. Kombination	$3 + 7 = 10$	$2 + 8 = 10$
5. Kombination	$3 + 8 = 11$	$2 + 7 = 9$
6. Kombination	$7 + 8 = 15$	$2 + 3 = 5$

Nehmen wir zunächst an, die eingangs aufgestellte Nullhypothese (kein Unterschied zwischen den Phasen: H_0: $A = B$) soll zweiseitig getestet werden, d. h., die Alternativhypothese bzw. H_1 lautet „$A \neq B$“ bzw. „$A > B$ oder $B > A$“. Um zu beurteilen, ob das gefundene Ergebnis eher mit der H_0 oder der H_1 zu vereinbaren ist, werden alle Ergebnisse gemäß H_1 betrachtet, die dem gefundenen oder einem extremeren Ergebnis entsprechen. Das sind im vorliegenden Fall die Kombinationen 1 ($A = 5$, $B = 15$) und 6 ($A = 15$, $B = 5$), die zusammen mit einer Wahrscheinlichkeit von $2/6 = 1/3$ auftreten, d. h., die Wahrscheinlichkeit dieser Ergebnisse bei Gültigkeit der H_0 (Irrtums-/α-Fehler-Wahrscheinlichkeit p, ▶ Abschn. 12.5.1) beträgt $p = .33$. Eine Irrtumswahrscheinlichkeit von $p = .33$ für das gefundene oder ein extremeres Ergebnis liegt weit über den üblichen Signifikanzgrenzen (α-Fehler-Niveau) von 5 % bzw. 1 % ($\alpha = .05$ bzw. $\alpha = .01$), d. h., das Ergebnis ist **statistisch nicht signifikant**.

Bevor wir uns angesichts eines statistisch nicht-signifikanten Ergebnisses für die H_0 – die Belohnungen haben keinen Effekt – entscheiden, müssen wir noch sicherstellen, dass der Test über **ausreichende Teststärke** verfügte, um überhaupt ein signifikantes Ergebnis liefern zu können. Dies ist hier nicht der Fall: Bei zweiseitigem Test kann ein *A*-*B*-*A*-*B*-Plan (und natürlich auch jeder Plan mit weniger als 4 Phasen) im Randomisierungstest niemals zu einem Ergebnis führen, dessen Irrtumswahrscheinlichkeit kleiner als 33.3 % ist, weil eben bei vier Phasenmittelwerten nur 6 Kombinationen möglich sind und das Auftreten der zwei extremsten Kombinationen (zweiseitiger Test) zwangsweise zu einer Irrtumswahrscheinlichkeit von $p = 2/6$ führt. Dasselbe gilt für den einseitigen Test einer gerichteten Hypothese, die hier eigentlich angebracht ist: Die Logopädin hat hinreichend

Gründe anzunehmen, dass Belohnungen das Sprechverhalten des Kindes verbessern und wird deshalb eine **gerichtete** $\boldsymbol{H_1}$: $A < B$ formulieren. In diesem Falle gibt es nur ein Ergebnis, das im Sinne der H_1 extremer oder gleich dem empirisch gefundenen Ergebnis ist, nämlich die 1. Kombination. Dieses Ergebnis tritt bei Gültigkeit der H_0 mit einer Wahrscheinlichkeit von $p = 1/6 = .16$ auf, d. h., die Irrtumswahrscheinlichkeit von 16 % für den einseitigen Test übersteigt noch immer das übliche Signifikanzniveau von 5 % bzw. 1 %. Auch dieses nichtsignifikante Ergebnis darf nicht als Beleg für die H_0 gewertet werden, sondern ist **nicht interpretierbar**, da dem Test die notwendige Teststärke fehlt.

Der Randomisierungstest hat bei einem **4-Phasen-Plan** generell **keine ausreichende Teststärke** und ist somit nicht aussagekräftig. Ein signifikantes Ergebnis ist erstmals dann möglich, wenn ein Untersuchungsplan mit **8-Phasen** vorliegt (z. B. ein A-B-A-B-A-B-A-B-Plan), so dass 8 Phasenmittelwerte resultieren. Die 4 A- und die 4 B-Phasen werden zusammengefasst, so dass $\binom{8}{4} = 70$ Kombinationen möglich sind, von denen die zwei extremsten (zweiseitiger Test) dann mit einer Wahrscheinlichkeit von $p = 2/70 = 2.8\,\%$ die Signifikanzschwelle von 5 % unterschreiten würden. Mit wachsender Zahl an Phasen im Untersuchungsplan (und somit an Messwerten) gewinnt der Randomisierungstest an Teststärke, allerdings steigt gleichzeitig der Aufwand für die Studie. Eine weitere Strategie zur Erhöhung der Teststärke besteht darin, anstelle einer einzelnen Person eine sehr kleine Stichprobe von z. B. drei oder vier einander sehr ähnlichen Personen zu untersuchen (Levy & Lemeshow, 1999).

Bei einer Einzelfalluntersuchung mit kardinalskalierter abhängiger Variable lässt sich die Alternativhypothese, dass Unterschiede zwischen zusammengefassten A- und B-Phasen nicht zufällig, sondern systematisch ausfallen (z. B. H_1: $A < B$), mit dem *Randomisierungstest* prüfen. Für ausreichende Teststärke sind dabei mindestens 8 Untersuchungsphasen bzw. mindestens 8 Phasenmittelwerte (jeweils gebildet aus mindestens zehn Messwerten) notwendig.

Trendtests Die gerichtete Alternativhypothese zu den Phasenmittelwerten A und B (H_1: $A < B$) sagt nichts über mögliche Unterschiede zwischen den beiden A-Phasen bzw. zwischen den beiden B-Phasen aus. Wenn wir jedoch davon ausgehen, dass die Sprechtherapie erfolgreich ist, ließe sich auch die weitergehende Hypothese rechtfertigen, dass $A_1 < A_2$ und dass $B_1 < B_2$, bzw. dass zusammengenommen $\boldsymbol{A_1 < A_2 < B_1 < B_2}$ gilt (d. h. die durchschnittliche Zahl der richtig gesprochenen Wörter steigt an). Diese **monotone Trendhypothese** gilt als bestätigt, wenn die 4 Phasenmittelwerte genau diese Reihenfolge aufweisen. Erfolgt die Verteilung der Mittelwerte auf die 4 Phasen gemäß der Nullhypothese zufällig, treten alle $4! = 24$ möglichen Reihenfolgen mit gleicher Wahrscheinlichkeit auf ($p = 1/24 = .042$). Entspricht – wie im Beispiel (A_1: 2 < A_2: 3 < B_1: 7 < B_2: 8) – die vorhergesagte Reihenfolge der empirisch gefundenen Reihenfolge, gilt die monotone Trendhypothese auf dem 5 %-Signifikanzniveau als bestätigt.

Man beachte, dass die H_0: „Die Reihenfolge der Mittelwerte ist zufällig“ theoretisch durch jede beliebige Reihenfolge auf dem $\alpha = 5\,\%$-Niveau verworfen wird, denn jede Reihenfolge tritt mit einer Wahrscheinlichkeit von $p = .042$ auf. Die Alternativhypothese impliziert jedoch eine abgestufte Treatmentwirkung der Form $A_1 < A_2 < B_1 < B_2$, d. h., jede hiervon abweichende Reihenfolge steht im Widerspruch zu dieser Alternativhypothese. Damit kann die H_0 nur mit dieser einen Reihenfolge verworfen werden. Tritt diese auf, dann unterschreitet die Irrtumswahrscheinlichkeit das Signifikanzniveau von 5 % – somit kann der Effekt prinzipiell statistisch signifikant werden bzw. verfügt der Test über ausreichende Teststärke.

Wie würde der Test ausfallen, hätte man die **Alternativhypothese weniger restriktiv** formuliert? Eine weniger restriktive H_1 kann schließlich inhaltlich sinnvoll sein: Zwar ist man sicher, dass die zweite Treatment-Phase wirksamer ist als die erste ($B_1 < B_2$) und dass unter Treatment-Bedingungen insgesamt mehr Wörter richtig gesprochen werden als unter Baseline-Bedingungen ($A < B$); über Unterschiede zwischen den beiden Baseline-Phasen will man jedoch vielleicht keine Vorhersagen machen. Damit umfasst die Alternativhypothese 2 Rangreihen, nämlich $A_1 < A_2 < B_1 < B_2$ und $A_2 < A_1 < B_1 < B_2$. Die Wahrscheinlichkeit, dass eine diese Alternativhypothese bestätigende Rangreihe auftritt, beträgt dann $p = 2/24 = .08$. Bei nur 4 Phasen hätte der Trendtest mit dieser weniger restriktiven Trendhypothese somit wiederum eine zu geringe Teststärke: Selbst ein hypothesenkonformes empirisches Ergebnis wäre nicht in der Lage, ein statistisch signifikantes Ergebnis zu produzieren.

Auch bei dieser Vorgehensweise gilt die H_1 von vorne herein (d. h. ohne jegliche statistische Berechnung) als nicht bestätigt, wenn empirisch eine Rangreihe auftritt, die nicht als Alternativhypothese vorhergesagt wurde (bei einer gerichteten H_1 ist die Hypothese generell – ohne weitere Berechnungen – bereits dann widerlegt, wenn das empirische Ergebnis der Hypothesenrichtung widerspricht). Ob diese Rangreihe nur geringfügig oder sehr deutlich von der oder den durch die Hypothesen vorhergesagten Rangreihen abweicht, ist hierbei unerheblich. Für

diese Entscheidungsstrategie ist es also ohne Belang, ob z. B. die Alternativhypothese H_1: $A_1 < A_2 < B_1 < B_2$ durch die Rangreihe $A_2 < A_1 < B_1 < B_2$ oder durch die Rangreihe $B_2 < B_1 < A_2 < A_1$ verworfen wird, obwohl letztere zur Alternativhypothese in deutlicherem Widerspruch steht als erstere.

Bei einer Einzelfalluntersuchung mit kardinalskalierter abhängiger Variable lässt sich die Alternativhypothese, dass die Reihenfolge der einzelnen *A*- und *B*-Phasen einem monotonen Trend folgt und diesem nicht zuwiderläuft (z. B. H_1: $A_1 < A_2 < B_1 < B_2$), mit dem *Trendtest* prüfen. Für ausreichende Teststärke sind dabei mindestens vier Untersuchungsphasen und somit mindestens vier Phasenmittelwerte (gebildet jeweils aus mindestens 10 Messwerten) notwendig.

Exakter Permutationstest Die Schwäche, dass der Grad der Abweichung von der H_1 keine Berücksichtigung findet, wird beseitigt, wenn man die gefundenen **Phasenmittelwerte mit allen Permutationen der möglichen Rangplätze gewichtet** und für jede Permutation die Summe der so gewichteten Mittelwerte berechnet. Der Permutationstest lässt sich als Spezialform des Randomisierungstests verstehen. Bezogen auf das Beispiel resultieren die in Tab. 12.28 wiedergegebenen $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ Produktsummenwerte (der Wert der 1. Permutation resultiert aus $1 \cdot 2 + 2 \cdot 7 + 3 \cdot 3 + 4 \cdot 8 = 57$, für die 2. Permutation ergibt sich $1 \cdot 2 + 2 \cdot 7 + 4 \cdot 3 + 3 \cdot 8 = 52$ etc.). Die kleinste Produktsumme (39) und die größte Produktsumme (61) sind fett gedruckt.

In Tab. 12.29 sind die Wahrscheinlichkeiten der nach ihrer Größe geordneten Produktsummen (*PS*) aufgeführt.

Es wird deutlich, dass die Rangreihe $A_1 < A_2 < B_2 < B_1$ (5. Permutation mit PS = 60) und die Rangreihe $A_2 < A_1 < B_1 < B_2$ (9. Permutation mit PS = 60) mit der vorhergesagten (und auch aufgetretenen) Rangreihe $A_1 < A_2 < B_1 < B_2$ (3. Permutation mit PS = 61) am wenigsten im Widerspruch stehen. Umfasst die Alternativhypothese nicht nur eine, sondern auch in diesem Sinne ähnliche Rangreihen, wird die H_1 angenommen, wenn die empirische Rangreihe zu denjenigen extremen Rangreihen zählt, die zusammengenommen bei Gültigkeit von H_0 eine Wahrscheinlichkeit $p \leq 5\,\%$ (1 %) haben. Bei 4 Phasen haben 2 Rangreihen bereits eine Irrtumswahrscheinlichkeit von $p > 5\,\%$ ($p = 2/24 = .083$), d. h., auch bei dieser Vorgehensweise muss die empirische Rangreihe exakt der vorhergesagten entsprechen ($p = 1/24 = .042 < .05$). Dies ändert sich natürlich, wenn mehr als 4 Phasen untersucht werden, weil dann wiederum die Teststärke des Verfahrens zunimmt.

Tabelle 12.28 Produktsummen aus Phasenmittelwerten und permutierten Rangplätzen. (Mod. nach Levin et al., 1978) Fettdruck: kleinste (39) und größte (61) Produktsumme

Permutation	Mittelwerte				Produktsumme
	2.00	7.00	3.00	8.00	(PS)
1	1	2	3	4	57
2	1	2	4	3	52
3	**1**	**3**	**2**	**4**	**61**
4	1	3	4	2	51
5	1	4	2	3	60
6	1	4	3	2	55
7	2	1	3	4	52
8	2	1	4	3	47
9	2	3	1	4	60
10	2	3	4	1	45
11	2	4	1	3	59
12	2	4	3	1	49
13	3	1	2	4	51
14	3	1	4	2	41
15	3	2	1	4	55
16	3	2	4	1	40
17	3	4	1	2	53
18	3	4	2	1	48
19	4	1	2	3	45
20	4	1	3	2	40
21	4	2	1	3	49
22	**4**	**2**	**3**	**1**	**39**
23	4	3	1	2	48
24	4	3	2	1	43

Die Gewichtung der Mittelwerte mit den Rangnummen 1 bis 4 impliziert die Hypothese gleicher Abstände zwischen den Phasenmittelwerten **(lineare Trendhypothese)**. Diese Gewichte können – wenn entsprechende Vorkenntnisse vorliegen – durch beliebige andere Gewichtszahlen ersetzt werden, die den hypothetisch vor-

Tabelle 12.29 Wahrscheinlichkeiten der Produktsummen aus Tab. 12.28

Produktsumme	Wahrscheinlichkeit
39	1/24
40	2/24
41	1/24
43	1/24
45	2/24
47	1/24
48	2/24
49	2/24
51	2/24
52	2/24
53	1/24
55	2/24
57	1/24
59	1/24
60	2/24
61	1/24

hergesagten Größenverhältnissen der Mittelwerte entsprechen (Einzelheiten hierzu erläutern Levin et al., 1978, S. 185, für ein Rechenbeispiel ▶ Nikotinentzug durch Selbstkontrolle. Die Überprüfung von Hypothesen in einer Einzelfalluntersuchung).

Asymptotischer Permutationstest Die Anzahl möglicher Permutationen wird mit wachsender Phasenzahl schnell sehr groß. Für 5 Phasen ergeben sich bereits $5! = 120$ verschiedene Abfolgen, d. h., eine auf dem $\alpha = 5\,\%$-Niveau bestätigte Alternativhypothese kann sechs im Sinne von Tab. 12.28 ähnliche Abfolgen umfassen. Bei Einzelfallunterschungen mit 6 Phasen (z. B. $ABABAB$) sind 720 Abfolgen der Phasenmittelwerte möglich; hier kann eine wenig restriktive Alternativhypothese aus 36 einander ähnlichen Abfolgen bestehen und – sofern ein Ergebnis davon empirisch auftritt – immer noch auf dem 5 %-Niveau bestätigt werden.

Für **mehr als 8 Phasen** (für 8 Phasen sind $8! = 40\,320$ Abfolgen möglich) geht die Wahrscheinlichkeitsverteilung der Produktsummen (PS) in eine Normalverteilung über (asymptotischer Test). Der Erwartungswert (Mittelwert) und die Varianz dieser normalverteilten Produktsummen sind berechenbar über folgende Formeln:

$$E(\mathrm{PS}) = \frac{1}{n} \cdot \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \bar{x}_{\mathrm{i}} \right) \cdot \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} y_{\mathrm{i}} \right) \tag{12.44}$$

$$\mathrm{VAR}(\mathrm{PS}) = \frac{1}{n-1} \cdot \left[\sum_{\mathrm{i}=1}^{\mathrm{n}} (\bar{x}_{\mathrm{i}} - \bar{\bar{x}})^2 \right] \cdot \left[\sum_{\mathrm{i}=1}^{\mathrm{n}} (y_{\mathrm{i}} - \bar{y})^2 \right] \tag{12.45}$$

mit

$\bar{x}_{\mathrm{i}}$ = Mittelwert der i-ten Phase
$\bar{\bar{x}}$ = Mittelwert der Phasenmittelwerte
y_{i} = Gewicht des i-ten Phasenmittelwertes
n = Anzahl der Phasen
$\bar{y}$ = durchschnittliches Gewicht.

Eine empirisch gefundene Produktsumme

$$\mathrm{PS} = \sum_{\mathrm{i}=1}^{\mathrm{n}} y_{\mathrm{i}} \bar{x}_{\mathrm{i}} \tag{12.46}$$

lässt sich dann nach der schon bekannten z-Transformation (Gl. (12.15) aus ▶ Abschn. 12.4.2)

$$z = \frac{\mathrm{PS} - E(\mathrm{PS})}{\sqrt{\mathrm{VAR}(\mathrm{PS})}} \tag{12.47}$$

in einen empirischen z-Wert der Standardnormalverteilung (Tab. 1, ▶ Webangebot zu diesem Buch) überführen. Schneidet dieser bei einseitigem Test von der Standardnormalverteilungsfläche weniger als 5 % ($z = 1.65$) bzw. weniger als 1 % ($z = 2.33$) ab, gilt die Alternativhypothese, die durch die Wahl der Gewichte y_{i} festgelegt ist, als bestätigt. Dieser Ansatz wird in ▶ Nikotinentzug durch Selbstkontrolle. Die Überprüfung von Hypothesen in einer Einzelfalluntersuchung anhand eines Zahlenbeispiels verdeutlicht.

Die bisherigen Ausführungen zeigen, dass derselbe empirische Befund je nach Art der Hypothese sowie Teststärke des Tests statistisch signifikant oder nicht-signifikant sein kann. Je präziser eine Hypothese das beschreibt, was empirisch auch eintritt, desto größer ist die Wahrscheinlichkeit, dass dieses Ergebnis statistisch signifikant wird. Allerdings wächst mit zunehmender Präzision der Hypothese auch die Anzahl möglicher Ergebnisse, die der Hypothese widersprechen. Diesen Sachverhalt haben wir bereits beim Vergleich eines einseitigen Tests mit einem zweiseitigen Test im Kontext der Logik des klassischen Signifikanztests kennengelernt (▶ Abschn. 12.5.1).

Hier wird nun die Notwendigkeit, Hypothesen **vor** der Datenerhebung zu formulieren, noch deutlicher. Es ist

Nikotinentzug durch Selbstkontrolle. Die Überprüfung von Hypothesen in einer Einzelfalluntersuchung

Ein starker Raucher will zeigen, dass es ihm gelingt, seinen Zigarettenkonsum durch Selbstdisziplin („bewusstes" Rauchen) deutlich zu reduzieren. Er beabsichtigt, abwechselnd 14 Tage „normal" zu rauchen (Baselinephase) und 14 Tage „bewusst" zu rauchen („Therapiephase") mit insgesamt 5 Baselinephasen und 5 Therapiephasen. Während dieser insgesamt 140 Tage wird täglich sorgfältig die Anzahl geraucht Zigaretten registriert.

Dieses Material soll drei Hypothesen unterschiedlicher Präzision überprüfen. (Selbstverständlich stellt man üblicherweise nur diejenige Hypothese auf, die die vermutete Veränderung am präzisesten wiedergibt. Zu Demonstrationszwecken sei jedoch im Folgenden dasselbe Material zur Überprüfung von drei unterschiedlich genauen Hypothesen verwendet.)

1. Hypothese H_1: In den Baselinephasen wird mehr geraucht als in den Therapiephasen ($A > B$).
2. Hypothese H_2: Der Zigarettenkonsum sinkt sowohl in den Baselinephasen als auch in den Therapiephasen kontinuierlich; dennoch wird in der letzten Baselinephase noch mehr geraucht als in der ersten Therapiephase ($A_1 > A_2 > A_3 > A_4 > A_5 > B_1 > B_2 > B_3 > B_4 > B_5$).
3. Hypothese H_3: Wie 2., jedoch werden in jeder Therapiephase mindestens 10 Zigaretten weniger geraucht als in der jeweils vorangegangenen Baselinephase (für die Abfolge $A_1 - B_1 - A_2 - B_2 - A_3 - B_3$ etc. wären dann z. B. die Gewichte 15, 5, 14, 4, 13, 3, 12, 2, 11, 1 zu verwenden).

Für die 10 Phasen registriert der Raucher die folgenden Tagesdurchschnitte:

- 1. Baselinephase A_1: 45 Zigaretten
- 1. Therapiephase B_1: 25 Zigaretten
- 2. Baselinephase A_2: 41 Zigaretten
- 2. Therapiephase B_2: 28 Zigaretten
- 3. Baselinephase A_3: 38 Zigaretten
- 3. Therapiephase B_3: 21 Zigaretten
- 4. Baselinephase A_4: 32 Zigaretten
- 4. Therapiephase B_4: 19 Zigaretten
- 5. Baselinephase A_5: 33 Zigaretten
- 5. Therapiephase B_5: 14 Zigaretten

Auf die A-Phasen entfallen damit 189 Zigaretten (Summe der fünf A-Phasen-Mittelwerte) und auf die B-Phasen 107 Zigaretten (Summe der fünf B-Phasen-Mittelwerte). Diese Aufteilung der Mittelwerte auf A- und B-Phasen ist eine unter $\binom{10}{5} = \frac{10\cdot9\cdot8\cdot7\cdot6}{5\cdot4\cdot3\cdot2\cdot1} = 252$ möglichen Aufteilungen. Da keine dieser Aufteilungen für die H_1 günstiger wäre als die empirisch ermittelte (bei jeder anderen Aufteilung resultiert ein kleinerer Unterschied zwischen A und B), kann die Alternativhypothese $A > B$ mit einer Irrtumswahrscheinlichkeit von $p = 1/252 = .004 \leq \alpha = .05$ akzeptiert werden.

Bei mehr als 8 Phasen ($n > 8$) wählen wir zur Überprüfung der 2. Hypothese die Normalverteilungsapproximation nach Gl. (12.47). Da in dieser Hypothese keine Angaben über die Größe des Unterschiedes zweier Phasen gemacht wurden, wählen wir als Gewichte y_i die einfachsten Zahlen, die dem in der Hypothese behaupteten monotonen Trend genügen. Dies sind die Zahlen 1, 2 ... 10; sie repräsentieren einen linearen Trend. Diejenige Phase, die hypothesengemäß den höchsten Wert erzielen sollte, erhält das Gewicht 10, die Phase, für die man den zweithöchsten Wert erwartet, das Gewicht 9 etc.

Für die Produktsumme (PS) ergibt sich nach Gl. (12.46):

$$\begin{aligned}\text{PS} &= 10\cdot45 + 9\cdot41 + 8\cdot38 + 7\cdot32 \\ &\quad + 6\cdot33 + 5\cdot25 + 4\cdot28 + 3\cdot21 \\ &\quad + 2\cdot19 + 1\cdot14 \\ &= 1\,897.\end{aligned}$$

Nach Gl. (12.44) ermitteln wir den Erwartungswert der Produktsumme:

$$\begin{aligned}E(\text{PS}) &= \frac{1}{n}\cdot\left(\sum_{i=1}^{n}\bar{x}_i\right)\cdot\left(\sum_{i=1}^{n}y_i\right) \\ &= \frac{1}{10}\cdot296\cdot55 = 1\,628\end{aligned}$$

und nach Gl. (12.45) die Varianz der Produktsumme:

$$\begin{aligned}\text{VAR(PS)} &= \frac{1}{n-1}\cdot\left[\sum_{i=1}^{n}(\bar{x}_i-\bar{\bar{x}})^2\right]\cdot\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \\ &= \frac{1}{9}\cdot908.4\cdot82.5 \\ &= 8\,327.\end{aligned}$$

Der resultierende empirische z-Wert lautet also:

$$z = \frac{1\,897 - 1\,628}{\sqrt{8\,327}} = 2.95.$$

Dieser empirische z-Wert schneidet nach der Tabelle zur Standardnormalverteilung (Tab. 1 ▶ Webangebot zu diesem Buch) $.0051 = 0.51\,\%$ von der Standardnormalverteilungsfläche ab, d. h., die Irrtumswahrscheinlichkeit ist kleiner als das übliche Signifikanzniveau von 5 %, so dass wir uns für die Alternativhypothese entscheiden.

Die 3. Hypothese überprüfen wir in gleicher Weise. Es werden lediglich statt der Zahlen 1–10 die in der Hypothese festgelegten Gewichte y_i eingesetzt. Es resultieren:

$$\begin{aligned}\text{PS} &= 15\cdot45 + 5\cdot25 + 14\cdot41 \\ &\quad + 4\cdot28 + \ldots + 11\cdot33 + 1\cdot14 \\ &= 2\,842,\end{aligned}$$

$E(\text{PS}) = 2\,368$ und $\text{VAR(PS)} = 27\,252$ und damit $z = 2.87$. Dieser Wert schneidet $.0065 = 0.65\,\%$ der Standardnormalverteilungsfläche ab, d. h., auch die spezifizierte Trendhypothese hat eine Irrtumswahrscheinlichkeit, die das Signifikanzniveau unterschreitet, so dass das Ergebnis signifikant ausfällt und man sich für die Annahme der Alternativhypothese entscheidet.

Offensichtlich sind die Veränderungen im Zigarettenkonsum nicht durch Zufall erklärbar:

- In allen Therapiephasen zusammengenommen wird signifikant weniger geraucht ($p = .004 \leq \alpha = .05$) als in den Baselinephasen (Hypothese 1).
- Zudem folgen die Mittelwerte der Phasen gemäß Hypothese 2 überzufällig einem linearen Trend $A_1 > A_2 > A_3 > A_4 > A_5 > B_1 > B_2 > B_3 > B_4 > B_5$ abnehmender Zigarettenzahl ($p = .0051 \leq \alpha = .05$)
- Darüber hinaus folgen die Mittelwerte der Phasen den in Hypothese 3 spezifizierten Effektgrößen von jeweils um die Zahl 10 reduziertem Zigarettenkonsum ($p = .0065 \leq \alpha = .05$).

nahezu unmöglich, eine Hypothese statistisch zu widerlegen, die erst nach Vorliegen der Daten dem Ergebnis entsprechend formuliert wurde. Entsprechende **Pseudo-Signifikanztests**, bei denen ein induktiv gefundener Effekt im Nachhinein als Bestätigung einer Hypothese dargestellt wird, sind deswegen wissenschaftlich unzulässig (► Abschn. 12.3.2).

Bei einer Einzelfalluntersuchung mit kardinalskalierter abhängiger Variable lässt sich die Alternativhypothese, dass die Reihenfolge der einzelnen A- und B-Phasenmittelwerte einem linearen (oder einem anderen polynomialen) Trend folgt und diesem nicht zuwiderläuft (z. B. H_1: $A_1 < A_2 < B_1 < B_2$), mit dem Permutationstest prüfen. Bei Untersuchungsplänen mit weniger als 8 Phasen wird der *exakte Permutationstest*, ab 8 Phasen der *asymptotische Permutationstest* (über die Normalverteilung) eingesetzt.

Nachzutragen bleibt, dass der hier behandelte Permutationstest auf der Annahme beruht, Baseline- und Interventionsphasen folgten zufällig aufeinander. Dies ist bei Einzelfallplänen üblicherweise nicht der Fall, denn A- und B-Phasen folgen einem bestimmten Muster (z. B. A-B-A-B-Plan). In Ermangelung voraussetzungsärmerer und dennoch teststarker Auswertungsverfahren für Einzelfalldaten stellen auf systematische Abfolgen angewandte Randomisierungstests jedoch eine angemessene Näherungslösung dar (Edgington, 1975, 1980, 1995, sowie Levin et al., 1978; weitere Anregungen zur statistischen Analyse quantitativer Einzelfalldaten geben Bortz, Lienert, & Boehnke, 2008, Kap. 11; zum Thema Randomisierungs- bzw. Permutationstests sei zusätzlich auf Good, 2000, verwiesen). Verbesserte Ansätze zur statistischen Analyse von Multiple-Baseline-Plänen oder A-B-A-B-...-Plänen werden von Koehler und Levin (1998) vorgestellt. In der Praxis wird im Rahmen der Single-Subject-Forschung auch häufig mit **grafischen Auswertungsmethoden** gearbeitet. Nähere Informationen zu diesen grafischen Verfahren sind z. B. Callahan & Barisa (2005) und Kazdin (2010) zu entnehmen.

▪ Nominalskalierte abhängige Variable

In Einzelfallanalysen fallen gelegentlich wiederholte Messungen eines dichotomen Merkmales (z. B. Symptom vorhanden – nicht vorhanden) bzw. eines mehrstufigen nominalen Merkmals (z. B. Art des Symptoms: z. B. Spannungskopfschmerz, Clusterkopfschmerz, Migräne) an. Tritt das in einer Untersuchung interessierende Ereignis häufig auf, empfiehlt es sich, Beobachtungszeiträume (Baseline- und Interventionsphasen) so festzulegen, dass mehrere Ereignisse in eine Phase fallen. Man erhält dann für die einzelnen Phasen unterschiedliche Häufigkeiten, die wie eine quantitative Zeitreihe (► Abschn. 12.5.4) behandelt werden.

Im Folgenden gehen wir davon aus, dass solche Zusammenfassungen nicht möglich oder sinnvoll sind, so dass bei einem dichotomen Merkmal eine Abfolge von Merkmalsalternativen (z. B. 001010 etc.) und bei einem mehrstufigen nominalen Merkmal eine Abfolge von Merkmalskategorien (z. B. $AACDBBCAB$ etc.) zu untersuchen sind. Dazu werden statistische Verfahren für dichotome und polytome nominalskalierte abhängige Variablen vorgestellt.

▪ Dichotome abhängige Variablen

Iterationshäufigkeitstest Bezeichnen wir die Merkmalsalternativen eines dichotomen Merkmals mit 0 und 1, sind z. B. die beiden folgenden Zeitreihen denkbar: 00001111 und 01010101. Beide Abfolgen scheinen nicht zufällig zustande gekommen zu sein. In der ersten Abfolge treten zunächst nur Nullen und dann nur Einsen auf, und in der zweiten Abfolge wechseln sich Nullen und Einsen regelmäßig ab. Weder die erste noch die zweite Abfolge stimmt mit unserer Vorstellung über eine zufällige Abfolge (die etwa für die Ereignisse „Kopf“ und „Zahl“ bei wiederholten Münzwürfen auftritt) überein. Für eine zufällige Durchmischung von Nullen und Einsen wechseln die Zahlen in der ersten Abfolge zu selten und in der zweiten Abfolge zu häufig.

Die Häufigkeit des Wechsels zwischen Nullen und Einsen in einer Zeitreihe bezeichnen wir als **Iterationshäufigkeit**. Nach dieser Definition weist die erste Zeitreihe 2 und die zweite Zeitreihe 8 Iterationen auf. Die erste Hypothese, die wir hier ausführlicher behandeln wollen, bezieht sich auf die Häufigkeit der Iterationen in Zeitreihen dichotomer Merkmale. Der Nullhypothese (zufällige Abfolge) steht die ungerichtete Alternativhypothese gegenüber, dass die Anzahl der Iterationen entweder zu groß oder zu klein ist. Diese Hypothese überprüft der Iterationshäufigkeitstest (Stevens, 1939), den das folgende Beispiel näher erläutert:

Exakter Iterationshäufigkeitstest Untersucht wird ein Kind, das unter Bettnässen leidet. Es soll überprüft werden, ob symptomfreie Nächte (0 = kein Einnässen) und Nächte mit Symptom (1 = Einnässen) zufällig aufeinander folgen (H_0) oder ob sich längere symptomfreie Phasen mit längeren Symptomphasen abwechseln (H_1), was dafür spräche, dass die das Einnässen auslösenden Faktoren nicht zufällig, sondern phasenweise wirksam sind. Die letztgenannte (gerichtete) Hypothese wird an folgender Zeitreihe von $n = 32$ Beobachtungen über-

prüft:

$$\overline{000}\,1\,\overline{00}\,\underline{11111}\,\overline{0000}\,1\,\bar{0}\,\underline{111}\,\overline{000000}\,1\,\overline{000}\,\underline{11}\,.$$

Insgesamt zählen wir $N_1 = 19$ Nächte ohne Symptom und $N_2 = 13$ Nächte mit Symptom. Die Anzahl der **Iterationen** (Zustandswechsel, markiert als über- bzw. unterstrichene Zahlengruppen) beläuft sich auf $r = 12$. Gemäß der H_0 (zufällige Abfolge) erwarten wir

$$\mu_r = 1 + \frac{2 \cdot N_1 \cdot N_2}{N} \qquad (12.48)$$
$$= 1 + \frac{2 \cdot 19 \cdot 13}{32} = 1 + 15.4 \approx 16$$

Iterationen, d. h., die Zahl empirischer Iterationen liegt hypothesengemäß unter der Zufallserwartung. Ob sie auch statistisch bedeutsam von ihr abweicht, entscheiden wir anhand der im ▶ Webangebot zu diesem Buch aufgeführte Tab. 6 Iterationshäufigkeitstest (auf die Berechnung der exakten Wahrscheinlichkeiten wollen wir verzichten. Hierfür findet sich eine ausführliche Anleitung bei Bortz et al., 2008, S. 545ff.). Danach dürfen bei einem Signifikanzniveau von $\alpha = 5\,\%$ höchstens $r = 11$ Iterationen auftreten. Diese Zahl wird von der Anzahl der Iterationen in der empirischen Zeitreihe überschritten, d. h., die H_0 kann nicht verworfen werden. Die aufgetretenen Phasen mit oder ohne Symptom sind also nicht überzufällig lang.

Asymptotischer Interationshäufigkeitstest Für $N_1 > 30$ und $N_2 > 30$ folgt die Prüfgröße r einer Normalverteilung mit dem in Gl. (12.48) angegebenen Erwartungswert (Mittelwert) und der Streuung

$$\sigma_r = \sqrt{\frac{2 \cdot N_1 \cdot N_2 \cdot (2N_1 \cdot N_2 - N)}{N^2 \cdot (N-1)}}. \qquad (12.49)$$

Der folgende z-Wert kann anhand der Standardnormalverteilungstabelle (▶ Webangebot zu diesem Buch) zufallskritisch bewertet werden:

$$z = \frac{r - \mu_r}{\sigma_r}. \qquad (12.50)$$

Obwohl das Beispiel die Erfordernisse einer brauchbaren Normalverteilungsapproximation nicht erfüllt ($N_1 = 19$, $N_2 = 13$), soll der in Gl. (12.50) angegebene asymptotische Test auch anhand der oben erwähnten Zahlen verdeutlicht werden.

Wir ermitteln

$$\sigma_r = \sqrt{\frac{2 \cdot 19 \cdot 13 \cdot (2 \cdot 19 \cdot 13 - 32)}{32^2 \cdot (32-1)}} = 2.68$$

und

$$z = \frac{12 - 16}{2.68} = -1.49.$$

Diesem z-Wert entspricht gemäß Standardnormalverteilungstabelle (▶ Webangebot zu diesem Buch) bei einseitigem Test eine Irrtumswahrscheinlichkeit von $6.81\,\% > 5\,\%$ (nicht signifikant).

Iterationshäufigkeitstest – Bei einer Einzelfalluntersuchung mit dichotomer abhängiger Variable lässt sich die Alternativhypothese, dass der Wechsel zwischen dem Auftreten beider Merkmalsausprägungen nicht zufällig, sondern systematisch erfolgt, mit dem Iterationshäufigkeitstest prüfen, wobei entweder der **exakte Iterationshäufigkeitstest** oder – bei größeren Messwertreihen (n_1 und $n_2 > 30$) – der **asymptotische Iterationshäufigkeitstest** (über die Normalverteilung) zu nutzen sind.

Rangsummentest Eine zweite, auf Zeitreihen binärer Daten bezogene Hypothese könnte lauten, dass die Zeitreihe einem **monotonen Trend** folgt, bzw. dass – auf das Beispiel bezogen – die Häufigkeit des Einnässens im Verlauf der Zeit abnimmt. Diese Hypothese überprüfen wir nach Meyer-Bahlburg (1969, zit. nach Lienert, 1978, S. 263f.) mit dem Rangsummentest.

Exakter Rangsummentest Hierzu nummerieren wir die untersuchten Nächte und notieren die Nummern des seltenen Ereignisses, also im Beispiel die Nummern derjenigen Nächte, in denen eingenässt wurde. Diese lauten 4, 7, 8, 9, 10, 11, 16 etc. Die Summe dieser Zahlen beträgt $T = 212$ und ihre Anzahl $N_1 = 13$. Je kleiner diese Summe ist, desto deutlicher wird unsere Hypothese eines monoton fallenden Trends für das seltenere Ereignis bestätigt (umgekehrt erwarten wir bei einem monoton steigenden Trend einen höheren Wert für T). Folgen die 0/1-Werte keinem Trend, sondern einer Zufallsabfolge, erwarten wir für T

$$\mu_T = \frac{N_1 \cdot (N+1)}{2} \qquad (12.51)$$
$$= \frac{13 \cdot (32+1)}{2} = 214.5.$$

Der beobachtete T-Wert ist kleiner als μ_T und spricht damit der Tendenz nach für unsere Hypothese. Ob T auch signifikant von μ_T abweicht, entscheiden wir anhand der im ▶ Webangebot zu diesem Buch wiedergegebenen Tab. 7 zum Rangsummentest. Für $N_1 = 13$, $N_2 = 19$ und $\alpha = 5\,\%$ lesen wir dort den Wert $T_{krit} = 171$ ab. Dieser Wert darf vom empirischen T-Wert nicht überschritten werden. Unser T-Wert ist erheblich größer als T_{krit}, d. h., wir müssen die H_0 beibehalten. Die Veränderungen der Symptomhäufigkeit folgen offenbar keinem abfallenden Trend. Idealerweise müsste anlässlich des nicht-signifikanten Ergebnisses wiederum die Teststärke bestimmt werden, um sicherzugehen, dass das Ergebnis

wirklich interpretierbar ist. Im Bereich der nonparametrischen Signifikanztests haben sich Teststärkebestimmungen bislang jedoch noch nicht so stark etabliert.

Überprüfen wir einen monoton steigenden Trend, ist statt des T-Wertes der Komplementärwert $T' = 2 \cdot \mu_T - T$ mit T_{krit} zu vergleichen. Bei zweiseitigem Test – der Trend ist entweder monoton steigend oder monoton fallend – muss der kleinere der beiden Werte T oder T' mit dem Tabellenwert verglichen und das α-Fehler-Niveau verdoppelt werden.

Asymptotischer Rangsummentest Wenn eines der beiden Ereignisse häufiger als 25-mal auftritt, ist die Prüfgröße T praktisch normal verteilt. Die Verteilung hat den nach Gl. (12.51) definierten Mittelwert und eine Streuung von

$$\sigma_T = \sqrt{\frac{N_1 \cdot N_2 \cdot (N+1)}{12}}. \qquad (12.52)$$

Der folgende z-Wert wird wiederum anhand der Standardnormalverteilungsfläche (Tab 1, ▶ Webangebot zu diesem Buch) zufallskritisch bewertet:

$$z = \frac{T - \mu_T}{\sigma_T}. \qquad (12.53)$$

Für unser Beispiel ermitteln wir zu Demonstrationszwecken (der asymptotische Test ist wegen $N_1 = 13$ und $N_2 = 19$ nicht indiziert):

$$\sigma_T = \sqrt{\frac{13 \cdot 19 \cdot (32+1)}{12}} = 26.06$$

und

$$z = \frac{212 - 214.5}{26.06} = -0.096.$$

Dieser empirische z-Wert ist nicht signifikant. (Hinweis: Der Rangsummentest entspricht formal dem sog. U-Test bzw. Mann-Whitney-U-Test; s. hierzu Bortz et al., 2008, S. 200ff.).

Rangsummentest – Bei einer Einzelfalluntersuchung mit dichotomer abhängiger Variable lässt sich die Alternativhypothese, dass im Sinne eines monotonen Trends eine Merkmalsalternative im Verlauf der Zeit immer häufiger auftritt, mit dem **Rangsummentest** prüfen, wobei entweder der **exakte Rangsummentest** oder – bei größeren Messwertreihen (n_1 oder $n_2 > 25$) – der **asymptotische Rangsummentest** (über die Normalverteilung) zu nutzen sind.

▪ Polytome abhängige Variablen

Multipler Iterationshäufigkeitstest Bisher gingen wir von Zeitreihen dichotomer Merkmale aus. Wir wollen nun die gleichen Hypothesen für **mehrkategorielle (polytome) nominale Merkmale** überprüfen. Zunächst wenden wir uns der Nullhypothese zu, dass die Anzahl der Iterationen in einer Zeitreihe einer zufälligen Abfolge entspricht. Die Überprüfung dieser Hypothese erfolgt mit dem multiplen Iterationshäufigkeitstest, der im Folgenden an einem Beispiel, das wir einer Anregung Lienerts (1978, S. 270) verdanken, verdeutlicht wird.

Exakter multipler Iterationshäufigkeitstest Für Zeitreihen mit höchstens 12 Ereignissen ermittelt man die exakte Wahrscheinlichkeit einer Abfolge nach den bei Bortz, Lienert und Boehnke (2008, S. 566 ff.) beschriebenen Rechenvorschriften.

Asymptotischer multipler Iterationshäufigkeitstest Angenommen, ein Student habe $n = 20$ gleich schwere Aufgaben eines Tests zu lösen. Jede Aufgabe hat drei nominale Antwortkategorien, sie kann gelöst (G), nicht gelöst (N) oder ausgelassen (A) werden. Folgende Zeitreihe zeigt das Resultat:

$$\overline{G\ G}\ \underline{A}\ \overline{N\ N\ N}\ \underline{G\ G\ G\ G}\ \overline{A\ A}\ \underline{G}\ \overline{N\ N}\ \underline{G\ G\ G\ G\ G}$$

Erneut fragen wir, ob die Mischung der $k = 3$ nominalen Kategorien G, N und A zufällig ist (H_0) oder ob die Wechsel zwischen je 2 Kategorien zu häufig oder zu selten auftreten (ungerichtete H_1). Mit letzterem wäre z. B. zu rechnen, wenn zwischen aufeinander folgenden Aufgaben Übertragungseffekte auftreten. Wir stellen zunächst fest, dass die Ereignisabfolge mit $n = 20$ Ereignissen $r = 8$ Iterationen aufweist. Die für den (asymptotischen) multiplen Iterationshäufigkeitstest benötigte Prüfgröße v lautet

$$v = N - r = 20 - 8 = 12. \qquad (12.54)$$

Ihr steht gemäß der H_0 ein Erwartungswert von

$$\mu_v = \frac{\sum_{i=1}^{k} N_i \cdot (N_i - 1)}{N} = \frac{12 \cdot 11 + 5 \cdot 4 + 3 \cdot 2}{20} = 7.9 \qquad (12.55)$$

gegenüber (mit N_i = Häufigkeiten des Auftretens der Kategorie i: $N_1 = 12$, da G 12-mal auftritt, $N_2 = 5$ und $N_3 = 3$). Der Unterschied zwischen v und μ_v spricht also für eine zu kleine Anzahl von Iterationen (man beachte, dass $v = N - r$).

Für $N > 12$ ist die Prüfgröße v des multiplen Iterationshäufigkeitstests approximativ normalverteilt mit einer

Varianz von

$$\sigma_v^2 = \frac{\sum_{i=1}^{k} (N_i \cdot (N_i - 1)) \cdot (N - 3)}{N \cdot (N - 1)} + \frac{\left[\sum_{i=1}^{k} N_i \cdot (N_i - 1)\right]^2}{N^2 \cdot (N - 1)} - \frac{2 \cdot \sum_{i=1}^{k} N_i \cdot (N_i - 1) \cdot (N_i - 2)}{N \cdot (N - 1)}. \quad (12.56)$$

Setzen wir die Werte des Beispiels ein, resultiert

$$\sigma_v^2 = \frac{(12 \cdot 11 + 5 \cdot 4 + 3 \cdot 2) \cdot (20 - 3)}{20 \cdot 19} + \frac{(12 \cdot 11 + 5 \cdot 4 + 3 \cdot 2)^2}{20^2 \cdot 19} - \frac{2 \cdot (12 \cdot 11 \cdot 10 + 5 \cdot 4 \cdot 3 + 3 \cdot 2 \cdot 1)}{20 \cdot 19}.$$

$$\sigma_v = \sqrt{7.07 + 3.28 - 3.65} = \sqrt{6.70} = 2.59.$$

Damit ergibt sich für z:

$$z = \frac{v - \mu_v}{\sigma_v} = \frac{12 - 7.9}{2.59} = 1.58.$$

Dieser empirische z-Wert schneidet 5.71 % der Fläche der Standardnormalverteilungsfläche ab. Da wir gemäß der Alternativhypothese entweder zu viele oder zu wenige Iterationen erwarten, testen wir zweiseitig, d. h., die H_0 ist mit einer Irrtumswahrscheinlichkeit von $p = 11.42\,\%$ beizubehalten. Die Reihenfolge der Ereignisse G, N und A ist zufällig. Auch hier sei jedoch darauf verwiesen, dass bei der Interpretation eines nicht-signifikanten Ergebnisses eigentlich noch auf die Teststärke einzugehen wäre, die – wie bereits erwähnt – im Bereich der nonparametrischen Verfahren bislang oft vernachlässigt wird.

Bei einer Einzelfalluntersuchung mit polytomer abhängiger Variable lässt sich die Alternativhypothese, dass der Wechsel zwischen dem Auftreten der k verschiedenen Merkmalsausprägungen nicht zufällig, sondern systematisch erfolgt, mit dem *multiplen Iterationshäufigkeitstest* prüfen, wobei entweder der *exakte multiple Iterationshäufigkeitstest* oder – bei größeren Messwertreihen ($N > 13$) – der *asymptotische multiple Iterationshäufigkeitstest* (über die Normalverteilung) zu nutzen sind.

Trendtests Der multiple Iterationshäufigkeitstest erfasst beliebige Abweichungen einer k-kategoriellen Zeitreihe von einer entsprechenden Zufallsabfolge. Interessiert jedoch als spezielle Art der Abweichung ein **monoton steigender oder monoton fallender Trend** für die Wahrscheinlichkeiten des Auftretens der einzelnen Kategorien, ist ein spezieller Trendtest indiziert.

Für diesen Test ist es erforderlich, dass hypothetisch festgelegt wird, in welcher Reihenfolge die Häufigkeiten der Merkmalskategorien im Verlauf der Zeitreihe zunehmen oder abnehmen. Werden z. B. die Kategorien A, B und C untersucht, könnte die Alternativhypothese lauten: $A > B > C$. Man beachte, dass mit dieser Hypothese nicht behauptet wird, dass A häufiger als B und B häufiger als C auftritt, sondern dass die Wahrscheinlichkeit für A im Verlaufe der Zeitreihe am meisten wächst, gefolgt von den Wahrscheinlichkeitszuwächsen für B und C.

Ermüdungs- und Sättigungseffekte lassen es im genannten Beispiel plausibel erscheinen, dass die Anzahl nicht gelöster Aufgaben (N) am meisten, die Anzahl ausgelassener Aufgaben (A) am zweitmeisten, und die Anzahl gelöster Aufgaben (G) am wenigsten zunimmt (bzw. am stärksten abnimmt). Die Alternativhypothese lautet damit $N > A > G$.

Eine Beschreibung des hier einschlägigsten Verfahrens (**Trendtest von Jonckheere**) findet man bei Bortz, Lienert und Boehnke (2008, S. 569f.). Ein weiteres Verfahren zur Überprüfung von Verläufen für mehrkategorielle Merkmale wurde von Noach und Petermann (1982) vorgeschlagen.

Bei einer Einzelfalluntersuchung mit polytomer abhängiger Variable lässt sich die Alternativhypothese, dass der Wechsel zwischen dem Auftreten der k verschiedenen Merkmalsausprägungen einem monoton steigenden oder monoton fallenden Trend entspricht (und nicht zufällig oder trendkonträr verläuft), mit entsprechenden Trendtests für polytome abhängige Variablen prüfen.

Signifikanztests für die Einzelfalldiagnostik

Ging es in den letzten Abschnitten um Hypothesen über den Verlauf individueller Zeitreihen, wenden wir uns nun Fragen zu, die die Bewertung einmalig erhobener Testergebnisse einer Person betreffen. Erhebungsinstrumente sind hierbei die in der psychologischen Diagnostik gängigen *psychologischen Tests* (▶ Abschn. 10.4) bzw. andere standardisierte Messinstrumente, deren testtheoretische Eigenschaften bekannt sind (◘ Abb. 12.69).

Viele psychologische Tests sind im Grunde Testbatterien (bzw. mehrdimensionale Tests), die mehrere separat auswertbare Untertests umfassen, d. h., das Testergebnis besteht häufig nicht nur aus einem Gesamttestergebnis, sondern aus mehreren Teilergebnissen (Untertestergebnissen), die zusammengenommen ein individuelles Testprofil ergeben. Die Gestalt eines Testprofils liefert wich-



Abb. 12.69 Die alltägliche Einzelfalldiagnostik ist selten völlig zufriedenstellend. (Aus Goldmanns Großer Cartoonband [1989] Schweine mit Igeln. München: Goldmann, S. 172. Mit freundlicher Genehmigung von Freimut Woessner)

tige Hinweise über die geprüfte Person, wenn davon auszugehen ist, dass die Differenzen zwischen den Untertestergebnissen nicht zufällig sind, sondern „wahre" Merkmalsunterschiede abbilden. Aufgabe der **Einzelfalldiagnostik** ist es, die **Zufälligkeit** bzw. **Bedeutsamkeit individueller Testergebnisse** abzuschätzen, da auf der Basis diagnostischer Befunde meist weitreichende Entscheidungen in der Praxis getroffen werden (z. B. Bewilligung einer Psychotherapie, Personalauswahl, Berentung, vorzeitige Entlassung aus dem Strafvollzug etc.).

Die Einzelfalldiagnostik betrachtet jeden individuellen **Testwert als eine Realisierung einer Zufallsvariablen**, deren Verteilung man erhielte, wenn eine Person beliebig häufig unter identischen Bedingungen mit demselben Test untersucht wird. Je kleiner die (Fehler-)Varianz dieser Verteilung, desto verlässlicher (reliabler) wäre eine Einzelmessung und desto unbedenklicher könnte man auch geringfügige Unterschiede zweier Testergebnisse interpretieren. Diese Verteilung auf empirischem Wege ermitteln zu wollen, ist nicht nur für die Testperson unzumutbar, sondern auch aus inhaltlichen Gründen fragwürdig, denn in der Regel dürfte sich die eigentlich interessierende „wahre" Merkmalsausprägung im Laufe der wiederholten Messungen durch Lern-, Übungs- und ähnliche Effekte verändern. Außerdem würde dieses Ansinnen die Praktikabilität einer Testanwendung erheblich infrage stellen.

Man ist deshalb darauf angewiesen, die Fehlervarianz bzw. Reliabilität einer individuellen Messung indirekt zu schätzen. Wie Huber (1973, S. 55ff.) zeigt, ist dies möglich, wenn man annimmt, dass die individuellen, auf einen Test bezogenen Fehlervarianzen zwischen den Individuen einer bestimmten Population nur geringfügig differieren. Zieht man eine repräsentative Stichprobe aus dieser Population, kann die Varianz der Testwerte zwischen den Personen **(Gruppenfehlervarianz)** als Schätzwert der **individuellen Fehlervarianzen** der Individuen dieser Population verwendet werden. Damit wären dann auch die anhand repräsentativer Stichproben ermittelten Reliabilitäten (die nach einem der in ▶ Abschn. 10.4.4 beschriebenen Verfahren geschätzt werden müssen) auf einzelne Individuen der Referenzpopulation übertragbar.

Psychologische Tests und vergleichbare Untersuchungsinstrumente, bei denen hohe Messgenauigkeit nachgewiesen ist, eignen sich somit, unter der Annahme annähernd gleich großer individueller Fehlervarianzen, auch für die Einzelfalldiagnostik (zur Problematik dieser Annahme s. Krauth, 1995, S. 208ff.; zit. nach Bühner, 2011, S. 197).

Im Folgenden behandeln wir sechs in der Einzelfalldiagnostik häufig gestellte Fragen, die mit inferenzstatistischen Verfahren zu bearbeiten sind:

1. Wie genau ist der Testwert einer Person?
2. Unterscheiden sich zwei Testwerte aus verschiedenen Tests bei einer Person statistisch bedeutsam?
3. Besteht zwischen einem Untertestwert und dem Gesamttestwert einer Person ein signifikanter Unterschied?
4. Sind die Schwankungen innerhalb eines individuellen Testprofils zufällig oder bedeutsam?
5. Hat sich der Testwert oder das Testprofil einer Person nach einer Intervention (z. B. einer Behandlung) signifikant geändert?
6. Weicht ein Individualprofil signifikant von einem Referenzprofil ab?

Wir begnügen uns damit, die Verfahren zur Überprüfung der **Hypothesen**, die diese Fragen implizieren, jeweils kurz an einem Beispiel zu demonstrieren. Auf die Diskussion der Teststärke wird bei der Einzelfalldiagnostik verzichtet, da sie in der Praxis kaum eine Rolle spielt. Für Einzelheiten verweisen wir auf Huber (1973). Zudem sei zur zufallskritischen Bewertung von Testergebnissen im Rahmen der Einzelfalldiagnostik auf Bühner (2011, Abschn. 4.8) verwiesen.

In der *Einzelfalldiagnostik* werden Testergebnisse einer einzelnen Person zufallskritisch bewertet und miteinander verglichen. Für einzelne Testwerte können im Sinne der *Parameterschätzung* Konfidenzintervalle bestimmt werden, zudem können erwartete Unterschiede zwischen Testwerten durch spezielle *Signifikanztests* geprüft werden. Die zufallskritische Bewertung einzelner Testwerte spielt in der Einzelfalldiagnostik als *praktischem Einsatzfeld* psychologischer Tests eine wichtige Rolle (z. B. Eignungsdiagnostik, klinische Diagnos-

tik). Sie kann aber auch im Rahmen *wissenschaftlicher Studien* (z. B. Evaluation von Therapie- oder Trainingsprogrammen anhand von Einzelfällen) von Bedeutung sein.

1. **Genauigkeit eines Testwertes.** Die Frage nach der Genauigkeit eines Testwertes beantworten wir durch die Berechnung eines Konfidenzintervalls (▶ Abschn. 12.4.2). Geht man davon aus, dass der beobachtete Testwert (y) eine Schätzung des wahren Wertes (T) darstellt (Äquivalenzhypothese; s. Bühner, 2011, Abschn. 4.8.1), ergibt sich das **Konfidenzintervall** wie folgt:

$$KI_T = y \pm z_{(\alpha/2)} \cdot SE_x \qquad (12.57)$$

$z_{(\alpha/2)} = 1.96$ (2.58) für das 95 %ige (99 %ige) Konfidenzintervall

$$SE_x = \sigma \cdot \sqrt{1 - r} \quad \text{(Standardmessfehler)} \qquad (12.58)$$

σ = Standardabweichung der Testwerte
r = Reliabilität

Beispiel (nach Fisseni, 1997, S. 91): Im Intelligenzstrukturtest (IST; Amthauer, 1971) hat eine Testperson einen IQ von 107 erzielt. Mit einer Standardabweichung von $\sigma = 10$ und einer Reliabilität von $r = .83$ ermittelt man ein 95 %iges Konfidenzintervall von

$$KI_T = 107 \pm 1.96 \cdot 10 \cdot \sqrt{1 - .83} = 107 \pm 8.1.)$$

Mit einer Konfidenz von 95 % befindet sich der „wahre" IQ-Wert im Bereich 98.9–115.1.

2. **Vergleich zweier Testwerte.** Die Intelligenzuntersuchung einer 21-jährigen Frau mit dem Intelligenzstrukturtest (IST; Amthauer, 1971) führte in den Untertests „ Gemeinsamkeiten" (GE) und „Figurenauswahl" (FA) zu den Testwerten GE = 118 und FA = 99. Es interessiert die Frage, ob diese Testwertedifferenz statistisch signifikant und damit diagnostisch verwertbar ist.
Dem IST Manual entnehmen wir, dass jeder Untertest auf einen Mittelwert von $\mu = 100$ und eine Streuung von $\sigma = 10$ normiert ist und dass die hier angesprochenen Untertests Reliabilitäten von $r_{GE} = .93$ und $r_{FA} = .84$ aufweisen. Der deutliche Reliabilitätsunterschied legt eine Normierung der Testwerte nahe, die die unterschiedlichen Reliabilitätskoeffzienten berücksichtigt (sog. τ-Normierung; siehe Huber, 1973, Abschn. 4.5). Dies geschieht, indem die beiden Testwerte nach folgender Gleichung transformiert werden:

$$\tau = \frac{y}{\sqrt{r}} + \mu \cdot \left(1 - \frac{1}{\sqrt{r}}\right) \qquad (12.59)$$

τ = normierter Testwert
y = Testwert
r = Reliabilität des Tests
μ = Erwartungswert (Mittelwert) des Tests.

Nach dieser Beziehung errechnen wir für die beiden Testwerte

$$\tau_1 = \frac{118}{\sqrt{.93}} + 100 \cdot \left(1 - \frac{1}{\sqrt{.93}}\right) = 118.67$$
$$\tau_2 = \frac{99}{\sqrt{.84}} + 100 \cdot \left(1 - \frac{1}{\sqrt{.84}}\right) = 98.91.$$

Den Unterschied der beiden τ Werte überprüfen wir nach Gl. (12.60)

$$z = \frac{\tau_1 - \tau_2}{\sigma \cdot \sqrt{\left(\frac{1 - r_1}{r_1}\right) + \left(\frac{1 - r_2}{r_2}\right)}} \qquad (12.60)$$

Mit $\sigma = 10$, $r_1 = .93$ und $r_2 = .84$ resultiert in unserem Beispiel:

$$z = \frac{118.67 - 98.91}{10 \cdot \sqrt{\left(\frac{1 - .93}{.93}\right) + \left(\frac{1 - .84}{.84}\right)}} = 3.83.$$

Dieser Wert schneidet von der Standardnormalverteilungsfläche (Tab. 1, ▶ Webangebot zu diesem Buch) weniger als 0.5 % ab, d. h., die Differenz ist bei zweiseitigem Test (also nach Verdopplung des Flächenwertes) auf dem $\alpha = 1$ % Niveau signifikant.
Weitere Hinweise zum Vergleich zweier Subtestwerte findet man bei Cahan (1989)

3. **Vergleich eines Untertestwertes mit dem Gesamttestwert.** Gelegentlich möchte man wissen, ob sich die Leistung in einem einzelnen Untertest deutlich bzw. statistisch signifikant von der Gesamttestleistung unterscheidet. Im Rahmen einer Umschulungsberatung führte eine Intelligenzprüfung mit dem Wechsler-Intelligenztest für Erwachsene (WIE; Aster, Neubauer, & Horn, 2006, frühere Bezeichnung HAWIE: Hamburg-Wechsler-Intelligenztest für Erwachsene, Abschn. 10.4.3 zu Intelligenztests) bei einem Angestellten im Untertest „Zahlen Nachsprechen" (ZN) zu einem Testwert von 14. Als Gesamt-IQ ergab sich ein Wert von 96. Man interessiert sich nun für die Frage, ob diese Abweichung auf eine spezielle Begabung hinweist oder ob sie zufällig zustande kam. Die Beantwortung dieser Frage setzt voraus, dass die Korrelation zwischen dem Untertest und dem Gesamt-IQ bekannt ist. Sie lautet im Beispiel $r_{ZN-G} = .63$. Mithilfe dieser Korrelation lässt sich regressionsanalytisch ermitteln, welcher Untertestwert bei einem IQ

von 96 zu erwarten ist. Die Regressionsgleichung lautet:

$$\hat{y}_1 = \mu_1 + \frac{\sigma_1}{\sigma_G} \cdot r_{1G} \cdot (y_G - \mu_G) \qquad (12.61)$$

μ_1 = Mittelwert des Untertests
σ_1 = Streuung des Untertests
μ_G = Mittelwert des Gesamttests
σ_G = Streuung des Gesamttests
r_{1G} = Korrelation zwischen Untertest und Gesamttest
y_G = Gesamttestwert.

Für $\mu_1 = 10$, $\sigma_1 = 3$, $\mu_G = 100$, $\sigma_G = 15$, $r_{1G} = .63$ (diese Werte sind dem jeweiligen Testhandbuch zu entnehmen) und $y_G = 96$ errechnen wir für $\hat{y}_1$:

$$\hat{y}_1 = 10 + \frac{3}{15} \cdot .63 \cdot (96 - 100) = 9.50.$$

Die statistische Bedeutsamkeit der Differenz zwischen dem erwarteten und dem erzielten Untertestwert überprüfen wir in folgender Weise:

$$z = \frac{y_1 - \hat{y}_1}{\sigma_1 \cdot \sqrt{1 - r_{1G}^2}} \qquad (12.62)$$
$$= \frac{14 - 9.5}{3 \cdot \sqrt{1 - .63^2}} = 1.93.$$

Dieser Wert ist bei zweiseitigem Test gem. Standardnormalverteilungstabelle (Tab. 1 im ▶ Webangebot zu diesem Buch) nicht signifikant.

4. **Bewertung von Schwankungen im Profilverlauf.** Nicht nur die Höhe eines Testprofils, sondern auch dessen Verlauf liefert oftmals wichtige diagnostische Hinweise. Bevor man jedoch aus einem Profilverlauf diagnostische Schlüsse zieht, sollte man sich vergewissern, dass die Schwankungen der Untertestwerte tatsächlich vorhandene Merkmalsunterschiede abbilden und nicht zufällig sind.
Nehmen wir an, eine Testperson habe in einem Persönlichkeitstest mit 6 Untertests die folgenden Werte erhalten:

$$y_1 = 38; \quad y_2 = 44; \quad y_3 = 42;$$
$$y_4 = 49; \quad y_5 = 35; \quad y_6 = 51.$$

Alle Untertests seien auf den Mittelwert $\mu = 50$ und die Streuung $\sigma = 5$ normiert. Als Reliabilitäten der Untertests werden berichtet:

$$r_1 = .72; \quad r_2 = .64; \quad r_3 = .80;$$
$$r_4 = .78; \quad r_5 = .67; \quad r_6 = .76.$$

Die unterschiedlichen Reliabilitäten lassen eine τ Normierung der Testwerte ratsam erscheinen. Wir ermitteln nach Gl. (12.59)

$$\tau_1 = 35.86; \quad \tau_2 = 42.50; \quad \tau_3 = 41.06;$$
$$\tau_4 = 48.87; \quad \tau_5 = 31.67; \quad \tau_6 = 51.15.$$

Über die H_0, dass die Differenzen zufällig sind, entscheidet die folgende Chi-Quadrat-verteilte Prüfgröße:

$$\chi^2 = \frac{1}{\sigma^2} \cdot \sum_{j=1}^{m} \frac{r_j}{1 - r_j} \cdot (\tau_j - \bar{\tau})^2 \qquad (12.63)$$

σ^2 = Varianz der Untertests
r_j = Reliabilität des Untertests j
τ_j = normierter Testwert im Untertest j
$\bar{\tau}$ = Durchschnitt der τ_j-Werte
m = Anzahl der Untertests.

Unter der Annahme, dass die Fehleranteile der Testwerte in den einzelnen Untertests voneinander unabhängig und normalverteilt sind, ist diese Prüfgröße mit $m - 1$ Freiheitsgraden χ^2-verteilt.
Für das Beispiel resultiert

$$\chi^2 = \frac{1}{5^2} \cdot \left[\frac{.72}{1 - .72} \cdot (35.86 - 41.85)^2 + \frac{.64}{1 - .64} \right.$$
$$\cdot (42.50 - 41.85)^2 + \ldots + \frac{.76}{1 - .76}$$
$$\left. \cdot (51.15 - 41.85)^2 \right] = \frac{754.52}{25} = 30.18.$$

Mit $6 - 1 = 5$ Freiheitsgraden ist dieser empirische χ^2-Wert auf dem $\alpha = 1\,\%$-Niveau gem. Tab. 8 im ▶ Webangebot zu diesem Buch signifikant, d. h., wir können davon ausgehen, dass die Profilgestalt nicht zufällig zustande kam, sondern tatsächliche Merkmalsunterschiede wiedergibt.
Ein Verfahren zur Überprüfung der Reliabilität eines Testprofils findet man bei Rae (1991) bzw. Yarnold (1984). Die statistischen Probleme, die sich bei mehreren Vergleichen von Untertestwerten einer Testperson ergeben, behandelt Bird (1991).

5. **Vergleich von Testwerten bei wiederholter Testanwendung.** Psychologische Tests werden nicht nur zu diagnostischen Zwecken, sondern z. B. auch zur Kontrolle therapeutischer oder anderer Maßnahmen eingesetzt. Es stellt sich dann die Frage, ob die mit der Intervention einhergehenden Merkmalsveränderungen zufällig oder bedeutsam sind.

In einem Test über berufliche Interessen erhielt ein Abiturient in $m = 5$ Untertests die folgenden Werte:

$$y_{11} = 18; \quad y_{21} = 22; \quad y_{31} = 25;$$
$$y_{41} = 20; \quad y_{51} = 19.$$

Die Testskalen sind auf $\mu = 20$ und $\sigma = 3$ normiert. Ihre Reliabilitäten lauten

$$r_1 = .72; \quad r_2 = .89; \quad r_3 = .81;$$
$$r_4 = .90; \quad r_5 = .85.$$

Nach der ersten Testvorgabe arbeitet der Abiturient Informationsmaterial über einige ihn interessierende Berufe durch. Danach lässt er seine Berufsinteressen erneut prüfen und erzielt diesmal folgende Werte:

$$y_{12} = 17; \quad y_{22} = 22; \quad y_{32} = 20;$$
$$y_{42} = 22; \quad y_{52} = 18.$$

Sind die aufgetretenen Veränderungen mit den nicht perfekten Reliabilitäten der Untertests erklärbar oder hat die Auseinandersetzung mit den tatsächlich anfallenden Tätigkeiten und Aufgaben in den geprüften Berufen das Interessenprofil des Abiturienten verändert? Diese Frage beantwortet folgender Test:

$$\chi^2 = \frac{1}{2 \cdot \sigma^2} \cdot \sum_{j=1}^{m} \frac{(y_{j1} - y_{j2})^2}{1 - r_j} \qquad (12.64)$$
$$= \frac{1}{2 \cdot 3^2} \cdot \left[\frac{(18-17)^2}{1-.72} + \frac{(22-22)^2}{1-.89} + \frac{(25-20)^2}{1-.81} + \frac{(20-22)^2}{1-.90} + \frac{(19-18)^2}{1-.85} \right]$$
$$= \frac{1}{18} \cdot 181.82 = 10.10.$$

Mit $m = 5$ Freiheitsgraden ist dieser empirische χ^2 Wert auf dem $\alpha = 5\,\%$-Niveau nicht signifikant (Tab. 8 im ▶ Webangebot zu diesem Buch). Die in den einzelnen Untertests festgestellten Veränderungen liegen im Zufallsbereich.

Weitere Informationen zum Vergleich individueller Testwerte bei mehrfacher Testanwendung findet man bei Maassen (2000) oder Yarnold (1988)

6. **Vergleich eines Individualprofils mit einem Referenzprofil.** Von vielen Tests, die in der Praxis häufig benötigt werden, sind Profile bestimmter Subpopulationen bekannt, wie z. B. die einer bestimmten Alterspopulation, Berufspopulation oder Patientenpopulation. Der Vergleich eines Individualprofils mit derartigen Referenzprofilen informiert über die mutmaßliche Zugehörigkeit der untersuchten Personen zu einer der infrage kommenden Referenzpopulationen.

Bezogen auf das zweite Interessenprofil des im letzten Beispiel erwähnten Abiturienten soll überprüft werden, wie gut dieses Profil mit den durchschnittlichen Interessen von Steuerberatern übereinstimmt. Man entnimmt dem Testhandbuch, dass eine Stichprobe von $n = 60$ Steuerberatern folgendes Durchschnittsprofil erzielte:

$$\bar{y}_1 = 18; \quad \bar{y}_2 = 23; \quad \bar{y}_3 = 21;$$
$$\bar{y}_4 = 24; \quad \bar{y}_5 = 16.$$

Mit Gl. (12.65) wird die Zufälligkeit der Abweichung eines Individualprofils von einem Referenzprofil überprüft

$$\chi^2 = \frac{n}{(1+n) \cdot \sigma^2} \cdot \sum_{j=1}^{m} \frac{(y_j - \bar{y}_j)^2}{1 - r_j}. \qquad (12.65)$$

Unter der Voraussetzung, dass sich die zu einem Durchschnittsprofil zusammengefassten Einzelprofile nur zufällig unterscheiden und dass die Messfehler voneinander unabhängig und normalverteilt sind, ist diese Prüfgröße mit m Freiheitsgraden χ^2-verteilt. Für das Beispiel errechnen wir:

$$\chi^2 = \frac{60}{(60+1) \cdot 3^2} \cdot \left[\frac{(18-17)^2}{1-.72} + \frac{(22-23)^2}{1-.89} + \frac{(20-21)^2}{1-.81} + \frac{(22-24)^2}{1-.90} + \frac{(18-16)^2}{1-.85} \right]$$
$$= 0.11 \cdot 84.59 = 9.30.$$

Dieser empirische Wert ist nach Tab. 8 zu χ^2 Verteilungen (▶ Webangebot zu diesem Buch) bei 5 Freiheitsgraden auf dem 5 %-Niveau nicht signifikant ($9.30 < 11.07 = \chi^2_{crit}$).

Der hier beschriebene Vergleich eines Individualprofils mit einem Referenzprofil sollte mit allen infrage kommenden Referenzpopulationen durchgeführt werden. Der Vergleich mit dem kleinsten χ^2-Wert signalisiert dann die bestmögliche Übereinstimmung (ein anderes Zuordnungsverfahren wird bei Bortz & Schuster, 2010, in Kap. 27 beschrieben).

Lernquiz

1. Was ist unter einer hypothesenprüfenden Einzelfalluntersuchung zu verstehen?
2. Welche Vor- und Nachteile haben Einzelfalluntersuchungen gegenüber Stichprobenuntersuchungen?
3. Eine Schülerin erreicht in einem Intelligenztest einen IQ von 96. Heißt dies, dass sie unterdurchschnittlich intelligent ist?

4. Was unterscheidet den A-BC-B-BC-Plan von dem A-B-A-B-Plan?
5. Sind die Messfehler der Daten aus Einzelfallstudien voneinander unabhängig oder abhängig – und wie ist damit umzugehen?
6. In einem validen Test zur Messung von Sozialkompetenz (0–100 Punkte) erzielt ein Polizist einen Testwert von 56. Nach einem Kompetenztraining wird derselbe Test erneut durchgeführt und es resultiert ein Testwert von 67. Lässt sich aus diesen Daten schlussfolgern, dass das Training die Sozialkompetenz verbessert hat?
7. Auf welche Weise kann bei dem Randomisierungstest die Teststärke erhöht werden?
8. Für die Überprüfung von Einzelfallhypothesen stehen meist exakte und asymptotische Varianten des gleichen Signifikanztests zur Verfügung. Wann wird die jeweilige Testvariante eingesetzt?
9. In einem Artikel, der den Einfluss von Gartenarbeit auf die Depressivität eines Patienten mit einer Minor Depression untersucht, lesen Sie, dass Gartenarbeit zu keiner Verbesserung der Depressivität führt. Es wurde ein A-B-A-B-Plan eingesetzt und die Auswertung der Daten erfolgte mit dem Randomisierungstest. Ist die Beibehaltung der Nullhypothese korrekt? Bitte begründen Sie Ihre Antwort!
10. Wie ist im Rahmen der Einzelfalldiagnostik bei einem Vergleich von einem Untertestwert mit dem Gesamttestwert vorzugehen?

Übungsaufgaben

1. An einer Schule wird die monatliche Zahl der erteilten Verwarnungen bzw. schriftlichen Missbilligungen registriert und pro Jahr gemittelt. Es ergeben sich folgende ganzzahlig gerundeten Daten:

   2007: 1 – 2008: 3 – 2009: 5 – 2010: 8 – 2011: 12

   Prüfen Sie die Hypothese, dass die durchschnittliche Anzahl der Verwarnungen pro Monat bei der Schule als Einzelfall über die betrachteten Jahre hinweg überzufällig angestiegen ist.
2. Könnte man die Trendhypothese, dass die Anzahl der monatlichen Verwarnungen an der Schule monoton ansteigt, auch annehmen, wenn nur die Daten der letzten 3 Jahre berücksichtigt würden? (Begründung)
3. Im Zuge einer Diagnose zur Aufmerksamkeitsstörung ADHS werden einem Kind unterschiedliche Konzentrationsaufgaben vorgelegt, die sich in drei Schwierigkeitsstufen einteilen lassen: geringe (g), mittlere (m) oder hohe (h) Schwierigkeit. Wenn das Kind bei einer Aufgabe nicht weiterkommt, kann es eine selbst gewählte andere Aufgabe aus dem Aufgabenpool bearbeiten. Insgesamt wurden $n = 15$ Aufgaben bearbeitet, wobei $r = 12$ Iterationen stattfanden. Die nachfolgende Zeitreihe zeigt die Abfolge der Aufgaben:

   $$g\,g\;m\;h\;m\,m\;g\;h\;g\,g\;m\;g\;h\;m\;g$$

   Man möchte nach Abschluss des Tests wissen, ob das Kind überzufällig oft zwischen den Aufgabentypen gewechselt hat. Führen Sie den entsprechenden Signifikanztest durch!

Weiterführende Literatur

- Barker, J. (2011). *Single Case Research Methods in Sport and Exercise Psychology*. New York: Routledge.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer
- Kazdin, A. (2010) *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford: Oxford University Press.
- Morgan, D. L. & Morgan, R. K. (2008) *Single-case research methods for the behavioral and health sciences*. Thousand Oaks: Sage.

Literatur

Amthauer, R. (1971). *Intelligenz-Struktur-Test (I-S-T 70)*. Göttingen: Hogrefe.

Aster, M. von, Neubauer, A., & Horn, R. (2006). *WIE – Wechsler Intelligenztest für Erwachsene*. Frankfurt: Pearson.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*(1), 91–97.

Barker, M., Richards, C., Jones, R., & Monro, S. (2011). BiReCon: An international academic conference on bisexuality. *Journal of Bisexuality, 11*(2), 157–170.

Barlow, D. H. & Hersen, M. (1973). Single case experimental designs. *Archives of General Psychiatry, 29*(33), 319–325.

Barlow, D. H. & Hersen, M. (Eds.). (1984). *Single case experimental designs: Strategies for studying behaviour change*. New York: Pergamon.

Bird, K. D. (1991). Exploratory $n = 1$ profile analysis. *Educational and Psychological Measurement, 51*(3), 523–530.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin Heidelberg: Springer

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson Studium.

Cahan, S. (1989). A critical examination of the „relibability" and „abnormality" approaches to the evaluation of subtest score differences. *Educational and Psychological Measurement, 49*(4), 807–814.

Callahan, C. D. & Barisa, M. T. (2005). Statistical process control and rehabilitation outcome: The single-subject design reconsidered. *Rehabilitation Psychology, 50*(1), 24–33.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology, 90*(1), 57–68.

Edgington, E. S. (1980). Overcoming obstacles to single subject experimentation. *Journal of Educational and Behavioral Statistics, 5*(3), 261–267.

Edgington, E. S. (1995). *Randomization tests*. New York: Dekker.

Fichter, M. M. (1979). Versuchsplanung experimenteller Einzelfalluntersuchungen in der Psychotherapieforschung. In F. Petermann & F. J. Hehl (Hrsg.), *Einzelfallanalyse* (S. 140–158). München: Urban & Schwarzenberg.

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik*. Göttingen: Verlag für Psychologie.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1996). *Design and analysis of single-case research*. Mahwah: Erlbaum.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder: University Press.

Good, P. (2000). *Permutation tests* (2nd ed.). New York: Springer.

Gottman, J. M. (1973). N-of-one and N-of-two research in psychotherapy. *Psychological Bulletin, 80*(2), 93–105.

Huber, H. P. (1973). *Psychometrische Einzelfalldiagnostik*. Weinheim: Beltz.

Kazdin, A. E. (1976). Statistical analysis for single-case experimental designs. In M. Hersen & D. H. Barlow (Eds.), *Single case experimental designs: strategies for studying behavior change* (pp. 265–316). New York: Pergamon.

Kazdin, A. E. (1978). Methodological and interpretative problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46*(4), 629–642.

Kazdin, A. E. (1982). *Single case research designs: Methods for Clinical and Applied Settings*: Oxford University Press.

Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings*. Oxford: Oxford University Press.

Koehler, M. J. & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple baseline design. *Psychological Methods, 3*(2), 206–217.

Kratochwill, T. R. (1978). *Single subject research. Strategies for evaluating change*. Waltham: Academic Press.

Kratochwill, T. R., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arnston, P., et al. (1974). A further consideration in the application of an analysis–of–variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 7*(4), 629–633.

Kratochwill, T. R. & Levin, J. R. (Eds.). (1992). *Single case research design and analysis*. Hillsdale: Erlbaum.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Beltz.

Levin, J. R., Marascuilo, L. A., & Hubert, L. J. (1978). $N = 1$. Nonparametric randomization tests. In T. R. Kratochwill (Ed.), *Single subject research* (pp. 167–196). New York: Academic Press.

Levin, J. R. & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly, 14*(1), 59–93.

Levy, P. S. & Lemeshow, S. (1999). *Sampling of populations: methods and applications*. New York: Wiley.

Lienert, G. A. (1978). *Verteilungsfreie Methoden in der Biostatistik* (Bd. II). Meisenheim: Hain.

Maassen, G. H. (2000). Keley's formula as a basis for the assessment of reliable change. *Psychometrika, 65*(2), 187–197.

Meyer-Bahlburg, H. F. L. (1969). Spearmans rho als punktbiserialer Rangkorrelationskoeffizient. *Biometrische Zeitschrift, 11*(1), 60–66.

Morgan, D. L. & Morgan, R. K. (2008). *Single-case research methods for the behavioral and health sciences*. Thousand Oaks: Sage.

Nicolich, M. J. & Weinstein, C. S. (1977). *Time series analysis of behavioral changes in an open class-room.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Noach, H. & Petermann, F. (1982). Die Prüfung von Verlaufsannahmen in der therapeutischen Praxis. *Zeitschrift für personenzentrierte Psychologie und Psychotherapie, 1*, 9–27.

Parsonson, B. S. & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single subject design* (pp. 101–165). New York: Academic Press.

Petermann, F. (1981). Möglichkeiten der Einzelfallanalyse in der Psychologie. *Psychologische Rundschau, 32*(1), 31–48.

Petermann, F. (1982). *Einzelfalldiagnose und klinische Praxis*. Stuttgart: Kohlhammer.

Petermann, F. (1996). *Einzelfalldiagnostik in der klinischen Praxis*. Weinheim: Psychologie Verlags Union.

Rae, G. (1991). Another look at the reliability of a profile. *Educational and Psychological Measurement, 51*(1), 89–93.

Revenstorf, D. & Keeser, W. (1979). Zeitreihenanalyse von Therapieverläufen. Ein Überblick. In F. Petermann & F. J. Hehl (Hrsg.), *Einzelfallanalyse* (S. 183–228). München: Urban & Schwarzenberg. Steinmeyer, E. M. (1976). Zufallskritische Einzelfalldiagnostik im psychiatrischen Feld, dargestellt am Beispiel der Hebephrenie. *Zeitschrift für experimentelle und angewandte Psychologie, 23*(2), 271–283.

Stevens, W. L. (1939). Distribution of groups in a sequence of alternatives. *Annals of Human Genetics, 9*(1), 10–17.

Westmeyer, H. (1979). Wissenschaftstheoretische Grundlagen der Einzelfallanalyse. In F. Petermann & F. J. Hehl (Hrsg.), *Einzelfallanalyse* (S. 17–34). München: Urban & Schwarzenberg.

Yarnold, P. R. (1984). The reliability of a profile. *Educational and Psychological Measurement, 44*(1), 49–59.

Yarnold, P. R. (1988). Classical test theory methods for repeated measures $n = 1$ research designs. *Educational and Psychological Measurement, 48*(4), 913–919.

Ergebnispräsentation

Nicola Döring, Jürgen Bortz

13.1 Zeitschriftenartikel – 787

13.2 Konferenzvortrag – 792

13.3 Wissenschaftliches Poster – 799

13.4 Öffentlichkeitsarbeit – 801

Literatur – 804

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_13, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Verschiedene Formen der Präsentation wissenschaftlicher Ergebnisse voneinander abgrenzen und in ihren Besonderheiten charakterisieren können.
- Wissen, wie man einen Zeitschriftenartikel schreibt und in einer wissenschaftlichen Fachzeitschrift publiziert.
- Wissen, wie man einen wissenschaftlichen Fachvortrag für eine Konferenz vorbereitet.
- Wissen, wie man ein wissenschaftliches Poster erstellt und auf einer Konferenz präsentiert.
- Verschiedene Möglichkeiten der Kommunikation wissenschaftlicher Ergebnisse an die breite Öffentlichkeit kennen.

Die Ergebnisse qualitativer und quantitativer empirischer Studien gilt es im letzten Schritt des empirischen Forschungsprozesses in geeigneter Weise der Scientific Community zu präsentieren. Dies ist im Sinne der Qualitätssicherung wissenschaftlicher Forschung notwendig, denn nur **publizierte Ergebnisse** können fachlicher Kritik unterzogen werden. Sich auf wissenschaftliche Ergebnisse zu berufen, die nicht publiziert und somit nicht detailliert nachvollziehbar sind, widerspricht den Regeln guter wissenschaftlicher Praxis (▶ Abschn. 4.2.1), denn **intersubjektive Nachvollziehbarkeit** ist ein zentrales Kriterium der Wissenschaftlichkeit (▶ Abschn. 3.1.1). Darüber hinaus ist die Ergebnispräsentation für den wissenschaftlichen Erkenntnisfortschritt eines Faches unabdingbar: Denn zukünftige Studien sollen ja möglichst am bereits erarbeiteten Kenntnisstand ansetzen (kumulativer wissenschaftlicher Erkenntnisgewinn).

Als besonders hochwertig gelten Publikationen, die einer **Qualitätskontrolle durch den Peer-Review-Prozess** unterliegen, d. h. eingereichte Beiträge werden vor einer möglichen Publikation durch Fachkollegen kritisch begutachtet und zwar zur Sicherung der Neutralität in einem doppelblinden Verfahren (Gutachtende und Begutachtete bleiben füreinander anonym). Die wichtigsten referierten (d. h. durch einen Peer-Review-Prozess geprüften) wissenschaftlichen Fachpublikationen sind der **Zeitschriftenartikel** (▶ Abschn. 13.1), der **Konferenzvortrag** (▶ Abschn. 13.2) und schließlich das **wissenschaftliche Poster** (▶ Abschn. 13.3). Die Verantwortung dafür, dass Peer-Review-Prozesse sorgfältig und fair ablaufen, liegt bei den einzelnen Gutachtenden sowie bei den für die Durchführung von Peer-Review-Prozessen jeweils Verantwortlichen (z. B. Herausgebergremium einer Zeitschrift, Organisationsteam einer Konferenz). Der Ablauf von Peer Reviews unterliegt Richtlinien der Wissenschaftsethik (▶ Abschn. 4.2) und ist relativ standardisiert, so dass Gutachtende und Begutachtete wissen, worauf sie sich einlassen (Lovejoy, Revenson, & France, 2011). Die Grenzen und Schwächen von Peer-Review-Verfahren (z. B. mangelnde Sorgfalt) werden empirisch untersucht und sind auch Gegenstand öffentlicher Debatten (z. B. anlässlich aufgedeckter Plagiate und Fälschungen in wissenschaftlichen Publikationen).

Neben der Ergebnispräsentation in der wissenschaftlichen Fachöffentlichkeit durch referierte Fachpublikationen wird von Forschenden zunehmend eine Kommunikation ihrer Theorien, Methoden und Ergebnisse an die breite Öffentlichkeit gefordert und angestrebt. Dies gilt für die staatlich finanzierte Forschung an Hochschulen, die z. B. unter dem Druck steht, sich vor den Steuerzahlenden zu legitimieren, aber auch für die private Markt- und Meinungsforschung, deren Institute um Kundschaft werben. Websites, Pressekonferenzen oder populärwissenschaftliche Veröffentlichungen sind Beispiele für Instrumente der **Öffentlichkeitsarbeit**, auf die Forschungseinrichtungen sowie einzelne Forschende zurückgreifen (▶ Abschn. 13.4).

Studierende sind zunächst gehalten, die Ergebnisse ihrer Studien in Form von **Qualifikationsarbeiten** (Hausarbeiten, Bachelor- und Masterarbeiten) schriftlich niederzulegen sowie bei Abschlussprüfungen mündlich in Vortragsform darzustellen. Die dabei geltenden Regeln (z. B. Seitenzahl, Vortragslänge, Gliederung, Zitationsregeln) sind durch die jeweils lokalen **Prüfungsordnungen** vorgegeben und werden oft durch **interne Handreichungen** ergänzt bzw. präzisiert. Es ist empfehlenswert, sich hier im Vorfeld genau über die Regularien an der eigenen Hochschule zu informieren und diese exakt einzuhalten. Generell zeichnet sich der Trend ab, die an Studierende gestellten Anforderungen denen des üblichen Wissenschaftsbetriebs anzugleichen. Wenn etwa gefordert wird, dass eine Bachelorarbeit im Aufbau einem Zeitschriftenartikel entspricht oder die mündliche Verteidigung der Bachelorarbeit wie ein Konferenzvortrag gestaltet ist, fördert dies die **Professionalisierung des akademischen Nachwuchses**. Teilweise wird dies von Studierenden selbst vorangetrieben, die z. B. zunehmend freiwillig ihre Ergebnisse in der internationalen Wissenschaftssprache Englisch präsentieren. Darüber hinaus hat es sich in vielen Fächern eingebürgert, dass wissenschaftlich ambitionierte Studierende ihre Ergebnisse – meist zusammen mit den Betreuerinnen und Betreuern – auf wissenschaftlichen Konferenzen vortragen oder in Fachzeitschriften veröffentlichen. In der Promotionsphase wird heute üblicherweise gefordert, dass **Promovierende** Zwischen- oder Teilergebnisse ihrer wissenschaftlichen Arbeit auf Konferenzen vorstellen oder in Fachzeitschriften publizieren. **Drittmittelfinanzierte Forschungsprojekte** stehen ebenfalls unter hohem Publikationsdruck. Die folgenden Abschnitte wollen praktische Hinweise und Orientierungshilfen zur wissenschaftlichen Ergebnis-

präsentation in Form von Veröffentlichungen bieten und sind größtenteils auch auf studentische Qualifikationsarbeiten übertragbar.

13.1 Zeitschriftenartikel

Wissenschaftliche Publikationen sind die wichtigste **„Währung" in der Wissenschaft**: über Publikationen gewinnen Personen und Institutionen Reputation. Denn Anzahl und Hochrangigkeit von Publikationen gelten heute als zentrale Indikatoren für den Beitrag zum wissenschaftlichen Erkenntnisfortschritt. Die Anzahl von Publikationen geht deswegen in die Evaluation von Forschungsgruppen ein und hat ernsthafte (auch finanzielle) Konsequenzen. Zahlreiche und hochrangige Publikationen sind zudem notwendig für die Akquise von Drittmitteln. Deswegen heißt es so kernig: **„Publish or Perish"**. Ein regelmäßiger Publikationsoutput ist nach heutigen Leistungsstandards im Wissenschaftsbetrieb für alle Beteiligten nicht freiwillige „Kür", sondern schlicht „Pflicht". Doch das Publizieren ist zum Glück auch mehr als eine lästige Pflicht: Das Verfassen wissenschaftlicher Publikationen bietet die Möglichkeit, die eigenen Forschungsergebnisse einem interessierten Fachpublikum mitzuteilen und Feedback zu erhalten, die eigenen Gedankengänge zu schärfen, am wissenschaftlichen Diskurs teilzunehmen und etwas „Bleibendes" zu schaffen. Trotz (oder wegen) aller Mühen bietet das Publizieren auch viele Freuden, etwa Spaß am Schreibfluss oder Stolz über einen vollendeten, gedruckten und von anderen zitierten Artikel.

Der wachsende **Publikationsdruck** in der Wissenschaft bezieht sich sowohl auf die Quantität als auch die Qualität der Beiträge. Am höchsten angesehen sind Artikel in den führenden internationalen Fachzeitschriften mit doppelblindem Peer-Review-Verfahren, denn hier werden die strengsten Qualitätsmaßstäbe angelegt und die meisten Einreichungen abgelehnt. Hilfe bei der Identifikation der relevantesten Zeitschriften bietet unter anderem der **Impact-Factor**, ein einheitlicher Vergleichsstandard, der die Reichweite und Relevanz der jeweiligen Zeitschrift in der Forschungscommunity angibt. Eine Publikation in einer hochrangigen Zeitschrift wird zusätzlich durch eine große Zahl von Zitierungen in anderen Arbeiten aufgewertet, weil dadurch belegt ist, dass es sich um einen Beitrag von besonderer inhaltlicher Relevanz handelt. Doch diese **bibliometrischen Bewertungskriterien** sind auch zu relativieren. Sie hängen stark vom Forschungsgebiet und der Größe der jeweiligen Fachcommunity ab. So wird ein Methodenbeitrag naturgemäß sehr viel häufiger zitiert als ein inhaltlicher Beitrag zu einem seltenen Spezialthema, auch wenn der betreffende Methodenartikel vielleicht nur mittelmäßige und der inhaltliche Beitrag zu dem Spezialthema herausragende Qualität hat.

Während es für Nachwuchswissenschaftlerinnen und -wissenschaftler zunächst darum geht, überhaupt einen referierten Fachartikel zu schreiben und unterzubringen, verfolgen etablierte Forschende meist eine dezidierte **Publikationsstrategie** und planen im Vorfeld, in welchen Zeitschriften sie welche ihrer Arbeiten mit welchen Mitautorinnen und Mitautoren platzieren wollen, um den größten Reputationsnutzen aus den Artikeln zu ziehen.

Wir konzentrieren uns im Folgenden auf Zeitschriftenartikel als dem einschlägigen wissenschaftlichen Publikationsformat mit Qualitätssicherung durch das Peer-Review-Verfahren. Auch wenn **Bücher** eine eigene Aura haben und es für Studierende schmeichelhaft ist, ihre Abschlussarbeit als Buch gedruckt zu sehen, so ist das wissenschaftliche Ansehen dieser Druckerzeugnisse geringer. Diejenigen Buchverlage, die aktiv auf Studierende zugehen, um deren Abschlussarbeiten zu publizieren, sind primär am Verkaufserlös und dem zu zahlenden Druckkostenbeitrag interessiert, nicht an wissenschaftlicher Qualität. Dementsprechend gering ist ihr Ansehen. Anders ist es bei den hochrangigen Wissenschaftsverlagen, die nur ausgewählte Autorinnen und Autoren bzw. wissenschaftlich hochwertige Manuskripte drucken.

13.1.1 Auswahl der Zeitschrift

Sollen die Ergebnisse eines größeren Forschungsprojektes oder die Befunde einer sehr guten Abschlussarbeit als wissenschaftlicher Zeitschriftenartikel veröffentlicht werden, so muss zunächst das geeignete Publikationsorgan identifiziert werden. Dabei ist als erstes auf die **thematische Passung** zu achten. Jede Fachzeitschrift listet auf ihrer Website in einer Selbstbeschreibung die von ihr abgedeckten Themenfelder auf, zudem liefern die ebenfalls online zugänglichen Inhaltsverzeichnisse der Zeitschriften Hinweise darauf, ob die zur Publikation vorgesehene eigene Studie in den Kontext der anderen Beiträge der Zeitschrift inhaltlich, methodisch, disziplinär etc. gut hineinpasst. Nicht zuletzt sollte beachtet werden, in welchen Publikationsorganen Artikel erschienen sind, auf die man sich in der eigenen Arbeit stützt.

Hat man auf diese Weise einige prinzipiell passende Zeitschriften identifiziert, so ist als zweites Auswahlkriterium die **Hochrangigkeit des Publikationsorgans** („journal ranking") zu beachten. Führende internationale Journals bringen einerseits die höchste Anerkennung, sofern eine Publikation dort gelingt, sind aber gleichzeitig mit dem höchsten Arbeitsaufwand sowie einer sehr hohen Ablehnungswahrscheinlichkeit verbunden. Je origineller und umfangreicher eine Studie ist, je mehr zeitliche und

personelle Ressourcen zur Verfügung stehen und je mehr Publikationserfahrung vorliegt, umso eher können internationale Journals als Publikationsort anvisiert werden. Andernfalls ist eine Einreichung bei einer deutschsprachigen Fachzeitschrift zu bevorzugen. Dies gilt auch, wenn Forschungsthemen bearbeitet werden, die einen starken lokalen oder nationalen Bezug haben und deswegen für ein internationales Publikum nur begrenzt von Interesse sind.

13.1.2 Autorenrichtlinien

Hat man sich für eine Zeitschrift entschieden, so sind deren **Richtlinien für Autoren**, die strikte Vorgaben hinsichtlich Umfang des Artikels, Format von Abbildungen, Zitationsweisen etc. machen, zu konsultieren und einzuhalten. Viele sozialwissenschaftliche Journals berufen sich mittlerweile auch auf die jeweils aktuelle Auflage des **Publication Manual** der **American Psychological Association** (American Psychological Association, 2009), das ausgesprochen anschaulich und detailliert Formalitäten des wissenschaftlichen Schreibens – von der Wortwahl über die Formatierung von Tabellen bis zu Zitationsregeln – erläutert.

Parallel zu den formalen Autorenrichtlinien sichtet man im betreffenden Journal bereits erschienene **Artikel zu ähnlichen Themen**. Dies bietet einen Einblick in die Konventionen der Zeitschrift und gibt Anregungen für die eigene Darstellungsweise (bis hin zu Details der Formulierung und Formalisierung, Gliederung und Formatierung).

13.1.3 Gliederung des Artikels

Bei einem empirischen Artikel ist die Gliederung relativ standardisiert in **fünf Hauptabschnitte**: 1. Einleitung („introduction"), 2. Theorie/Forschungsstand („theory"/„state of research"), 3. Methode („method"), 4. Ergebnisse („results") und 5. Diskussion („discussion"). In vielen empirisch ausgerichteten internationalen Journalen wird der Theorieteil in die Einleitung integriert. Sie beinhaltet dann nicht nur die Hinführung zum Thema, sondern auch die Darstellung des Forschungsstandes und den theoretischen Rahmen (► Kap. 6), so dass sich der Artikel dann auf **vier Hauptabschnitte** plus Literaturverzeichnis reduziert.

Die **Diskussion** ist neben dem **Abstract** (d. h. der Kurzzusammenfassung des Beitrags) der wichtigste Teil des Artikels. Denn sie wird vom kundigen Publikum – neben dem Abstract – zuerst gelesen, um die Quintessenz des Beitrags herauszufiltern. Entsprechend ist eine langatmige Wiederholung von Detailbefunden im Diskussionsteil eines Beitrags wenig sinnvoll. Stattdessen werden die zentralen Ergebnisse resümiert und in ihrer Bedeutung interpretiert und eingeordnet. Dabei werden zukünftige Forschungsperspektiven aufgezeigt, zudem wird auf die praktische Relevanz der Ergebnisse eingegangen. Jede gute Diskussion geht auch ausdrücklich auf die Grenzen der Aussagekraft der Studie ein und übt somit Selbstkritik. Die Diskussion sollte in der Regel nicht mehr als 2–4 Seiten umfassen, einen klaren roten Faden und möglichst keine oder nur sehr wenig Literaturzitierungen aufweisen (Bem, 2004).

Weitere Hinweise zu den Inhalten und den Funktionen der einzelnen Artikelabschnitte sind im APA Publication Manual (2009) zu finden. Auf der Basis der Standardgliederung und abgestimmt mit Beispielartikeln aus der Zielzeitschrift kann die eigene Gliederung entworfen und bis zur zweiten oder dritten Gliederungsebene verfeinert werden.

13.1.4 Auswahl der Inhalte

Nach dem Erstellen der Gliederung sollten die Inhalte des zu schreibenden Artikels **ausgewählt** und zunächst stichpunktartig den Gliederungspunkten zugeordnet werden. So stellt man meist in empirischen Artikeln im Ergebnisteil nicht alle Befunde in derselben Ausführlichkeit dar, sondern trifft eine sinnvolle Auswahl. Dabei sind v. a. drei Auswahlkriterien zu beachten:

1. **Gesamtlänge des Artikels**: Die in der Autorenrichtlinie der Zeitschrift vorgegebene maximale Länge des Beitrags darf keinesfalls überschritten werden. Daraus ergibt sich in der Regel ein genereller Selektionszwang.
2. **Angemessenes Verhältnis der einzelnen Teile des Artikels**: Die einzelnen Teile des Artikels sollten in ausgewogenem Verhältnis zueinander stehen. So wirkt ein 20-seitiges Paper zu einer empirischen Studie, das 12 Seiten lang Theorien referiert und nur auf 2 Seiten Ergebnisse präsentiert, unausgewogen. Anhand der Gliederung, der Vorgabe zur maximalen Länge sowie dem Anspruch einer ausgewogenen Binnenstruktur kann den einzelnen Gliederungspunkten jeweils ein entsprechender Umfang zugeordnet werden (z. B. Einleitung: 1 Seite, Theorieteil mit Forschungsstand und theoretischem Hintergrund, Forschungsfragen bzw. Forschungshypothesen: 4 Seiten, Methodenteil: 2 Seiten, Ergebnisteil: 6 Seiten, Diskussion: 3 Seiten, Literaturverzeichnis: 4 Seiten).
3. **Relevanz der Inhalte für das Zielpublikum**: Ein wissenschaftlicher Zeitschriftenartikel ist für ein vorgebildetes Fachpublikum gedacht. Textpassagen, in

denen das eigene Wissen unter Beweis gestellt wird (z. B. ausführliche Erklärung, was eine Quotenstichprobe ist), sind zwar in Qualifikationsarbeiten teilweise notwendig, in Fachartikeln allerdings nicht üblich. Bei der Auswahl der empirischen Ergebnisse sollte das Augenmerk auf Befunde gelegt werden, deren theoretische und/oder praktische Implikationen für das Zielpublikum des Journals besonders relevant sind. Wissenschaftstheoretisch sowie wissenschaftsethisch besonders wichtig sind methodische Hintergrundinformationen, die intersubjektive Nachvollziehbarkeit (Objektivität) des Erkenntnisgewinns sowie empirische Replizierbarkeit sicherstellen und vor Wissenschaftsfälschungen schützen. Da entsprechende Detailinformationen (z. B. Fragebögen und Interviewleitfäden im Wortlaut) aus Platzgründen oft nicht im Artikel untergebracht werden können, kann man die interessierte Leserschaft darauf hinweisen, dass und wo entsprechendes Material zum Download im Internet hinterlegt ist oder auf Anfrage per E-Mail zugeschickt wird.

13.1.5 Schreibprozess

Der Schreibprozess bzw. das inhaltliche Füllen der Gliederung basiert bis zu einem gewissen Grad auf Kreativität, vor allem aber auf der Anwendung von **handwerklichen Fertigkeiten** des wissenschaftlichen Schreibens (Booth, Colomb, & Williams, 2008). Scherzhaft spricht man deswegen von „10 % Inspiration und 90 % Transpiration". Je mehr man bereits geschrieben hat, desto leichter fällt es. Durch die weitestgehend vorgegebene Struktur ist der Zeitaufwand für einen empirischen Artikel (sofern Literatur und empirische Daten bereits vorliegen) im Grunde recht gut abschätzbar. Voraussetzung ist, dass man das eigene Schreibtempo kennt. Es ist deswegen empfehlenswert, die eigene Schreibgeschwindigkeit zu protokollieren, um **realistische Arbeits- und Zeitpläne** erstellen und dann auch planmäßig umsetzen zu können. Erfahrungsgemäß erleichtert es das zügige Vorankommen, wenn mehrere Tage am Stück konzentriert am selben Text gearbeitet werden kann (vorzugsweise im Home Office). Nach jeder Unterbrechung muss man sich nämlich neu eindenken und verliert dadurch relativ viel Zeit.

Schreibblockaden können dadurch entstehen, dass man beim Schreiben einen perfekten Stil anstrebt und deswegen z. B. lange über Formulierungen grübelt und oft umformuliert. Da der Artikel ohnehin mehrfach überarbeitet und korrigiert werden muss, gibt es jedoch später immer noch genügend Gelegenheit für stilistischen Feinschliff. Dementsprechend kann man sich zunächst auf die Inhalte konzentrieren und einfach spontan formulieren. Manchmal hilft es, nicht mit der Einleitung anzufangen, sondern z. B. mit dem Methodenteil (dieser ist sehr konkret, die Struktur der einzelnen Teile ist klar vorgegeben). In der Regel ist es so, dass sich irgendwann ein **Schreibfluss** einstellt. Förderlich kann es auch sein, bereits vorliegende Textteile durchzulesen, um in das Schreiben hineinzukommen. Es kann auch helfen, die Hauptinhalte des nächsten Absatzes zunächst mündlich zu formulieren, um sich den roten Faden klarzumachen (zum Umgang mit und zur Prävention von Schreibblockaden s. auch Silvia, 2002).

Es ist darauf zu achten, nicht zu salopp umgangssprachlich oder zu journalistisch zu schreiben, sondern sich am **wissenschaftlichen Sprachstil** zu orientieren und z. B. Fachbegriffe korrekt zu verwenden. Dennoch muss gewährleistet werden, dass der Text anschaulich und verständlich ist. Auch hier finden sich Hinweise zur Umsetzung im APA Publication Manual (2009) oder in entsprechenden Praxisbüchern (z. B. Wallwork, 2011). Bei sozialwissenschaftlichen Veröffentlichungen sollte die **gendergerechte Sprache** berücksichtigt werden, d. h. eine ausschließliche Verwendung des generischen Maskulinums, um Menschen aller Geschlechter anzusprechen bzw. zu repräsentieren, sollte vermieden werden. Diese Forderung ist mittlerweile in vielen Autorenrichtlinien wissenschaftlicher Zeitschriften und auch im APA Publication Manual explizit aufgeführt. Der flexible Umgang mit unterschiedlichen Alternativen zum generischen Maskulinum führt auch keineswegs automatisch dazu, dass Texte unlesbar werden. Dieses Problem tritt nur bei einer schematischen und starren Umsetzung auf.

Wichtig für die Lesbarkeit ist außerdem eine klare und konsistente **Struktur des Textes**. Die Hauptgliederungsebenen sind standardisiert, aber auch Unterpunkte sollten in sinnvolle Einheiten gegliedert sein. Teilabschnitten des Artikels ist jeweils ein Ausblick auf die im Folgenden dargestellten Inhalte („Advanced Organizer") voranzustellen. „Textwüsten" sollten vermieden werden. Stattdessen müssen sinnvolle inhaltliche Einheiten gebildet und voneinander abgegrenzt werden (Zwischenüberschriften, Tabellen, Abbildungen, Anstrichlisten etc.). Wichtig für die Qualität des Textes ist die **Meta-Reflexion**, d. h. die Inhalte sollten nicht nur dargestellt, sondern in ihrer Bedeutung, ihren Querbezügen, ihrer Aktualität oder Relevanz immer wieder mit Abstand und Blick auf das Forschungsproblem betrachtet und eingeordnet werden.

13.1.6 Literaturverwaltung und Zitation

Ein Literaturverwaltungsprogramm (z. B. Endnote, BibTeX etc.) hilft bei der Organisation der Quellen und der Erstellung des Literaturverzeichnisses. Idealerweise wer-

den bereits bei der Literaturrecherche alle gefundenen Quellen in das Literaturverwaltungsprogramm eingelesen, so dass eine für das bearbeitete Forschungsthema einschlägige Literaturdatenbank vorliegt (zu Literaturverwaltungsprogrammen ▶ Abschn. 6.1.3). Viele dieser Programme verfügen über eine „**Cite while you write**"-**Funktion** und ermöglichen ein unkompliziertes Einfügen von Quellenangaben beim Schreiben und erstellen bzw. aktualisieren parallel das Literaturverzeichnis. Die Vorgaben der Autorenrichtlinien zur Gestaltung des Literaturverzeichnisses können dabei im Programm eingestellt werden.

13.1.7 Formatierung

Es empfiehlt sich, von Anfang an die **Formatvorlagen** in den Textverarbeitungsprogrammen zu benutzen. Hier kann man die Formatierung von Titeln, Untertiteln, Überschriften, Textkörper und Aufzählungszeichen noch vor dem Schreiben definieren. So kann der Text bereits während des Schreibens durch Anwendung der entsprechenden Formatvorlagen grob durchformatiert werden. Sonderformate wie Hervorhebungen oder Tabellenbestandteile werden dann im Arbeitsprozess nach Bedarf angelegt. Das Inhaltsverzeichnis kann automatisch erstellt werden, wenn man die Überschriften mit den Formatvorlagen als solche formatiert hat.

Eine andere Arbeitsvariante verlagert die Formatierung auf das Ende des Schreibprozesses. So sehen z. B. viele Autorenrichtlinien einen Zeilenabstand von 1.5 Zeilen vor. Dies kann während des Schreib- und Korrekturprozesses stören, bei dem es für einen besseren Überblick meist hilfreicher ist, mit mehr Text pro Seite zu arbeiten. In diesem Fall sind entsprechende Formatierungen erst nach Fertigstellung des Manuskripts sinnvoll.

13.1.8 Tabellen und Abbildungen

Besonders wichtige Inhalte werden visuell vom Fließtext abgehoben (z. B. Grafik des Theoriemodells, Tabelle der Hauptergebnisse, Foto der Versuchsanordnung, Screenshot von Beispielen aus einer Online-Inhaltsanalyse etc.). Eine grafische Darstellung von Nebensächlichkeiten (z. B. Altersverteilung der Stichprobe) ist nicht sinnvoll, ebenso sollten dieselben Inhalte nicht doppelt und ggf. dreifach (im Fließtext und in einer Tabelle und in einer Abbildung) dargestellt werden – eine einzige Darstellungsform ist jeweils ausreichend. Abbildungen (z. B. Fotos) müssen eine druckfähige Qualität/Auflösung besitzen und der Zeitschrift in separaten Dateien zur Verfügung gestellt werden. Jede Abbildung und jede Tabelle muss

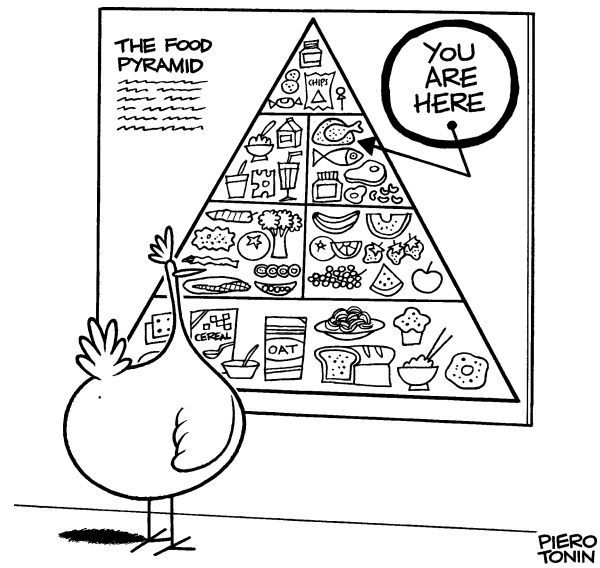


Abb. 13.1 Zentrale Inhalte sollten in selbsterklärenden Abbildungen dargestellt werden. (© Piero Tonin/Search ID: pton296, Rights Available from CartoonStock.com)

losgelöst vom Fließtext verständlich sein Abb. 13.1, weshalb eine aussagekräftige Legende notwendig ist (zur Gestaltung von Tabellen siehe Nicol & Pexman, 2010).

13.1.9 Korrektur des Manuskriptes

Das Korrekturlesen von Manuskripten durchläuft meist mehrere Schleifen (zunächst liest man selbst mehrmals Korrektur, dann Bekannte und/oder Kollegen). Es ist völlig normal, z. B. beim dritten Korrekturdurchgang noch einmal größere Änderungen vorzunehmen. Es kann sogar sein, dass eine Änderung vom Anfang des Korrekturzyklus später wieder zurückgeändert wird, damit sich ein stimmiges Gesamtbild ergibt. Erst durch diesen geduldigen und sorgfältigen „Feinschliff" kommt ein Artikel zur Reife.

Wichtig ist, dass die Korrekturlesenden **jegliche kritische Anmerkung** tatsächlich einfließen lassen. Hier bieten sich die Verwendung der Korrekturmodi und Kommentarfunktionen des jeweiligen Textverarbeitungsprogramms und gegebenenfalls auch mündliche Rückmeldungen an. Eine Schonung ist nicht angemessen, denn das Peer-Review-Verfahren wird in der Regel noch viel strenger sein. Angemerkt werden sollten: Inhaltliche Fehler, inhaltliche Brüche und unlogische Bezüge, Auslassungen, nicht nachvollziehbare Interpretationen, Rechtschreib- und Grammatikfehler, verbesserungswürdige Formulierungen etc. Alles, was auch nur im Mindesten negativ auffallen könnte, sollte auch im Feedback mitgeteilt wer-

den. Die Korrigierten nehmen das Feedback neutral auf und **prüfen jeden Hinweis** hinsichtlich seiner Relevanz. Es werden meist nicht alle Anmerkungen umgesetzt. Jedoch sollte man dann begründen können, warum man entsprechende Änderungsvorschläge nicht aufgreift. Wer ein Manuskript nicht mehrfach gründlich überarbeitet hat, ist noch nicht fertig. Neulinge unterschätzen erfahrungsgemäß den großen Aufwand, der mit **mehreren grundlegenden Überarbeitungsschleifen** verbunden ist, in deren Verlauf ggf. ganze Abschnitte neu geschrieben, zusätzliche Literaturquellen verarbeitet, einzelne Datenanalysen revidiert, Interpretationen relativiert und zusätzliche Abbildungen erstellt werden müssen.

13.1.10 Artikeleinreichung und Peer Review

Die Einreichung des Beitrags zusammen mit einem Begleitschreiben erfolgt heute meist elektronisch: Entweder per E-Mail-Attachment an die Redaktion der Zeitschrift oder über eine Einreichungs-Website. Erfüllt der Artikel die formalen Voraussetzungen gemäß Autorenrichtlinien (z. B. korrekte Formatierung, Einhaltung der maximalen Seiten- bzw. Zeichenzahl), so wird er in das doppelblinde Peer-Review-Verfahren gegeben. Andernfalls wird er gar nicht zur Begutachtung angenommen. Die anonyme Begutachtung durch mindestens zwei Fachkollegen sowie die auf der Basis der Gutachten zu treffende Herausgeberentscheidung beanspruchen in der Regel eine Zeitspanne von mehreren Monaten. In dieser Phase befindet sich das Paper im Status „**under review**“. Es darf als solches zitiert werden (z. B. in Drittmittel-Anträgen), sofern das Manuskript beigelegt wird. Während des Review-Prozesses darf das Manuskript jedoch keiner weiteren Zeitschrift zur Begutachtung vorgelegt werden, um doppelte Arbeit von den Gutachtenden abzuwenden.

Das **Ergebnis der ersten Runde des Peer-Review-Prozesses** besteht in einer Mitteilung der Herausgebenden der Zeitschrift, wonach der Artikel entweder

- zur Veröffentlichung angenommen ist – mit ggf. kleineren Änderungsauflagen („accept“; „accept with minor revisions“),
- nach einer grundlegenden Überarbeitung erneut zur Begutachtung eingereicht werden darf („revise and resubmit“) oder
- definitiv von der betreffenden Zeitschrift abgelehnt wurde („reject“).

Dem Herausgeberbrief mit der entsprechenden Mitteilung sind üblicherweise die Gutachten beigefügt. Im Falle einer Ablehnung sollte man die Gutachterhinweise beachten und eine überarbeitete Fassung des Manuskripts einer anderen, vielleicht thematisch passenderen oder weniger hochrangigen Zeitschrift anbieten. Wird die Möglichkeit zur Wiedereinreichung einer revidierten Fassung bei derselben Zeitschrift geboten, so sollte diese genutzt werden. Dafür ist das Manuskript akribisch gemäß den Gutachteranmerkungen sowie dem **Herausgeberbrief** („action letter“) zu überarbeiten. Dass und wie die einzelnen Anmerkungen umgesetzt wurden, muss bei der Einreichung des überarbeiteten Manuskriptes in einem **Begleitschreiben** („cover letter“) detailliert dokumentiert werden. Zum Umgang mit Review-Anmerkungen und dem Verhalten während des Review-Prozesses gibt es zahlreiche Praxisbücher (z. B. Wager, Godlee, & Jefferson, 2002). Auch auf den Websites der Wissenschaftsverlage sind mitunter Hinweise und Tipps zum Review-Prozess zu finden (z. B. die How-to-Guides auf http://www.emeraldinsight.com/authors/guides/). Üblicherweise wird in dem Begleitschreiben zu einem erneut eingereichten, überarbeiteten Manuskript jeder einzelne Kritikpunkt aus den Gutachten aufgegriffen (z. B. wörtlich zitiert) und explizit beantwortet, und zwar mit einem Verweis auf entsprechende Änderungen im Manuskript oder triftiger Begründung, warum die gewünschte Änderung nicht vorgenommen wurde. Die Einreichung des überarbeiteten Manuskripts zusammen mit dem Begleitschreiben muss zeitnah erfolgen; meist wird eine Frist von 4–6 Wochen eingeräumt.

An der **zweiten Runde des Peer-Review-Verfahrens** nehmen üblicherweise dieselben Gutachtenden wie in der ersten Runde teil. Wurden die monierten Schwächen und Fehler des Manuskriptes sorgfältig beseitigt, so wird der revidierte Artikel in der Regel angenommen. Es können aber auch eine zweite oder dritte Nachbesserung notwendig sein, wenn die Gutachtenden noch nicht von der Qualität des Manuskriptes überzeugt sind.

Liegt schließlich die Herausgeber-Mitteilung vor, dass der überarbeitete Beitrag („revised paper“) angenommen wurde, so befindet sich der Artikel nicht mehr im Status „under review“, sondern im Status **„in press“** (im Druck). Unmittelbar vor der Publikation erhält man noch die Korrekturfahnen und muss eine Druckfreigabe erteilen. Da wissenschaftliche Fachzeitschriften gedruckt nur zwei- bis viermal pro Jahr erscheinen, sind die Hefte schnell „ausgebucht“, so dass ein angenommener Artikel zuweilen erst nach Monaten oder Jahren erscheinen kann (eine Vorab-Publikation – **„online first“** - auf der Website der Fachzeitschrift ist üblich, um die Verzögerung der Print-Publikation auszugleichen). Insgesamt verstreichen von der Ersteinreichung eines Manuskriptes bis zur Veröffentlichung der finalen Fassung nicht selten ein bis drei Jahre.

13.1.11 Verbreitung des Artikels

Wurde der Artikel veröffentlicht, gibt es einige Strategien, um dessen Verbreitung und damit auch die Zitationswahrscheinlichkeit zu erhöhen. Zuweilen stellen die Verlage den Autorinnen und Autoren **Sonderdrucke** („reprints") kostenfrei oder kostenpflichtig zu Verfügung. Diese können pro-aktiv an im selben Bereich tätige Kolleginnen und Kollegen versendet werden. Ebenso können solche Sonderdrucke oder Ausdrucke des elektronischen Artikels auf Konferenzen ausgelegt werden. Dies bietet sich insbesondere dann an, wenn man selbst auf der entsprechenden Konferenz mit einem Vortrag oder Poster zum Thema vertreten ist.

Eine **Manuskriptfassung** des Artikels, die auf den publizierten Artikel verweist, kann auch auf der eigenen Internetpräsenz oder in Open-Access-Repositorien zum freien Download angeboten werden. Auf diese Weise ist der Beitrag leicht auffindbar, was wiederum seine Rezeption und Zitation fördert. Die Rechte an der gedruckten Endfassung des Artikels liegen in der Regel, zumindest für begrenzte Zeit, beim Verlag, so dass die Endfassung nur bedingt frei verbreitet werden darf.

Manche Verlage bieten den Autorinnen und Autoren jedoch mittlerweile **Open-Access-Optionen** an, d. h. die Autoren bezahlen dafür, dass ihr Artikel kostenfrei im Internet abrufbar sein darf. Zudem existieren zunehmend mehr Open-Access-Fachzeitschriften, bei denen alle Beiträge frei zugänglich sind. In diesem Fall kann man den entsprechenden Online-Link auf den Volltext Interessierten gezielt per individueller E-Mail, Mailingliste oder über Web-2.0-Kanäle (z. B. Weblogs, Microblogs, Social Networking Sites) zukommen lassen (zur Wissenschaftskommunikation Online ► Abschn. 13.4.1).

13.2 Konferenzvortrag

Neben Zeitschriftenartikeln haben mündliche Ergebnispräsentationen in Form von Fachvorträgen auf Konferenzen einen hohen Stellenwert in der Scientific Community. Die Ergebnisse einer empirischen Studie werden im Rahmen eines **Forschungsreferats** („research paper") präsentiert. Diese Vortragsform unterscheidet sich vom **Positionsreferat** („position paper"), das nicht eine einzelne Studie, sondern ein ganzes Forschungsfeld behandelt. Neben Forschungsreferaten und Positionsreferaten als begutachteten Konferenzvorträgen werden auf Konferenzen auch **Grundsatzreferate** („key note speech") gehalten, zu denen die Konferenzleitung führende Forschungspersönlichkeiten einlädt. Bei kleineren Konferenzen wird ein einzelnes Grundsatzreferat als Eröffnungsvortrag eingeplant, bei größeren Konferenzen finden täglich Grundsatzreferate (z. B. in der Mittagszeit) statt.

Der wissenschaftliche Fachvortrag in Form eines **Forschungsreferats** dient der mündlichen Präsentation von neuen Forschungsergebnissen. Dafür sind meist 15–25 Minuten Präsentationszeit plus 5 Minuten Diskussion vorgesehen. Typischerweise werden wissenschaftliche Vorträge in den Human- und Sozialwissenschaften heute durch einen Satz an Präsentationsfolien begleitet (► Abschn. 13.2.5). Über die Annahme eines Vortrags zu einer wissenschaftlichen Konferenz entscheidet ein doppelblindes Peer-Review-Verfahren. Einzureichen ist dafür im Vorfeld entweder das **Abstract** (Kurzzusammenfassung), ein **Extended Abstract** (erweiterte Zusammenfassung) oder ein **Full Paper** (vollständig ausformulierter Vortrag; entspricht einem Zeitschriftenartikel). Nicht selten werden Forschungsergebnisse als erstes auf einer Konferenz präsentiert und danach als Zeitschriftenaufsatz publiziert, insbesondere wenn als Konferenzeinreichung bereits ein Full Paper erstellt wurde.

Die Abstracts oder Paper zu den Vorträgen werden in **Tagungs- bzw. Konferenzbänden**, den sog. **Proceedings**, publiziert. In den Technikwissenschaften sind referierte Vorträge wichtiger als Zeitschriftenpublikationen, da letztere meist einen Vorlauf von ein bis drei Jahren haben und innovative Technologien dann schon veraltet sein können. In den Technikwissenschaften enthalten die Proceedings üblicherweise Kurzpaper (ca. fünf Seiten pro Vortrag), die als zitierfähige Publikationen gelten. In den Sozialwissenschaften haben Konferenzvorträge einen geringeren Stellenwert als Zeitschriftenartikel, sind aber dennoch als wissenschaftlicher Output wichtig und dienen zudem ganz wesentlich der persönlichen Bekanntheit und Vernetzung in der Fachcommunity. Wer sich später im Wissenschaftsbetrieb auf eine Stelle bewirbt, sollte zumindest einigen Mitgliedern der Auswahlkommission von Konferenzen schon persönlich bekannt sein.

In den unterschiedlichen Wissenschaftsdisziplinen variiert nicht nur die Bedeutung von Vorträgen, sondern auch der dominierende **Vortragsstil**. Während in den Geistes- und Kulturwissenschaften bei Vorträgen häufig Manuskripte wortwörtlich vorgelesen werden, weil viel Mühe in die exakte Wortwahl und Formulierung investiert wurde, werden sozial- und technikwissenschaftliche Vorträge in der Regel frei gesprochen und stützen sich auf Präsentationsfolien oder andere Veranschaulichungen. Auf internationalen Konferenzen ist häufig ein nordamerikanisch geprägter, eher lockerer Stil verbreitet, bei dem durch persönliche Bemerkungen und Anekdoten Nähe zum Publikum hergestellt wird. Videoaufzeichnungen wissenschaftlicher Vorlesungen und Konferenzvorträge finden sich u. a. auf YouTube sowie auf Konferenz-Websites (z. B. TED [Technology, En-

tertainment, Design]: http://www.ted.com; International Communication Association [ICA] Virtual Conference: http://www.icavirtual.com) und auf Universitäts-Websites (z. B. des MIT [Massachusetts Institute of Technology]: http://watch.mit.edu). Vortragsmanuskripte und Foliensätze können nicht selten auch von den Internetpräsenzen der Forscherinnen und Forscher heruntergeladen werden.

13.2.1 Auswahl der Konferenz

Die Teilnahme an Konferenzen muss rechtzeitig im Vorfeld geplant werden. Das betrifft die inhaltliche Auswahl (welche nationalen und/oder internationalen Konferenzen sollen besucht werden), aber auch organisatorische Fragen (z. B. Finanzierungsquellen). Ebenso wie bei der Auswahl einer passenden Fachzeitschrift für einen Artikel müssen auch bei der Auswahl der passenden Konferenz die thematische Passung sowie die Hochrangigkeit der Veranstaltung berücksichtigt werden. Nicht zuletzt variieren auch Arbeitsaufwand (Anmeldung mittels Abstract oder Full Paper) und Kosten (z. B. nationaler Tagungsort vs. Tagungsort in Übersee) für Konferenzbesuche sehr stark.

Wichtigste Veranstalter wissenschaftlicher Konferenzen sind die nach Fachdisziplinen wissenschaftlichgetrennt organisierten **wissenschaftlichen Fachgesellschaften**, wobei sowohl nationale als auch internationale (v. a. nordamerikanische bzw. europäische) Gesellschaften und deren Konferenzreihen zu beachten sind. Im deutschsprachigen Raum sind etwa Fachgesellschaften wie die Deutsche Gesellschaft für Psychologie (DGPs), die Deutsche Gesellschaft für Soziologie (DGS), die Deutsche Gesellschaft für Erziehungswissenshaft (DGEW), die Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft (DGPuK) oder die Deutsche Gesellschaft für Online-Forschung (DGOF) zu nennen, die in der Regel alle ein bis zwei Jahre an wechselnden Konferenzorten eine **große Konferenz des gesamten Faches** durchführen. Auf internationaler Ebene sind Fachgesellschaften wie die APA (American Psychological Association), die ICA (International Communication Association) oder die IAMCR (International Asssociation for Media and Communication Research) zu nennen, auf europäischer Ebene z. B. die EASP (European Association of Social Psychology) oder die ECREA (European Communication Research and Education Association). Neben den großen Konferenzen sind **kleinere Fachkonferenzen einzelner Sektionen/Fachgruppen der Fachgesellschaften** (z. B. Fachgruppe „Medienpsychologie“ innerhalb der Deutschen Gesellschaft für Psychologie), die ebenfalls ein- bzw. zweijährlich stattfinden, relevant. Neben den Konferenzreihen der Fachgesellschaften existieren auch **Konferenzreihen zu einzelnen Themengebieten** z. B. die mLearn (World Conference on Mobile and Contextual Learning), die TeaP (Tagung experimentell arbeitender Psychologen), die Internet Research (Annual Conference of the Association of Internet Researchers) oder die Sunbelt (Conference of the International Network for Social Network Analysis).

Über die Websites, Newsletter und Mailinglisten der Fachgesellschaften wird auf Konferenzen mittels eines **Teilnahmeaufrufs** (Call for Contributions/Call for Abstracts/ Call for Papers; kurz: **Call**) aufmerksam gemacht. Eine weitere Möglichkeit, relevante Konferenzen zu identifizieren, ist die Recherche über die Vortragslisten von etablierten Forschenden des jeweiligen Wissenschaftsgebiets. Die Konferenzen, die von ihnen besucht werden, sind in vielen Fällen auch für eigene Vortragseinreichungen relevant. Jede Konferenz hat eine **Deadline** (Einsendeschluss) für die Beitragseinreichung. Diese liegt oft bis zu einem Jahr vor dem eigentlichen Konferenztermin und muss unbedingt im Auge behalten werden, da verspätete Einreichungen nicht akzeptiert werden.

Auf welchen Konferenzen man mit eigenen Beiträgen vertreten sein möchte, muss rechtzeitig geplant werden. Dabei spielen auch **Kosten** eine wichtige Rolle: So schwanken die Teilnahmegebühren teilweise deutlich und lassen sich durch eine Mitgliedschaft in entsprechenden Fachgesellschaften (Vollmitgliedschaft ab Doktorgrad, vorher assoziierte Mitgliedschaft) sowie durch den Frühbucherrabatt reduzieren. Zudem hängen die Gesamtkosten maßgeblich vom nationalen oder internationalen **Konferenzort** ab. Man sollte es sich nicht entgehen lassen, einschlägige internationale Konferenzen zu besuchen, wenn diese einmal in Europa oder gar im deutschsprachigen Raum stattfinden. Es empfiehlt sich, möglichst frühzeitig eine Unterkunft zu suchen. Die Erfahrung zeigt, dass günstige Hotels in der Nähe des Tagungsortes für den Konferenzeitraum schnell ausgebucht sind. Bei einem Besuch von Konferenzen im Ausland kann eine Bezuschussung durch den DAAD (Deutscher Akademischer Austauschdienst) im Rahmen der Förderlinien „Kongressreisen ins Ausland“ und „Vortragsreisen ins Ausland“ beantragt werden. Hierbei sollten die Fristen (ca. vier Monate vor der Reise) zur Bewerbung unbedingt eingehalten werden. Alternativ bieten viele Stiftungen und Unternehmen **Reisekostenstipendien** oder „Travel Grants“ für Nachwuchswissenschaftlerinnen und -wissenschaftler an.

13.2.2 Abstract

Je nach Konferenz ist für die Anmeldung eines Vortrags ein **Abstract** (einfache Kurzzusammenfassung; z. B. bei der DGPs-Tagung 150–300 Wörter), ein **Extended Abstract** (erweiterte Zusammenfassung; z. B. bei der DGPuK-Jahrestagung 4 000–6 000 Zeichen inklusive Leerzeichen) oder ein **Full Paper** (vollständiges Vortragsmanuskript in Form eines Zeitschriftenartikels; z. B. bei der ICA-Jahrestagung ca. 25 Seiten) einzureichen. Das einzureichende Format des Abstracts ist dem jeweiligen Teilnahmeaufruf der Konferenz (Call) zu entnehmen.

In den meisten Fällen werden Abstracts und Extended Abstracts verlangt, die später dann auch im Konferenzband veröffentlicht werden. Bei englischsprachigen Einreichungen durch Nicht-Muttersprachler ist eine Vorabkorrektur durch Native Speaker empfehlenswert, da im Falle einer Annahme das Abstract veröffentlicht wird und im Nachhinein meist nicht mehr korrigiert werden kann.

Inhaltlich sollten Abstracts der klassischen Gliederung folgen und mit jeweils etwa einem knappen Satz auf Forschungsproblem, theoretischen Hintergrund/Forschungsstand, Forschungsfragen und Hypothesen, Methode, Ergebnisse und Diskussion eingehen. Weitere Hinweise zu Aufbau und Inhalten verschiedener Formen von Abstracts finden sich im APA Publication Manual (2009).

13.2.3 Beitragseinreichung und Peer Review

Die Einreichung eines Konferenzvortrags muss vor der **Deadline** erfolgen. Es hat sich etabliert, Beitragseinreichungen über die Konferenz-Websites mit speziellen Konferenzverwaltungsprogrammen zu organisieren. Mit der Einreichung sollte man nicht bis kurz vor Ablauf der Deadline warten, da dann die Server erfahrungsgemäß oft überlastet sind.

Auf der Basis der eingereichten Abstracts, Extended Abstracts oder Full Papers erfolgt der Peer-Review-Prozess, der meist mehrere Wochen in Anspruch nimmt. Im Unterschied zum Peer Review bei Zeitschriftenartikeln, die häufig mehrstufig verlaufen und oft eine gründliche Überarbeitung des Manuskripts verlangen, ist das Review von Konferenzvorträgen in der Regel **einstufig**: Der eingereichte Konferenzvortrag wird entweder als solcher **angenommen**, vom Vortrag in ein **Poster umgewandelt** (▶ Abschn. 13.3) oder **abgelehnt**. Nachbesserungen und Überarbeitungen sind hier untypisch und könnten aufgrund der Fülle der Beiträge – auf den großen Konferenzen der Fachgesellschaften werden Hunderte von Vorträgen gehalten – sowie des Zeitdrucks im Hinblick auf den näher rückenden Konferenztermin kaum geleistet werden.

13.2.4 Vortragsplanung

Die Vorbereitung eines wissenschaftlichen Vortrags beginnt genau wie die Vorbereitung eines Artikels mit der **Gliederung**. Sie folgt dem klassischen Schema der Darstellung empirischer Studien. Den einzelnen Gliederungspunkten sind dann die ausgewählten Inhalte zuzuordnen. Die Obergrenze bildet das Zeitlimit, zudem ist auf eine Ausgewogenheit der einzelnen Teile zu achten. So sollte z. B. ein lebendiger Einstieg geplant werden, dennoch sollte die Einleitung nicht zu viel Zeit beanspruchen, da die Ergebnisse der inhaltlich wichtigste Teil sind. Von herausgehobener Bedeutung ist beim Vortrag das Fazit („take home message", „wrap up"), das die Hauptbefunde und daraus abzuleitenden Schlussfolgerungen auf den Punkt bringt und dem Publikum hoffentlich im Gedächtnis bleibt (Kuzbari & Ammer, 2006).

Als Grundgerüst für die Planung des Vortrags und die Vorbereitung der Präsentationsfolien (▶ Abschn. 13.2.5) seien für den Vortragstyp **Forschungsreferat** (Darstellung einer empirischen Studie) folgende Hinweise gegeben:

1. **Lebendiger Einstieg in das Thema**: Der Vortragseinstieg sollte Interesse am Thema wecken und das Publikum „abholen". Dazu kann eine Anekdote, eine provokante These, eine Frage an das Publikum, ein Bild, ein Fallbeispiel, eine aktuelle Nachricht und ggf. auch ein Witz oder Cartoon genutzt werden. Neben einer guten Motivation für das Thema sollte auch eine anschauliche Gegenstandsbeschreibung (z. B. anhand von Beispielen, Fotos, Zitaten etc.) nicht fehlen.
2. **Vortragsstruktur darstellen**: Wurden die Bedeutung des Vortragsthemas erläutert und der Untersuchungsgegenstand beschrieben, sollte die Struktur des Vortrags anhand einer Gliederungsfolie verdeutlicht werden.
3. **Hauptinhalte in ausgewogener Weise präsentieren**: Die einzelnen Aspekte der im Konferenzvortrag vorgestellten Studie sollten in zeitlich ausgewogener Form präsentiert werden. Dies erfordert eine strikte Reduktion auf das Wesentliche. Zudem sollte immer klar sein, wo in der Gliederung man sich gerade befindet.
4. **Studienergebnisse ins Zentrum rücken**: Im Zentrum eines Forschungsreferats sollten die Studienergebnisse stehen, die dann auch in tabellarischer und/oder grafischer Form darzustellen und ausführlich zu erläutern sind. Typische Vortragsfehler bestehen darin, dass a) mit Einleitung, Theorie und Methode

die meiste Zeit verbraucht wird und die Ergebnispräsentation dann sehr knapp und hektisch ausfällt oder dass b) das Publikum mit einer Überfülle an Detailergebnissen konfrontiert wird (z. B. riesige Tabellen, zahlreiche statistische Kennwerte, Darstellung jeder einzelnen Kontrollvariable oder Teilgruppe) und somit der Überblick verloren geht.
5. **Mit einprägsamem Fazit enden**: Die theoretischen und praktischen Schlussfolgerungen aus der Studie sollten deutlich herausgearbeitet und zu wenigen Kernpunkten, einem Appell oder einer Pointe verdichtet werden.

Forschungsreferate werden auf Konferenzen zu thematisch passenden Referatsblöcken (**Panels** bzw. **Sessions**) zusammengefasst. Es ist empfehlenswert, bei der Vorbereitung des eigenen Vortrags die Einordnung in das Panel zu berücksichtigen, etwa um Dopplungen mit benachbarten Vorträgen zu vermeiden oder umgekehrt inhaltliche Querbezüge herzustellen. Im Zweifelsfall kann man sich mit anderen Referierenden aus dem Panel im Vorfeld abstimmen.

13.2.5 Präsentationsfolien und Zusatzmaterial

Die Verwendung von Präsentationsfolien hat sich in den Human- und Sozialwissenschaften eingebürgert (da sich zur Erstellung von Präsentationsfolien die Software **Microsoft Powerpoint** aus dem Office-Paket so stark etabliert hat, wird synonym oft von Powerpoint-Folien gesprochen). Der routinemäßige Einsatz von Folien bei wissenschaftlichen Vorträgen ist nicht ohne Kritik geblieben: Tatsächlich können schlechte Präsentationsfolien sowie eine ungünstige Kombination von mündlichem Vortrag mit begleitenden Folien ein Forschungsreferat abwerten. Besonders negativ fällt es auf, wenn in schneller Folge zu viele Folien gezeigt werden, die vom Publikum nicht mehr aufgenommen werden können und den Eindruck von Hektik vermitteln („Folienschleuder"). Als Daumenregel gilt: **Maximal 1 Folie pro Minute** einplanen. Der zweite häufige Fehler besteht darin, dass Folien überladen, kaum lesbar und unübersichtlich wirken, weil zu viel Text untergebracht wurde, die Schrift zu klein ist und ungünstige Farbkontraste gewählt wurden. Hier gilt: **Reduktion auf das Wesentliche:** Weniger mit Text, lieber mit grafischen Mitteln arbeiten. Die im Foliensatz angestrebte Anschaulichkeit sollte jedoch nicht zu einem Ablenkungseffekt führen: Bilder, die nur dekorativen Charakter haben, lenken vom eigentlichen Inhalt ab. Eine angemessene Schriftgröße ist zu wählen (je nach Schriftart mindestens 18 oder 20 Punkt). Auf klare und angenehme Farbkontraste ist zu achten, die auch bei Farbverzerrungen durch den lokalen Beamer noch funktionieren. Generell sollte auf spielerische Elemente wie auffällige visuelle oder gar akustische Folienübergänge verzichtet werden. Ebenso wirken die folientypischen **Anstrichlisten** („bullet point lists") auf Dauer sehr ermüdend. Dies gilt insbesondere, wenn sie von den Vortragenden dann wörtlich vorgelesen werden (Franck & Stary, 2006).

Generell gilt: Der **mündliche Vortrag** sollte in sich rhetorisch gestaltet werden, mit entsprechenden Erzählungen und Spannungsbögen, die Folien sollen **nur begleitend** und unterstützend wirken. Dementsprechend muss auch nicht jedes Foliendetail mündlich wiederholt werden. Insbesondere sollte der Fehler vermieden werden, dass die Vortragenden sich auf die Folien konzentrieren, anstatt Blickkontakt mit dem Publikum zu halten. Die Folien können in einer Voransicht ausgedruckt werden und die Folienbedienung kann per Fernsteuerung erfolgen, so dass man sich beim Referieren auch frei im Raum bewegen kann, anstatt am Präsentationsrechner zu verharren.

Das **Foliendesign** sollte schlicht gehalten werden, ggf. kann das Corporate Design der eigenen Forschungseinrichtung genutzt werden. Im Internet stehen kostenpflichtige, aber auch kostenlose Powerpoint-Templates von unterschiedlichen Anbietern zur Verfügung (für Online-Ressourcen zur Erstellung von Präsentationsfolien ► Online-Ressourcen zur Erstellung von Präsentationsfolien). Wichtig ist, dass der Foliensatz mit einer **Titelfolie** beginnt (Konferenzname, Titel des Vortrags, Namen und Institutszugehörigkeiten der Autorinnen und Autoren, ggf. Logo der Einrichtungen), dann eine **Gliederungsfolie** über den Ablauf des Vortrags liefert und mit einer **Abschlussfolie** endet (z. B. Dank für die Aufmerksamkeit und Kontaktadressen der Vortragenden). Günstig ist es oft, wenn jeder Folie zu entnehmen ist, wo in der Gliederung man sich aktuell befindet. Zudem sollten die Folien durchnummeriert sein, damit bei Rückfragen die konkrete Folie adressiert werden kann. In der Fußzeile der Folien lassen sich zudem der Vortrags(kurz)titel sowie die Autoreninformation unterbringen. Digitales Bildmaterial lässt sich online recht einfach beschaffen, allerdings muss – auch bei nicht-kommerzieller wissenschaftlicher Verwendung – das Urheberrecht beachtet und zudem stets eine Quellenangabe gemacht werden. Manche Vortragende antizipieren bereits Diskussionsfragen (z. B. wenn im Vortrag auf bestimmte methodische Details oder einzelne Befunde nicht näher eingegangen wurde) und halten dafür weitere Folien bereit, die wahlweise bei Rückfragen zum Einsatz kommen.

Die Kritik am Einsatz von Präsentationsfolien bezieht sich nicht zuletzt auf häufige **technische Probleme**. Diesen sollte durch eine entsprechend sorgfältige Vorbe-

Online-Ressourcen zur Erstellung von Präsentationsfolien

Ein professioneller Foliensatz wertet einen Konferenzvortrag auf, sofern er inhaltlich auf den mündlichen Vortrag abgestimmt ist. Ein Ersatz für gute Vortragsrhetorik sind Folien nicht. Zur Vorbereitung des Foliensatzes, der meist mit Microsoft Powerpoint erstellt wird, können folgende Ressourcen hilfreich sein.

Vorlagen (Templates) für Microsoft Powerpoint

- Powerpoint-Templates speziell für wissenschaftliche Präsentationen sind im Internet teils kostenlos teils kostenpflichtig von unterschiedlichen Anbietern zu finden (Suchbegriffe: „scientific powerpoint templates“, „research powerpoint templates“ etc.)
- http://office.microsoft.com/de-de/templates/ (offizielle Website von Microsoft mit Powerpoint-Templates zu verschiedenen Anlässen)

Leistungsstarke Alternativen zu Microsoft Powerpoint

- Impress (http://de.openoffice.org/product/impress.html)
- KeyNote (http://www.apple.com/keynote)
- Prezi (http://www.prezi.com)
- SlideRocket (http://www.sliderocket.com)

Forschung und Diskussion zum Einsatz von Präsentationsfolien

Für eine Reflexion des Einsatzes von Powerpoint- bzw. Präsentationsfolien sei exemplarisch auf aktuelle Forschungsliteratur verwiesen, die anstelle des Konferenzvortrags allerdings oft die Vorlesung in den Mittelpunkt rückt, um Vor- und Nachteile des Folieneinsatzes bei mündlichen Präsentationen zu beleuchten (Beyer, 2011; Burke & James, 2008; Gabriel, 2008; James, Burke, & Hutchins, 2006; Koeber, 2005; Levasseur & Sawyer, 2006).

reitung entgegengewirkt werden: Folien können im PDF-Format mitgebracht werden, dann sind z. B. Schriftverzerrungen durch unterschiedliche Powerpoint-Versionen auf dem eigenen und dem Präsentationsrechner ausgeschlossen. Sollen neben einem Foliensatz weitere digitale Inhalte medial präsentiert werden (Videos, Animationen, Audiodateien von Interviews etc.), so kann das Arbeiten mit dem eigenen Rechner und mitgebrachten Lautsprecherboxen vorteilhaft sein. Eine rechtzeitige Anwesenheit im Vortragsraum erlaubt ein Austesten der Folienpräsentation, der Lautsprecher, der Beleuchtungs- und Verdunklungsmöglichkeiten etc. Bei Bedarf können neben der Computerpräsentation auch weitere Medien genutzt werden: Tafel, Flipchart, Handout, Anschauungsmaterial etc. Dabei ist jedoch zu bedenken, dass die genauen räumlichen Bedingungen vor Ort vorab zu klären sind. Häufig ist es sinnvoll, bei einem Vortrag **Zusatzmaterial zum Mitnehmen** anzubieten, etwa Visitenkarten, Flyer von Forschungsprojekten, Sonderdrucke von Publikationen, die im Vortragsraum ausgelegt werden können.

13.2.6 Üben des Vortrags

Nach der Fertigstellung des Vortrags (Konzept des mündlichen Vortrags, Stichpunktkarten, Foliensatz) empfiehlt es sich, den Vortrag ausreichend zu üben, insbesondere wenn er nicht in der eigenen Muttersprache gehalten werden soll. Durch das Training und die daran anschließende (selbst-)kritische Reflektion können grundlegende Fehler vermieden werden, Sprechangst abgebaut sowie die Qualität des Vortrags erhöht werden. Zudem wird man selbst vertrauter mit den Vortragsinhalten und kann insbesondere das Zeitmanagement optimieren. Wer die Möglichkeit hat, sollte den Vortrag einem Testpublikum präsentieren und dessen Feedback aufnehmen. Folgende Aspekte sind beim Üben und Halten eines Vortags zu beachten:

- **Zahl der Vortragenden**: Wenn eine von mehreren Personen erarbeitete Studie vorzustellen ist, können ggf. auch mehrere Personen im Wechsel vortragen. Dies kann dramaturgisch günstig sein, weil es Abwechslung schafft. Allerdings kann ein Referentenwechsel auch ablenkend wirken. Hier ist fallweise zu entscheiden und entsprechend im Vorfeld zu üben. In jedem Fall sollte immer nur die aktuell referierende Person vorne stehen, wartende Mitreferenten sollten sich im Hintergrund halten, um das Publikum nicht abzulenken.
- **Kleidung**: Ein etwas formalerer Kleidungsstil ist auf Konferenzen üblicher als ein legerer Freizeitlook. Generell sollte das Publikum möglichst nicht durch Äußerlichkeiten vom Vortragsinhalt abgelenkt werden.
- **Mimik und Gestik**: Durch Mimik und Gestik kann das Gesagte unterstrichen und der Vortrag lebendiger gestaltet werden. Hände und Arme sollten v. a. eingesetzt werden, um Zusammenhänge, Relationen und Größenverhältnisse optisch zu unterstreichen. Die Hände sollten dabei stets offen gehalten werden. Das Ballen zu einer Faust oder das Zeigen mit einzelnen Fingern auf bestimmte Personen im Auditorium sollten vermieden werden. Dauerndes Gestikulieren wirkt ebenso unvorteilhaft wie das Verstecken der Hände in den Hosentaschen. Auch ein ständiges Herumspielen am „rettenden“ Kugelschreiber, der womöglich zwischendurch noch herunterfällt, oder das Herumzupfen an den Vortragsnotizen lenken das Auditorium unnötig ab.
- **Blickkontakt**: Idealerweise sollte Blickkontakt zu den Anwesenden gehalten werden, anstatt auf den Monitor, die Projektionswand oder die eigenen Auf-

zeichnungen zu schauen. Der Blick kann dabei z. B. in einem M-förmigen Kurs über die Zuhörerschaft wandern, um das gesamte Publikum einzubeziehen.

- **Stimme**: Neben Mimik und Gestik ist die Stimme ein weiteres wichtiges Gestaltungsmittel für den Vortrag. Abgesehen von der richtigen Lautstärke (ggf. unterstützt durch Mikrofon und Lautsprecheranlage) sind Sprechtempo und Betonung wichtig für die Verständlichkeit des Vortrags. Am Ende jedes Satzes sollte die Stimme gesenkt und eine kurze Pause (1–2 Sekunden) gemacht werden. Dadurch wirkt der Vortrag lebendiger und die Struktur des Gesagten wird deutlicher.
- **Getränk**: Im Falle von Husten oder Heiserkeit kann ein Getränk helfen, wobei dieses bei Konferenzen nicht immer bereitgestellt wird und deswegen mitgebracht werden sollte.
- **Laserpointer**: Mithilfe eines Laserpointers kann auf einzelne Details der projizierten Präsentationsfolien gezeigt werden. Dabei sollte ein lichtstarker Pointer genutzt und im Vorfeld getestet werden. Laserpointer werden bei Konferenzen in der Regel nicht bereitgestellt und sollten deswegen ebenfalls mitgebracht werden.
- **Fernsteuerung**: Eine Fernsteuerung erlaubt es, sich von der Computertastatur bzw. -maus zu entfernen und während des Vortrags zu bewegen, was in der Regel den Publikumskontakt erleichtert.
- **Stichpunktkarten**: Handzettel mit den Folien – ergänzt um Stichpunkte – dienen als Gedächtnisstütze. Der Vortrag sollte in den Sozialwissenschaften möglichst frei gehalten werden. Anstelle ausformulierter Vortragsmanuskripte sind deswegen Karten mit Stichpunkten nützlicher.
- **Wortwahl**: Ein wissenschaftlicher Sprachstil, der nicht zu sehr in saloppe Umgangssprache oder Jargon abgleitet, ist die Norm. Abkürzungen sollten bei der ersten Verwendung erläutert werden. Eine direkte Ansprache des Publikums oder auch Fragen an das Publikum erhöhen die Aufmerksamkeit.
- **Vortragseinstieg**: Der Vortrag sollte mit einer freundlichen Begrüßung und einem anschaulichen Beispiel, einer interessanten Abbildung oder einer Anekdote beginnen, um die Aufmerksamkeit des Publikums zu wecken (▶ Abschn. 13.2.4). Notfalls kann der Vortragseinstieg auswendig gelernt werden, um die Anfangsnervosität zu überwinden.
- **Vortragsabschluss**: Man sollte den Vortag mit einer Abschlussformel wie „Ich danke Ihnen für Ihre Aufmerksamkeit!“ oder „Vielen Dank, ich freue mich auf die Diskussion mit Ihnen!“ schließen.

13.2.7 Vortrag und Diskussion

Trotz umfangreichen Übens wird am Tag des Vortrags innere Unruhe auftreten. Diese Spannung vor dem öffentlichen Auftritt ist normal und erwünscht, fördert sie doch die Konzentration der Vortragenden. Allerdings sollte darauf geachtet werden, dass das **Lampenfieber** erträglich bleibt. Hierzu gibt es folgende Empfehlungen:

- **Die eigenen Ansprüche nicht zu hoch schrauben**. Das Publikum erwartet die verständliche und sichere Präsentation der Ergebnisse einer Forschungsarbeit, keine Jahrhundertrede.
- **Gute Vorbereitung**. Das Publikum hat, auch bedingt durch eigene Erfahrungen, Verständnis für Nervosität bei guter Vorbereitung. Nervosität gepaart mit schlechter Vorbereitung hinterlässt hingegen einen schlechten Eindruck. Gute Vorbereitung bedeutet v. a. auch, den Vortrag im Vorfeld oft genug geübt zu haben, um den Gedankengang flüssig frei formulieren zu können.
- **Stichpunkte auf Karteikarten anlegen.** Meist werden die Stichpunkte gar nicht benötigt, sie geben aber die Sicherheit, im Notfall auf eine gedankliche Blockade vorbereitet zu sein.
- **Mit der Präsentationstechnik (Bedienung des Projektors, Laserpointers, Mikrofon) vertraut machen**. Dadurch fällt ein Unsicherheitsfaktor weg und man ist vor dem Vortrag beschäftigt. Deswegen wichtig: Rechtzeitige Anwesenheit im Vortragsraum, um die Technik auszutesten.
- **Einige Worte mit der Panelleitung, anderen Vortragenden und/oder den Zuhörern wechseln**. Dadurch verliert das anonyme Auditorium im Vorfeld des Vortrags etwas von seiner „Bedrohlichkeit“ und es entwickelt sich vielleicht sogar Vorfreude darauf, Interessierten die eigenen Forschungsergebnisse vorstellen zu dürfen.

Während des Vortrags sollte man die Zeit im Auge behalten und gleichzeitig die Reaktionen des Publikums beachten, um eine möglichst optimale **Vortragssteuerung** zu realisieren. Bei Störungen (z. B. Sitzplatzmangel, stickige Luft, blendendes Sonnenlicht) sollte man ggf. eine kurze Pause einlegen, um die notwendigen Voraussetzungen für ein konzentriertes Zuhören zu schaffen. Zwischenfragen werden in der Regel nicht zugelassen, sondern auf die Diskussion nach Abschluss des Vortrags verschoben. Sollte man während des Vortrags feststellen, dass man sich doch in der Zeit verkalkuliert hat, so ist eine Verdopplung des Redetempos in der Regel nicht die beste

Lösung. Günstiger ist es, souverän zum Fazit zu springen, dieses in Ruhe zu erläutern und für offen gebliebene und nicht behandelte Punkte auf Gesprächsmöglichkeiten nach dem Vortrag oder auf Publikationen zu verweisen. Meist sorgt die Panelleitung für die Einhaltung der Redezeit, indem fünf, drei und/oder eine Minute vor Ablauf der **Vortragszeit** entsprechende Signale gegeben werden.

Die auf den Vortrag folgende **Diskussion** mit dem Publikum findet auf Konferenzen häufig moderiert statt (z. B. durch die Sessionleitung = Session Chair). Es ist aber auch eine Selbstmoderation durch die Referierenden möglich. Die **Moderation** umfasst die Regulation des Rederechts (bei vielen Wortmeldungen ggf. Liste erstellen), die Sicherstellung der Diskussionskultur (z. B. kein Ins-Wort-Fallen, keine Beleidigungen) sowie die Kontrolle der Diskussionszeit. Fragen aus dem Publikum können dabei jeweils einzeln oder gebündelt beantwortet werden. Oft empfiehlt es sich für Referierende, die Fragen mitzuschreiben, um sie umfassend beantworten zu können.

Eine **gute Antwort** hat folgende Struktur: 1. Kurze Wiederholung und Einordnung der Frage (z. B. „Ihre Frage zielt auf die Details des Forschungsdesigns …“), 2. knappe und zielgerichtete Beantwortung (nicht abschweifen oder ablenken), dabei möglichst handfeste Begründungen und Fakten liefern (z. B. Verweis auf Daten, Studien, Theorien, bekannte Autorinnen und Autoren).

Was tun, wenn man bei **Detailfragen** eine derart konkrete Antwort nicht parat hat? In diesem Fall möglichst nicht spekulieren, die Frage ignorieren oder auf andere Themen ausweichen, sondern offensiv damit umgehen: Referierende müssen nicht alles wissen. Im Zweifelsfall kann man Wissenslücken zugeben und begründen (z. B. „Diese Frage betrifft die Wirkungen beim Publikum, diese haben wir nicht untersucht, sondern nur die Medieninhalte.“; „Die exakten Ergebnisse dazu habe ich leider nicht im Kopf, aber insgesamt kann ich sagen, dass keine oder nur geringe Geschlechtsunterschiede auftraten“; „Danke für Ihr Interesse an den Details, ich schicke Ihnen gern unsere Projektpublikation zu, da wird genauer auf die von Ihnen genannten Punkte eingegangen.“).

Neben Detailfragen sind **konfrontative und provokative Fragen und Kommentare** eine Herausforderung für die Referierenden. In manchen Konferenzpanels gleitet kritischer, aber kollegialer akademischer Diskurs in mehr oder minder unterschwellige persönliche Angriffe ab (z. B. „Ich frage mich die ganze Zeit, welchen Sinn diese Studie überhaupt hat“, „Die Ergebnisse sind doch völlig trivial“). Hier gilt es, Ruhe zu bewahren, kurz und sachlich zu antworten und v. a. einen öffentlichen Zweier-Disput mit einem einzelnen Zuhörer zu vermeiden (Abb. 13.2). Im Zweifelsfall kann bei öffentlicher Grundsatzkritik auch ein ausführlicheres Gespräch angeboten werden (z. B. „Ich sehe, Sie haben ganz grundsätzliche Zweifel an unserer Studie, die sich im Rahmen dieser kurzen Diskussionsrunde sicher nicht klären lassen. Vielleicht setzen wir uns in der Mittagspause zusammen, dann erläutere ich Ihnen gern noch einmal ausführlich den Hintergrund.“). Zu sicherem Umgang mit Fragen in der Diskussion mit dem Publikum sowie zum wissenschaftlichem Vortragsstil gibt es zahlreiche Ratgeber (z. B. Franck, 2001; Hierhold, 2002; Nölke, 1998).



Abb. 13.2 Vortragende sollten darauf vorbereitet sein, dass es in der Diskussion nach dem Vortrag zu teilweise konfrontativen und provozierenden Fragen und Kommentaren kommen kann. (© Bradford Veley/Search ID: bve0011, Rights Available from CartoonStock.com)

13.2.8 Networking

Ein wichtiger Bestandteil eines Konferenzbesuchs ist das „Networking“ in der Scientific Community. Konferenzen bieten eine gute Gelegenheit, direkt mit anderen Forschenden in Kontakt zu treten (für Tipps zum professionellen Networking Hahn, 2009; Lochner, 2011; Lutz, 2009; Mai, 2008; Öttl & Härter, 2004). Wichtig ist es, sich bereits im Vorfeld anhand des Konferenzprogramms bzw. der Teilnehmerliste zu überlegen, welche Personen man ansprechen möchte. Der informelle Erfahrungsaustausch im Kollegenkreis ist eine wertvolle Informationsquelle zu aktuellen Entwicklungen im Forschungsbereich, relevanten Ereignissen oder möglichen Forschungskooperationen (gemeinsame Projekte, Gastaufenthalte etc.).

13.2.9 Auszeichnungen für Konferenzvorträge

Üblicherweise wird auf nationalen und internationalen Fachkonferenzen ein Vortragspreis („best paper award“) verliehen, über den eine entsprechende Fachjury entscheidet. Diese akademische Ehrung ist mit einer Urkunde verbunden und kann als wissenschaftliche Leistung im Lebenslauf vermerkt werden. Preiswürdige Vorträge berichten in einem brillanten Vortragsstil über eine besonders interessante Studie, liefern innovative Befunde und/oder zeichnen sich durch eine hervorragende Forschungsmethodik aus. Die Regularien für die Teilnahme am Wettbewerb unterscheiden sich je nach Konferenz und sollten bei Bedarf im Vorfeld eruiert werden.

13.3 Wissenschaftliches Poster

Das **wissenschaftliche Poster** („scientific poster“, „research poster“) dient der prägnanten grafischen und textuellen Präsentation von Forschungsergebnissen auf Konferenzen (Hoffmann, 2004; Hoffmann & Helmle, 2009). Im Unterschied zum Werbe- oder Veranstaltungsplakat enthält das wissenschaftliche Poster sehr viel mehr Inhalte, es bleibt dennoch eine primär visuelle Darstellungsform: Ein Poster ist kein an die Wand gehängter Kurzaufsatz, sondern muss einen ansprechenden Blickfang darstellen und primär grafische Mittel zur Informationsvermittlung nutzen (Tabellen, Charts, Fotos etc.). Poster werden auf Konferenzen in sog. **Poster-Ausstellungen** präsentiert.

Üblicherweise kann zu wissenschaftlichen Konferenzen anstelle eines Vortrags von vorne herein ein Poster eingereicht werden. Der **Peer-Review-Prozess** stützt sich dann auf eine Kurzzusammenfassung (Abstract) des Posters. Bei einem Poster sind die **Annahmechancen bei einer Konferenz** in der Regel höher als bei einem Vortrag, weshalb diese Einreichungsform manchmal bevorzugt wird. Zudem können im Rahmen des Review-Prozesses eingereichte Vorträge in Poster umgewandelt werden. Insbesondere bei Studien mit einem begrenzteren Erkenntnisgewinn (z. B. kleine Untersuchung, leichte Qualitätsmängel) wird die Posterform oft der Vortragsform vorgezogen. Aus organisatorischer Sicht haben Poster den Vorteil, dass sie das bei einer Konferenz meist sehr knapp bemessene Zeitkontingent nicht belasten und trotzdem weitere Konferenzbeiträge zugelassen werden können. Die Poster-Einreichung ist gerade auch für Studierende sowie für den Erstkonferenzbesuch empfehlenswert, denn im Vergleich zum Vortrag ist nicht nur die Erfolgschance beim Peer Review höher, sondern auch die Anforderung vor Ort geringer als beim Vortrag.

13.3.1 Inhalte des Posters

Die inhaltliche **Gliederung des Posters** gleicht der von Zeitschriftenartikeln und Konferenzvorträgen: In einer Einleitung muss das Forschungsproblem bzw. die Zielsetzung der Studie benannt und deren Relevanz begründet werden. Der theoretische Hintergrund bzw. Forschungsstand muss erläutert und die Methodik der Studie beschrieben werden. Im Zentrum stehen sollte dann die Darstellung der wichtigsten Befunde. Schließlich dürfen eine Diskussion und ein Literaturverzeichnis nicht fehlen. Die große Kunst bei der Poster-Erstellung ist die **rigorose Kürzung** nach dem Motto „Weniger ist mehr“. Teilweise kann auf vollständige Sätze verzichtet und z. B. das methodische Vorgehen **stichpunktartig** verdeutlicht werden. Bei der Darstellung von Methoden (z. B. experimentelles Setting) und v. a. von Ergebnissen sollte verstärkt mit **grafischen Mitteln** gearbeitet werden (Tabellen, Charts, Fotos, Diagramme etc.). Ein inhaltlich ansprechender und aussagekräftiger **Titel des Posters** ist von großer Wichtigkeit, um auch aus der Ferne das Interesse des Publikums zu wecken. Unter dem Titel sollten Autorinnen und Autoren sowie deren Forschungseinrichtungen (ggf. mit Logos) deutlich sichtbar genannt werden. Am Fuß des Posters sind **Kontaktdaten** einzufügen (z. B. Post-, Web-, E-Mail-Adresse).

13.3.2 Gestaltung des Posters

Zunächst sollte man die formalen Gestaltungsvorgaben der jeweiligen Konferenz zugrunde legen (z. B. hinsichtlich Posterformat, Begrenzung des Textumfanges, Formatvorlage). Das typische Posterformat ist **DIN A0** (841 × 1 189 mm), wobei meist mit dem Hochformat gearbeitet wird, um die Poster in der Ausstellung platzsparend auf Stellwände hängen zu können. Ganz entscheidend ist die Lesbarkeit: Der Titel des Posters sollte aus ca. drei bis fünf Meter Entfernung gut lesbar sein, der übrige Inhalt aus ca. ein bis zwei Meter Entfernung (je nach Schriftart ist eine Schriftgröße um 36 Punkt zu wählen). Empfehlenswert ist eine serifenlose **Schrift** (z. B. Arial) und eine eher kurze **Zeilenlänge** sowie ein deutlicher Farbkontrast der Schrift zum Hintergrund. Es sollte nur mit einer einzigen Schriftart und mit maximal vier verschiedenen Schriftgrößen gearbeitet werden. Der Zeilenabstand sollte großzügig bemessen sein (z. B. 1.5-zeilig). Die einzelnen inhaltlichen Gliederungspunkte sollten auch optisch klar voneinander abgehoben werden (Zwischenüberschriften, Spaltenlayout etc.). Dabei sollte die **Freifläche** auf dem Poster bewusst genutzt werden, um die inhaltlichen Elemente zu strukturieren und den Blick des Publikums zu



Abb. 13.3 Umsetzungsbeispiel für ein Poster der Universität Hohenheim. (Aus Hoffmann & Helmle, 2009. Mit freundlicher Genehmigung des Ökom-Verlags, München)

lenken (Abb. 13.3). Auf Postern wird mit **Farben** gearbeitet, wobei der Farbeinsatz die inhaltlichen Aussagen unterstreichen (ähnliche Farben für inhaltlich Zusammenhängendes) und gleichzeitig zum ästhetischen Wert des Posters beitragen sollte. Eine Orientierung am **Corporate Design** der eigenen Forschungseinrichtung wird oft als wünschenswert angesehen. Allerdings kann dies einer besonders inhaltsangemessenen oder originellen grafischen Aufbereitung im Wege stehen.

13.3.3 Zusatzmaterial zum Poster

Ein Poster kann durch unterschiedliches Zusatzmaterial zum Mitnehmen aufgewertet werden, das dem Publikum über eine am unteren Rand des Posters befestigte Box zur Verfügung gestellt wird: z. B. Visitenkarten, Projektflyer, ein DIN A4-Ausdruck des Posters oder auch ein Sonderdruck einer thematisch passenden Publikation.

13.3.4 Erstellung des Posters

Erste Entwürfe des Posters lassen sich computergestützt, aber auch bequem mit Papier und Stift erstellen. Wenn das Grobkonzept steht, sollte die Feinarbeit am PC erfolgen. Poster werden meist mit dem Präsentationsprogramm Microsoft **Powerpoint** oder mit Grafikprogrammen (z. B. Adobe Photoshop) erstellt. Oft empfiehlt es sich, das Poster in DIN A4 zu erstellen und erst beim Druck auf A0 zu vergrößern (University of Leicester, 2011). Zudem können die einzelnen Text- und Grafikelemente flexibel erstellt und platziert werden. Für MS Powerpoint existieren von unterschiedlichen Anbietern eine Reihe von kostenlosen Templates für wissenschaftliche Poster (z. B. http://www.posterpresentations.com/html/free_poster_templates.html).

Die erste Fassung des Posters verlangt mehrere **Überarbeitungen**: Nach Fertigstellung des Posters sollte man selbst „eine Nacht darüber schlafen" und dann noch einmal selbstkritisch nachbessern. Schließlich ist eine Präsentation vor Testpublikum sinnvoll, um die Qualität des Posters zu prüfen und ggf. noch Verbesserungsvorschläge aufnehmen zu können.

Die Rechenzentren vieler Hochschulen bieten die Möglichkeit, Poster bis zur Größe von DIN A0 in Farbe **ausdrucken** und bei Bedarf auch **laminieren** zu lassen. Ebenso können dafür Copyshops in Anspruch genommen werden. In welchem Format die Datei einzureichen ist, wie viel Zeit das Drucken und Laminieren in Anspruch nimmt und welche Kosten dabei anfallen, sollte rechtzeitig geklärt werden.

13.3.5 Präsentation des Posters

Poster hängen während der Konferenz in der **Poster-Ausstellung**. Für individuelle Betrachterinnen und Betrachter, die in Konferenzpausen auf das Poster stoßen, sind deswegen Zusatzmaterialien zum Mitnehmen hilfreich (▶ Abschn. 13.3.3). Zudem finden meist **Poster-Sessions** statt, bei denen die Autorinnen und Autoren zu vorangekündigten Zeiten bei ihrem Poster für Fragen zur Verfügung stehen und/oder ihre Poster jeweils in zwei bis drei Minuten nacheinander dem Publikum erläutern. Im Unterschied zum Konferenzvortrag bietet die Posterpräsentation häufig bessere Möglichkeiten, um mit interessierten Fachkollegen, die vor dem Poster verweilen, direkt in Kontakt zu treten und Networking zu betreiben. Es ist somit im Sinne des kollegialen Austauschs sinnvoll, sich auch außerhalb offizieller Poster-Sessions bei seinem Poster aufzuhalten.

13.3.6 Auszeichnungen für wissenschaftliche Poster

Ähnlich dem Best Paper Award für wissenschaftliche Vorträge (▶ Abschn. 13.2.9) werden auf Fachkonferenzen

von einer Fachjury auch **Best Poster Awards** vergeben. Preiswürdige Poster berichten besonders interessante Befunde und stellen diese dem Poster-Format entsprechend in formal und inhaltlich angemessener Weise dar, wobei sie meist auf den ersten Blick sehr ansprechend wirken und sich positiv von den übrigen Postern abheben müssen. Preisgekrönte Poster können z. B. über die Google-Bildersuche („Poster-Preis", „poster award") recherchiert werden. Zudem werden wissenschaftliche Poster auf der Plattform „ePosters – The Online Journal of Scientifc Posters" (http://www.eposters.net) gesammelt und kommentiert.

13.3.7 Weiterverwertung des Posters

Nach dem Konferenzbesuch werden Poster häufig im Flur des Forschungsinstituts aufgehängt. Eine digitale Fassung kann auch über das Online-Publikationsverzeichnis von Forschenden oder über die Websites von Forschungsprojekten verbreitet werden.

13.4 Öffentlichkeitsarbeit

Mit einem wissenschaftlichen Fachzeitschriftenartikel, Konferenzvortag oder Poster werden eigene Forschungsergebnisse der Scientific Community präsentiert. Doch auch die breite Öffentlichkeit kann eine relevante Zielgruppe sein: Nicht selten wird der Wissenschaft vorgeworfen, sich in den „akademischen Elfenbeinturm" zurückzuziehen und zu wenig in Kontakt mit der Gesellschaft zu treten. Doch alle Hochschulen, wissenschaftlichen Fachgesellschaften und teilweise auch einzelne Forschungsprojekte beschäftigen mittlerweile Fachleute für **Öffentlichkeitsarbeit**, um die Kommunikation mit unterschiedlichen Bezugsgruppen (z. B. Presse, Wirtschaftsunternehmen, Politik, Studieninteressierte, internationale Fachkollegen, Drittmittelgeber etc.) zu organisieren und dadurch Akzeptanz und Zusammenarbeit zu fördern (Brauer, 2005). Im Rahmen der Öffentlichkeitsarbeit sorgen Universitäten z. B. dafür, dass Erfolgsstatistiken über ihre Absolventenzahlen, ihre Drittmitteleinwerbungen, ihre Patente etc. in der Presse erscheinen. Wenn Öffentlichkeitsarbeit auf die Anwerbung von Studierenden (z. B. durch Give-Aways und Werbespots von Universitäten) oder von Auftraggebern (z. B. durch Messestände von Forschungseinrichtungen) hinausläuft, spricht man von **Wissenschaftsmarketing**. Wenn jedoch konkrete Forschungsergebnisse der breiten Öffentlichkeit vermittelt werden sollen, handelt es sich um **Wissenschaftskommunikation** als einer speziellen Variante der Öffentlichkeitsarbeit.

Die Wissenschaftskommunikation von akademischen Einrichtungen und einzelnen Forschenden hat unterschiedliche Anspruchsgruppen und Zielsetzungen: So strebt man häufig Bekanntheit und Anerkennung in der breiten Öffentlichkeit an, möchte insbesondere Multiplikatoren wie Journalisten auf dem Laufenden halten, aber auch den wissenschaftlichen Nachwuchs, nationale und internationale Fachkollegen, Kunden, Auftraggeber und Kooperationspartner in der Wirtschaft sowie Politiker und Forschungsförderer von der eigenen Forschungsarbeit überzeugen. Einige verbreitete **Instrumente der Wissenschaftskommunikation** werden im Folgenden kurz aufgeführt.

13.4.1 Wissenschaftskommunikation im Internet

Von großer Bedeutung ist heute eine ansprechende und informative, deutsch- und englischsprachige Präsenz im Internet in Form einer **Website**, die möglichst professionell gestaltet und hinsichtlich ihrer Auffindbarkeit über Suchmaschinen optimiert sein sollte. Fest etabliert haben sich auch einschlägige Mailinglisten (meist organisiert durch die wissenschaftlichen Fachgesellschaften), über die Fachkollegen auf aktuelle Publikationen, Konferenzen, Großprojekte oder Stellenausschreibungen aufmerksam gemacht werden können. Nicht selten dienen Mailinglisten auch als Plattformen für den wissenschaftlichen Diskurs innerhalb der Fachcommunity. Darüber hinaus werden mittlerweile verstärkt Web 2.0-Technologien von Forschungseinrichtungen genutzt (z. B. Weblogs, Microblogs wie Twitter, Profile auf Social Networking Sites, Podcasts etc.), was jedoch entsprechende Medienkompetenz und personelle Ressourcen zur Erarbeitung und kontinuierlichen Pflege der Internet-Angebote voraussetzt (Kaden, 2009).

So kann z. B. theoretisch jeder und jede Forschende **Wikipedia-Beiträge** neu anlegen oder editieren, um dort Verweise auf eigene Forschungsarbeiten unterzubringen. Entsprechende Aktivitäten werden aber nur dann von Erfolg gekrönt sein, wenn man sich detailliert mit den Regularien und Gepflogenheiten der Wikipedia-Community auskennt, entsprechend viel Zeit investiert und inhaltlich wertvolle Beiträge liefert.

Ebenso ist für einen **Wissenschaftsblog** („science blog") ein nicht unerheblicher Aufwand einzukalkulieren, um immer wieder neue, interessante Inhalte zu generieren und sich mit dem Leserfeedback sowie den Reaktionen in der Blogosphäre auseinander zu setzen (für eine

Übersicht von Wissenschaftsblogs siehe z. B. http://www.scienceblogging.org). Einige Forscherinnen und Forscher sehen Blogs trotz des Aufwandes als eine fruchtbare Plattform, um ihre wissenschaftlichen Arbeiten in einem informelleren Stil interessierten Publika auch jenseits des Wissenschaftsbetriebs zu präsentieren und mit diesen in einen Austausch zu treten.

13.4.2 Umgang mit Journalisten

Wenn größere Forschungsprojekte starten oder wichtige Befunde vorliegen, kann dies über **Pressemeldungen** oder sogar **Pressekonferenzen** mitgeteilt werden. Die Pressestellen bzw. Referate für Öffentlichkeitsarbeit der Hochschulen verfügen über entsprechende regionale und überregionale Presseverteiler und helfen bei Bedarf bei der Überarbeitung einer Pressemitteilung oder der Vorbereitung einer Pressekonferenz. Zudem haben viele wissenschaftliche Fachgesellschaften und Hochschulen mittlerweile im Web öffentliche **Expertendatenbanken** eingerichtet (z. B. die „Spezialisten-Börse" der Deutschen Gesellschaft für Psychologie: http://www.dgps.de/presse oder die Expertendatenbank der Humboldt-Universität zu Berlin: http://www.hu-berlin.de/pr/mitarbeiter/experten), so dass Journalisten mit ihren Anfragen zielgerichtet auf einzelne Forschende zugehen können.

Dabei ist aus Sicht der Forschenden zu beachten, dass journalistische Anfragen oft sehr kurzfristig eintreffen und umfassende Hintergrundgespräche sich später oft nur in ein oder zwei zitierten Sätzen niederschlagen (◘ Abb. 13.4). Während bei Printbeiträgen um eine Autorisierung gebeten werden kann, ist dies bei Radio- und TV-Beiträgen nicht der Fall, so dass für die Forschenden nicht kontrollierbar ist, in welchem Kontext ihre Aussagen später stehen. Es ist durchaus nicht selbstverständlich, dass die Übersetzung wissenschaftlicher Inhalte in journalistische Beiträge zur Zufriedenheit der Forschenden tatsächlich gelingt. Darüber hinaus sind Wissenschaftler gehalten, im Sinne der Wissenschaftsethik (▶ Abschn. 4.2) nur zu Themen Stellung zu nehmen, zu denen sie tatsächlich forschen. In der Psychologie gilt es insbesondere als unseriös, für die Medien „Ferndiagnosen" zu stellen und z. B. angesichts aktueller Schlagzeilen über den psychischen Gesundheitszustand von Prominenten zu spekulieren, denen sie persönlich nie begegnet sind. Auch sollten Forschungsergebnisse im Sinne der Qualitätssicherung zunächst in referierten Fachzeitschriften publiziert werden, bevor sie an die Presse geleitet werden. Weitere Hinweise zum professionellen Umgang von Forschenden mit journalistischen Anfragen stellt die APA auf ihrer Website bereit (http://www.apa.org/pubs/authors/media/tips.aspx).



◘ **Abb. 13.4** Forschende sollten auf die Besonderheiten journalistischer Interviews vorbereitet sein. (© Douglas Pike/Search ID: dpin233, Rights Available from CartoonStock.com)

13.4.3 Beiträge in Transferzeitschriften

Im Unterschied zu wissenschaftlichen Fachzeitschriften, die sich an ein akademisches Fachpublikum wenden, verfolgen Transferzeitschriften das Ziel eines Wissensaustausches zwischen Forschung und Praxis. Praxis meint hierbei die Entscheidungsträger einer bestimmten Branche. Beiträge in Transferzeitschriften sind kürzer und allgemeinverständlicher geschrieben und unterliegen meist keinem Peer Review, sondern werden nur redaktionell geprüft. Beiträge in Transferzeitschriften werden in der Wissenschaft als Indikator für Praxisrelevanz und Anwendungsbezüge der jeweiligen Forschung betrachtet und sind zudem wichtig, um Kooperationsbeziehungen mit der Wirtschaft zu fördern. Beispiele für Transferzeitschriften sind z. B. „transfer – Werbeforschung und Praxis" (http://www.transfer-zeitschrift.net), „projektMANAGEMENT aktuell" (http://www.pmaktuell.org), „Forschung & Lehre" (http://www.forschung-und-lehre.de), „f&w – führen und wirtschaften im Krankenhaus" (http://www.bibliomed.de).

13.4.4 Populärwissenschaftliche Publikationen

Bei populärwissenschaftlichen Publikationen handelt es sich um Darstellungen wissenschaftlicher Ergebnisse, die sich an ein breites, nicht-wissenschaftliches Publikum – darunter manchmal auch Kinder und Jugendliche – richten. Populärwissenschaftliche Veröffentlichungen stammen teils von Wissenschaftsjournalisten, oft aber auch von Wissenschaftlerinnen und Wissenschaftlern selbst.

Beiträge in populärwissenschaftlichen Magazinen (z. B. Spektrum der Wissenschaft, Bild der Wissenschaft, Psychologie heute, Gehirn und Geist) sowie Bücher (z. B. Ratgeberliteratur) kommen hier infrage.

13.4.5 Eingeladene Vorträge

Bei Forschungsthemen von breiterem Interesse sowie entsprechender Bekanntheit erhalten Forschende auch Einladungen zu Vorträgen jenseits wissenschaftlicher Tagungen und Konferenzen. Diese sog. eingeladenen Vorträge („invited talks“) auf öffentlichen Veranstaltungen werden in der Wissenschaft als ein Indikator der Relevanz des Forschungsgebietes sowie des Renommees der Forschenden betrachtet. Hier findet kein Peer Review statt und oft werden Reisekosten vom Veranstalter übernommen. Teilweise wird – je nach Veranstaltungsrahmen – über eingeladene Vorträge auch in der Presse berichtet.

13.4.6 Präsentation auf Messen

Messen sind Marketingveranstaltungen, bei denen Produkte und Dienstleistungen einem Fachpublikum vorgestellt werden. Angewandte Forschungsprojekte, die auf die Entwicklung von Produkten hinauslaufen, lassen sich auch auf Messen präsentieren, z. B. Lernprogramme auf der Bildungsmesse „didacta“ (http://www.didacta-koeln.de). Aufbau und Betreuung eines professionellen Messestandes sowie Standmiete sind jedoch mit entsprechenden Kosten verbunden.

Lernquiz

1. Was versteht man unter doppelblindem Peer Review?
2. Worin unterscheidet sich der Peer-Review-Prozess bei Zeitschriftenartikeln von dem bei Konferenzvorträgen?
3. Was ist der Impact-Faktor und wann steigt er?
4. Nennen Sie Unterschiede im Umgang mit Konferenzvorträgen in den Geistes-, Sozial- und Technikwissenschaften!
5. Welches Format haben wissenschaftliche Poster üblicherweise?
6. Welche Vortragsformen auf Konferenzen sind zu unterscheiden?
7. Warum sind die Erfolgschancen beim Peer Review bei wissenschaftlichen Postern meist höher als bei Konferenzvorträgen?
8. Worin unterscheiden sich wissenschaftliche Fachzeitschriften und Transferzeitschriften?
9. Welche Ergebnisse kann die Begutachtung eines bei einer Fachzeitschrift eingereichten Artikels haben?
10. Welche Einreichungsformen für Vorträge fordern Calls wissenschaftlicher Fachkonferenzen?

Übungsaufgaben

1. Identifizieren Sie passende und relevante wissenschaftliche Zeitschriften für eine Publikation zum Thema „Entwicklung, Validierung und Normierung einer Skala zur Qualität der Mensch-Roboter-Interaktion“. Recherchieren Sie anschließend die Autorenrichtlinien der jeweiligen Journale!
2. Recherchieren Sie eine passende Konferenz, für die Sie einen Beitrag mit dem Thema „Prävalenz von Essstörungen – Ein Vergleich zwischen Adoleszenz und frühem Erwachsenenalter“ einreichen könnten. Beurteilen Sie die Passung anhand der Calls der Konferenzen. Infomieren Sie sich über die Deadlines für die Beitragseinreichung und die Vorgaben für die Einreichung von Forschungsreferaten, Positionsreferaten und Postern.
3. Suchen Sie im Internet ein aktuelles Poster, das auf einer Fachkonferenz den Posterpreis gewonnen hat. Diskutieren Sie das Poster hinsichtlich seiner inhaltlichen und formalen Gestaltung anhand der Hinweise in diesem Kapitel.
4. Finden Sie heraus, ob über die Website Ihrer Hochschule Blogs mit wissenschaftlichen Inhalten betrieben werden, welche Hochschuleinrichtungen oder Hochschulmitglieder hier aktiv sind und welche Themen behandelt werden!
5. Wie hoch ist jeweils der Impact-Faktor der Zeitschriften a) „Nature“, b) „Psychological Bulletin“, c) „Journal of Applied Psychology“ und d) „ Swiss Journal of Psychology“?

Weiterführende Literatur

- American Psychological Association (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington: American Psychological Association.
- Bem, D. J. (2004). Writing the Empirical Journal Article. In J. M. Darley, M. P. Zanna & H. L. Roediger (Eds.), *The compleat academic: A*

career guide (2nd ed., pp. 185–219). Washington, DC: American Psychological Association.
- Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed.). Chicago: University of Chicago Press.
- Hierhold, E. (2002). *Sicher präsentieren – wirksamer vortragen. Tipps und Tricks für die Praxis, visuelle und verbale Techniken, Überzeugungsstrategie und Argumentationstaktik von Flip-Chart bis Power-Point* (6. Aufl.). Frankfurt/Main: Redline Wirtschaft
- Hoffmann, V. (2004). Gestaltung und Bewertung von wissenschaftlichen Postern. In: Deutscher Fachjournalisten-Verband (Hrsg.), *Fachjournalismus. Expertenwissen professionell vermitteln* (S. 153–157). Konstanz: UVK

Literatur

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington: American Psychological Association.

Bem, D. J. (2004). Writing the Empirical Journal Article. In J. M. Darley, M. P. Zanna & H. L. Roediger (Eds.), *The compleat academic: A career guide* (2nd ed., pp. 185–219). Washington, DC: American Psychological Association.

Beyer, A. M. (2011). Improving student presentations. *Teaching of Psychology, 38*(2), 122.

Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed.). Chicago: University of Chicago Press.

Brauer, G. (2005). *Presse- und Öffentlichkeitsarbeit: Ein Handbuch*. Konstanz: UVK.

Burke, L. A. & James, K. E. (2008). PowerPoint-based lectures in business education: An empirical investigation of student-perceived novelty and effectiveness. *Business Communication Quarterly, 71*(3), 277–296.

Franck, N. (2001). *Rhetorik für Wissenschaftler. Selbstbewusst auftreten, selbstsicher reden*. München: Vahlen.

Franck, N. & Stary, J. (2006). *Gekonnt visualisieren. Medien wirksam einsetzen*. Paderborn: Schöningh, UTB.

Gabriel, Y. (2008). Against the tyranny of PowerPoint: Technology-in-use and technology abuse. *Organization Studies, 29*(2), 255–276.

Hahn, T. (2009). *77 Irrtürmer des Networkings erfolgreich vermeiden: So bauen Sie Kontakte auf, die Sie weiterbringen*. München: Finanz-Buch Verlag.

Hierhold, E. (2002). *Sicher präsentieren – wirksamer vortragen. Tipps und Tricks für die Praxis, visuelle und verbale Techniken, Überzeugungsstrategie und Argumentationstaktik von Flip-Chart bis Power-Point* (6. Aufl.). Frankfurt/Main: Redline Wirtschaft.

Hoffmann, V. (2004). Gestaltung und Bewertung von wissenschaftlichen Postern. In Deutscher Fachjournalisten–Verband (Hrsg.), *Fachjournalismus. Expertenwissen professionell vermitteln* (S. 153–157). Konstanz: UVK.

Hoffmann, V. & Helmle, S. (2009). Hinweise zur Gestaltung von wissenschaftlichen Postern. In A. Gerber, V. Hoffmann, & A. Thomas (Hrsg.), *Transdisziplinäre Umweltforschung: Methodenhandbuch* (S. 246–251). München: Ökom-Verlag.

James, K. E., Burke, L. A., & Hutchins, H. M. (2006). Powerful or pointless? Faculty versus student perceptions of powerpoint use in business education. *Business Communication Quarterly, 69*(4), 374–396.

Kaden, B. (2009). *Library 2.0 und Wissenschaftskommunikation*. Berlin: Simon.

Koeber, C. (2005). Introducing multimedia presentations and a course website to an introductory sociology course: how technology affects student perceptions of teaching effectiveness. *Teaching Sociology, 33*(3), 285–300.

Kuzbari, R. & Ammer, R. (2006). *Der wissenschaftliche Vortrag*. Wien: Springer.

Levasseur, D. G. & Sawyer, J. K. (2006). Pedagogy meets PowerPoint: A research review of the effects of computer-generated slides in the classroom. *Review of Communication, 6*(1), 101–123.

Lochner, L. (2011). Networkingtipps. Abgerufen 22. August, 2013, unter http://www.networkingscout.de/netzwerktipps.php.

Lovejoy, T., Revenson, T., & France, C. (2011). Reviewing manuscripts for peer-review journals: A primer for novice and seasoned reviewers. *Annals of Behavioral Medicine, 42*(1), 1–13.

Lutz, A. (2009). *Praxisbuch Networking. Einfach gute Beziehungen aufbauen – Von Adressmanagement bis Xing.com*. Wien: Linde.

Mai, J. (2008). Netter worken – 45 Tipps für besseres Netzwerken. Abgerufen 22. August, 2013, unter http://karrierebibel.de/netter-worken-45-tipps-fuer-besseres-netzwerken/.

Nicol, A. A. M. & Pexman, P. M. (2010). *Presenting your findings: A practical guide for creating tables* (6th ed.). Washington: American Psychological Association.

Nölke, C. (1998). *Präsentieren*. Planegg: STS.

Öttl, C. & Härter, G. (2004). *Networking: Kontakte gekonnt knüpfen, pflegen und nutzen*. Hamburg: Hoffmann & Campe.

Silvia, P. J. (2002). *How to write a lot. A practical guide to productive academic writing*. Washington: American Psychological Association.

University of Leicester. (2011). Designing an academic poster. Retrieved August 22, 2013, from http://www2.le.ac.uk/offices/ld/resources/presentation/designing-poster/designing-poster.

Wager, E., Godlee, F., & Jefferson, T. (2002). *How to survive peer review*. London: BMJ Books.

Wallwork, A. (2011). *English for writing research papers*. New York: Springer.

Vertiefung

Kapitel 14 **Bestimmung von Teststärke, Effektgröße und optimalem Stichprobenumfang – 807**
Nicola Döring, Jürgen Bortz

Kapitel 15 **Minimum-Effektgrößen-Tests – 867**
Nicola Döring, Jürgen Bortz

Kapitel 16 **Metaanalyse – 893**
Nicola Döring, Jürgen Bortz

Kapitel 17 **Strukturgleichungsmodelle – 945**
Christina S. Werner, Karin Schermelleh-Engel, Carla Gerhard, Jana C. Gäde

Kapitel 18 **Evaluationsforschung – 975**
Nicola Döring, Jürgen Bortz

Bestimmung von Teststärke, Effektgröße und optimalem Stichprobenumfang

Nicola Döring, Jürgen Bortz

14.1 Teststärkeanalysen – 809

14.2 Ergebnisse quantitativer Studien in Form von standardisierten Effektgrößen berichten – 816

14.3 Quantitative Studien mit optimalen Stichprobenumfängen planen – 840

14.4 Beispiele für die Planung optimaler Stichprobenumfänge und das Berichten von Effektgrößen – 849

Literatur – 864

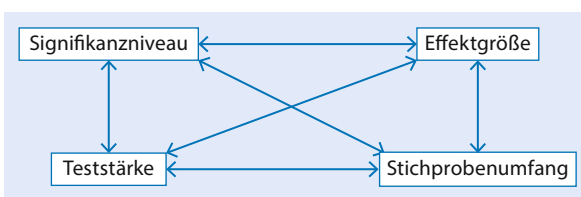
N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_14, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Die Teststärke definieren und Post-hoc- sowie A-priori-Teststärkeanalysen voneinander abgrenzen können.
- Wissen, was man unter der Effektgröße versteht und wie man sie berechnet.
- Verschiedene standardisierte Effektgrößenmaße unterscheiden und hinsichtlich ihrer Ausprägung als kleine, mittlere oder große Effekte einordnen können.
- Das Konzept des optimalen Stichprobenumfangs erläutern können.
- Wissen, wie man den optimalen Stichprobenumfang für Studien mit unterschiedlichen Signifikanztests im Zuge der Untersuchungsplanung festlegt.

In der quantitativen empirischen Sozialforschung sind hypothesenprüfende (explanative) Studien der Regelfall. Sie werden meist mittels **klassischem Nullhypothesen-Signifikanztest** ausgewertet, der auch als **Nil-Nullhypothesen-Signifikanztest** bezeichnet wird, um zum Ausdruck zu bringen, dass gemäß Nullhypothese ein Populationseffekt exakt vom Betrag Null („Nil“) oder in Gegenrichtung des postulierten Effekts erwartet wird (▶ Abschn. 12.5.1). Es wurde bereits dargestellt, dass und warum es unbefriedigend und nicht selten sogar irreführend ist, ein empirisches Studienergebnis allein daraufhin zu betrachten, ob es statistisch signifikant ist oder nicht, d. h. ob die Irrtumswahrscheinlichkeit p das Signifikanzniveau α unterschreitet oder nicht (▶ Abschn. 12.5.1) und somit die Nil-Nullhypothese zurückzuweisen ist. Denn jeder statistische Signifikanztest basiert auf vier – wechselseitig voneinander abhängigen – Größen (◘ Abb. 14.1): 1. dem Signifikanzniveau α (konventionell fixiert auf 5 %), 2. der Teststärke $1 - \beta$ (konventionell fixiert auf 80 %), 3. dem Stichprobenumfang N sowie 4. der Effektgröße ES in der Population. Wenn drei dieser Parameter festgelegt sind, lässt sich innerhalb des gewählten statistischen Rahmenmodells (z. B. Allgemeines Lineares Modell und t-Test für unabhängige Stichproben auf der Basis der t-Verteilung mit entsprechenden Freiheitsgraden) die vierte Größe bestimmen.

Durch die Verknüpfung der vier Größen kann es vorkommen, dass ein statistisch signifikantes Ergebnis auf einem theoretisch und/oder praktisch **bedeutungslosen Effekt** (z. B. einem sehr kleinen Effekt) basiert – das ist v. a. bei großem Stichprobenumfang bzw. großer Teststärke der Fall. Ebenso kann es vorkommen, dass ein **theoretisch und/oder praktisch bedeutsamer Effekt** (z. B. ein sehr großer Effekt) statistisch dennoch nicht signifikant wird – das ist besonders bei kleinem Stichprobenumfang bzw. geringer Teststärke der Fall. Um ein Signifikanztestergebnis richtig zu deuten, ist es somit notwendig, neben der Signifikanzaussage (Vergleich der Irrtumswahrscheinlichkeit mit dem Signifikanzniveau) immer auch Teststärke und Effektgröße zu betrachten (▶ Abschn. 12.5.1).



◘ **Abb. 14.1** Wechselseitige Beziehungen im Signifikanztest

Vereinfacht gesagt interessiert man sich bei einem **nicht-signifikanten Ergebnis** primär dafür, ob die Teststärke ausreichend war, um den interessierenden Populationseffekt tatsächlich aufdecken zu können. Denn nur bei ausreichender Teststärke kann ein nicht-signifikantes Ergebnis als Hinweis auf die Gültigkeit der Nullhypothese interpretiert werden. Diese nachträgliche Bestimmung der Teststärke wird als **Post-hoc-Teststärkeanalyse** bezeichnet und in ▶ Abschn. 14.1 neben anderen Formen der Teststärkeanalyse behandelt, wobei auch auf die Problematik dieses Vorgehens eingegangen wird.

Liegt ein **signifikantes Ergebnis** vor, war die Teststärke offenbar ausreichend, um einen Effekt aufzudecken, sofern er in der Population vorliegt (zu beachten ist, dass ein signifikantes Ergebnis natürlich auch durch einen α-Fehler zustande kommen kann, wenn kein Populationseffekt existiert – diese Möglichkeit muss gerade bei geringer Teststärke in Betracht gezogen werden). Bei einem signifikanten Ergebnis interessiert primär, wie groß der Effekt ist. Hierzu sollte er in der Regel in Form eines **standardisierten Effektgrößenmaßes** angegeben werden, welcher von der Messeinheit der Variable sowie vom Stichprobenumfang der jeweiligen Studie unabhängig und dadurch direkt mit anderen Studien vergleichbar ist. Anhand der standardisierten Effektgröße wird der Effekt hinsichtlich seiner theoretischen und/oder praktischen Bedeutsamkeit im Kontext anderer Studien inhaltlich eingeordnet und interpretiert. Die vollständige Angabe von standardisierten Effektgrößenmaßen trägt auch dazu bei, dass Ergebnisse von Einzelstudien später in Metaanalysen (▶ Kap. 16) problemlos zusammengefasst werden können. Es stehen – je nach Signifikanztest und Datenlage – verschiedene standardisierte Effektgrößenmaße zur Verfügung, von denen die wichtigsten in ▶ Abschn. 14.2 vorgestellt werden. Da das aus den Stichprobendaten errechnete Effektgrößenmaß nur eine (Punkt-)Schätzung der Populationseffektgröße darstellt, werden **Konfidenzintervalle für Effektgrößen** eingesetzt, um mittels Intervallschätzung die Schätzgenauigkeit anzugeben.

Teststärke und Effektgröße sind nicht nur wichtige statistische Konzepte, um die Ergebnisse einer bereits durchgeführten Studie umfassend interpretieren zu kön-

nen. Sie dienen auch dazu, Studien sachgerecht zu planen, indem der **optimale Stichprobenumfang** festgelegt wird: Vor Durchführung der Studie wird anhand des aufzudeckenden Populationseffektes derjenige Mindeststichprobenumfang bestimmt, der notwendig ist, um einen Effekt dieser Größenordnung mit hoher Wahrscheinlichkeit anhand der Daten aufdecken zu können, wenn er denn vorliegt. Wie man im Zuge einer sog. **A-priori-Teststärkeanalyse** die optimalen Stichprobenumfänge für unterschiedliche Signifikanztests bestimmt, zeigt ▶ Abschn. 14.3. Abschließend wird in ▶ Abschn. 14.4. noch einmal die Bestimmung optimaler Stichprobenumfänge sowie die Schätzung von Populationseffektgrößen anhand unterschiedlicher **Studienbeispiele** rechnerisch demonstriert.

Dass bei einem statistisch signifikanten Effekt auch auf den Betrag der Mittelwertdifferenz oder des Korrelationskoeffizienten zu achten ist (Effektgröße), dass es möglicherweise an der zu geringen Stichprobengröße lag, dass ein Signifikanztest knapp die Signifikanzschwelle verfehlte (Teststärke) oder dass eine Studie ausreichend groß anzulegen ist, um differenzierte inferenzstatistische Aussagen treffen zu können (optimaler Stichprobenumfang) – all dies wird argumentativ im Zusammenhang mit statistischen Signifikanztests durchaus angeführt. Die in den letzten Jahren an Bedeutung gewinnenden Konzepte Effektgröße, Teststärke und optimaler Stichprobenumfang formalisieren diese „weichen" Argumente durch statistische Größen und Normen und schärfen den Blick für die tatsächliche Aussagekraft von Signifikanztests. Eine besondere Herausforderung für die Sozialwissenschaften besteht dabei darin, sich mit Effektgrößen und deren theoretischer, praktischer oder klinischer Bedeutung zu beschäftigen, anstatt sich damit zufrieden zu geben, dass ein hypothesenkonformer Effekt ungleich Null vorliegt (statistische Signifikanz). Die Interpretation von Effektgrößen ist dabei gebunden an die Einordnung der Ergebnisse einer einzelnen Studie in den entsprechenden Forschungsstand bzw. den Vergleich mit thematisch ähnlichen Studien und deren Effektgrößen.

14.1 Teststärkeanalysen

Man spricht von **Teststärkeanalysen**, wenn von den vier definierenden Parametern eines statistischen Signifikanztests (Signifikanzniveau/α-Fehler-Niveau, Teststärke, Stichprobenumfang und Effektgröße) jeweils eine Größe aus den anderen drei Größen berechnet wird. Dementsprechend werden auch verschiedene Typen von Teststärkeanalysen differenziert, wobei die **Post-hoc-Teststärkeanalyse**, bei der nachträglich die erreichte Teststärke einer Studie bestimmt wird, sowie die **A-priori-Teststärkeanalyse**, bei der vorab der optimale Stichprobenumfang für eine Studie ermittelt wird, am meisten verbreitet sind (Cohen, 1988; Erdfelder, Faul, Buchner, & Cüpper, 2010). Beide Varianten sowie weitere Typen der Teststärkeanalyse werden in diesem Abschnitt vorgestellt.

Teststärke – Die Teststärke („power") hat das Symbol $1 - \beta$ und ist definiert als die Wahrscheinlichkeit, mit der ein Signifikanztest einen in der Population vorhandenen **„wahren" Effekt bestimmter Größe** durch ein statistisch **signifikantes Ergebnis** tatsächlich aufdeckt.
Bei hoher Teststärke weist ein Signifikanztest eine hohe **Sensitivität** auf, kann also auch kleine Populationseffekte mit großer Sicherheit aufdecken. Konventionell wird in Anlehnung an Cohen (1988) eine Teststärke von **mindestens 80 %** gefordert. Das heißt umgekehrt, es wird eine β-Fehlerwahrscheinlichkeit (Wahrscheinlichkeit, die H_0 fälschlich anzunehmen bzw. die H_1 fälschlich zu verwerfen) von maximal 20 % akzeptiert. Man beachte, dass die Konventionen für das α-Fehler-Niveau (.05), das β-Fehler-Niveau (.20) sowie die Teststärke (.80) nicht mechanisch anzuwenden, sondern inhaltlich begründet festzulegen sind. Wenn also z. B. inhaltlich ein α-Fehler nicht 4-mal gravierender ist als ein β-Fehler, sollten α- und β-Fehler-Niveau entsprechend ausgewogen festgelegt werden (▶ Abschn. 12.5.1).

Die Teststärke eines Signifikanztests ist umso größer (▶ Mangelnde Teststärken und geringe Effektgrößen quantitativer empirischer Studien – ein von der Scientific Community lange ignoriertes Problemfeld),

- je größer das **Signifikanzniveau** α („significance criterion"; denn α- und β-Fehler-Wahrscheinlichkeit verhalten sich gegenläufig, ▶ Abschn. 12.5.1; d. h. bei $\alpha = .05$ ist die β-Fehler-Wahrscheinlichkeit geringer und somit die Teststärke größer als bei $\alpha = .01$; bei einseitigem Test mit $\alpha = .05$ ist die Teststärke höher als bei zweiseitigem Test mit $\alpha/2 = .025$, wenn das Vorzeichen des Effekts hypothesenkonform ist),
- je größer der **Stichprobenumfang** N („sample size") und
- je größer die **Effektgröße** ES in der Population („effect size").

14.1.1 Die Post-hoc-Teststärkeanalyse

Bei der Post-hoc-Teststärkeanalyse wird nachträglich („post hoc") die Teststärke einer bereits durchgeführten Studie bestimmt. Im Kontext der Post-hoc-Teststärkeanalyse unterscheiden wir zwischen einer Ausgangslage mit nicht-signifikantem und mit signifikantem Studienergebnis.

Post-hoc-Teststärkeanalyse – Bei einer Post-hoc-Teststärkeanalyse („post hoc power analysis") wird nach Durchführung einer quantitativen empirischen Studie anhand des festgelegten Signifikanzniveaus, des realisierten Stichprobenumfanges sowie der interessierenden Populationseffektgröße die **Teststärke** geschätzt. Dabei wird die interessierende Populationseffektgröße entweder anhand der Stichprobeneffektgröße festgesetzt (retrospektive Teststärkeanalyse, „retrospective power analysis"), was methodisch problematisch ist, oder anhand theoretischer bzw.

Mangelnde Teststärken und geringe Effektgrößen quantitativer empirischer Studien – ein von der Scientific Community lange ignoriertes Problemfeld

Der US-amerikanische Psychologe und Statistiker **Jacob Cohen** (1923–1998) hatte bereits 1962 auf die Problematik unzureichender Teststärke der verwendeten Signifikanztests in vielen psychologischen und anderen verhaltenswissenschaftlichen Studien hingewiesen. 1969 publizierte er ein anwendungsorientiertes, in der zweiten Auflage noch heute gebräuchliches Handbuch zur Teststärkeanalyse (Cohen, 1988) und hoffte, die Scientific Community für das Thema zu sensibilisieren.

Tatsächlich wurden **kritische Stimmen** an einer allein auf die Irrtumswahrscheinlichkeit des Nil-Nullhypothesen-Signifikanztests orientierten statistischen Hypothesenprüfung lauter (Cohen, 1990, 1994; Gigerenzer, 1993; Nickerson, 2000; Sedlmeier & Gigerenzer, 1989).

Die **alltägliche Forschungs- und Publikationspraxis** zeigte sich jedoch weitgehend unbeeindruckt und hielt an der alleinigen Beachtung der Irrtumswahrscheinlichkeit fest, obwohl auch konkrete Empfehlungen zur verbesserten Gestaltung der Ergebnispräsentation statistischer Befunde vorlagen (z. B. Thompson, 1994). Dreißig Jahre nach seinem ersten Beitrag zur Teststärkeproblematik klagte Cohen (1992) weiterhin über die geringe Beachtung des Konzeptes im Kontext der statistischen Hypothesenprüfung.

Es liegen eine Reihe von **Übersichtsartikeln** vor, die für einen bestimmten Zeitraum und für ausgewählte Fachzeitschriften alle Artikel daraufhin analysieren, ob sie a) eine Teststärkeanalyse beinhalten und b) welche Teststärken jeweils realisiert wurden (zum Überblick Ellis, 2010, S. 73ff.). Diese Studien kamen und kommen nicht selten zu dem ernüchternden Ergebnis, dass die Teststärkeproblematik in der Fachliteratur kaum diskutiert wird und dass anhand des jeweiligen Studiendesigns (Stichprobenumfang, festgelegtes Signifikanzniveau) sowie der erwarteten oder berichteten Effektgrößen auf geringe Teststärke geschlossen werden muss („**underpowered studies**“). Manche Reviews berichteten **durchschnittliche Teststärken um 50 %** (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Das bedeutet, die Studien waren so angelegt, dass sie von vorne herein nur etwa eine 50 : 50-Chance hatten, einen vorhandenen Populationseffekt tatsächlich durch ein signifikantes Stichprobenergebnis aufzudecken. Anstelle dieser empirischen Studien hätte man also ebenso gut auf der Basis eines Münzwurfes über Annahme bzw. Ablehnung der Hypothesen entscheiden können. In vielen sozialwissenschaftlichen Disziplinen liegen die Teststärken publizierter Studien bei kleinen Effekten sogar noch deutlich niedriger, nämlich um 20 %, so dass umgekehrt die Wahrscheinlichkeit 80 % beträgt, dass kleine Effektgrößen durch die Studien unaufgedeckt bleiben (Ellis, 2010, S. 76).

Eine Durchsicht referierter psychologischer und erziehungswissenschaftlicher Zeitschriftenartikel aus den Jahren 2005 bis 2010 ergab, dass in 3.5 % der Beiträge über eine Post-hoc-Teststärkeanalyse und in 2 % der Artikel über eine A-priori-Teststärkeanalyse berichtet wurde; das heißt **94.50 % der Beiträge ignorierten das Teststärkeproblem** (Peng, Long, & Abaci, 2012). Welche fatalen Konsequenzen mangelnde Teststärken für den kumulativen Erkenntnisgewinn empirischer Wissenschaften haben, wird eindrucksvoll bei Maxwell (2004) gezeigt. In der klinischen Forschung wird gelegentlich sogar behauptet, dass Studien mit zu geringer Teststärke gegenüber den Patientinnen und Patienten ethisch nicht zu vertreten seien (Halpern, Karlawish, & Berlin, 2002; für eine andere Position s. Janosky, 2002; Lilford & Stevens, 2002).

14

praktischer Argumente festgelegt, was die empfehlenswerte Vorgehensweise ist.

Bei einem **nicht-signifikanten Ergebnis** ist die Post-hoc-Teststärkeanalyse relevant, um mangelnde Teststärke bzw. eine nicht aussagekräftige Studie zu identifizieren.

Bei einem **signifikanten Ergebnis** ist die Post-hoc-Teststärkeanalyse wichtig, um übermäßige Teststärke zu identifizieren bzw. ein signifikantes Ergebnis als theoretisch und/oder praktisch bedeutungslos zu identifizieren, wobei Effektgrößenmaße zum Einsatz kommen. Aber auch bei einem signifikanten Ergebnis kann zu geringe Teststärke ein Problem sein und als Hinweis auf das mögliche Vorliegen eines α-Fehlers dienen.

Nicht-signifikantes Ergebnis und Teststärke

Ein **nicht-signifikantes Signifikanztestergebnis** darf nicht vorschnell zum Anlass genommen werden, eine Forschungshypothese zu verwerfen und die Nullhypothese anzunehmen. Denn ein nicht-signifikantes Ergebnis kommt nicht nur zustande, wenn kein oder nur ein vernachlässigbar kleiner Effekt in der Population existiert, sondern sehr häufig auch einfach dadurch, dass die Studie auf einem **zu geringen Stichprobenumfang** und damit auf **zu geringer Teststärke** fußt, um überhaupt einen Populationseffekt aufdecken zu können.

Wenn ein nicht-signifikantes Ergebnis vorliegt, muss deswegen zunächst geprüft werden, ob denn eine ausreichende Teststärke (konventionell mindestens $1-\beta = .80$) erreicht wurde. Nur wenn die Teststärke ausreichend war, darf ein nicht-signifikantes Ergebnis dazu veranlassen, die Nullhypothese beizubehalten und die spezifische Forschungshypothese zu verwerfen. Stellt sich jedoch bei der **Post-hoc-Teststärkeanalyse** heraus, dass die Teststärke unzureichend war („underpowered study“), so muss die gesamte Studie als nicht aussagekräftig im Hinblick auf den angezielten Hypothesentest eingeordnet werden. Denn bei nicht-signifikantem Ergebnis darf man sich nicht für die H_1 entscheiden, bei mangelnder Teststärke jedoch auch nicht für die H_0, so dass letztlich gar keine Entscheidung im Hinblick auf das aus H_1 und H_0 bestehende Hypothesenpaar möglich ist. Man beachte, dass im Kontext der Teststärkeanalyse eine **spezifische H_1** formuliert werden muss, welche die erwartete Effektgröße in der Population spezifiziert (▶ Abschn. 12.5.1).

Die Post-hoc-Teststärkeanalyse erfolgt auf der Basis von **Tafelwerken** (Cohen, 1988) oder Software wie **G*Power** (▶ Hinweise zur Durchführung von Teststärkeanalysen) und wird in der Praxis in zwei verschiedenen

Hinweise zur Durchführung von Teststärkeanalysen

Um der Ignoranz gegenüber dem Teststärkeproblem in der Scientific Community entgegenzuwirken (▶ Mangelnde Teststärken und geringe Effektgrößen quantitativer empirischer Studien – ein von der Scientific Community lange ignoriertes Problemfeld), setzte die American Psychological Association (APA) Ende der 1990er-Jahre eine **„Task Force on Statistical Inference"** ein, die Richtlinien zur statistischen Hypothesenprüfung entwickelte, die dezidiert auch die Betrachtung von Effektgröße und Teststärke beinhalten (Wilkinson & Inference, 1999).

Inzwischen verlangen das **APA Publication Manual** (APA, 2001, 2009) sowie die **Autorenrichtlinien vieler Fachzeitschriften** (z. B. *Educational and Psychological Measurement, Journal of Applied Psychology, Journal of Consulting and Clinical Psychology, Health Psychology*) neben Angaben zu Teststatistiken und Irrtumswahrscheinlichkeiten auch immer das Berichten von Effektgrößen und Teststärken. Es liegen mehrere aktuelle englischsprachige **Handbücher zur Teststärkeanalyse** vor (Aberson, 2010; Bausell & Li, 2003; Cumming, 2012; Ellis, 2010; Grissom & Kim, 2011; Murphy, Myors, & Wolach, 2008).

Die praktische Durchführung von Teststärkeanalysen kann einerseits in der Tradition von Cohen (1988) anhand von **Tafelwerken** erfolgen (eine Reihe von einschlägigen Tafeln werden in diesem Kapitel präsentiert), andererseits **analytisch mithilfe entsprechender Software**. So enthält z. B. das Statistikprogrammpaket SPSS ein Modul zu „Sample Power" und „Observed Power" (für einen Vergleich verschiedener Softwarelösungen s. Peng et al., 2012).

- Ein international anerkanntes Computerprogramm zur Durchführung von **Teststärkeanalysen** steht mit **G*Power** kostenlos zur Verfügung (http://www.gpower.hhu.de/; Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007; zur Bewertung s. auch Peng et al., 2012). Bei der Benutzung von G*Power ist es von großer Bedeutung, dass man a) die korrekte Testprozedur wählt und b) die Effektgrößen unter H_1 und sonstige Eingabeparameter korrekt spezifiziert. Dazu ist anzuraten, G*Power erst nach gründlicher Lektüre der von G*Power verwendeten Effektgrößen und sonstigen Eingabeparameter – wie in Faul et al. (2007, 2009) beschrieben – zu benutzen.

Aus methodischer Sicht ist heute unstrittig, dass zur statistischen Hypothesenprüfung mittels Nil-Nullhypothesen-Signifikanztest eine **Teststärkeanalyse** dazugehört, um zu aussagekräftigen Resultaten zu gelangen. In der Forschungspraxis werden zudem Angaben zu **unstandardisierten oder standardisierten Effektgrößen** häufiger gefordert und geliefert, nicht zuletzt im Zuge der wachsenden Bemühungen um Forschungssynthese mittels statistischer Metaanalysen, die auf vollständige Darstellung der Effektgrößen in Primärstudien angewiesen sind (▶ Kap. 16). Die 6. Auflage des APA Publication Manual mahnt:

> » For the reader to appreciate the magnitude or importance of a study's findings, it is **almost always** necessary to include some **measures of effect size** in the Results section. Whenever possible, provide a **confidence interval for each effect size** reported to indicate the precision of estimation of the effect size. Effect sizes may be expressed in the **original units** (e. g., the mean number of questions answered correctly; kg/month for a regression slope) and are often most easily understood when reported in original units. It can often be valuable to report an effect size not only in original units but also in some **standardized or units-free unit** (e. g., as a Cohen's d value) or a standardized regression weight. Multiple degree-of-freedom effect size indicators are often less useful than effect-size indicators that decompose multiple degree-of-freedom tests into meaningful **one degree-of-freedom effects** – particularly when the latter are the results that inform the discussion. The general principle to be followed, however, is to provide the reader with enough information to assess the magnitude of the observed effect. (APA, 2009, S. 34; Hervorhebungen durch Autoren hinzugefügt)

Hinweise zu Berechnung und Interpretation von standardisierten Effektgrößenmaßen inklusive Konfidenzintervallen für Effektgrößen sind den genannten einschlägigen Handbüchern zur Teststärke zu entnehmen, zudem sind für den Einstieg forschungspraktische Übersichtsartikel empfehlenswert (z. B. Ferguson, 2009; Fritz, Morris, & Richler, 2012). Während **Punktschätzungen für Effektgrößen** von gängigen Statistikprogrammpaketen wie SPSS sowie von G*Power vorgenommen werden, muss für die Bestimmung von **Effektgrößen-Konfidenzintervallen** auf spezielle Software zurückgegriffen werden, da hier nichtzentrale Prüfverteilungen genutzt werden (Fritz et al., 2012, S. 13). Eine umfassende und komfortable Softwarelösung zur Bestimmung von Effektgrößen-Konfidenzintervallen liegt aktuell nicht vor, vielmehr kann nur auf unterschiedliche kostenlose Tools mit jeweils begrenztem Funktionsumfang und Komfort zurückgegriffen werden:

- **ESCI** („Exploratory Software for Confidence Intervals") von Cumming (Cumming, 2012; Download: http://www.thenewstatistics.com)
- **Skripts/Makros von Smithson** (2003), die mit Excel bzw. SPSS oder anderen Statistik-Programmpaketen laufen (Download: http://dl.dropbox.com/u/1857674/CIstuff/CI.html).
- **Tools von Steiger** (2004) zur Berechnung von Nichtzentralitätsparametern sowie zur direkten Bestimmung von Effektgrößen-Konfidenzintervallen (z. B. des multiplen Determinationskoeffizienten; Download: http://www.statpower.net/Software.html).
- **Online-Rechner von Wilson** („Practical Meta-Analysis Effect Size Calculator"), basierend auf Lipsey und Wilson (2001), der Punkt- und Intervallschätzer unterschiedlicher Gruppendifferenz-, Zusammenhangs- und Risiko-Effektgrößenmaße berechnet. Er wird über die Website der Campbell-Collaboration (▶ Abschn. 16.1.2) angeboten (http://www.campbellcollaboration.org/resources/research/effect_size_calculator.php) sowie über die Website des Autors (http://gemini.gmu.edu/cebcp/EffectSizeCalculator/)
- **SAS-Syntax** zur Bestimmung von Effektgrößen-Konfidenzintervallen (▶ Anhang zu diesem Buch).

Studienbeispiel

Sind Frauen gesprächiger als Männer? Beispiel einer Post-hoc-Teststärkeanalyse bei nicht-signifikantem Ergebnis

„Frauen reden und Männer schweigen" – so ein gängiges Geschlechterstereotyp (Abb. 14.2). Auf rund **20 000 Wörter pro Tag kämen Frauen** im Durchschnitt – **dreimal so viel wie die Männer**, die im Mittel nur 7 000 Wörter produzierten, behauptete die Medizinerin Louann Brizendine in ihrem vielzitierten Buch „The female brain" (2006). Legt man eine Sprechgeschwindigkeit von 100 Wörtern pro Minute zugrunde, wurde also behauptet, dass Frauen im Durchschnitt gut **2 Stunden** (bzw. **13 000 Wörter**) mehr als Männer täglich sprechen. Allerdings stützte die Autorin sich bei dieser von den Massenmedien weltweit aufgegriffenen statistischen Angabe auf eine einzige, zudem nicht-wissenschaftliche Quelle. In der Neuauflage des Buches wurde die fragliche Behauptung eliminiert.

Eine Forschergruppe um den Psychologen Matthias Mehl von der Arizona University prüfte daraufhin die Hypothese, dass Frauen gesprächiger als Männer seien, anhand von objektiven (d. h. mithilfe von Aufzeichnungsgeräten durchgeführten) Wortzählungen bei $n = 210$ **Frauen** und $n = 186$ **Männern**, allesamt nordamerikanische oder mexikanische Studierende zwischen 17 und 29 Jahren (Mehl, Vazire, Ramirez-Esparza, Slatcher & Pennebaker, 2007). Es stellte sich heraus, dass die untersuchten Frauen durchschnittlich **16 215** (**SD = 7 301**) Wörter pro Tag äußerten und die Männer **15 669** (**SD = 8 633**).

Der interessierende Effekt ist in diesem Studienbeispiel die Geschlechterdifferenz in der Anzahl der täglich gesprochenen Wörter (verhältnisskalierte Variable mit absolutem Nullpunkt). Quantitativ handelt es sich um eine unstandardisierte Effektgröße von $M_{\text{Frauen}} - M_{\text{Männer}} = 546$ **Wörter** (bzw. gut **5 min** Redezeit). Relativiert man die absolute durchschnittliche Wortdifferenz an der Streuung, so ergibt sich eine standardisierte Effektgröße von $d = 0.07$ (▶ Abschn. 12.4.1) als Punktschätzung bzw. $\text{KI}(95\,\%) = -0.13 \leq d \leq 0.27$ als Intervallschätzung (▶ Abschn. 12.4.2) – also ein nach gängigen Effektgrößenklassifikationen vernachlässigbar kleiner Effekt (▶ Abschn. 14.2.2). Der einseitige Signifikanztest (t-Test für unabhängige Stichproben) zeigte ein nicht-signifikantes Resultat: $t(\text{df} = 394) = 0.682$, $p = .248$, so dass die in der Stichprobe gefundene Mittelwertdifferenz als reine Zufallsschwankung interpretierbar und die Nullhypothese (kein Geschlechtsunterschied in der Gesprächigkeit) anzunehmen ist.

Die in der renommierten Zeitschrift „Science" publizierte Studie kam somit zu dem Ergebnis, dass sich – entgegen des Geschlechterklischees von der gesprächigen Frau und dem schweigsamen Mann – beide Gruppen **nicht** systematisch in der Redemenge unterschieden, beide äußerten durchschnittlich **rund 16 000 Wörter pro Tag**, was einer Rededauer von ca. 2.7 Stunden entspricht. Dabei gab es jedoch deutliche interindividuelle Variation, etwa männliche Vielredner, die auf mehr als 43 000 Wörter (7 Stunden) pro Tag kamen und Schweigsame, die sich mit insgesamt 700 Wörtern (7 Minuten) begnügten.

Angesichts der Tatsache, dass in der Studie ein nicht-signifikantes Ergebnis zum Anlass genommen wurde, die Nullhypothese (kein Geschlechtsunterschied) anzunehmen, ist eine **Post-hoc-Teststärkeanalyse** indiziert. Schließlich könnte es sein, dass der Test nur deswegen nicht signifikant wurde, weil der Stichprobenumfang zu gering gewesen ist, um einen Populationseffekt aufzudecken (Erdfelder et al., 2010, S. 359).

Aber welcher Populationseffekt hätte aufgedeckt werden sollen? Da die Studie selbst auf einer unspezifischen gerichteten Alternativhypothese basierte (H_1: $\mu_{\text{Frauen}} > \mu_{\text{Männer}}$), muss für die Post-hoc-Bestimmung der Teststärke eine **Festlegung der aufzudeckenden Populationseffektgröße** vorgenommen werden. Dazu gibt es die bereits beschriebenen zwei Wege:

Post-hoc-Teststärkeanalyse mit empirisch festgelegter Populationseffektgröße

Der in der **Stichprobe beobachtete Effekt** ($d = 0.07$) wird als Schätzer des Populationseffekts verwendet und für diesen wird im Zuge der retrospektiven Teststärkeanalyse die erreichte Teststärke („observed power") bestimmt. Sie beläuft sich hier laut G*Power auf $1 - \beta = .17$. Um einen dermaßen kleinen standardisierten Populationseffekt von $d = 0.07$ aufdecken zu können, ist ein extrem großer Stichprobenumfang notwendig ($n = 5\,050$ bei gleichgroßen Gruppen), der in der vorliegenden Studie ($n = 396$) nicht gegeben war, so dass die beobachtete Teststärke mit 17 % weit unter 80 % liegt. Die Studie wäre nach dieser Argumentation als nicht aussagekräftig einzustufen: Die Nullhypothese dürfte nicht angenommen werden. Denn ein Populationseffekt von gut 500 Wörtern bzw. **5 Minuten** hätte in der vorliegenden Studie mangels zu geringer Teststärke statistisch gar kein signifikantes Resultat liefern können.

Post-hoc-Teststärkeanalyse mit theoretisch festgelegter Populationseffektgröße

Anstatt die Stichprobeneffektgröße zu nutzen (was wegen der Gleichsetzung von Stichprobe und Population fragwürdig ist), sollte bei der Post-hoc-Teststärkeanalyse besser eine aus theoretischer oder praktischer Sicht als relevant erachtete Populationseffektgröße herangezogen werden. So argumentieren z. B. Erdfelder et al. (2010, S. 361), eine **unstandardisierte Effektgröße von 10 % der Wortdifferenz** (was anhand der vorliegenden Stichprobendaten 1 600 Wörtern oder rund 16 Minuten Rededauer entspricht) könne bei der vorliegenden Hypothese „sicherlich als erheblich angesehen werden". Der unstandardisierte Effekt von 10 % Wortdifferenz entspricht – wenn man die Daten von Mehl et al. (2007) zugrunde legt – einer standardisierten Effektgröße von $d = 0.20$. Mithilfe von G*Power lässt sich für $d = 0.20$ für die realisierten Gruppengrößen eine Teststärke von $1 - \beta = .63$ bestimmen, die unterhalb der konventionell geforderten Teststärke von 80 % liegt, so dass der Test auch nach dieser Argumentation als nicht aussagekräftig und die Annahme der Nullhypothese als unzulässig einzuordnen wäre: Eine Rededifferenz von rund einer **Viertelstunde** in der Population wäre mit der vorliegenden Studie nicht mit ausreichender Sicherheit aufzudecken gewesen, weshalb ein solcher Effekt *nicht* als widerlegt gelten darf.

Eine andere inhaltliche Argumentation könnte bei dem Umstand ansetzen, dass in der Öffentlichkeit ein extrem großer Effekt von rund 13 000 Wörtern Differenz diskutiert wird, was – die Streuungswerte von Mehl et al. (2007) zugrunde gelegt – einer standardisierten Effektgröße von etwa $d = 1.63$ entsprechen würde. Vor diesem Hintergrund kann es als relevant erachtet werden, auch mittlere Effekte empirisch zu testen. Man könnte z. B. einen schwachen mittleren Effekt für bedeutsam erklären, also eine **standardisierte Effektgröße von** $d = 0.40$ zugrunde legen (dies entspräche

einem unstandardisierten Effekt von ca. 3 000 Wörtern bzw. 30 Minuten Redezeit). In diesem Fall wäre die Studie laut Post-hoc-Teststärkeanalyse mit ausreichender Teststärke ausgestattet. Denn bereits ab einer standardisierten Effektgröße von $d = 0.25$ erreicht der Test laut G*Power für die realisierten Gruppengrößen eine Teststärke von **mindestens** $1 - \beta = .80$. Somit wäre dann die Annahme der Nullhypothese gerechtfertigt: Die vorliegende Studie hatte ausreichende Teststärke, um eine Rededifferenz von **30 min** (oder auch von **20 min**) aufzudecken. Das nicht-signifikante Ergebnis deutet also darauf hin, dass eine Differenz dieser Größenordnung in der Population nicht vorliegt und wenn überhaupt nur ein kleinerer Populationseffekt existiert. Die Alltagsvorstellung einer sehr großen Geschlechterdifferenz in der Gesprächigkeit hat sich empirisch nicht bestätigt.

Das Beispiel zeigt nicht nur die Vorgehensweise bei der Post-hoc-Teststärkeanalyse auf, sondern auch deren Limitationen rund um die überzeugende Festlegung der interessierenden bzw. relevanten Effektgröße. Da die abhängige Variable im vorliegenden Beispiel verhältnisskaliert gemessen wurde, lassen sich die anschaulichen unstandardisierten Effektgrößen (Anzahl der Wörter bzw. daraus abgeleitete Rededauer) hier vergleichend mit den standardisierten Effektgrößen (Cohens d-Werten) betrachten. Unabhängig davon, ob mit standardisierten oder unstandardisierten Effektgrößen operiert wird, bleibt es immer eine Frage der **inhaltlichen Interpretation**, Effekte welcher Größenordnung in welchem theoretischen oder praktischen Kontext als bedeutsam groß oder vernachlässigbar gering zu bewerten sind.

Nicht zuletzt ist zu beachten, dass das Ergebnis einer einzelnen Studie immer auch vor dem Hintergrund des Forschungsstandes bzw. hinsichtlich seiner **Replizierbarkeit** zu beurteilen ist. Für die Nullhypothese vernachlässigbarer Geschlechterdifferenz in der Gesprächigkeit spricht in diesem Fall auch, dass andere Studien ebenfalls keine Geschlechtsunterschiede oder sogar eine größere Gesprächigkeit der Männer in formalen wie informellen Gesprächssituationen nachgewiesen haben (James & Drakich, 1993; zur theoriebasierten Interpretation empirischer Geschlechtsunterschiede im interpersonalen Kommunikationsverhalten s. auch Reedera, 1996).

Varianten durchgeführt, die sich darin unterscheiden, wie die zur Spezifizierung der H_1 notwendige Effektgröße festgelegt wird (dazu auch ▶ Studienbeispiel „Sind Frauen gesprächiger als Männer? Beispiel einer Post-hoc-Teststärkeanalyse bei nicht-signifikantem Ergebnis“):

- **Post-hoc-Teststärkeanalyse mit empirisch festgelegter Populationseffektgröße**: Hier wird die interessierende Populationseffektgröße auf der Basis der in der Studie beobachteten Stichprobeneffektgröße festgelegt. Die Stichprobeneffektgröße geht als Schätzer der Populationseffektgröße zusammen mit dem Stichprobenumfang der realisierten Studie sowie dem Signifikanzniveau in die Analyse sein. Mit dieser **retrospektiven Teststärkeanalyse** („retrospective power analysis“) wird die sog. **beobachtete Teststärke** („observed power“) berechnet, die Auskunft darüber geben soll, ob eine Populationseffektgröße entsprechend der Stichprobeneffektgröße mit dem vorliegenden Stichprobenumfang hätte aufgedeckt werden können. Diese Vorgehensweise ist beliebt, weil sie von mehr oder minder anspruchsvollen Vorüberlegungen und Argumentationen zur Festlegung der relevanten Populationseffektgröße entlastet und einfach „mechanisch“ die in der Stichprobe gefundene Effektgröße eingesetzt werden kann. Das Problem dieser Vorgehensweise liegt aber genau in der Gleichsetzung von Stichprobe und Population, die zu einer Überschätzung der Teststärke führen kann (Gillett, 1994). Obwohl retrospektive Teststärkeanalysen verbreitet und in Statistikprogrammpaketen wie SPSS implementiert sind, ist von dieser Vorgehensweise aus methodischer Sicht **abzuraten** (Hoenig & Heisey, 2001).
- **Post-hoc-Teststärkeanalyse mit theoretisch festgelegter Populationseffektgröße**: Auf der Basis theoretischer und/oder praktischer Überlegungen sowie des publizierten Forschungsstandes wird diejenige Populationseffektgröße festgelegt, die durch die Studie mindestens aufgedeckt werden soll, um von einem



Abb. 14.2 Sind Frauen gesprächiger als Männer? (© Isabella Bannerman/Search ID: ibn0041, Rights Available from www.CartoonStock.com)

inhaltlich bedeutsamen Effekt ausgehen zu können. Diese argumentativ festgelegte Populationseffektgröße geht zusammen mit dem Stichprobenumfang der realisierten Studie sowie dem festgelegten Signifikanzniveau in die Post-hoc-Teststärkeanalyse ein. Diese Vorgehensweise ist empfehlenswert.

Stellt sich bei einem nicht-signifikanten Ergebnis heraus, dass die durchgeführte Studie mit dem realisierten Stichprobenumfang gar nicht in der Lage gewesen wäre, den interessierenden Populationseffekt aufzudecken, muss die Studie nachträglich als **underpowered** (d. h. mit unzureichender Teststärke ausgestattet) und somit nicht aussagekräftig eingeordnet werden.

Signifikantes Ergebnis und Effektgröße

Ein **statistisch signifikantes Ergebnis** kann beim Nil-Nullhypothesen-Signifikanztest theoretisch und/oder praktisch völlig bedeutungslos sein, weil es auf einem für die zugrunde liegende Theorie oder für praktische Anwendungszwecke irrelevant kleinen Effekt basiert, der nur durch übergroße Teststärke als statistisch signifikant ausgewiesen wird. Deswegen muss bei einem statistisch signifikanten Resultat ergänzend immer auch die Teststärke bzw. alternativ die Effektgröße angegeben und interpretiert werden.

Vor allem mithilfe eines **standardisierten Effektgrößenmaßes**, das von der Messeinheit der Variable und vom Stichprobenumfang der konkreten Studie unabhängig ist, lässt sich das Studienergebnis hinsichtlich seiner theoretischen und/oder praktischen Bedeutung („theoretical/practical significance"; „theoretical/practical importance") durch Fachleute bewerten, in den bisherigen Forschungsstand einordnen und kann bei Bedarf auch im Zuge einer Metaanalyse (▶ Kap. 16) mit vergleichbaren anderen Studien zu einem Gesamteffekt verrechnet werden. Zudem ist zu beachten, dass die Stichprobeneffektgröße nur eine Schätzung der Populationseffektgröße darstellt und die Schätzgenauigkeit mit einem Konfidenzintervall für die Effektgröße angegeben werden sollte.

Die Forschungscommunity bewegt sich in den letzten Jahren verstärkt von einer bloßen Signifikanzbetrachtung hin zu einer umfassenden Darstellung von Effektgrößen, zu der auch wir dringend raten. Es existieren eine Reihe von **unterschiedlichen Effektgrößenmaßen** („measures of effect sizes") für die verschiedenen statistischen Signifikanztests bzw. Untersuchungsdesigns. Das vorliegende Kapitel wird wichtige Effektgrößenmaße für unterschiedliche Signifikanztests vorstellen (▶ Abschn. 14.2).

Effektstärke bzw. Effektgröße – Die Effektstärke bzw. Effektgröße („effect size", ES) ist definiert als die Größe des im Zuge einer Hypothesenprüfung interessierenden Populationsparameters. Bei dem interessierenden Populationsparameter kann es sich z. B. um einen Prozentwert oder eine Prozentwertdifferenz, einen Mittelwert oder eine Mittelwertdifferenz, einen bivariaten oder multivariaten Korrelations- oder Regressionskoeffizienten oder eine Korrelationsdifferenz handeln. Die **Populations-Effektgröße** ist in der Regel unbekannt und wird auf der Basis der **Stichproben-Effektgröße** geschätzt, sei es als **Punktschätzung** oder als **Intervallschätzung** mit einem Konfidenzintervall.

Zwei Arten von Effektgrößenmaßen sind zu unterscheiden:

Unstandardisierte bzw. absolute Effektgrößen – Unstandardisierte bzw. absolute Effektgrößen („unstandardized effect size measures") werden in der Messeinheit der gemessenen Variablen angegeben. Unstandardisierte Effektgrößenmaße sind nur bei **verhältnisskalierten Variablen** mit absolutem Nullpunkt sinnvoll interpretierbar. So kann der Effekt von Interventionsmaßnahmen direkt beurteilt werden, wenn man z. B. erfährt, um welchen Absolutbetrag diese die Anzahl der gerauchten Zigaretten pro Tag oder das Körpergewicht in Kilogramm reduziert haben.

Standardisierte bzw. relative Effektgrößenmaße – Standardisierte bzw. relative Effektgrößenmaße („standardized effect size measures") sind durch Relativierung an der Merkmalsstreuung unabhängig von der Messeinheit und vom Stichprobenumfang. Standardisierte Effektgrößenmaße sind notwendig bei **nominal-, ordinal-** oder **intervallskalierten Variablen**, die keinen natürlichen Nullpunkt haben (z. B. Ratingskala). Sie erlauben a) eine direkte Vergleichbarkeit verschiedener Studien zum selben Phänomen (auch wenn zur Operationalisierung verschiedene Messeinheiten verwendet werden), b) eine Verrechnung der Effekte vergleichbarer Studien zu einem standardisierten Gesamteffekt im Rahmen der Metaanalyse (▶ Kap. 16) und c) die Einstufung der inhaltlichen Bedeutsamkeit des Effektes. Es existieren verschiedene standardisierte Effektgrößenmaße für unterschiedliche Effekte bzw. unterschiedliche Signifikanztests (▶ Abschn. 14.2.3). Man beachte, dass die Standardisierung stets anhand der **Streuung der Werte** erfolgt und somit relativ zur Unterschiedlichkeit der erhobenen Werte zu betrachten ist, welche wiederum von der Stichprobenzusammensetzung sowie den Untersuchungsbedingungen – etwa der Dosierung der unabhängigen Variablen – abhängt (▶ Abschn. 14.3.1).

Eine (un)standardisierte Effektgröße von Null signalisiert die Abwesenheit des Effekts, mit wachsendem Betrag steigt die Effektgröße, wobei je nach Effektgrößenmaß unterschiedliche **Klassifikationen für kleine, mittlere und große Effekte** vorliegen (▶ Abschn. 14.2.2).

14.1.2 Die A-priori-Teststärkeanalyse

Wünschenswert ist es, empirische Studien, die inferenzstatistisch ausgewertet werden sollen, von vorne herein (a priori) so zu planen, dass sie über eine ausreichende Teststärke verfügen und somit aussagekräftige Resultate liefern. Diese Zielsetzung kann erreicht werden, wenn bereits im Zuge der Untersuchungsplanung nicht nur wie üblich das Signifikanzniveau α (z. B. .05), sondern auch die erwartete Effektgröße ES (klein, mittel, groß) und die Teststärke $1 - \beta$ (z. B. .80) festgelegt werden. Mit einer solchen **A-priori-Teststärkeanalyse**

(„a priori/prospective power analysis“) lässt sich nämlich der **mindestens notwendige Stichprobenumfang** bestimmen, der ausreichende Teststärke und somit statistisch aussagekräftige Resultate für den erwarteten Effekt sichert.

A-priori-Teststärkeanalyse – Bei einer A-priori-Teststärkeanalyse („a priori/prospective power analysis“) werden vor Durchführung einer Untersuchung Signifikanzniveau, Teststärke und interessierende Effektgröße festgelegt, um auf dieser Basis den **optimalen“ Stichprobenumfang** („sufficient sample size“) zu bestimmen. Dies ist der notwendige Mindeststichprobenumfang, um anhand eines statistischen Nil-Nullhypothesen-Signifikanztests mit ausreichender Teststärke eine aussagekräftige Entscheidung über das Hypothesenpaar zu treffen.
Die notwendige Festlegung der interessierenden Effektgröße ist gleichbedeutend mit der Anforderung, eine (in der Regel gerichtete) **spezifische** H_1 zu formulieren, d. h. bei der Untersuchungsplanung nicht nur zu postulieren, dass ein Effekt existiert, sondern auch seine Größe bzw. zumindest seine Größenordnung anzugeben. In der Praxis wird hier häufig mit groben **Effektgrößenklassifikationen** gearbeitet, d. h. ein kleiner, mittlerer oder großer Effekt postuliert (▶ Abschn. 14.2.2).

Der für einen aussagekräftigen Signifikanztest mindestens notwendige Stichprobenumfang wird hier als **„optimaler“ Stichprobenumfang** bezeichnet, da eine Orientierung an ihm forschungsökonomisch optimal ist: Es werden eben genau so viele Untersuchungseinheiten (z. B. Personen) in die Studie einbezogen, wie mindestens notwendig sind, um über die Hypothese entscheiden zu können. Ein überflüssiger Mehraufwand für Forschende und Untersuchungsteilnehmende wird vermieden. Gleichzeitig lässt sich statistisch argumentieren, dass mit wachsendem Stichprobenumfang die Schätzgenauigkeit für die interessierenden Parameter steigt und somit im Grunde ein möglichst großer Stichprobenumfang stets optimal ist. Dagegen ist jedoch das bereits angeführte Argument der Forschungsökonomie einzuwenden (Umgang mit knappen Ressourcen) sowie das Problem, dass mit sehr großen Stichproben bei konventionell fixiertem Alpha-Fehler-Niveau schließlich winzig kleine Effekte nachweisbar werden, die praktisch bedeutungslos sind. Im vorliegenden Kapitel werden Hinweise zur Bestimmung des sog. optimalen Stichprobenumfanges differenziert für unterschiedliche Signifikanztests präsentiert (▶ Abschn. 14.3).

Es ist jedoch zu beachten, dass bei den in den Sozialwissenschaften oft interessierenden kleinen Effekten die optimalen Stichprobenumfänge nicht selten unrealisierbar hoch ausfallen, d. h. die Ressourcen eines gängigen akademischen (und insbesondere studentischen) Forschungsprojektes übersteigen. Neben dem Hochsetzen des Stichprobenumfangs sind zur Erhöhung der Teststärke deswegen auch noch mindestens **vier weitere untersuchungsplanerische Maßnahmen** zu bedenken (▶ Abschn. 14.3.1).

14.1.3 Weitere Typen der Teststärkeanalyse

Nach Cohen (1988) werden neben der Post-hoc- und der A-priori-Teststärkeanalyse noch zwei weitere Typen von Teststärkeanalysen unterschieden:

1. **Sensitivitäts-Teststärkeanalyse**: Hier wird die mindestens **aufdeckbare Populationseffektgröße** („detectable effect size“) berechnet für einen Signifikanztest, bei dem α-Niveau, Teststärke und Stichprobenumfang festliegen. Die Sensitivitäts-Teststärkenanalyse kann zur **Bewertung publizierter Studienergebnisse** eingesetzt werden. Für die Eckdaten des bereits durchgeführten Signifikanztests wird ermittelt, welche Effektgröße überhaupt hätte aufgedeckt werden können. Stellt sich dabei heraus, dass der Test nur große Effekte aufgedeckt hätte – also eine geringe Sensitivität aufwies – ist sein Beitrag zum Erkenntnisgewinn gering, wenn er ein nicht-signifikantes Resultat erzielt hat und auch mittlere oder kleine Effekte inhaltlich relevant sind.
2. **Kriteriums-Teststärkeanalyse**: Hier wird das Signifikanzkriterium bzw. Signifikanzniveau α bestimmt, das notwendig ist, um einen interessierenden Populationseffekt mit vorgegebenem Stichprobenumfang und festgelegter Teststärke in einem Signifikanztest aufdecken zu können. Soll z. B. eine Populationskorrelation von $\rho = .30$ anhand einer Stichprobe von $n = 50$ mit einer Teststärke von $1 - \beta = .75$ aufgedeckt werden, so ist bei einseitigem Test ein α-Niveau von rund 8 % und bei zweiseitigem Test von 15 % notwendig (Cohen, 1988, S. 15). Die Kriteriums-Teststärkeanalyse ist in der Praxis ungebräuchlich, da Signifikanztests, die nicht mit dem konventionalisierten 5 %- bzw. 1 %-Niveau operieren, auf wenig Akzeptanz stoßen. Die Festlegung des **optimalen Signifikanzniveaus** im Rahmen der Kriteriums-Teststärkeanalyse kann jedoch nützlich sein, wenn bei einem Signifikanztest die Nullhypothese die Wunschhypothese ist und somit die Fixierung der Teststärke auf einen hohen Wert Vorrang hat (Erdfelder et al., 2010, S. 367). Die Kriteriums-Teststärkeanalyse dient dabei der **Planung eigener Studien**, bei denen der Stichprobenumfang aus forschungsökonomischen Gründen fixiert ist und somit durch Verschiebung des Signifikanzkriteriums die gewünschte Teststärke zu sichern ist.

Darüber hinaus hat Erdfelder (1984) eine fünfte Variante vorgeschlagen, die Kompromiss-Teststärkeanalyse, die auch in das Computerprogramm G*Power (▶ Hinweise zur Durchführung von Teststärkeanalysen) integriert ist:

Kompromiss-Teststärkeanalyse: Hier werden **sowohl das α-Niveau als auch das β-Niveau** (und somit gleichzeitig die Teststärke) eines Tests bestimmt, die not-

wendig sind, um eine interessierende Effektgröße mit einem definierten Stichprobenumfang aufzudecken, wobei gleichzeitig noch das Verhältnis bzw. der Quotient von α- und β-Fehler festgelegt wird ($q = \beta/\alpha$). Bei der konventionellen Festlegung von $\alpha = 5\,\%$ und $\beta = 20\,\%$ (bzw. $1 - \beta = 80\,\%$) stehen beide Fehlerwahrscheinlichkeiten im Verhältnis $q = 4 : 1 = 4$, d. h. ein α-Fehler wird als viermal gravierender eingestuft als ein β-Fehler. Die Kompromiss-Teststärkeanalyse erlaubt nun z. B. die Planung von Studien, bei denen α- und β-Fehler gleich gewichtet werden ($q = 1/1 = 1$). Die Kompromiss-Teststärkeanalyse stellt eine Ergänzung der Kriteriums-Teststärkeanalyse dar und soll die Untersuchungsplanung unterstützen. Sie kommt selten zum Einsatz, unter anderem, da auch sie ein Abweichen von den üblichen Konventionen von $\alpha = 5\,\%$ sowie $\beta = 20\,\%$ verlangt (Erdfelder et al., 2010, S. 366).

14.2 Ergebnisse quantitativer Studien in Form von standardisierten Effektgrößen berichten

Nur zu berichten, dass ein Signifikanztest signifikant geworden und somit die Forschungshypothese H_1 anzunehmen sei, ist angesichts der Teststärkenproblematik nicht ausreichend. Eine umfassende Darstellung eines signifikanten Studienergebnisses muss eine Beschreibung und Interpretation der Effektgröße beinhalten. Dabei ist in der Regel auf standardisierte Effektgrößenmaße zurückzugreifen. Denn verhältnisskalierte Variablen, bei denen unstandardisierte Effektgrößenmaße vorzuziehen wären, sind seltener als intervallskalierte Variablen wie z. B. Ratingskalen (▶ Abschn. 8.4.4). Standardisierte Effektgrößenmaße lassen sich in unterschiedliche Typen einteilen (▶ Abschn. 14.2.1). In der Praxis hat es sich eingebürgert, standardisierte Effektgrößenmaße bestimmter Größenordnung als „kleine", „mittlere" oder „große" Effekte zu klassifizieren (▶ Abschn. 14.2.2). Wie die standardisierten Effektgrößen der wichtigsten Signifikanztests samt ihrer Konfidenzintervalle zu bestimmen sind, erläutert ▶ Abschn. 14.2.3.

14.2.1 Verschiedene Typen von Effektgrößenmaßen

Wir haben schon festgehalten, dass es sich bei den interessierenden Effekten, auf die sich Forschungshypothesen beziehen, um ganz verschiedene Populationsparameter handelt, etwa um einen Prozentwert oder eine Prozentwertdifferenz, einen Mittelwert oder eine Mittelwertdifferenz, einen bivariaten oder multivariaten Korrelationskoeffizienten oder eine Korrelationsdifferenz etc. Anhand der Stichprobendaten lässt sich für den unbekannten Populationsparameter **(nicht-standardisierte Effektgröße)** eine Punkt- oder Intervallschätzung auf der Basis des entsprechenden Stichprobenkennwertes vornehmen (▶ Abschn. 12.4).

Zum Zweck der besseren Vergleichbarkeit und Interpretierbarkeit werden Effektgrößen in der Regel mithilfe eingeführter **standardisierter Effektgrößenmaße** standardisiert. Beispiel: Man entnimmt einer Evaluationsstudie mit experimentellem Kontrollgruppendesign ($n = 100$), dass ein bestimmtes Verhaltenstraining die Prüfungsangst von Studierenden auf einer 10-stufigen Skala (1: sehr geringe Angst bis 10: sehr starke Angst) um 1.20 Skalenpunkte reduziert. Bei Vorliegen eines statistisch signifikanten Ergebnisses ist auf den ersten Blick schwer zu entscheiden, ob das Training einen großen, einen mittleren oder einen kleinen Effekt hatte. Denn bei der **Mittelwertdifferenz** von $M_{KG} - M_{EG} = 1.6$ Skalenpunkte handelt es sich um die **unstandardisierte Effektgröße** einer Skala ohne absoluten Nullpunkt. Auch ist nicht auf den ersten Blick ersichtlich, ob ein anderes Training über- oder unterlegen ist, wenn dieses laut einer Studie mit $n = 40$ Probanden auf einer 8-stufigen Skala zu einer Angstreduktion um 1.10 Punkte führte. Indem die Mittelwertdifferenz hier an der anhand der Stichprobendaten geschätzten Populationsstreuung relativiert wird, entsteht ein **standardisiertes Effektgrößenmaß**, nämlich **Cohens *d***, das von der Messeinheit und vom Stichprobenumfang unabhängig ist. Konventionell werden Cohens d-Werte um 0.20 als kleiner Effekt, um 0.50 als mittlerer Effekt und um 0.80 als großer Effekt eingeordnet (▶ Abschn. 14.2.2).

Das ***d*-Maß von Cohen** ist eines der bekanntesten und gebräuchlichsten standardisierten Effektgrößenmaße für Mittelwertdifferenzen bei zwei Gruppen. Alternative Maße sind das **Δ-Maß von Glass**, bei dem die Mittelwertdifferenz nur an der Streuung der Kontrollgruppe relativiert wird oder das ***g*-Maß von Hedges**, bei dem die Mittelwertdifferenz an der gepoolten Streuung beider Gruppen relativiert wird (zur Berechnung Gl. 14.3.). Bei **Gruppendifferenz-Effektgrößenmaßen** als standardisierten Effektgrößen („effect sizes in standardized units of difference/group difference effect size estimates") handelt es sich also jeweils um die Mittelwertdifferenz beider Gruppen relativiert an unterschiedlichen Schätzungen der Gesamtstreuung. Ihr Wertebereich liegt beim Fehlen eines Effektes bei 0 und ist nach oben bzw. unten (je nach Vorzeichen der Differenz) nicht begrenzt.

Neben denjenigen Effektgrößenmaßen, die sich auf Gruppendifferenzen in der zentralen Tendenz beziehen, existieren noch **Zusammenhangs-, Varianzaufklärung- und Risiko-Effektgrößenmaße** (Ferguson, 2009;

Tabelle 14.1 Vier verschiedene Typen von Effektgrößenmaßen und wichtige Vertreter. (Mod. nach Ferguson, 2009. Copyright © 2009 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

Vier Typen von Effektgrößenmaßen	Wichtige Vertreter (Auswahl)
Gruppendifferenz-Effektgrößenmaße (2 Gruppen)	d-Maß von Cohen Δ-Maß von Glass g-Maß von Hedges
Zusammenhangs-Effektgrößenmaße	Bivariater Korrelationskoeffizient/Produkt-Moment-Korrelation r Punktbiserialer Korrelationskoeffizient r_{pbis} Partialkorrelationskoeffizient $r_{xy.z}$ Rangkorrelationskoeffizient Spearmans ρ (rho) Rangkorrelationskoeffizient Kendalls τ (tau) Vierfelder-Korrelationskoeffizient φ (phi) Kontingenzkoeffizient C Cramérs V Goodman und Kruskalls λ (lambda) Multipler Korrelationskoeffizient R Standardisierter Regressionskoeffizient b (Beta-Gewicht)
Varianzaufklärungs-Effektgrößenmaße	Quadrierte Zusammenhangsmaße: – Bivariater Determinationskoeffizient r^2 – Multipler Determinationskoeffizient R^2 – Kanonischer Determinationskoeffizient R_C^2 Korrigierte quadrierte Zusammenhangsmaße: – Adjustierter multipler Determinationskoeffizient $R_{adj.}^2$ Effektgrößenmaße der Varianzanalyse: – Cohens f – Eta-Quadrat η^2 – partielles Eta-Quadrat η_p^2 – Eta-Quadrat gesamt η_{ges}^2 Korrigierte Effektgrößenmaße der Varianzanalyse: – Omega-Quadrat ω^2 – Epsilon-Quadrat ε^2
Risiko-Effektgrößenmaße	Odds Ratio OR Relative Risk RR

Tab. 14.1, zur Berechnung Tab. 14.6, ►Risiko-Effektgrößenmaße: Odds Ratio und Relative Risk und ►Abschn. 14.2.3). Es sei jedoch erwähnt, dass die Fülle der verschiedenen Effektgrößenmaße – allein Kirk (1996) berichtet von 41 verschiedenen Maßen – in der Literatur in unterschiedlicher Weise in Gruppen oder Typen eingeteilt wird. So differenziert z. B. Ellis (2010, S. 13ff.) nur zwischen den Effektgrößenmaßen der ***d*-Familie** (Gruppendifferenz- und Risiko-Effektgrößenmaße) einerseits und den Maßen der ***r*-Familie** (Zusammenhangs- und Varianzaufklärungsmaße) andererseits.

Bei den **Zusammenhangs-Effektgrößenmaßen** („strength of association effect size estimates“, „strength of relationship effect size measures“) handelt es sich um unterschiedliche Korrelations- und Regressionskoeffizienten, die direkt als standardisierte Effektgrößenmaße nutzbar sind, weil sie bereits an der jeweiligen Streuung relativiert sind. Ihr Wertebereich liegt meist zwischen 0 und ± 1.

Zu den **Varianzaufklärungs-Effektgrößenmaßen** („variance-accounted-for effect size measures“) zählen die quadrierten Zusammenhangsmaße, d. h. die Determinationskoeffizienten, die den durch den Merkmalszusammenhang aufgeklärten Varianzanteil beziffern. In der Forschungspraxis noch wichtiger sind innerhalb dieser Gruppe jedoch die auf der Varianzanalyse basierenden Varianzaufklärungsmaße, von denen das partielle Eta-Quadrat mit Abstand am häufigsten zum Einsatz kommt; vermutlich deswegen, weil es von Statistikprogrammpaketen wie SPSS automatisch berechnet wird (Fritz et al., 2012). Dabei ist das partielle Eta-Quadrat ein Effektgrößenmaß, das für den Vergleich unterschiedlicher Studien oft ungeeignet ist, da es sich je nach Studiendesign verändert, weshalb das generalisierte Eta-Quadrat vorzuziehen ist (►Abschn. 14.2.3; Olejnik & Algina, 2003). Die Effektgrößenmaße der Varianzanalyse berechnen sich, indem die Quadratsumme bzw. Varianz des Effekts durch Schätzungen der relevanten Gesamtvarianz dividiert wird.

Risiko-Effektgrößenmaße: Odds Ratio und Relative Risk

Wie viel höher ist das Lungenkrebsrisiko bei Rauchern im Vergleich zu Nichtrauchern? Anhand dieses Beispiels sollen Berechnung und Interpretation von Risiko-Effektgrößenmaßen illustriert werden. Ausgangspunkt ist eine 2 × 2-Kreuztabelle (Tab. 14.2), deren 4 Zellen alphabetisch bezeichnet sind, um die Anzahl der Personen in den Zellen zu beziffern (Ferguson, 2009). Es handelt sich hier um Daten aus einer **Fall-Kontroll-Studie** (Held, 2010, S. 635), d. h. es wurden zunächst 1 357 Patienten mit Lungenkarzinom identifiziert (Fälle; $A + C$), denen dann gezielt dieselbe Zahl an Nicht-Erkrankten (Kontrollen; $B + D$) mit ähnlichen Eigenschaften (z. B. hinsichtlich Alter, Blutdruck, Vorerkrankungen etc.) gegenübergestellt wurden. (Man beachte, dass bei diesem Design der Faktor Rauchen nicht experimentell manipuliert, sondern als Gruppenfaktor vorgefunden wurde, wobei durch das Matching der Erkrankten mit vergleichbaren Nicht-Erkrankten potenzielle bekannte Störvariablen – nicht jedoch unbekannte Störvariablen – kontrolliert wurden. Tab. 14.2)

Odds Ratio (OR; relative Chance/Chancenverhältnis)

Als Wahrscheinlichkeit bezeichnen wir gemäß frequentistischen Wahrscheinlichkeitsmodell die relative Häufigkeit, d. h. die Anzahl der interessierenden Fälle dividiert durch alle Fälle. Die Wahrscheinlichkeit, als Raucher an Lungenkrebs zu erkranken, errechnet sich dementsprechend als Anzahl der erkrankten Raucher dividiert durch alle Raucher, in unserem Zahlenbeispiel also: $p_1 = p_R = 1\,350/2\,646 = 0.51\,\%$. Zu jeder Wahrscheinlichkeit lässt sich nun **der Odds bzw. die Chance** bestimmen, und zwar nach der Formel $\mathbf{Odds}(p) = p/(1-p)$. In unserem Beispiel: $\text{Odds}(p_R) = .51/(1-.51) = .51/.49 = 1.04$. Bei einer Wahrscheinlichkeit von 50 % ist der Odds 1 bzw. umgangssprachlich formuliert, steht die Chance 1 : 1, dass das Ergebnis eintritt. Odds größer als 1 weisen auf eine höhere Chance bzw. ein höheres Risiko hin. Im Datenbeispiel stellt es sich so dar, dass Raucher zu etwas mehr als 50 % bzw. mit einer Chance knapp über 1 an Lungenkrebs erkranken als dass sie nicht erkranken.

Um den Effekt des Rauchens abzuschätzen, muss nun der Vergleich mit den Nichtrauchern erfolgen. Die Wahrscheinlichkeit für Lungenkrebs bei Nichtrauchern ist $p_0 = p_{NR} = 7/68 = .10$; die entsprechende Chance ist $\text{Odds}(p_{NR}) = .10/.90 = .11$.

Die Odds Ratio (OR) bzw. das Chancenverhältnis setzt nun beide Odds in Relation zueinander. $\mathbf{OR = Odds}(p_R)/\mathbf{Odds}(p_{NR}) = \mathbf{1.04/.11 = 9.45}$. Für Raucher ist das Risiko an Lungenkrebs zu erkranken also rund 9.5-mal größer als das Risiko für Nichtraucher, was als sehr starker Effekt einzuordnen ist. Allgemein ist der OR auf der Basis einer Kontingenztafel folgendermaßen zu berechnen:

$$\begin{aligned}\mathbf{OR = Odds}(p_1)/\mathbf{Odds}(p_0) &= [(p_1/(1-p_1)]/[p_0/(1-p_0)] \\ &= (A/B)/(C/D) = A \times D/B \times C\end{aligned}$$

Eine OR von 1 deutet auf keinen Effekt hin, etwa 2 auf einen kleinen, 3 auf einen mittleren und 7 auf einen starken Effekt (Chen, Cohen, & Chen, 2010; Tab. 14.4).

Relative Risk (RR)

Beim Effektgrößenmaß des relativen Risikos (RR) werden nicht die Odds bzw. die Chancen, sondern direkt die Wahrscheinlichkeiten bzw. relativen Häufigkeiten zueinander ins Verhältnis gesetzt.

$$\begin{aligned}\mathbf{RR} = p_1/p_0 &= [(A/(A+B)]/[C/(C+D)]\end{aligned}$$

Im Zahlenbeispiel beträgt die Wahrscheinlichkeit für Raucher an Lungenkrebs zu erkranken 51 % und die Wahrscheinlichkeit für Nichtraucher 10 %. Das rechnerische relative Risiko beläuft sich auf $\mathbf{RR = .51/.10 = 5.10}$.

RR-Werte zwischen 1 und 2 werden als vernachlässigbar gering eingestuft, danach ist von kleinen, mittleren oder großen Effekten zu sprechen, wobei das Basisrisiko zu berücksichtigen ist. Eine RR ist umso bedeutungsvoller, je größer das Basisrisiko ist, auf das sie sich bezieht (Ferguson, 2009).

Vergleich von OR und RR

Odds Ratio und Relative Risk sind ähnliche Maße, die sich umso mehr annähern, je kleiner die betrachteten Wahrscheinlichkeiten sind. Werden also Phänomene untersucht, die insgesamt sehr selten auftreten, sind OR und RR vergleichbar, da dann $1 - p_1$ und $1 - p_0$ jeweils nahe 1 liegen.

Man beachte, dass das relative Risiko als Effektgrößenmaß nur sinnvoll heranzuziehen ist, wenn die **Zahl der Kontrollen den Populationsverhältnissen entspricht** und nicht – wie beim Beispiel der Fall-Kontroll-Studie – von den Forschenden festgelegt werden. Die obige Berechnung von RR = 5.1 ist für das vorliegende Beispiel einer Fall-Kontroll-Studie also zwar rechnerisch möglich, aber inhaltlich nicht sinnvoll!

Der Wertebereich dieser Effektgrößenmaße liegt zwischen 0 und 1 bzw. 0 % und 100 %. Die **korrigierten Maße** ($R^2_{\text{adjustiert}}$, ω^2, ε^2) sind hauptsächlich indiziert bei a) kleinen Effekten, b) vielen Variablen und c) geringen Gruppengrößen; sie sollen Überschätzungen der Effektgröße entgegenwirken.

Die **Risiko-Effektgrößenmaße** schließlich sind in der medizinischen und klinischen sowie epidemiologischen Forschung verbreitet. Sie quantifizieren den Unterschied in der Wahrscheinlichkeit bzw. im Risiko des Auftretens eines bestimmten Ergebnisses (z. B. einer Erkrankung) unter zwei oder mehr verschiedenen Bedingungen (z. B. Behandlung/keine Behandlung; ▶ Risiko-Effektgrößenmaße: Odds Ratio und Relative Risk).

In der Praxis sollte dasjenige Effektgrößenmaß bestimmt und berichtet werden, das dem jeweils verwendeten Signifikanztest entspricht. Will man die Effekte mehrerer Studien vergleichen oder im Zuge einer Metaanalyse (▶ Kap. 16) zusammenfassen, ist es sinnvoll und möglich, verschiedene Effektgrößenmaße zu vereinheitli-

Tabelle 14.2 Kreuztabelle als Grundlage zur Bestimmung von Risiko-Effektgrößenmaßen inklusive Datenbeispiel (mod. nach Held, 2010, S. 635) einer Fall-Kontroll-Studie

	Erkrankte (Lungenkrebs)	Nicht-Erkrankte (kein Lungenkrebs)	Gesamt
Treatmentgruppe/Exponierte (Rauchen)	A (1 350) p_1	B (1 296) $1 - p_1$	$A + B$ (2 646) 1
Kontrollgruppe/Nicht-Exponierte (Nichtrauchen)	C (7) p_0	D (61) $1 - p_0$	$C + D$ (68) 1
Gesamt	$A + C$ 1 357	$B + D$ 1 357	$A + B + C + D$ 2 714

Tabelle 14.3 Formeln zur Bestimmung und wechselseitigen Umrechnung wichtiger standardisierter Effektgrößenmaße (gültig für etwa gleichgroße Gruppen; wenn eine Gruppe kleiner ist als ein Drittel des gesamten N ist r kleiner als der Umrechnungswert; Fritz et al., 2012, S. 9. Copyright © 2012 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

Umrechnung	IN dieses Effektgrößenmaß		
VON diesem Effektgrößenmaß	d-Maß von Cohen	Bivariater Korrelationskoeffizient r	η^2 (gleiche/ähnliche Gruppengrößen)
Direkte Formel	$\delta = \frac{\mu_A - \mu_B}{\sigma}$	$r = \sqrt{\frac{QS_{Effekt}}{QS_{Total}}}$	$\eta^2 = \frac{QS_{Faktor}}{QS_{Total}}$
d-Maß von Cohen	–	$r = \frac{\delta}{\sqrt{\delta^2 + 4}}$	$\eta^2 = \frac{\delta^2}{\delta^2 + 4}$
Bivariater Korrelationskoeffizient r	$\delta = \frac{2 \cdot r}{\sqrt{1 - r^2}}$	–	$\eta^2 = r^2$
η^2 mit ähnlichen Gruppengrößen	$\delta = \frac{2 \cdot \sqrt{\eta^2}}{\sqrt{(1 - \eta^2)}}$	$r = \sqrt{\eta^2}$	–
t-Wert mit ähnlichen Gruppengrößen	$\delta = \frac{2 \cdot t}{\sqrt{N - 2}}$	$r = \sqrt{\frac{t^2}{t^2 + df}}$	$\eta^2 = \frac{t^2}{t^2 + df}$

chen (Cohen, 1992). Tab. 14.3 gibt eine Übersicht zur Berechnung und wechselseitigen **Transformation der drei gebräuchlichsten Effektgrößenmaße** (das d-Maß von Cohen, die bivarialte Korrelation und Eta-Quadrat). Von einer Umrechnung von Risiko-Effektgrößenmaßen in Differenz- oder Zusammenhangsmaße wird abgeraten (Ferguson, 2009, S. 536).

14.2.2 Klassifikation von Effektgrößen: kleine, mittlere und große Effekte

Die Betrachtung von Effektgrößen soll dazu führen, dass neben der statistischen Signifikanz auch die **theoretische** und/oder **praktische Bedeutsamkeit** von Studienergebnissen diskutiert wird. Welche Effektgröße jedoch inhaltlich bedeutsam genug ist, um eine Theorie als gut belegt aufzufassen oder eine Interventionsmaßnahme als wichtig zu befürworten, kann **nicht** sinnvoll **pauschal** anhand numerischer Normen für kleine, mittlere oder große Effekte vorgegeben werden. Ein numerisch sehr kleiner Effekt kann praktisch sehr bedeutsam sein, etwa wenn es um Menschenleben und Gesundheit geht (z. B. Reduktion von Unfallzahlen, Operationsrisiken oder Rückfallquoten). In diesem Zusammenhang sei ein Doppelblindversuch zur Senkung des Herzinfarktrisikos erwähnt, der 1987 abgebrochen wurde. Aufgrund von Zwischenergebnissen erschien es ethisch nicht vertretbar, Patienten der Kontrollgruppe mit einem Placebo statt mit dem wirksamen Medikament zu behandeln, obwohl der Behandlungseffekt nur einer Korrelation von $r = .034$ bzw. 0.11 % erklärter Varianz entsprach (nach Westermann 2000, S. 365).

Umgekehrt kann ein großer Effekt praktisch bedeutungslos sein, etwa wenn für seine Realisierung unverhältnismäßiger und nicht leistbarer Aufwand zu betreiben wäre. Die Notwendigkeit, Studienergebnisse differenziert

Tabelle 14.4 Verschiedene Typen von Effektgrößenmaßen und verbreitete Referenzwerte zu ihrer Klassifikation in kleine, mittlere und große Effekte. (Mod. nach Cohen, 1988. Republished with permission of Taylor & Francis Ltd, from Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Permission conveyed through Copyright Clearance Center, Inc.; [a]Chen et al., 2010. Reprinted by permission of Taylor & Francis Ltd, http://www.tandf.co.uk/journals)

Vier Typen von Effektgrößenmaßen	Kleiner Effekt	Mittlerer Effekt	Großer Effekt
Gruppendifferenzmaße (2 Gruppen)			
– z. B. d Maß von Cohen	0.20	0.50	0.80
Zusammenhangsmaße			
– z. B. bivariater Korrelationskoeffizient r	.10	.30	.50
Varianzaufklärungsmaße			
– z. B. bivariater Determinationskoeffizient r^2	.01	.09	.25
– z. B. Eta-Quadrat η^2	.01	.06	.14
Risikomaße			
– z. B. Odds Ratio OR[a]	≈2.00	≈3.00	≈7.00

14

im Hinblick auf ihre theoretische und praktische Bedeutsamkeit zu interpretieren und sich nicht mit dem Hinweis auf „statistische Signifikanz“ zu begnügen, erfordert also immer in erster Linie fachliche Auseinandersetzungen und inhaltliche Argumente auf dem aktuellen Wissensstand des Forschungsfeldes.

Effektgrößenmaße und Konventionen zu ihrer Klassifikation als kleine, mittlere und große Effekte sind dementsprechend nur **pragmatische Hilfsmittel**. Sie können und dürfen die **inhaltliche Auseinandersetzung** nicht ersetzen. Ein „mechanisches“ Vorgehen, bei dem man sich etwa damit zufrieden gibt, einen numerisch „mittleren Effekt“ belegt zu haben, bleibt ebenso unbefriedigend wie der pauschale Hinweis auf „statistische Signifikanz“, sofern die Bedeutsamkeit des statistischen Befundes nicht mit Rückbezug zu Theorie und Praxis argumentativ herausgearbeitet wird – in einer unter Fachleuten intersubjektiv nachvollziehbaren Art und Weise.

Es war Jacob Cohen (1988, 1992), der eine an der empirischen Forschungspraxis orientierte Klassifikation von Effektgrößen in **kleine, mittlere und große Effekte** vorgeschlagen hat, die inzwischen als Faustregel weitgehend akzeptiert ist (Tab. 14.4), aber wie ausgeführt die inhaltliche Interpretation nicht ersetzt. Eine ungefähre Vorstellung von der Bedeutung dieser Klassifikation vermittelt Cohen (1988, S. 26f.) anhand der Effektgröße des d-Maßes (Stichprobeneffekt) bzw. von δ (unbekannter Populationseffekt) bei Gruppendifferenzen: Danach werden $\delta = 0.2$ (klein), $\delta = 0.5$ (mittel) und $\delta = 0.8$ (groß) wie folgt charakterisiert:

- kleiner Effekt: durchschnittlicher Körpergrößenunterschied bei 15- versus 16-jährigen Mädchen,
- mittlerer Effekt: durchschnittlicher Körpergrößenunterschied bei 14- versus 18-jährigen Mädchen,
- großer Effekt: durchschnittlicher Körpergrößenunterschied bei 13- versus 18-jährigen Mädchen.

Weitere Anregungen zur Interpretation von Effektgrößen findet man bei Algina, Keselman und Penfield (2005), Rosenthal (1994, S. 241ff.) sowie Fritz, Morris und Richler (2012). Eine Veranschaulichung ist ▶ Studienbeispiel „Sind Frauen gesprächiger als Männer? Beispiel einer Post-hoc-Teststärkeanalyse bei nicht-signifikantem Ergebnis“ zu entnehmen.

Es ist zu beachten, dass die **konventionellen Referenzwerte** für kleine, mittlere und große Effekte je nach Typ von Effektgrößenmaß schwanken (Tab. 14.5). Zudem gibt es in der Methodenforschung auch Kontroversen darüber, welche Referenzwerte zu nutzen sind. **Effektgrößenkonventionen auf der Basis von Cohen** (1988, 1992) sind am bekanntesten; sie werden oft genutzt – auch in diesem Lehrbuch. Dabei sind auch die Grenzen und Probleme dieses Ansatzes zu nennen: Cohens Effektgrößenkonventionen sind **teilweise inkonsistent**, d. h. bei der Umrechnung von statistisch äquivalenten Effektgrößenmaßen (z. B. f für die Varianzanalyse und f^2 für die multiple Korrelation; Tab. 14.5) ergeben sich teilweise Abweichungen. Das kann in der Praxis zu Verwirrung führen, etwa wenn eine Varianzanalyse mit G*Power regressionsanalytisch ausgewertet wird. Auch innerhalb der varianzanalytischen Effektgrößenkonventionen sind die verschiedenen Effektgrößenmaße von Cohen (1988) nicht vergleichbar, etwa beim Übergang von Designs ohne Messwiederholungen hin zu Designs mit Messwiederholungen. Im ersten Fall zählt die interindividuelle Varianz zur Fehlervarianz, im letztgenannten Fall dagegen nicht. Das hat erhebliche Auswirkungen auf die inhaltliche Bedeutung von kleinen, mittleren und großen Effekten. In Messwiederholungsdesigns bedeuten sie bei Cohen (1988) etwas anderes als in Designs ohne Messwiederholungen. Abgesehen von Inkonsistenzen zwischen den Effektgrößenmaßen werden die Cohen-Konventionen auch hinsichtlich der angesetzten Normwerte kritisch hinterfragt. Sie werden teilweise als **zu niedrig angesetzt** betrachtet: So postuliert z. B. Ferguson (2009) für das d-Maß von Cohen ein $d = 0.41$ als kleinen Effekt, $d = 1.15$ als mittleren Effekt und $d = 2.70$ als großen Effekt. Entsprechende methodische Kontroversen und Probleme sollten nicht zum Anlass genommen werden, das Thema Teststärke in hypothesenprüfenden empirischen Studien zu vernachlässigen, sondern sind

Tabelle 14.5 Effektgrößen der wichtigsten Signifikanztests (Erläuterungen s. Text; Tabelle in Anlehnung an Cohen, 1992, S. 157. Copyright © 1992 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

Test	Effektgröße	Klassifikation der Effektgrößen		
		Klein	Mittel	Groß
1. t-Test	$\delta = \frac{\mu_A - \mu_B}{\sigma}$ gilt analog für Hedges g und Glass Δ	0.20	0.50	0.80
2. Korrelationstest	ρ ρ^2	.10 .01	.30 .09	.50 .25
3. Korrelationsdifferenz	$Q = Z_A - Z_B$	0.10	0.30	0.50
4. Differenz eines Anteilswertes von .50	$G = \pi - .50$	.05	.15	.25
5. Differenz zweier Anteilswerte	$H = \varphi_A - \varphi_B$	0.20	0.50	0.80
6. χ^2-Test	$W = \sqrt{\sum_{i=1}^{k} \frac{(\pi_{0i} - \pi_{1i})^2}{\pi_{0i}}}$ gilt analog für φ, V, C	.10	.30	.50
7. Varianzanalysen	$E = f = \frac{\sigma_\mu}{\sigma}$ $\eta^2 = \frac{QS_{Treatment}}{QS_{Total}}$ $\hat{\eta}_p^2 = \frac{QS_{Effekt}}{QS_{Effekt} + QS_{Prüf}}$	.10 .01	.25 .06	.40 .14
8. Multiple Korrelation	R^2 $K^2 = f^2 = \frac{R^2}{1 - R^2}$	.02 .02	.13 .15	.26 .35

eher Ausdruck der Notwendigkeit, sich diesem Themenfeld in der Methodenlehre und Methodenforschung noch intensiver zu widmen, um Teststärkeprobleme sachgerecht behandeln zu können.

Zusammenfassend bleibt festzuhalten, dass konventionelle Referenzwerte zur Einstufung von Effektgrößen nur grobe Anhaltspunkte liefern und z. B. im Kontext der A-priori-Teststärkeanalyse nützlich sind, um die Größenordnung der interessierenden Effektgröße festzulegen (sofern keine genaueren Informationen zur Spezifizierung der H_1 vorliegen). Zur Interpretation von empirischen Befunden sind als Bezugsgrößen **die in früheren, inhaltlich vergleichbaren Studien aufgedeckten Effektgrößen** meist aussagekräftiger. Ihnen ist zu entnehmen, ob die eigene Studie frühere Untersuchungen hinsichtlich der Effektgröße bekräftigt (repliziert) oder auf kleinere bzw. größere Effekte hinweist. In dieser relativen Betrachtung ist dann ein Effekt z. B. als „groß" einzustufen, wenn er in vergleichbaren früheren Studien aufgedeckte Effekte übertrifft.

Neben dem Konzept der **statistischen Signifikanz** und der – im Kontext der Interpretation von Effektgrößen relevanten – **theoretischen und/oder praktischen Signifikanz** (Bedeutsamkeit) wird zudem von **klinischer Signifikanz** gesprochen. Die klinische Bedeutsamkeit bewertet das Ausmaß, in dem eine Intervention (z. B. Therapiemaßnahme) dazu führt, dass diagnostizierte Fälle (z. B. Patientinnen und Patienten mit diagnostizierter Depression oder Aufmerksamkeitsstörung) nach der Maßnahme nicht länger die Diagnosekriterien erfüllen und somit keine weitere Behandlung benötigen (Thompson, 2002).

14.2.3 Standardisierte Effektgrößen der wichtigsten Signifikanztests und deren Konfidenzintervalle

In diesem Abschnitt wollen wir detailliert die Bestimmung von standardisierten Effektgrößenmaßen behandeln (Cohen, 1988, 1992; Kline, 2004). Dabei geht es im Kontext inferenzstatistischer Hypothesentests immer darum, die unbekannte Populationseffektgröße anhand der berechneten Stichprobeneffektgröße zu schätzen, und zwar als Punktschätzung sowie als Intervallschätzung mittels Konfidenzintervall.

Die Notation für Effektgrößenmaße ist uneinheitlich. Abweichend von der Cohen-Notation verwenden wir in Tab. 14.5 entweder **Großbuchstaben** oder **griechische Buchstaben**. Damit soll zum Ausdruck gebracht werden, dass es sich bei den Effektgrößen um Annahmen

für eine spezifische H_1, also um **Populationsparameter** handelt. Für eine empirisch anhand von Stichprobendaten ermittelte Effektgröße verwenden wir wie bei Stichprobenkennwerten generell **Kleinbuchstaben** oder – bei griechischen Buchstaben – ein ^ (für „geschätzt"). Das heißt, das ***d*-Maß von Cohen ist eigentlich als δ-Maß zu bezeichnen**, weil es zur Schätzung der standardisierten Populationsmittelwertdifferenz dient und wir Populationsparameter traditionell mit griechischen Buchstaben benennen.

Wir weisen darauf hin, dass die Verwendung statistischer Symbole in der Fachliteratur – und auch in diesem Buch – nicht vollkommen konsistent ist bzw. dass eine Reihe von synonymen Symbolen genutzt werden (z. B. ***d*-Maß = δ-Maß von Cohen**), deren Bedeutung jedoch aus dem Kontext erschlossen werden kann. Wir empfehlen im Zweifelsfall, sich bei der **Definition und Verwendung statistischer Symbole** auch an der aktuellen Auflage des APA Publication Manual zu orientieren (APA, 2009, S. 119ff.).

Eine Übersicht der standardisierten Effektgrößenmaße für wichtige Signifikanztests gibt ◘ Tab. 14.5. Nicht aufgeführt sind Effektgrößen für **mehrfaktorielle Varianzanalysen** (Haupteffekte, Interaktionen), die in ▶ Abschn. 14.2.3 „Mehrfaktorielle Varianzanalysen" gesondert behandelt werden. Die Tabelle ist folgendermaßen zu lesen: Differenzen zweier Mittelwerte aus unabhängigen Stichproben werden mit dem ***t*-Test für unabhängige Stichproben** auf Signifikanz geprüft. Die mit diesem Signifikanztest verbundene standardisierte Effektgröße heißt δ (bzw. d-Maß von Cohen) und berechnet sich über die Formel $\delta = \frac{\mu_A - \mu_B}{\sigma}$. Die in die Formel eingehenden Populationsparameter werden mit den entsprechenden Stichprobenkennwerten geschätzt. Zur Klassifikation von empirisch gefundenen oder theoretisch erwarteten δ-Maßen sind die Referenzwerte für kleine, mittlere und große Effekte angegeben (zu Problemen und Grenzen der Effektgrößenklassifikation ▶ Abschn. 14.2.2).

Im Folgenden wird erläutert, wie die Effektgrößen zu berechnen sind. Diese Erläuterungen kann ohne weiteres überspringen, wer nur daran interessiert ist, den **optimalen Stichprobenumfang** für eine eigene geplante Untersuchung herauszufinden. In diesem Falle ist lediglich anhand der Forschungsliteratur festzulegen, ob man für die geprüfte Maßnahme einen **kleinen, mittleren oder großen Effekt** erwartet (◘ Tab. 14.5); für diese Effektgröße ist dann aus ◘ Tab. 14.11 der optimale Stichprobenumfang für den gewählten Signifikanztest zu entnehmen. Wichtig ist jedoch der Hinweis, dass auf eine Schätzung der Effektgröße aufgrund der Untersuchungsergebnisse im Falle eines signifikanten Ergebnisses nicht verzichtet werden sollte (sog. **Ex-post-Bestimmung von Effektgrößen**). In diesem Falle ist es empfehlenswert, die Schätzgenauigkeit der Populationseffektgröße mit einem Konfidenzintervall für die Effektgröße anzugeben.

In den folgenden Ausführungen werden wir – im Vorgriff auf ▶ Kap. 16 – auch auf die Frage eingehen, welche Effektgrößen für **Metaanalysen** gut bzw. weniger gut geeignet sind. Diese Frage zu erörtern, ist immer dann erforderlich, wenn „traditionelle" Effektgrößenmaße für Metaanalysen Probleme bereiten. Untersuchungsbeispiele findet man in ▶ Abschn. 14.4 und eine Beschreibung der in ◘ Tab. 14.5 angesprochenen Signifikanztests z. B. bei Bortz und Schuster (2010).

Vergleich von zwei Mittelwerten

t-Test für unabhängige Stichproben

Die Bestimmung der Effektgröße δ bzw. des ***d*-Maßes von Cohen** setzt voraus, dass man eine Vorstellung darüber hat, wie stark sich 2 Populationen A und B (z. B. unter Experimental- und Kontrollbedingungen) angesichts der Merkmalsstreuung σ mindestens unterscheiden müssen, um von einem praktisch bedeutsamen Effekt sprechen zu können. Diese Schätzungen erübrigen sich, wenn man der Literatur entnehmen kann, welche Effekte im fraglichen Untersuchungsgebiet typischerweise erzielt werden. Will man die Effektgröße ex post, also nach Abschluss der Untersuchung, mit $\hat{\delta}$ schätzen, verwendet man $\overline{x}_A$ und $\overline{x}_B$ als Schätzwerte für μ_A und μ_B und die Streuung des Merkmals in den Stichproben als Schätzung für σ (zur Zusammenfassung von Streuungen Gl. 14.3).

$$\hat{\delta} = \frac{\overline{x}_A - \overline{x}_B}{\hat{\sigma}} . \qquad (14.1)$$

Bei kleineren Stichproben ($n_A = n_B = n < 20$) überschätzt $\hat{\delta}$ den Parameter δ geringfügig. Dieser Bias lässt sich nach **Hedges** (1982, zit. nach Kline 2004, Gl. 4.8) wie folgt korrigieren:

$$\hat{\delta}_{corr} = \left(1 - \frac{3}{4 \cdot (n_A + n_B - 2) - 1}\right) \cdot \hat{\delta} . \qquad (14.2)$$

Für die Festlegung einer Effektgröße δ **in der Planungsphase** lassen sich vergleichbaren Untersuchungen oftmals brauchbare Schätzwerte für σ entnehmen. Stehen entsprechende Angaben nicht zur Verfügung, stellt der **Range**, der einfacher zu schätzen ist als die Standardabweichung, eine geeignete Hilfsgröße dar (▶ Abschn. 12.4.2, Punktschätzung von Populationsstreuungen auf der Basis des Ranges). Bei normalverteilten Merkmalen ist die Differenz zwischen dem mutmaßlich größten Wert in der Population und dem mutmaßlichen kleinsten Wert zu bilden und durch 5.15 zu dividieren (genauer hierzu ▶ Abschn. 12.4.2). Es ist darauf zu achten, dass sich die Schätzung des Range auf die Messungen innerhalb der zu vergleichenden Populationen bezieht und nicht auf die beiden

zusammengefassten Populationen, denn die Streuung der zusammengefassten Populationen enthält – falls die H_1 gilt – auch Unterschiede zwischen den Populationen (s. hierzu jedoch Olejnik & Algina, 2003). Wenn bekannt oder damit zu rechnen ist, dass die Streuungen in den Populationen unterschiedlich sind, müssen zwei getrennte Streuungsschätzungen vorgenommen werden, die dann gemäß **Hedges *g*** gepoolt (d. h. zusammengefasst) werden. Die Zusammenfassung dieser Schätzungen erfolgt nach Gl. 14.3):

$$\hat{\sigma} = \sqrt{\frac{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}{2}} \quad (n_A = n_B) \ . \tag{14.3}$$

Sind die Stichprobenumfänge nicht gleich groß, ermittelt man $\hat{\sigma}$ nach Gl. 14.4:

$$\hat{\sigma} = \sqrt{\frac{(n_A - 1) \cdot \hat{\sigma}_A^2 + (n_B - 1) \cdot \hat{\sigma}_B^2}{(n_A - 1) + (n_B - 1)}} \ . \tag{14.4}$$

Diese beiden Gleichungen setzen homogene Varianzen voraus.

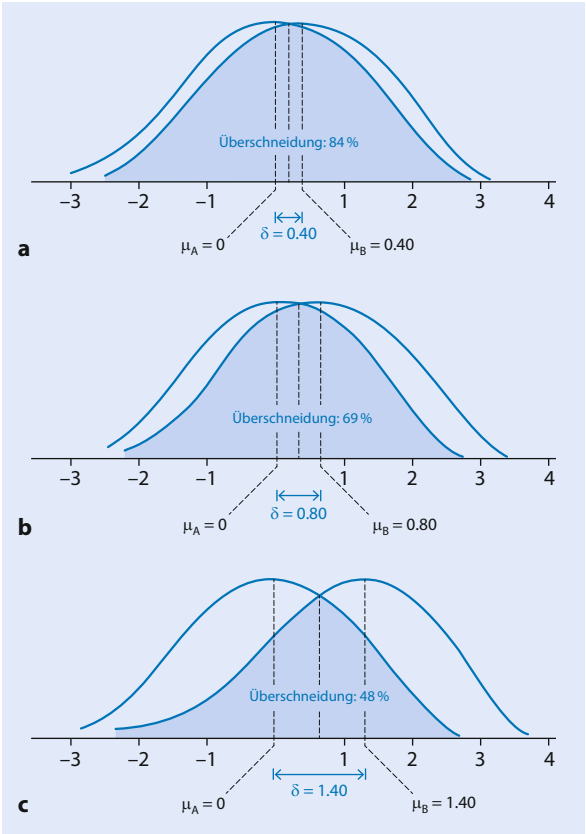
Kline (2004, S. 104) empfiehlt für heterogene Varianzen ($\hat{\sigma}_A^2/\hat{\sigma}_B^2 > 4$ mit $\hat{\sigma}_A^2 > \hat{\sigma}_B^2$) δ durch den **Δ-Koeffizienten (Δ: großes Delta) von Glass (1976)** zu ersetzen, bei dem $\mu_A - \mu_B$ nicht an der zusammengefassten Standardabweichung relativiert wird, sondern – z. B. beim Vergleich einer Kontrollgruppe mit einer Experimentalgruppe – lediglich an der **Standardabweichung der Kontrollgruppe**. Sind die Standardabweichungen in den zu vergleichenden Stichproben deutlich verschieden (Regel oben), sollte der Δ-Koeffizient sowohl für $\hat{\sigma}_A$ als auch für $\hat{\sigma}_B$ bestimmt und berichtet werden. Typischerweise tritt dieser Fall ein, wenn ein Treatment nicht nur die zentrale Tendenz, sondern auch die Variabilität in der Experimentalgruppe verändert (mit der Frage der Bestimmung von $\hat{\delta}$ bei heterogenen Varianzen befasst sich eine Übersichtsarbeit von Grissom & Kim, 2011, S. 29).

Nach Durchführung eines *t*-Tests errechnet man $\hat{\delta}$ wie folgt aus dem *t*-Wert (Westermann, 2000, S. 357):

$$\hat{\delta} = t \cdot \sqrt{\frac{n_A + n_B}{n_A \cdot n_B}} \ . \tag{14.5}$$

Das δ-Maß lässt sich durch den **Überschneidungsbereich** der beiden zu vergleichenden Verteilungen veranschaulichen (◘ Abb. 14.3).

Als Überschneidungsbereich zweier normalverteilter Verteilungen (mit $\sigma = 1$) definieren wir denjenigen Bereich, in dem sich sowohl Elemente der einen als auch der anderen Verteilung befinden. Die ◘ Abb. 14.3a zeigt, dass einem $\delta = 0.40$ ein Überschneidungsbereich von 84 % entspricht. In den Teilabbildungen b und c sind $\delta = 0.80$ mit einer Überschneidung von 69 % und



◘ **Abb. 14.3** Überschneidungsbereich und Effektgröße δ beim *t*-Test. **a** Überschneidung: 84 %, **b** Überschneidung: 69 %, **c** Überschneidung: 48 %

$\delta = 1.40$ mit einer Überschneidung von 48 % dargestellt. Allgemein lässt sich ein Überschneidungsbereich einfach anhand der Standardnormalverteilungstabelle (Tab. 1 im ▶ Webangebot zu diesem Buch) ermitteln: Wir lesen diejenige Fläche ab, den der Wert $\delta/2$ von der Standardnormalverteilungsfläche abschneidet und verdoppeln diese Fläche. Es resultiert der Überschneidungsbereich.

Beispiel: Für $\delta = 0.40$ mit $\delta/2 = 0.20$ schneidet der Wert $z = -0.2$ von der linken Seite der Standardnormalverteilung 42 % ab. Verdopplung führt zu 84 % – die Überschneidung, die in ◘ Abb. 14.3a dargestellt ist.

Für die konventionelle Effektgrößenklassifikation des *d*-Maßes gilt (◘ Tab. 14.4): **92 % Überschneidung für einen kleinen, 80 % für einen mittleren und 68 % für einen großen Effekt**.

Konfidenzintervalle In ▶ Abschn. 12.4.2 wurde das Konfidenzintervall des arithmetischen Mittelwertes eingeführt. Das Konfidenzintervall als Verfahren der Intervallschätzung gibt einen Wertebereich an, in dem sich ein gesuchter Populationsparameter mit hoher Wahrscheinlichkeit befindet. Typischerweise werden auf der Basis des kon-

ventionellen α-Niveaus von 5 % Konfidenzintervalle mit 95 %iger Konfidenz berechnet.

Das Konfindenzintervall-Konzept ist nun auf die Effektgrößenbestimmung zu übertragen. Mit $\hat{\delta} = (\overline{x}_A - \overline{x}_B)/\hat{\sigma}$ berechnen wir einen Stichprobenkennwert, der als Punktschätzer für den unbekannten Populationsparameter δ dient. Für die Bestimmung des Konfindenzintervalls von δ benötigt man die sog. **nichtzentrale *t*-Verteilung** (Cumming & Finch, 2001), die durch die Anzahl der Freiheitsgrade des *t*-Tests definiert ist sowie durch einen sog. **Nonzentralitätsparameter (NZ)**. Während die zentrale *t*-Verteilung um den Erwartungswert 0 streut, **streut die nichtzentrale *t*-Verteilung um die Effektgröße**. Den Nonzentralitätsparameter der nichtzentralen *t*-Verteilung – bezeichnet als NZ*t* – berechnet man wie folgt:

$$\mathrm{NZ}t = \delta \cdot \sqrt{\frac{n_A \cdot n_B}{n_A + n_B}} = t^* . \qquad (14.6)$$

Der Nonzentralitätsparameter NZ*t* entspricht – Varianzhomogenität vorausgesetzt – dem Parameter t^*, den man erhält, wenn man die *t*-Formel des *t*-Tests auf die zugrunde liegenden Populationsparameterwerte (Populationsmittelwert μ und Populationsstreuung σ) anwendet. Dieser Parameter t^* wird durch den auf der Basis der Stichprobenkennwerte berechneten empirischen *t*-Wert geschätzt. Da NZ*t* ein Populationsparameter, der empirische *t*-Wert aber ein Stichprobenkennwert ist, können beide formal nicht identisch sein bzw. nicht direkt gleichgesetzt werden.

Die Konfidenzintervallbestimmung erfolgt in 2 Schritten auf der Basis des sog. „**Confidence Interval Transformation Principle**“ (Steiger, 2004):

1. Man berechnet diejenigen *t*-Werte, die von der nichtzentralen *t*-Verteilung die unteren bzw. die oberen $\alpha/2 = 2.5\,\%$ der Fläche abschneiden (oder $\alpha/2 = 0.5\,\%$ für das 99 %ige Konfidenzintervall). Hierfür kann z. B. die im ▶ Anhang zu diesem Buch wiedergegebene SAS-Syntax verwendet werden (für andere Berechnungstools ▶ Hinweise zur Durchführung von Teststärkeanalysen). Genauer formuliert: Man bestimmt die **Nichtzentralitätsparameter nichtzentraler *t*-Verteilungen NZ*t***, von denen der empirische *t*-Wert entweder die oberen $\alpha/2\,\%$ (**NZt_o**) oder die unteren $\alpha/2\,\%$ (**NZt_u**) abschneidet. Diese Parameter sind die Grenzen, die gemäß Schritt 2 zu transformieren sind.
2. Diese Grenzwerte werden nach folgender Gleichung in δ-Einheiten transformiert:

$$\delta = \mathrm{NZ}t \cdot \sqrt{\frac{n_A + n_B}{n_A \cdot n_B}} . \qquad (14.7)$$

Es resultieren die untere (mit **NZt_u**) und die obere Grenze (mit **NZt_o**) des Konfidenzintervalls für δ.

Beispiel (nach Kline, 2004, S. 112): Ein *t*-Test über zwei unabhängige Stichproben (mit $n_A = n_B = 30$ bzw. df = 58) führte zu $t = 3.10$. Über Gl. 14.7 schätzen wir für NZ$t = t$ den Wert $\hat{\delta} = 0.80$. Mit $t = 3.10$, df = 58, $n_1 = 30$ und $n_2 = 30$ lassen sich als untere Grenze der Nichtzentralitätsparameter $\mathrm{NZ}t_u = 1.04844$ und als obere Grenze $\mathrm{NZ}t_o = 5.12684$ bestimmen für das 95 %ige Konfidenzintervall. $t = 3.10$ ist also das 97.5. Perzentil einer nichtzentralen *t*-Verteilung mit df = 58 und Nonzentralitätsparameter 1.04844 sowie das 2.5. Perzentil einer nichtzentralen *t*-Verteilung mit df = 58 und Nonzentralitätsparameter 5.12684. Diese Nichtzentralitätsparameter-Grenzwerte sind nun gem. Gl. 14.7 in Grenzwerte der δ-Skala zu überführen: $\delta_u = 0.27$ und $\delta_o = 1.32$. Das Konfidenzintervall heißt also: $0.27 \leq \delta \leq 1.32$.

Mit einer Konfidenz von 95 % können wir sagen, dass der „wahre“ δ-Wert bei einem $\hat{\delta} = d = 0.80$ im Bereich von $\delta_u = 0.27$ bis $\delta_o = 1.32$ liegt. Das entspricht nach gängiger Effektgrößenklassifikation (Tab. 14.4) einer kleinen bis sehr großen Populationseffektgröße. Mit dem **Wilson-Online-Rechner** (http://www.campbellcollaboration.org/resources/effect_size_input.php; ▶ Hinweise zur Durchführung von Teststärkeanalysen) bestimmen wir für eine Gesamtstichprobengröße von $n = 60$ (bei gleichgroßen Stichproben) und einem empirischen *t*-Wert von $t = 3.10$ eine Effektgröße von $d = 0.80$ und ein 95 %iges Konfidenzintervall von $0.27 \leq d \leq 1.33$.

▪ *t*-Test für abhängige Stichproben

Der *t*-Test für abhängige Stichproben überprüft die H_0, dass sich die Mittelwerte μ_1 und μ_2 einer zum Zeitpunkt t_1 und t_2 gemessenen abhängigen Variablen in einer Population nicht unterscheiden bzw. dass der Mittelwert der Einzeldifferenzen $\mu_D = 0$ ist. Der typische Anwendungsfall ist also gegeben, wenn eine Stichprobe zweimal untersucht wird und entschieden werden soll, ob sich der Stichprobenmittelwert $\overline{x}$ signifikant verändert hat. Dieses Verfahren kommt auch zum Einsatz, wenn „Matched Samples“ zu vergleichen sind (▶ Abschn. 12.5.4).

Die Effektgrößenklassifikation für den *t*-Test mit unabhängigen Stichproben gilt auch für den *t*-Test mit abhängigen Stichproben. Eine Differenz $\mu_1 - \mu_2$, die der halben Merkmalsstreuung entspricht, wird als mittlerer Effekt klassifiziert: $\mu_1 - \mu_2/\sigma = 0.50$. Wie wir in ▶ Abschn. 14.3.2 noch sehen werden, reicht für die Absicherung eines bestimmten Effektes beim *t*-Test für abhängige Stichproben in der Regel eine kleinere Stichprobe aus als beim *t*-Test für unabhängige Stichproben.

Für die **Effektgrößenbestimmung** beim *t*-Test für abhängige Stichproben werden in der Literatur 2 Varianten

diskutiert (Kline, 2004, S. 104ff.), die hier als Varianten a und b vorgestellt werden.

- Variante a:
  Die Differenz der Mittelwerte $\mu_1 - \mu_2 = \mu_D$ wird an der **Streuung der Differenzen σ_D** standardisiert (dieses Effektgrößenmaß wird zuweilen auch **Standardized Response Mean *SRM*** genannt):

$$\delta' = \frac{\mu_D}{\sigma_D} \tag{14.8}$$

  mit δ' = Effektgröße für zwei abhängige Stichproben. Kennt man die Größenordnung der Korrelation ρ zwischen den beiden Messwertreihen, kann δ' wie folgt bestimmt werden:

$$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{14.9}$$

  bzw. bei gleichen Varianzen ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$\sigma_D = \sqrt{2\sigma^2 - 2\rho\sigma^2} = \sigma \cdot \sqrt{2 \cdot (1 - \rho)}\,. \tag{14.10}$$

  Eingesetzt in Gl. 14.8 resultiert:

$$\delta' = \frac{\mu_D}{\sigma \cdot \sqrt{2 \cdot (1 - \rho)}}\,. \tag{14.11}$$

  δ' entspricht δ für $\rho = .50$. Höhere Korrelationen führen zu einer „Aufwertung" von δ'.
- Variante b:
  Die Differenz der Mittelwerte μ_D wird nicht an der Streuung der Differenzen, sondern an der **Merkmalsstreuung** σ standardisiert. Man behandelt also die abhängigen Stichproben wie unabhängige Stichproben und schätzt δ' über $\hat{\delta}$ analog zu Gl. 14.1. Variante b ist v. a. für **Metaanalysen** (▶ Kap. 16) vorteilhaft, wenn für eine Fragestellung Untersuchungen mit abhängigen, aber auch mit unabhängigen Stichproben vorliegen. Variante b sichert also die Vergleichbarkeit der entsprechenden Studien. Ein weiteres Argument, das für Variante b spricht, ist die mangelnde Anschaulichkeit von σ_D und auch die geringe Stabilität von σ_D über verschiedene vergleichbare Studien hinweg. σ auf der anderen Seite ist häufig eine bekannte Größe, mit der sich anschaulich operieren lässt. Wenn z. B. eine Unterrichtsmethode evaluiert werden soll und als abhängige Variable ein Schulleistungstest mit $\sigma = 10$ eingesetzt wird, ist es unmittelbar einleuchtend, dass ein $\hat{\delta}$ von 0.60 eine Verbesserung der Schulleistung um $\overline{x}_{pre} - \overline{x}_{post} = 6$ Testpunkte bedeutet. Eine Standardisierung dieser Differenz an $\hat{\sigma}_D$ hätte diese Anschaulichkeit nicht.
  Falls die Merkmalsstreuung unbekannt ist, sollte man die Mittelwertdifferenz an der Streuung der Pretestwerte ($\hat{\sigma}_{pre}$) standardisieren, insbesondere wenn damit zu rechnen ist, dass das Treatment neben der zentralen Tendenz auch die Variabilität verändert. Andernfalls wäre auch eine Zusammenfassung von $\hat{\sigma}_{pre}$ und $\hat{\sigma}_{post}$ über Gl. 14.3 möglich (jedoch hierzu auch die Ausführungen in ▶ Abschn. 14.2.3).

Konfidenzintervalle Auch bei der Konfidenzintervallbestimmung sind die Varianten a und b zu unterscheiden.

- Variante a:
  Wie oben ausgeführt wird bei dieser Variante μ_D an der Streuung der Differenzen σ_D standardisiert. Die Konfidenzintervallbestimmung für δ' ähnelt der für δ: Es werden zunächst die Nichtzentralitätsparameter derjenigen nichtzentralen t-Verteilungen mit $df = n - 1$ bestimmt, von denen der empirische t-Wert (d. h. der Wert des t-Tests für abhängige Stichproben) die oberen $\alpha/2\,\%$ (NZt_o) bzw. die unteren $\alpha/2\,\%$ abschneidet (NZt_u). Diese Grenzwerte werden sodann über folgende Gleichung in die Grenzwerte δ'_u und δ'_o transformiert (die SAS-Syntax zur Bestimmung von δ'_u und δ'_o ist im ▶ Anhang zu diesem Buch zu finden):

$$\delta' = \frac{NZt}{\sqrt{n}}\,. \tag{14.12}$$

  Beispiel: Angenommen, eine Pretest-Posttest-Untersuchung mit $n = 30$ führte zu $\overline{x}_{pre} = 36$, $\overline{x}_{post} = 20$ und $\hat{\sigma}_D = 20$. Daraus ergibt sich über Gl. 14.8 folgende Effektgrößenschätzung:

$$\hat{\delta}' = \frac{36 - 20}{20} = 0.80$$

  und ein t-Wert von (Bortz & Schuster, 2010, Gl. 8.1)

$$t = \frac{16}{20} \cdot \sqrt{30} = 4.38\,.$$

  Mit $t = 4.38$, $df = 30 - 1 = 29$ und $n = 30$ als Eingangsparameter ergeben sich $NZt_u = 2.09$, $NZt_o = 6.61$, $\delta'_u = 0.38$, $\delta'_o = 1.21$. Das 95 %ige Konfidenzintervall für $\hat{\delta}' = 0.80$ lautet also $0.38 \leq \hat{\delta}' \leq 1.21$ bzw. $\hat{\delta}' = 0.80 \pm 0.415$.
- Variante b:
  Bei der Variante b erfolgt die Standardisierung von μ_D nicht über σ_D, sondern über σ, die Merkmalsstreuung. Für diese Variante kann lediglich ein **approximatives Konfidenzintervall** bestimmt werden (Cumming & Finch, 2001, zit. nach Kline, 2004, S. 113). Hierfür benötigt man folgenden asymptotischen Standardfehler der standardisierten Mittelwertedifferenz (Kline, 2004, S. 108):

$$\hat{\sigma}_{\hat{\delta}} = \sqrt{\frac{\hat{\delta}^2}{2 \cdot (n - 1)} + \frac{2 \cdot (1 - r)}{n}}\,. \tag{14.13}$$

Mit $\hat{\delta} = \frac{\bar{x}_D}{\hat{\sigma}}$ erhält man das Konfidenzintervall (KI_δ) über

$$KI_\delta = \hat{\delta} \pm \hat{\sigma}_{\hat{\delta}} \cdot z_{(\frac{\alpha}{2})} \qquad (14.14)$$

mit $z_{(\alpha/2)} = 1.96$ für $\alpha = .05$ und $z_{(\alpha/2)} = 2.58$ für $\alpha = .01$.

Beispiel (nach Kline, 2004, S. 109): Eine Pretest-/Posttest-Untersuchung mit $n = 30$ hat $\hat{\delta} = 0.80$ und $r = .75$ ergeben. Damit erhält man als Standardfehler

$$\hat{\sigma}_{\hat{\delta}} = \sqrt{\frac{0.8^2}{2 \cdot (30-1)} + \frac{2 \cdot (1-0.75)}{30}} = 0.1664 .$$

Es resultiert also nach Gl. 14.14 folgendes 95 %ige Konfidenzintervall für $\hat{\delta} = 0.80$:

$$KI_\delta = 0.8 \pm 0.1664 \cdot 1.96 = 0.8 \pm 0.33 \quad \text{bzw.}$$
$$0.47 \leq \hat{\delta} \leq 1.13 .$$

Weitere Hinweise zur Effektgrößenbestimmung bei abhängigen Stichproben findet man bei Dunlop, Cortina, Vaslow und Burke (1996). So kann bei einer Vorher-Nachher-Messung die Mittelwertdifferenz z. B. auch mithilfe der Streuung zum Vorher-Messzeitpunkt standardisiert werden.

Bivariate Korrelation

Der Korrelationstest überprüft die Signifikanz einer Produkt-Moment-Korrelation. Die Effektgröße dieses Signifikanztests ist direkt der Korrelationskoeffizient r (bzw. genauer die Populationskorrelation ρ, die durch r geschätzt wird). Zur Veranschaulichung von r wird häufig der Determinationskoeffizient r^2 herangezogen, der dem Anteil gemeinsamer Varianz bzw. – bei Kausalmodellen mit X als Prädiktor- und Y als Kriteriumsvariablen – dem durch X erklärten Varianzanteil von Y entspricht.

Konfidenzintervalle Für die Berechnung von Konfidenzintervallen wird zunächst r in einen **Fisher-Z-Wert** transformiert. Dies geschieht am einfachsten unter Zuhilfenahme von Tab. 9 im ▶ Webangebot zu diesem Buch. Als nächstes wird der Standardfehler von Z berechnet (z. B. Bortz, 2005, Gl. 6.89):

$$\sigma_Z = \sqrt{\frac{1}{n-3}} . \qquad (14.15)$$

Bivariate Normalverteilung (oder große Stichproben) vorausgesetzt, unterliegt Z einer Normalverteilung mit einer Streuung von σ_Z, d. h., man ermittelt das Konfidenzintervall von Z über

$$KI_Z = Z \pm z_{(\alpha/2)} \cdot \sigma_Z \qquad (14.16)$$

mit $z_{(\alpha/2)} = 1.96$ für das 95 %ige Konfidenzintervall bzw. 2.58 für das 99 %ige Konfidenzintervall (Tab. 1; ▶ Webangebot zu diesem Buch). Schließlich werden die Z-Wertegrenzen von KI_Z über Tab. 9 (▶ Webangebot zu diesem Buch) in ρ-Werte transformiert. (Bei den Grenzwerten über Tab. 9 handelt es sich um gerundete Werte. Genauere Werte erhält man über Gl. 10.7 und 10.8 bei Bortz & Schuster, 2010.)

Beispiel: Eine Untersuchung mit $n = 57$ hat zu $r = .72$ geführt; Tab. 9 (▶ Webangebot zu diesem Buch) entnehmen wir $Z(r = .72) = 0.908$. Als Standardfehler ergibt sich

$$\sigma_Z = \sqrt{\frac{1}{57-3}} = 0.136$$

d. h., wir errechnen über Gl. 14.16 für das 95 %ige Konfidenzintervall

$$KI_Z = 0.908 \pm 1.96 \cdot 0.136 = 0.908 \pm 0.267 .$$

Das Konfidenzintervall von Z hat also die Grenzen 0.641 und 1.175. Wir transformieren diese Werte über Tab. 9 (▶ Webangebot zu diesem Buch) in Grenzen des Konfidenzintervalls für ρ: .57 und .83. Der durch $r = .72$ geschätzte Populationsparameter ρ hat also ein 95 %iges Konfidenzintervall von

$$.57 \leq \rho \leq .83 .$$

Dasselbe Ergebnis lässt sich mit dem **Wilson-Online-Rechner** (▶ Hinweise zur Durchführung von Teststärkeanalysen) bestimmen, wenn man bei Effect Size Type: „Correlation Coefficient“ und dann „Correlation and Sample Size“ anwählt und $r = .72$ sowie $n = 57$ eingibt.

Korrelationsdifferenz

Dieser Test überprüft, ob sich die für eine Stichprobe A ermittelte Korrelation zweier Variablen von der entsprechenden Korrelation in einer Stichprobe B signifikant unterscheidet. Zur Schätzung der Effektgröße Q werden die Korrelationen zunächst in sog. **Fisher-Z-Werte** transformiert. Diese Transformation wird einfachheitshalber anhand von Tab. 9 (▶ Webangebot zu diesem Buch) vorgenommen.

Konfidenzintervalle Die Differenz der Fisher-Z-Werte ($q = Z_A - Z_B$) zweier Korrelationen r_A und r_B ist asymptotisch normalverteilt mit einem Standardfehler von

$$\sigma_q = \sqrt{\frac{1}{n_A - 3} + \frac{1}{n_B - 3}} , \qquad (14.17)$$

wobei n_A und n_B die Stichprobenumfänge der beiden unabhängigen Stichproben kennzeichnen, für die r_A und r_B

berechnet wurden (Hays, 1994, S. 650f.). Das Konfidenzintervall von Q erhalten wir über

$$\mathrm{KI}_Q = q \pm z_{(\alpha/2)} \cdot \sigma_q \,. \tag{14.18}$$

Über Tab. 9 (▶ Webangebot zu diesem Buch) werden die Grenzen des Konfidenzintervalls für Q in Grenzen für Korrelationsdifferenzen transformiert. (Zur Erläuterung von $z_{(\alpha/2)}$ Gl. 14.14.)

Beispiel: Für eine Stichprobe A ($n_A = 48$) wurde $r_A = .52$ ermittelt und für eine Stichprobe B ($n_B = 69$) $r_B = .45$. Hierfür entnehmen wir Tab. 9 (▶ Webangebot zu diesem Buch) $Z_A(r_A = .52) = 0.576$ und $Z_B(r_B = .45) = 0.485$. Man errechnet also als Effektgrößenschätzung $q = 0.576 - 0.485 = 0.091$. Als Standardfehler ergibt sich

$$\sigma_q = \sqrt{\frac{1}{48-3} + \frac{1}{69-3}} = 0.1933$$

und damit

$$\mathrm{KI}_Q = 0.091 \pm 1.96 \cdot 0.1933 \,.$$

Die Grenzen dieses 95 %igen Konfidenzintervalls (−0.288 bis 0.470) werden über Tab. 9 (▶ Webangebot zu diesem Buch) **in Korrelationseinheiten transformiert**. (Da die Z- und die r-Werte symmetrisch um Null verteilt sind, gelten die Transformationen positiver Z/r-Werte analog für negative Z/r-Werte):

$$-.28 \leq \rho_A - \rho_B \leq .44 \,.$$

Differenz eines Anteilswertes von .50

Dieser Test wird bei kleineren Stichproben über die sog. Binomialverteilung und bei größeren Stichproben über die Standardnormalverteilung durchgeführt. Er findet z. B. Anwendung, wenn man erwartet, dass eine Maßnahme überwiegend positive Veränderungen bewirkt und die Nullhypothese behauptet, dass positive und negative Veränderungen zufällig auftreten bzw. gleich wahrscheinlich sind ($\pi = .50$). Die Effektgröße G wird hier über die Abweichung des Anteilswertes π von .50 geschätzt. Diese Effektgröße könnte z. B. beim McNemar-χ^2-Test oder beim Vorzeichentest eingesetzt werden (Bortz & Lienert, 2003, Abschn. 2.5.1 und 3.3.1).

Konfidenzintervalle Das Konfidenzintervall für π (KI_π) enthält mit einer Konfidenz von 95 % (99 %) den wahren π-Parameter in der Population. Liegt $\pi = .50$ in diesem Intervall, so weicht P nicht signifikant von π ab. Andernfalls, wenn KI_π den Wert .50 nicht umschließt, ist die Abweichung des Anteilwertes P von .50 statistisch bedeutsam.

Nach Kline (2004, Tab. 5.3) berechnen wir das Konfidenzintervall über die **Normalverteilungsapproximation** mit einem Standardfehler σ_p von:

$$\sigma_p = \sqrt{\frac{P \cdot (1-P)}{n}} \,. \tag{14.19}$$

Die Normalverteilungsapproximation wird als ausreichend angesehen, wenn $n \times P \times (1-P) \geq 9$ ist (Sachs, 2002, S. 228). Das Konfidenzintervall erhält man wie üblich über

$$\mathrm{KI}_\pi = P \pm z_{(\frac{\alpha}{2})} \cdot \sigma_P \,. \tag{14.20}$$

(Zur Erläuterung von $z_{(\alpha/2)}$ Gl. 14.14.) Ein genaueres Konfidenzintervall wird bei Hays (1994, S. 259) beschrieben.

Beispiel: Eine Untersuchung mit $n = 100$ hat zu $P = .6$ geführt (z. B. 60-mal Zahl bei 100 Münzwürfen). Für das Konfidenzintervall errechnen wir zunächst

$$\sigma_P = \sqrt{\frac{0.60 \cdot (1-0.60)}{100}} = 0.049$$

und damit über Gl. 14.20

$$\mathrm{KI}_\pi = 0.60 \pm 1.96 \cdot 0.049 = 0.60 \pm 0.096 \,.$$

Das 95 %ige Konfidenzintervall hat also die Grenzen

$$0.504 \leq \pi \leq 0.696 \,.$$

Das Konfidenzintervall umschließt nicht den Parameter $\pi = .50$, d. h., die Abweichung $P = .60$ von $\pi = .50$ ist bei zweiseitigem Test und $\alpha = .05$ statistisch signifikant (zum Vorzeichentest als entsprechendem Signifikanztest s. Bortz, Lienert & Boehnke, 2008, S. 256). Dieses Ergebnis veranlasst zu der Frage: Ging es bei den Münzwürfen mit rechten Dingen zu?

Differenz zweier Anteilswerte

Dieser Test wird benötigt, um zu überprüfen, ob eine bestimmte Merkmalsausprägung x in einer Stichprobe A signifikant häufiger vorkommt als in einer Stichprobe B. Für die Bestimmung der Effektgröße H müssen die erwarteten Anteilswerte in φ-*(Phi-)*Werte transformiert werden, wobei φ einer **Arkussinustransformation von** π entspricht ($\varphi = 2 \cdot \arcsin\sqrt{\pi}$). Auch diese Transformation findet man im ▶ Webangebot zu diesem Buch (Tab. 10). Für die Durchführung des Tests fertigt man sich einfachheitshalber eine Vierfeldertafel nach Art von ◘ Tab. 14.6 an.

Die unabhängigen Anteilswerte π_A und π_B ergeben sich mit a, b, c und d als Häufigkeiten für die 4 Felder zu $\pi_A = a/(a+c)$ und $\pi_B = b/(b+d)$ (zur Unterschiedsprüfung von P_A und P_B als Schätzwerte für π_A

Tabelle 14.6 Vierfeldertafel

	Stichprobe	
	A	*B*
x vorhanden	*a*	*b*
x nicht vorhanden	*c*	*d*
	$\pi_A = a/(a+c)$	
	$\pi_B = b/(b+d)$	

Tabelle 14.7 „Binomial Effect Size Display" (BESD) für $r = .20$

		Experimentalgruppe	Kontrollgruppe	
Behandlungserfolg	Ja	60	40	100
	Nein	40	60	100
		100	100	

und π_B Bortz & Schuster, 2010, Abschn. 9.1, oder – für den exakten Test – Bortz & Lienert, 2003).

Der Unterschiedshypothese (H_0: $\pi_A = \pi_B$) entspricht eine Zusammenhangshypothese (H_0: $\rho = 0$), wobei ρ über den Phi-Koeffizienten geschätzt wird. Der Phi-Koeffizient ist eine Produkt-Moment-Korrelation über zwei 0/1-kodierte Variablen (Bortz, Lienert, & Boehnke, 2008, S. 330f.). Damit stehen für die Effektgrößenklassifikation zwei verschiedene Parameter zur Verfügung: H für den Vergleich von Anteilswerten und ρ für den Zusammenhang zwischen zwei alternativen Merkmalen. Die Klassifikation der Effektgrößen (klein, mittel, groß) ist nur bedingt kompatibel, was damit zu erklären ist, dass die Größe eines Phi-Koeffizienten auch von den Randverteilungen der Vierfeldertafel abhängt (Einzelheiten hierzu bei Cohen, 1988, S. 184f.). Haddock, Rindskopf und Shadish (1998) empfehlen deshalb als Effektgröße für die Vierfeldertafeln die **Odds-Ratio (OR)**. Diese wird – in der Terminologie von Tab. 14.6 – als $OR = a \cdot d/b \cdot c$ errechnet (▶ Risiko-Effektgrößenmaße: Odds Ratio und Relative Risk).

Der Phi-Koeffizient (bzw. ein Korrelationskoeffizient r allgemein) lässt sich mit einer speziellen Vierfeldertafel veranschaulichen: Setzen wir die Zeilen- und Spaltensummen auf 100, entspricht die durch 100 dividierte Differenz $a-b$ der Korrelation zwischen der Stichprobenzugehörigkeit und dem Vorhanden- bzw. Nichtvorhandensein von x. Sind z. B. A und B eine Experimental- und eine Kontrollgruppe und kennzeichnet x einen Behandlungserfolg, lässt sich eine Korrelation von $r = .20$ zwischen den Merkmalen „Gruppenzugehörigkeit" und „Behandlungserfolg ja/nein" über die in Tab. 14.7 wiedergegebene Vierfeldertafel veranschaulichen.

Die Korrelation $r = .20$ ergibt sich wegen $(60 - 40)/100 = 0.20$. Oder anders formuliert: Wenn Experimental- und Kontrollgruppe sowie die Anzahl aller Misserfolge und Erfolge gleich groß sind, bedeutet $r = .20$, dass der Behandlungserfolg in der Experimentalgruppe gegenüber der Kontrollgruppe um 20 Prozentpunkte überlegen ist.

Will man mit diesem **„Binomial Effect Size Display"** (BESD; Rosenthal & Rubin, 1982) einen Korrelationseffekt veranschaulichen, fertigt man analog zu Tab. 14.6 eine Vierfeldertafel an mit $a = 50 + 100 \cdot r/2$ und $b = 50 - 100 \cdot r/2$, wobei die Randsummen mit jeweils 100 festgelegt sind. Die Vierfelderkorrelation für die so resultierende Tafel ist r (Kritik und Alternativen zum BESD findet man bei Hsu, 2004).

Konfidenzintervalle Wegen der besseren Anschaulichkeit wird im Folgenden das Konfidenzintervall für die Differenz $\pi_A - \pi_B$ erläutert und nicht für H (Kline, 2004, S. 519f.). Als Standardfehler der Differenz ergibt sich

$$\sigma_{(P_A - P_B)} = \sqrt{\frac{P_A \cdot (1 - P_A)}{n_A} + \frac{P_B \cdot (1 - P_B)}{n_B}}. \quad (14.21)$$

Man errechnet das Konfidenzintervall über

$$KI_{(\pi_A - \pi_B)} = (\pi_A - \pi_B) \pm z_{(\alpha/2)} \cdot \sigma_{(P_A - P_B)} \quad (14.22)$$

(zur Erläuterung von $z_{(\alpha/2)}$ Gl. 14.14).

Weitere Standardfehler, die im Zusammenhang mit dem Vergleich zweier Anteilswerte interessieren könnten, sind der Standardfehler des Phi-Koeffizienten (Fleiss, 1994, S. 249) oder der Standardfehler von OR (Kline, 2004, Tab. 5.3). Da Phi eine Produkt-Moment-Korrelation über zwei dichotome Merkmale darstellt, wäre – zumindest bei symmetrischen Randverteilungen in der Vierfeldertafel – auch das Konfidenzintervall von ρ zu erwägen.

Beispiel: Eine Behandlung A möge bei $n_A = 80$ Patienten eine Erfolgsrate von 75 % erzielen und eine Behandlung B bei $n_B = 120$ Patienten eine Erfolgsrate von 60 %. Mit $P_A = .75$ und $P_B = .60$ schätzen wir über Tab. 10 (▶ Webangebot zu diesem Buch) eine Effektgröße von $h = 2.0944 - 1.7722 = 0.3222$. Gefragt wird nach dem Konfidenzintervall für die „wahre" Differenz $\pi_A - \pi_B$.

Zunächst wird über Gl. 14.21 der Standardfehler berechnet

$$\sigma_{(P_A-P_B)} = \sqrt{\frac{.75\cdot(1-.75)}{80}+\frac{.60\cdot(1-.60)}{120}} = 0.0659\,.$$

Über Gl. 14.22 resultiert für das 95 %ige Konfidenzintervall

$$KI_{(\pi_A-\pi_B)} = (.75-.60)\pm 1.96\cdot 0.0659 = 0.15\pm 0.13$$

bzw. $.02 \le \pi_A - \pi_B \le .28$.

Mit einer Konfidenz von 95 % wäre die Überlegenheit von Behandlung A gegenüber Behandlung B durch eine Differenz zwischen 2 % und 28 % zu charakterisieren.

χ^2-Test

Mit diesem Test wird überprüft, ob zwischen zwei nominalskalierten Merkmalen ein Zusammenhang besteht (Kontingenztafeltest) oder wie gut sich die Verteilung eines Merkmals an einen bestimmten Verteilungstyp wie Gleichverteilung oder Normalverteilung anpasst (Goodness-of-Fit-Test). Beim Kontingenztafeltest entspricht k der Anzahl der Felder in der Kontingenztafel (für ein r-stufiges und ein c-stufiges Merkmal wäre $k = r \times c$) und beim Goodness-of-Fit-Test der Anzahl der Merkmalsausprägungen oder Kategorien.

π_{0i} steht für die gemäß H_0 erwarteten relativen Häufigkeiten (im Kontingenztest: Zeilensumme × Spaltensumme/n^2), und die π_{1i}-Werte sind Anteilswerte, die man bei Gültigkeit von H_1 erwartet. Die Festlegung von W in der Planungsphase setzt also voraus, dass man eine Vorstellung davon hat, wie die Kontingenztafel bei Gültigkeit von H_1 und H_0 besetzt bzw. wie das Merkmal verteilt ist.

Will man nach Durchführung der Untersuchung W über w schätzen, kann man von der folgenden Beziehung Gebrauch machen (Westermann, 2000, S. 363):

$$w = \sqrt{\frac{\chi^2}{n}}\,. \tag{14.23}$$

Ein Korrelationsäquivalent erhält man über den **Cramér-Index** (CI)

$$CI = \sqrt{\frac{\chi^2}{n\cdot(L-1)}} = \frac{w}{\sqrt{c-1}}\,. \tag{14.24}$$

Wir vereinbaren $c \le r$. Für $c = 2$ wird eine $r \times 2$- (oder $2 \times r$-) Tafel untersucht, bei der der Zusammenhang zwischen dem dichotomen Merkmal und dem r-fach gestuften Merkmal über φ' bestimmt wird

$$\varphi' = w = \sqrt{\frac{\chi^2}{n}}\,. \tag{14.25}$$

φ' entspricht einer multiplen Korrelation zwischen $p = r-1$ Kodiervariablen für das r-fach gestufte Merkmal und dem dichotomen Merkmal (zum Beweis Küchler, 1980; zur Kodierung eines nominalen Merkmals durch Indikatorvariablen ▶ Abschn. 12.5.2 (Kodierung eines polytomnominalen Merkmals durch dichotome Indikatorvariablen) oder ausführlicher zum Stichwort „ALM“ Bortz & Schuster, 2010, Kap. 22).

Dass auch CI ein Korrelationsäquivalent ist, wurde von Kshirsagar (1972, Kap. 9.6) bewiesen. Es gilt folgende Beziehung:

$$CI = \frac{1}{c-1}\cdot\sum_{i=1}^{c-1} CR_i^2\,. \tag{14.26}$$

CI^2 entspricht dem arithmetischen Mittel der quadrierten kanonischen Korrelationen zwischen $c-1$ und $r-1$ Indikatorvariablen für die beiden nominalen Merkmale (Bortz, Lienert & Boehnke, 2008, S. 355ff.; zur kanonischen Korrelationsanalyse Bortz & Schuster, 2010, Kap. 28). Der in Gl. 14.26 definierte CI-Wert heißt bei Cramer und Nicewander (1979) **„Trace Correlation“.**

Damit stehen auch für die Analyse von Kontingenztafeln einige Effektgrößenmaße zur Verfügung, die der „Korrelationsfamilie“ angehören. Allerdings gelten für φ' und für CI die gleichen Einschränkungen wie für den Phi-Koeffizienten: Die **Höhe des Koeffizienten hängt von den Randverteilungen** ab. CI hat einen Wertebereich von 0 bis 1, wenn die Randverteilungen so geartet sind, dass $\chi^2_{max} = n\cdot(c-1)$ theoretisch möglich ist.

Die Berechnung von **Konfidenzintervallen** für die Effektgröße W ist unüblich und auch wenig sinnvoll, da ein bestimmter W-Wert auf π_{0i}- und π_{1i}-Diskrepanzen unterschiedlichster Art zurückgeführt werden kann. Für **metaanalytische Zwecke** (▶ Kap. 16) kann es jedoch sinnvoll oder erforderlich sein, eine $r \times c$-Tafel auf eine 2×2-Tafel zu reduzieren, indem man Kategorien zusammenfasst (oder außer Acht lässt). Für diese Tafel wäre dann die Effektgröße h zu berechnen samt Konfidenzintervall.

Varianzanalysen

Einfaktorielle Varianzanalyse

Die einfaktorielle Varianzanalyse testet, ob sich mindestens zwei der Mittelwerte aus p unabhängigen Stichproben signifikant unterscheiden (Overall-Signifikanz). Zudem können die Mittelwerte der verschiedenen Gruppen paarweise durch Einzelvergleiche bzw. Kontraste inferenzstatistisch verglichen werden. Das wichtigste standardisierte Effektgrößenmaß der einfaktoriellen Varianzanalyse ist **Cohens f** bzw. **Eta-Quadrat** für den Gesamteffekt des Faktors (Tab. 14.5). Zudem lassen sich Eta-Quadrat-Werte für Einzelvergleiche bestimmen.

Die Effektgröße E bzw. Cohens f entspricht dem Quotienten aus σ_μ, der Streuung der gemäß H_1 erwarteten Populationsmittelwerte, und σ, der Streuung des Merkmals innerhalb der Populationen:

$$E = f = \frac{\sigma_\mu}{\sigma} . \tag{14.27}$$

Für die Bestimmung der Streuung σ innerhalb der Populationen übernehmen wir die Empfehlungen, die bereits im Zusammenhang mit der Effektgröße δ des t-Tests genannt wurden. Stehen keine vergleichbaren Untersuchungen, denen Streuungsschätzungen entnommen werden können, zur Verfügung, dividieren wir den vermuteten Range der Werte innerhalb der Populationen durch 5.15 und erhalten so für angenähert normalverteilte Merkmale eine brauchbare Schätzung von σ (für andere Verteilungsformen Abb. 12.18). σ entspricht im Kontext der einfaktoriellen Varianzanalyse der sog. **Fehlerstreuung**, die durch $\hat{\sigma}_{\text{Fehler}}$ geschätzt wird. In komplexeren Plänen ist σ^2 die **Prüfvarianz** des zu testenden Effektes (siehe hierzu jedoch auch die Ausführungen unten).

Für die Schätzung von σ_μ legen wir zunächst den Mindestrange der Mittelwerte fest, d. h., wir überlegen, wie groß der Unterschied zwischen dem kleinsten und dem größten Mittelwert mindestens sein sollte, damit er praktisch bedeutsam wird. Dividiert durch die Streuung innerhalb der Population σ resultiert folgende Größe δ_v

$$\delta_v = \frac{\mu_{max} - \mu_{min}}{\sigma} . \tag{14.28}$$

Damit ist σ_μ natürlich noch nicht eindeutig bestimmt, denn die Anordnung der mittleren μ-Werte, die σ_μ ebenfalls beeinflussen, bleibt unberücksichtigt. Theoretisch sind für die mittleren μ-Werte beliebig viele Anordnungen denkbar; für praktische Zwecke genügt es jedoch, vier typische Anordnungen zu unterscheiden:

1. Alle verbleibenden $p - 2$-Mittelwerte liegen genau in der Mitte von μ_{max} und μ_{min} (Beispiel: $\mu_{max} = 10$ und $\mu_{min} = 6$; alle übrigen μ-Werte haben den Wert 8). Für diesen Fall erhält man für die in Tab. 14.5 (oder mit Gl. 14.27) genannte Effektgröße

$$E_1 = \delta_v \cdot \sqrt{\frac{1}{2 \cdot p}} . \tag{14.29}$$

2. Die verbleibenden $p - 2$-Mittelwerte liegen in gleichen Abständen zwischen μ_{max} und μ_{min} (Beispiel: $p = 5$, $\mu_{max} = 9$, $\mu_{min} = 5$; die verbleibenden 3 Mittelwerte lauten dann 6, 7 und 8). Für diese Anordnung ergibt sich die Effektgröße

$$E_2 = \frac{\delta_v}{2} \cdot \sqrt{\frac{p+1}{3 \cdot (p-1)}} . \tag{14.30}$$

3. Bei gradzahligem p ist die eine Hälfte der verbleibenden $p - 2$ Mittelwerte mit μ_{max} und die andere mit μ_{min} identisch (Beispiel: $p = 6$, $\mu_{max} = 7$ und $\mu_{min} = 4$; zwei weitere Mittelwerte haben dann den Wert 7 und die beiden übrigen den Wert 4). Für diese Anordnung erhalten wir die Effektgröße

$$E_3 = \frac{1}{2} \cdot \delta_v . \tag{14.31}$$

4. Bei ungradzahligem p nehmen wir an, dass ein Extremwert einmal häufiger vertreten ist als der andere (z. B. für $p = 7$; $4 \times \mu_{max}$ und $3 \times \mu_{min}$ bzw. umgekehrt). Hierfür berechnen wir

$$E_4 = \delta_v \cdot \sqrt{\frac{p^2 - 1}{2 \cdot p}} . \tag{14.32}$$

Man wählt einen der vier E-Werte in Abhängigkeit vom erwarteten Verteilungsmuster für die p Mittelwerte.

Die Effektgröße E der einfaktoriellen Varianzanalyse lässt sich auch durch den Anteil der Gesamtvarianz, der auf die unabhängige Variable (Gruppenzugehörigkeiten) zurückgeht, veranschaulichen. Der entsprechende Kennwert η^2 (**Eta-Quadrat**) lautet:

$$\eta^2 = \frac{E^2}{1 + E^2} . \tag{14.33}$$

Will man die Effektgröße durch η^2 festlegen, erhält man E nach folgender Beziehung:

$$E = \sqrt{\frac{\eta^2}{1 - \eta^2}} . \tag{14.34}$$

Ungleichgroße Stichproben Bei ungleichgroßen Stichproben ist zu beachten, dass sich die Streuung des Mittelwertes σ_μ ändert. Sie lautet

$$\sigma_\mu = \sqrt{\frac{\sum_{i=1}^{p} n_i \cdot (\mu_i - \mu)^2}{N}} \tag{14.35}$$

mit $N = \sum_i n_i$.

Die in Tab. 14.5 genannte Effektgröße E wäre also mit diesem σ_μ zu berechnen.

Nach Durchführung der Untersuchung kann **Eta-Quadrat (η^2)** in der Terminologie von Bortz und Schuster (2010, Gl. 12.8) wie folgt geschätzt werden,

$$\hat{\eta}^2 = \frac{QS_{treat}}{QS_{tot}} . \tag{14.36}$$

η^2 (auch **Correlation Ratio**) ist ein deskriptives Maß für den gemeinsamen Varianzanteil von abhängiger Variable und unabhängiger Variable (Bortz & Schuster, 2010, S. 205f. und 221)

$\hat{\eta}^2$ als Schätzwert für η^2 unterliegt stichprobenbedingten Zufallsschwankungen, deren Ausmaß durch die Berechnung von **Konfidenzintervallen** deutlich wird. Eine SAS-Syntax zur Berechnung dieses Konfidenzintervalls findet man im ▶ Anhang zu diesem Buch und ein Zahlenbeispiel in ▶ Abschn. 14.4.7. Für diese Syntax muss $F > 1$ sein.

Einzelvergleiche Die Effektgröße E (oder auch η^2) ist für **metaanalytische Zwecke** wenig geeignet, da sie stark von Ausreißermittelwerten beeinflusst wird oder auch durch die Anordnung („Pattern") der Mittelwerte (Gl. 14.29–14.32). Besser geeignet sind Einzelvergleiche (Kontraste), mit denen hypothesenrelevante Mittelwerte miteinander verglichen werden. Ein **Einzelvergleich** ψ (sprich: psi) ist definiert als

$$\psi = \sum_{i=1}^{p} c_i \cdot \mu_i \tag{14.37}$$

mit der Bedingung $\sum_{i=1}^{p} c_i = 0$ (ausführlicher zu Einzelvergleichen Bortz & Schuster, 2010, 13.1). Die Überprüfung der H_0: $\psi = 0$ erfolgt über den F-Test mit $df_Z = 1$ und $df_N = N - p$.

$$F_\psi = \frac{QS_{\hat{\psi}}}{\hat{\sigma}^2_{\text{Fehler}}} = \frac{\hat{\sigma}^2_{\hat{\psi}}}{\hat{\sigma}^2_{\text{Fehler}}} \tag{14.38}$$

mit

$$QS_{\hat{\psi}} = \frac{\hat{\psi}^2}{\sum_i \frac{c_i^2}{n_i}} \tag{14.39}$$

Zur Prüfung von gerichteten Einzelvergleichshypothesen macht man von der Beziehung $t_n^2 = F_{(1,n)}$ Gebrauch und testet einseitig über die t-Verteilung.

ψ wird nach Division durch die Merkmalsstreuung σ vergleichbar mit der Effektgröße δ (Ziffer 1 in Tab. 14.5),

$$\delta_\psi = \frac{\psi}{\sigma} . \tag{14.40}$$

σ kann über die Fehlervarianz der Varianzanalyse ($\hat{\sigma}_{\text{Fehler}}$) geschätzt werden oder z. B. über die Streuung in einer Kontrollgruppe.

Ein kleines Beispiel (mod. nach Kline, 2004, S. 172f.) soll die Berechnungen verdeutlichen; Tab. 14.8 zeigt die Daten einer einfaktoriellen Varianzanalyse mit $p = 3$ Faktorstufen und $n_i = 5$ Versuchspersonen.

Das Ergebnis der Varianzanalyse enthält Tab. 14.9.

Über Gl. 14.37 (mit $\overline{A}_i$ als Schätzer für μ_i) werden 2 Einzelvergleiche mit den c-Koeffizienten $(1, 0, -1)$ und $(1/2, -1, 1/2)$ berechnet. Hierbei handelt es sich um orthogonale Einzelvergleiche (hierzu z. B. Bortz & Schuster, 2010, S. 222ff.). Außerdem stellen die c-Koeffizienten jeweils einen sog. **Standardsatz** von Koeffizienten dar, bei dem die Summe der Absolutwerte der Koeffizienten zwei ergibt ($\sum_i |c_i| = 2$). Standardsätze gewährleisten den Vergleich von Mittelwerten („Mean Difference Scaling" nach Bird, 2002, zit. nach Kline, 2004, S. 165). Im Beispiel werden nach Gl. 14.37 die folgenden Mittelwerte verglichen:

$$\begin{aligned}\hat{\psi}_1 &= (1) \cdot \overline{A}_1 + (0) \cdot \overline{A}_2 + (-1) \cdot \overline{A}_3 = \overline{A}_1 - \overline{A}_3 \\ &= 13.00 - 12.00 = 1.00 \\ \hat{\psi}_2 &= (1/2) \cdot \overline{A}_1 + (-1) \cdot \overline{A}_2 + (1/2) \cdot \overline{A}_3 \\ &= \left(\overline{A}_1 + \overline{A}_3\right)/2 - \overline{A}_2 \\ &= (13.00 + 12.00)/2 - 11.00 = 1.50 .\end{aligned}$$

Als Quadratsummen für die Einzelvergleiche errechnet man über Gl. 14.39

$$QS_{\hat{\psi}_1} = \frac{1.00^2}{1^2/5 + 0^2/5 + 1^2/5} = 2.50 ;$$

$$QS_{\hat{\psi}2} = \frac{1.50^2}{(1/2)^2/5 + 1^2/5 + (1/2)^2/5} = 7.50 .$$

Tabelle 14.8 Beispiel für eine einfaktorielle Varianzanalyse mit zwei Einzelvergleichen. (Mod. nach Kline, 2004, S. 173, Tab. 6.3. Copyright © 2004 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

	a_1	a_2	a_3
	9	8	10
	12	12	11
	13	11	13
	15	10	11
	16	14	15
$\overline{A}_i$	13.00	11.00	12.00

Tabelle 14.9 Ergebnistabelle der Varianzanalyse über die Daten von Tab. 14.8. (Mod. nach Kline, 2004, S. 173, Tab. 6.4. Copyright © 2004 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

Q. d. V.	QS	df	$\hat{\sigma}^2$	F	$\hat{\delta}_{\hat{\psi}}$	$\hat{\eta}^2$	$\hat{\eta}^2_p$
A	10.00	2	5.00	0.91		.13	
$\hat{\psi}_1$	2.50	1	2.50	0.45	0.43	.03	.04
$\hat{\psi}_2$	7.50	1	7.50	1.36	0.64	.10	.10
Fehler	66.00	12	5.50				
Total	76.00	14					

Q. d. V. Quelle der Varianz

Wegen ihrer Orthogonalität addieren sich die beiden Einzelvergleiche zur QS_{treat}.

Die F-Tests für die Einzelvergleiche wurden über Gl. 14.38) berechnet und die an der Fehlervarianz standardisierten Einzelvergleiche über Gl. 14.40) geschätzt. Für die Berechnung der $\hat{\eta}^2$-Werte wurde Gl. 14.36 eingesetzt, wobei für die beiden Einzelvergleiche QS_{treat} durch $QS_{\hat{\psi}}$ nach Gl. 14.39 zu ersetzen ist.

Abzugrenzen von Eta-Quadrat ist das **partielle Eta-Quadarat** $\hat{\eta}_p^2$, bei dem $QS_{\hat{\psi}}$ nicht an der QS_{total}, sondern an $QS_{Fehler} + QS_{\hat{\psi}}$ relativiert wird (Gl. 14.49 oder z. B. Bortz, 2005, Gl. 8.20). Wie im Beispiel wird empfohlen, sowohl $\hat{\eta}^2$ als auch $\hat{\eta}_p^2$ zu berichten. Das partielle Eta-Quadrat gibt als Varianzaufklärungs-Effektgrößenmaß nur den erklärten Varianzanteil an derjenigen Varianz an, die **nicht** durch andere Faktorstufen (bei Kontrasten in der einfaktoriellen Varianzanalyse) bzw. durch andere Faktoren (bei der mehrfaktoriellen Varianzanalyse) erklärt wird. Das partielle Eta-Quadrat wird umso größer, je mehr Erklärungsfaktoren neben dem betrachteten Effekt in das Design aufgenommen wurden, da sich dadurch der Nenner verkleinert. (Auf $\hat{\eta}_p^2$ werden wir im Zusammenhang mit mehrfaktoriellen Varianzanalysen noch einmal eingehen.). Eine kritische Betrachtung der Aussagekraft des partiellen Eta-Quadrat als Effektgrößenmaß ist besonders wichtig, da dieses Maß u. a. von dem Statistikprogrammpaket SPSS berechnet und deswegen sehr oft in der Literatur berichtet wird (Ferguson, 2009).

Konfidenzintervalle Exakte Konfidenzintervalle für Einzelvergleiche werden wie folgt ermittelt: Man transformiert den F-Wert eines Einzelvergleiches über $t_{df} = \sqrt{F_{(1,df)}}$ in einen t-Wert mit df Freiheitsgraden und berechnet das Konfidenzintervall über die nichtzentrale t-Verteilung mit t als Nichtzentralitätsparameter und $df = df_{Fehler}$ (▶ Anhang zu diesem Buch). Das weitere Procedere entspricht der Konfidenzintervallbestimmung für δ.

Die Grenzen des Konfidenzintervalls in der nichtzentralen t-Verteilung werden in Grenzen der standardisierten Einzelvergleiche transformiert:

$$\delta_\psi = NZt \cdot \sum_i \frac{c_i^2}{n_i} . \qquad (14.41)$$

Für den ersten Einzelvergleich im oben genannten Beispiel ($\hat{\delta}_{\hat{\psi}_1} = 0.43$) ist wie folgt zu operieren: Zunächst wird der F-Wert in einen t-Wert überführt: $t = \sqrt{0.45} = 0.674$. Mit diesem Wert und mit $df = 12, n_1 = 5, n_2 = 5, n_3 = 5, c_1 = 1, c_2 = 0, c_3 = -1$ als Input verwendet und (mit $\alpha = .05$) folgendes Konfidenzintervall bestimmt:

$$NZt_u = -1.31766 \quad \text{und} \quad NZt_o = 2.63840 .$$

Über Gl. 14.41 erhält man dann:

$$\hat{\delta}_{\hat{\psi}_{1(u)}} = -0.83336 ,$$
$$\hat{\delta}_{\hat{\psi}_{1(o)}} = 1.66867 .$$

Somit ist das 95 %ige Konfidenzintervall von $\hat{\delta}_{\hat{\psi}_1}$ anzugeben als:

$$-0.83 \leq \delta_{\psi_1} \leq 1.67 .$$

Die Breite des Intervalls ist mit den kleinen Stichproben ($n = 5$) zu erklären. Hierzu analog wird das 95 %ige Konfidenzintervall für δ_{ψ_2} bestimmt. Es lautet

$$-0.48 \leq \delta_{\psi_2} \leq 1.73 .$$

Ein Beispiel für $p = 4$ Faktorstufen bzw. Gruppen findet man unten.

Zu Demonstrationszwecken wenden wir uns auch dem Konfidenzintervall für η^2 zu. Im oben genannten Beispiel kann dieses Konfidenzintervall nicht berechnet werden, da $F = 0.91 < 1$. Stattdessen verwenden wir als Beispiel die bei Bortz und Schuster (2010, S. 211) genannte Ergebnistabelle mit $\hat{\eta}^2 = .70$. Hieraus entnehmen wir für die im ▶ Anhang zu diesem Buch genannte SAS-Syntax folgende Eingangsparameter $F = 12.41$, $df_1 = 3$, $df_2 = 16$ für die Berechnung (deren Details hier nicht dargestellt werden). Das Konfidenzintervall lautet:

$$.614 \leq \eta^2 \leq .936 .$$

Die „wahre" Varianzaufklärung ($\alpha = .05$) befindet sich also zwischen 61.4 % und 93.6 %.

Zusätzlich wollen wir dieses Beispiel nutzen, um die SAS-Syntax (▶ Anhang zu diesem Buch) für die Konfidenzintervallbestimmung für Einzelvergleiche mit $p = 4$ zu demonstrieren. In diesem Beispiel ergeben sich folgende Mittelwerte:

$$\overline{A}_1 = 2 ; \quad \overline{A}_2 = 3 ; \quad \overline{A}_3 = 7 ; \quad \overline{A}_4 = 4 .$$

Wenn wir $\overline{A}_1$ mit den restlichen 3 Mittelwerten kontrastieren, ergibt sich als Einzelvergleich:

$$\hat{\psi} = (1) \cdot 2 + (-1/3) \cdot 3 + (-1/3) \cdot 7 + (-1/3) \cdot 4 = 2.67.$$

Nach Gl. 14.39 ermittelt man $QS_{\hat{\psi}}$ (mit $n_i = 5$) zu

$$QS_{\hat{\psi}} = \frac{2.67^2}{1^2/5 + 3 \cdot (-1/3)^2/5} = \frac{7.13}{0.267} = 26.70$$

und über Gl. 14.38 (mit $\hat{\sigma}^2_{Fehler} = 1.88$)

$$F = \frac{26.70}{1.88} = 14.20$$

bzw.

$$t = \sqrt{14.20} = 3.77 .$$

Der standardisierte Einzelvergleich lautet gem. Gl. 14.40:

$$\hat{\delta}_{\hat{\psi}} = \frac{2.67}{\sqrt{1.88}} = 1.95 .$$

Für die Konfidenzintervallbestimmung benötigen wir folgende Eingangsparameter: $t = 3.77$; $df = 16$; $n_1 = n_2 = n_3 = n_4 = 5$; $c_1 = 1$, $c_2 = -1/3$, $c_3 = -1/3$, $c_4 = -1/3$. Über die SAS-Syntax (▶ Anhang zu diesem Buch) erhält man das folgende Konfidenzintervall ($\alpha = .05$):

$$0.71 \leq \delta_{\psi} \leq 3.14 .$$

Einfaktorielle Varianzanalyse mit Messwiederholungen

Wird eine Stichprobe p-fach untersucht, so lässt sich mit der einfaktoriellen Varianzanalyse mit Messwiederholungen überprüfen, ob sich die p Mittelwerte signifikant verändert haben. Das Verfahren dient auch dem Vergleich der Mittelwerte aus p abhängigen Stichproben („Matched Samples“).

Effektgrößen werden für Varianzanalysen mit Messwiederholungen im Prinzip genauso bestimmt wie für Varianzanalysen ohne Messwiederholungen. Problematisch ist lediglich die Streuung σ, die in einer Varianzanalyse ohne Messwiederholungen die Streuung innerhalb der Populationen bzw. die Fehlerstreuung angibt, an deren Quadrat die Treatmentvarianz getestet wird.

In der einfaktoriellen Varianzanalyse mit Messwiederholungen wird die Treatmentvarianz an einer sog. Residualvarianz (σ^2_{res}) getestet, die der Varianz der „ipsativen“ Messwerte entspricht (Bortz, 2005, S. 335). Diese Varianz ist in der Regel kleiner als die Varianz innerhalb der Populationen. Wie auch beim t-Test für abhängige Stichproben hängt ihre Größe von den Korrelationen der zu p Zeitpunkten erhobenen Messungen ab. Mit wachsender Korrelation wird die Residualvarianz kleiner.

Leider bedarf es erheblicher Erfahrungen, die Residualvarianz vor Durchführung der Untersuchung verlässlich zu schätzen. Im Zweifelsfall verwendet man statt der Residualvarianz auch für Messwiederholungsanalysen die Varianz innerhalb der Populationen (d. h. die durchschnittliche Varianz der Messungen zu den p Messzeitpunkten), obwohl diese die Residualvarianz überschätzt.

Eine Schätzung der Residualvarianz (bzw. der entsprechenden Streuung) erhält man auch nach folgender Gleichung, die allerdings voraussetzt, dass man eine Vorstellung von der durchschnittlichen Korrelation $\overline{\rho}$ zwischen den Messungen zu den verschiedenen Messzeitpunkten hat (Winer, Brown, & Michels, 1991, S. 237ff.):

$$\sigma_{res} = \sigma \cdot \sqrt{1 - \overline{\rho}} . \qquad (14.42)$$

Hieraus ergibt sich

$$E' = \frac{\sigma_{\mu}}{\sigma_{res}} . \qquad (14.43)$$

In Analogie zu Gl. 14.28 resultiert ferner

$$\delta'_{v} = \frac{\mu_{max} - \mu_{min}}{\sigma \cdot \sqrt{1 - \overline{\rho}}} . \qquad (14.44)$$

Dieser δ'_v-Wert ersetzt den δ_v-Wert in den Gl. 14.29–14.32.

Auf die Effektgrößenklassifikation sind die Ausführungen zum t-Test für abhängige Stichproben analog anzuwenden. Sie ändert sich nicht gegenüber einer Varianzanalyse ohne Messwiederholung. Allerdings wird für die Absicherung eines kleinen, mittleren oder großen Effektes meist ein kleinerer Stichprobenumfang benötigt als in der Varianzanalyse ohne Messwiederholungen. Dieser Stichprobenumfang entspricht dem Stichprobenumfang, den man benötigen würde, um in der Varianzanalyse ohne Messwiederholungen einen um den Faktor $1/\sqrt{1 - \overline{\rho}}$ vergrößerten Effekt abzusichern. Ein mittlerer Effekt ($E = 0.25$) wird also durch eine Korrelation von $\overline{\rho} = .40$ zu $E' = 0.25/\sqrt{1 - 0.4} = 0.32$ „aufgewertet“. Welche Stichprobenersparnis damit verbunden ist, werden wir in ▶ Abschn. 14.3.2 erfahren.

Wie bereits zur einfaktoriellen Varianzanalyse ohne Messwiederholungen ausgeführt, sind hypothesenrelevante **Einzelvergleiche** auch in der Varianzanalyse mit Messwiederholungen besser geeignet, das Ergebnis einer Varianzanalyse zu verdeutlichen als der Overall-F-Wert. Zur Berechnung eines Einzelvergleichs wird erneut Gl. 14.37 eingesetzt und zur Bestimmung der Quadratsumme Gl. 14.39. Wegen $df = 1$ gilt $\hat{\sigma}^2_{\hat{\psi}} = QS_{\hat{\psi}}$.

Der Signifikanztest relativiert $\hat{\sigma}^2_{\hat{\psi}}$ an der Residualvarianz $\hat{\sigma}^2_{Res}$ (mit $df_Z = 1$ und $df_N = (p - 1) \cdot (n - 1)$

$$F = \frac{\hat{\sigma}^2_{\hat{\psi}}}{\hat{\sigma}^2_{Res}} . \qquad (14.45)$$

Allerdings setzt dieser Test voraus, dass die sog. **Zirkularitätsannahme** zutrifft (Bortz & Schuster, 2010, Abschn. 18.4). Ist das nicht der Fall, wird empfohlen, nur die am Einzelvergleich beteiligten Stichproben zur Berechnung der Prüfvarianz heranzuziehen. Dies ist – wie beim t-Test für abhängige Stichproben – der Standardfehler der Differenzwerte ($\sqrt{\hat{\sigma}^2_{D_{\psi}}/n}$) (zur Berechnung s. unten).

$$t_{\hat{\psi}} = \frac{\hat{\psi}}{\sqrt{\frac{\hat{\sigma}^2_{D_{\hat{\psi}}}}{n}}} \quad (\text{mit } df = n - 1) . \qquad (14.46)$$

Für die Berechnung von standardisierten Kontrasten stehen – ebenfalls wie beim t-Test für abhängige Stichproben – 2 Varianten zur Verfügung. Entweder man standardisiert $\hat{\psi}$ an der Merkmalsstreuung (die über Gl. 14.3 oder über Gl. 14.4 bzw. über $\hat{\sigma}_{\text{Fehler}}$ geschätzt wird)

$$\hat{\delta}_{\hat{\psi}} = \frac{\hat{\psi}}{\hat{\sigma}}, \tag{14.47}$$

oder man standardisiert an der Streuung der Differenzen $\hat{\sigma}_{D_{\hat{\psi}}}$

$$\hat{\delta}'_{\hat{\psi}} = \frac{\hat{\psi}}{\hat{\sigma}_{D_{\hat{\psi}}}} . \tag{14.48}$$

Eine Standardisierung an der $\hat{\sigma}_{\text{Res}}$ wird nicht empfohlen, da die Metrik dieser Standardisierung mit der ursprünglichen Merkmalsmetrik nichts mehr zu tun hat. **Metaanalytische Zusammenfassungen** von Kontrasten aus Varianzanalysen mit unabhängigen Stichproben und mit abhängigen Stichproben bereiten am wenigsten Probleme, wenn über Gl. 14.47 standardisiert wird.

Beispiel (nach Kline, 2004, S. 173ff.): Zur numerischen Erläuterung des Gesagten soll das in Tab. 14.8 genannte Zahlenbeispiel erneut verwendet werden, mit der Annahme, die 3 Stichproben seien abhängig (z. B. Messwiederholungen über 3 Zeitpunkte). Zusätzlich sollen die beiden bereits oben genannten Einzelvergleiche geprüft werden:

$$\hat{\psi}_1 = \overline{A}_1 - \overline{A}_3 = 1 \quad \text{und}$$
$$\hat{\psi}_2 = \left(\overline{A}_1 + \overline{A}_3\right)/2 - \overline{A}_2 = 1.5 .$$

Nach den z. B. bei Bortz und Schuster (2010, Abschn. 18.1) genannten Regeln führt die Varianzanalyse zu den in Tab. 14.10 (Spalte 1–5) genannten Ergebnissen. Der Treatmentfaktor A und die beiden Einzelvergleiche sind nicht signifikant.

Da die Zirkularitätsannahme nicht geprüft wurde, berechnen wir sicherheitshalber auch t-Werte nach Gl. 14.46. Hierfür werden die Varianzen der Differenzen der an den beiden Einzelvergleichen beteiligten Stichproben benötigt. Die Differenzen für den ersten Vergleich sind $(-1, 1, 0, 4, 1)$ mit einer Varianz von $\hat{\sigma}^2_{D_{\psi_1}} = 3.50$; für den zweiten Vergleich ergeben sich Differenzen von $(1.5, -0.5, 2, 3, 1.5)$ mit einer Varianz von $\hat{\sigma}^2_{D_{\psi_2}} = 1.625$. (Die Differenzen ergeben sich hier wie folgt: $(9+10)/2 - 8 = 1.5$, $(12+11)/2 - 12 = -0.5$ etc.). Dieselben Varianzen resultieren auch nach Gl. 14.10.

Als t-Werte errechnet man über Gl. 14.46

$$t_{\hat{\psi}_1} = \frac{1}{\sqrt{\frac{3.5}{5}}} = 1.20 \quad (\text{df} = 4) ,$$
$$t_{\hat{\psi}_2} = \frac{1.5}{\sqrt{\frac{1.625}{5}}} = 2.63 .$$

Auch die t-Werte sind (bei zweiseitigen Tests) nicht signifikant.

Die standardisierten Einzelvergleiche ergeben sich in Abhängigkeit von der Art der Standardisierung (Gl. 14.47 und 14.48) zu

$$\hat{\delta}_{\hat{\psi}_1} = \frac{1}{\sqrt{5.5}} = 0.43 ,$$
$$\hat{\delta}_{\hat{\psi}_2} = \frac{1.5}{\sqrt{5.5}} = 0.64$$

(hier wurde an $\sqrt{\hat{\sigma}^2_{\text{Fehler}}} = \sqrt{5.5}$ standardisiert; Tab. 14.9) oder zu

$$\hat{\delta}'_{\hat{\psi}_1} = \frac{1}{\sqrt{3.5}} = 0.53 ,$$
$$\hat{\delta}'_{\hat{\psi}_2} = \frac{1.5}{\sqrt{1.625}} = 1.18 .$$

Wie zu erwarten, sind die an der Merkmalsstreuung standardisierten Einzelvergleichseffekte $\hat{\delta}'_{\hat{\psi}_1}$ kleiner als die an der Differenzenstreuung standardisierten Einzelvergleichseffekte $\hat{\delta}'_{\hat{\psi}_1}$ (wg. $r_{12} = .735$, $r_{13} = .730$ und $r_{23} = .839$).

Zusätzlich wurden noch die Varianzaufklärungen nach Gl. 14.36 berechnet. Auch diese Werte sollte man für **Metaanalysen** berichten, die mit Korrelationsäquivalenten bzw. Varianzaufklärungen operieren. Die partiellen Varianzaufklärungen ($\hat{\eta}^2_{\text{p}}$) wurden nach der allgemeinen Regel

$$\hat{\eta}^2_{\text{p}} = \frac{\text{QS}_{\text{Effekt}}}{\text{QS}_{\text{Effekt}} + \text{QS}_{\text{Prüf}}} \tag{14.49}$$

berechnet (z. B. Bortz, 2005, Gl. 8.20). $\text{QS}_{\text{Prüf}}$ ist hierbei die für die Prüfvarianz benötigte Quadratsumme, also QS_{Res}, und $\text{QS}_{\text{Effekt}}$ bezeichnet im Beispiel QS_{Treat} oder $\text{QS}_{\hat{\psi}} \cdot \hat{\eta}^2_{\text{p}}$-Werte sind für Vergleiche mit Effekten in mehrfaktoriellen Varianzanalysen sinnvoll (s. hierzu jedoch auch die Ausführungen zur mehrfaktoriellen Varianzanalyse). Sie empfehlen sich auch für die Varianzanalyse mit Messwiederholungen, da hier die totale Quadratsumme **nicht** der Summe aus QS_{Treat} $(= \text{QS}_{\text{Effekt}})$ + $\text{QS}_{\text{Fehler}}$ $(= \text{QS}_{\text{Prüf}})$ entspricht, sondern der Summe aus QS_{inVpn} und QS_{zwVpn} (Bortz & Schuster, 2010, Kap. 18.1). Die QS_{zwVpn} ist für die Effektprüfung ohne Bedeutung.

Vergleichen wir Tab. 14.9 mit Tab. 14.10 (also die Ergebnisse der Varianzanalysen mit bzw. ohne Messwiederholungen über dieselben Daten) wird – wegen der hohen positiven Korrelationen zwischen den Messwertreihen – der Teststärkevorteil der Varianzanalyse mit Messwiederholungen deutlich. Dies zeigen die F-Werte und auch diejenigen Effektgrößenmaße, die von der Messwiederholung „profitieren“ ($\hat{\delta}'_{\psi}$ und $\hat{\eta}^2_{\text{p}}$).

Tabelle 14.10 Ergebnistabelle der einfaktoriellen Varianzanalyse mit Messwiederholungen über die Daten der Tab. 14.8. (Mod. nach Kline, 2004, S. 175, Tab. 6.5. Copyright © 2004 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

Q. d. V.	QS	df	$\hat{\sigma}^2$	F	t	$\hat{\delta}_{\hat{\psi}}$	$\hat{\delta}'_{\hat{\psi}}$	$\hat{\eta}^2$	$\hat{\eta}^2_p$
Zwischen Vpn	54.67	4	13.67						
Innerhalb Vpn	21.33	10							
Treatment A	10.00	2	5.00	3.53				.13	.47
$\hat{\psi}_1 = 1.0$	2.50	1	2.50	1.76/1.43[a]	1.20	0.43	0.53	.03	.18/.26[a]
$\hat{\psi}_2 = 1.5$	7.50	1	7.50	5.28/6.93[a]	2.63	0.64	1.18	.10	.40/.63[a]
Residual	11.33	8	1.42						
Total	76.00	14							

Q. d. V. Quelle der Varianz; [a] Werte in SPSS abweichend, da SPSS für die Berechnung der F-Werte der Einzelvergleiche den Standardfehler für Differenzwerte und nicht die Residualvarianz heranzieht. Siehe Zirkularitätsannahme, ▶ Einfaktorielle Varianzanalyse mit Messwiederholungen Kontraste wurden über das Menü Kontraste/Polynomial berechnet.

Konfidenzintervalle Konfidenzintervalle sollten für Einzelvergleiche bestimmt werden, die an der Merkmalsstreuung (σ) standardisiert sind (δ_ψ). Die an der Differenzstreuung (σ_D) standardisierten Einzelvergleiche (δ'_ψ) haben den Nachteil, dass sie mit anderen Einzelvergleichen (z. B. für unabhängige Stichproben) nur schwer vergleichbar sind. Für die (approximative) Konfidenzintervallbestimmung geht man folgendermaßen vor:

Man berechnet zunächst den Standardfehler des Einzelvergleiches ($\hat{\sigma}_D/\sqrt{n}$) und kann dann in üblicher Weise das Konfidenzintervall des Einzelvergleiches bestimmen. Die Grenzen dieses Konfidenzintervalls werden an der Merkmalsstreuung standardisiert, die üblicherweise über die Fehlervarianz (Tab. 14.9) geschätzt wird ($\hat{\sigma}_{\text{Fehler}}$). Zusammengefasst erhält man

$$\text{KI}_{\delta_\psi} = \hat{\delta}_\psi \pm \frac{t_{(n-1,\alpha/2)} \cdot \frac{\hat{\sigma}_D}{\sqrt{n}}}{\hat{\sigma}_{\text{Fehler}}} . \qquad (14.50)$$

Im oben genannten Beispiel resultiert als 95 %iges Konfidenzintervall für ψ_1 (mit $t_{(5-1;.975)} = 2.776$ gem. Tab. 3, ▶ Webangebot zu diesem Buch)

$$\text{KI}_{\delta_{\psi_1}} = 0.43 \pm \frac{2.776 \cdot \sqrt{\frac{3.5}{5}}}{\sqrt{5.50}} = 0.43 \pm 0.99$$

bzw.

$$-0.56 \leq \delta_{\psi_1} \leq 1.42 .$$

Für ψ_2 ergibt sich

$$\text{KI}_{\delta_{\psi_2}} = 0.64 \pm \frac{2.776 \cdot \sqrt{\frac{1.625}{5}}}{\sqrt{5.50}} = 0.64 \pm 0.67$$

bzw.

$$-0.03 \leq \delta_{\psi_2} \leq 1.31 .$$

Multiple Korrelation

Die multiple Korrelation R prüft die H_0, dass zwischen p Prädiktorvariablen $X_1, X_2, \ldots, X_p$ und einer Kriteriumsvariablen Y kein Zusammenhang besteht. Die Überprüfung dieser H_0 erfolgt über den F-Test. Eine spezifische H_1 legt fest, welcher Zusammenhang zwischen den Prädiktoren und dem Kriterium mindestens erwartet wird. Der Determinationskoeffizient R^2 als derjenige Varianzanteil, den die Prädiktorvariablen zusammengenommen an der Kriteriumsvarianz aufklären, dient auch hier – wie bei der Produkt-Moment-Korrelation – als Interpretationshilfe.

Die Effektgröße K^2 ist definiert als Quotient aus erklärtem Varianzanteil (R^2) und nicht erklärtem Varianzanteil ($1 - R^2$). Dies ist gleichzeitig die Effektgröße für **Partialkorrelationen** (= partielle Korrelationen).

Bei multiplen Korrelationen unterscheidet man feste („fixed") Prädiktoren (wie z. B. die Faktorstufen eines varianzanalytischen Faktors) und zufällige („random") Prädiktoren. Letztere sind – wie üblicherweise auch bei der bivariaten Korrelation – Zufallsvariablen, die in Abhängigkeit von der Art der gezogenen Stichprobe unterschiedlich ausfallen (Alter, Intelligenz, Einkommen etc.).

Die Höhe des Zusammenhanges zwischen einer Kriteriumsvariablen und mehreren **festen Prädiktorvariablen** (z. B. Indikatorvariablen, ▶ Abschn. 12.5.2) wird üblicherweise über η^2 beschrieben. Über dieses Maß und dessen Konfidenzintervalle wurde bereits berichtet.

Hat man **zufällige Prädiktoren** untersucht, wird der Zusammenhang über R^2 (Varianzaufklärung) charakterisiert. R^2 hat bei zufälligen Prädiktoren ein breiteres Konfidenzintervall bzw. eine niedrigere Teststärke als η^2 für feste Prädiktoren (Gatsonis & Sampson, 1989). Zur Berechnung von Konfidenzintervallen für R^2 wird auf Stei-

ger (2004) verwiesen bzw. auf dessen Webseite (http://www.statpower.net/Software.html), die ein entsprechendes Tool „R2“ anbietet. Ausführliche Informationen zu dieser Thematik findet man auch bei Mendoza und Stafford (2001). Die dort genannten Tabellen für Konfidenzintervalle von R^2 (Random Model) sind im ▶ Webangebot zu diesem Buch als Tab. 12 wiedergegeben.

Für die **metaanalytische Integration** von Untersuchungen ist die multiple Korrelation nur bedingt geeignet. Das Quadrat einer multiplen Korrelation zeigt an, wie viel Varianz einer Kriteriumsvariablen durch einen (idealerweise theoriegeleiteten) Satz von Prädiktorvariablen erklärt wird. Diese Varianzaufklärung hängt natürlich von den ausgewählten Prädiktorvariablen ab, so dass eine für metaanalytische Zwecke erforderliche Vergleichbarkeit nur gegeben ist, wenn in verschiedenen Untersuchungen identische Prädiktoren zur Vorhersage derselben Kriteriumsvariablen eingesetzt werden. Auch das Herauslösen einer einzelnen Prädiktorvariablen (oder einer bestimmten Teilmenge von Prädiktorvariablen) ist problematisch, wenn die Bedeutung dieser Variablen für das Kriterium über das β-Gewicht in der multiplen Regression bestimmt werden soll. Multikollinearität der Prädiktoren bringt es mit sich, dass die Höhe eines β-Gewichtes in starkem Maße vom Kontext bzw. von den gleichzeitig geprüften Prädiktorvariablen abhängt. Kontextunabhängig und damit für metaanalytische Zwecke geeignet ist letztlich nur die **bivariate Korrelation**.

Die Situation ist vergleichbar mit einer mehrfaktoriellen Varianzanalyse, für die bereits die partiellen $\hat{\eta}_p^2$-Werte für metaanalytische Zwecke problematisiert und statt dessen die mit einfaktoriellen Plänen vergleichbaren $\hat{\eta}^2$-Werte präferiert werden. Dennoch haben selbstverständlich mehrfaktorielle Untersuchungspläne ihren eigenen Stellenwert, wenn es darum geht, unter mehreren Plänen mit identischer abhängiger Variable das beste Erklärungsmodell ausfindig zu machen. Dies gilt auch für die multiple Korrelation. Mit ihrer Hilfe findet man heraus, welches von verschiedenen „rivalisierenden“ Modellen am meisten Varianz einer bestimmten Kriteriumsvariablen erklärt. Dieses Modell wäre allerdings durch Replikationsstudien mit identischen Prädiktoren zu bestätigen, womit dann auch eine angemessene Basis für weiterführende Metaanalysen geschaffen wäre.

Mehrfaktorielle Varianzanalyse

Eta-Quadrat η^2 als standardisiertes Effektgrößenmaß für die Varianzaufklärung wurde bereits mehrfach angesprochen. Im Kontext der mehrfaktoriellen Varianzanalyse wird teilweise das **einfache Eta-Quadrat η^2** (wie in der einfaktoriellen Varianzanalyse) genutzt, sehr oft aber das **partielle Eta-Quadrat η_p^2**, da dieses von Statistikprogrammen wie SPSS automatisch berechnet wird. Das partielle Eta-Quadrat relativiert den interessierenden Effekt nicht wie das einfache Eta-Quadrat an der Gesamtvarianz, sondern nur an der Restvarianz, die nicht durch andere Faktoren im Design erklärt wird. Das partielle Eta-Quadrat für einen Faktor A wird somit größer, wenn im Design noch ein Faktor B berücksichtigt wird und steigt weiter an, wenn noch ein Faktor C hinzukommt etc. Die Effektgrößenschätzung für Faktor A wird durch das partielle Eta-Quadrat-Maß je nach Anzahl der Faktoren im varianzanalytischen Design mehr oder minder „aufgebläht“ und ist deswegen nicht mit anderen Studien vergleichbar. Vorzuziehen ist deswegen ein **generalisiertes Eta-Quadrat-Maß η_G^2**, das eine über unterschiedliche Studiendesigns vergleichbare Effektgrößenschätzung für den interessierenden Faktor liefert (Olejnik & Algina, 2003).

Da das Eta-Quadrat-Maß als Varianzaufklärungs-Effektgrößenmaß in allen drei Varianten (einfach, partiell, generalisiert) im Vergleich zum Determinationskoeffizienten eine leichte Überschätzung des Effektes beinhaltet, kann das **Omega-Quadrat-Maß** als korrigiertes Effektgrößenmaß eingesetzt werden. Das Omega-Quadrat-Maß existiert analog in den drei Varianten: **einfaches Omega-Quadrat ω^2**, **partielles Omega-Quadrat ω_p^2** und **generalisiertes Omega-Quadrat ω_G^2** (Olejnik & Algina, 2003).

Anhand eines Zahlenbeispiels einer dreifaktoriellen Varianzanalyse werden bei Fritz, Morris und Richler (2012) die verschiedenen Effektgrößenmaße verglichen. Das einfache Eta-Quadrat beträgt z. B. für Faktor A im Beispiel $\eta^2 = 8\,\%$ (korrigiert: $\omega^2 = 7\,\%$), das partielle Eta-Quadrat $\eta_p^2 = 14\,\%$ (korrigiert: $\omega_p^2 = 12\,\%$) und das generalisierte Eta-Quadrat $\eta_G^2 = 17\,\%$. Die Abweichungen zwischen Eta-Quadrat und Omega-Quadrat können jedoch deutlich größer sein, v. a. bei kleinen Stichprobenumfängen, kleinen Effekten und vielen Faktorstufen bei den untersuchten unabhängigen Variablen.

Die Varianzaufklärung im Rahmen einer Varianzanalyse wird über Gl. 14.36 (▶ Abschn. 14.2.3 „Einfaktorielle Varianzanalyse“) geschätzt. Man erhält den Anteil gemeinsamer Varianz aufgrund der linearen Beziehung zweier Variablen auch über den Determinationskoeffizienten ρ^2, d. h., $\rho^2 \cdot 100\,\%$ gibt an, wie viel Prozent einer Kriteriumsvariablen durch eine Prädiktorvariable erklärt wird. Beide Varianzaufklärungsmaße – η^2 und ρ^2 – weichen bzgl. der Größenklassifikation leicht voneinander ab, so steht z. B. als mittlerer Effekt $\rho^2 = .09$ einem $\eta^2 = .06$ gegenüber ◘ Tab. 14.5; zur Kompatibilität von E und η^2 gem. Gl. 14.33 s. Cohen, 1988, S. 284.

▪ Effektgrößen für zweifaktorielle Pläne

Mehrfaktorielle Pläne werden mit mehrfaktoriellen Varianzanalysen ausgewertet. Eine zweifaktorielle Varianzanalyse z. B. prüft mit F-Tests drei voneinander unabhän-

gige Nullhypothesen (hier und im Folgenden gehen wir davon aus, dass unter allen Faktorstufenkombinationen gleich große Stichprobenumfänge untersucht werden):

Faktor A: Die den Stufen eines Faktors A zugeordneten Populationen unterscheiden sich nicht

$$H_0\colon \mu_{1.} = \mu_{2.} = \ldots = \mu_{p.} \quad \text{oder} \quad H_0\colon \sigma_A^2 = 0\,.$$

Faktor B: Die den Stufen eines Faktors B zugeordneten Populationen unterscheiden sich nicht

$$H_0\colon \mu_{.1} = \mu_{.2} = \ldots = \mu_{.q} \quad \text{oder} \quad H_0\colon \sigma_B^2 = 0\,.$$

Interaktion $A \times B$: Die Mittelwerte der den Faktorstufenkombinationen zugeordneten Populationen ergeben sich additiv nach der Gleichung

$$H_0\colon \mu_{ij} = \mu_{i.} + \mu_{.j} - \mu_{..} \quad \text{oder} \quad H_0\colon \sigma_{A\times B}^2 = 0\,.$$

Für jede dieser Nullhypothesen können durch Effektgrößen spezifizierte Alternativhypothesen formuliert werden. Wir beginnen mit den Effektgrößen für die Faktoren A und B (kurz: Haupteffekte) und behandeln anschließend die Effektgröße der Interaktion.

Haupteffekte Für die Haupteffekte einer zweifaktoriellen Varianzanalyse werden Effektgrößen genauso spezifiziert wie in einer einfaktoriellen Varianzanalyse, d. h., wir schätzen die Streuung σ innerhalb der den Faktorstufenkombinationen zugewiesenen Populationen und ermitteln eine Effektgröße E, wie unter Ziffer 7 in ◘ Tab. 14.5 beschrieben.

Wie in der einfaktoriellen Varianzanalyse kann die Effektgröße E bzw. Cohens f eines Haupteffekts (oder auch eines Interaktionseffekts) in einer mehrfaktoriellen Varianzanalyse ebenfalls nach Gl. 14.33 in ein η^2 transformiert werden. η^2 gibt in mehrfaktoriellen Plänen jedoch nicht den Anteil an der Gesamtvarianz an, sondern an einer Varianz, die sich aus der Varianz innerhalb der Populationen sowie der Varianz des zu prüfenden Effektes zusammensetzt (hierzu auch Keren & Lewis, 1979; zum Vergleich der relativen Bedeutung verschiedener Haupteffekte Fowler, 1987). Dieser η^2-Wert wurde in ▶ Abschn. 14.2.3 „Einfaktorielle Varianzanalyse“ (Gl. 14.49) als **partielles Eta-Quadrat** (η_p^2) bezeichnet.

Nun wenden wir uns der Frage zu, wie die Effekte nach Durchführung der Untersuchung dargestellt werden sollten. Hierbei wird $\hat{\eta}_p^2$ insbesondere im Hinblick auf **Metaanalysen** für problematisch gehalten (zusammenfassend die Diskussion bei Kline, 2004, S. 221ff.). Ein $\hat{\eta}^2$-Wert auf der Basis einer einfaktoriellen Varianzanalyse stellt den Varianzanteil des untersuchten Merkmals dar, der durch eine unabhängige Variable (ein Faktor A oder allgemein ein Treatment) erklärt wird. Ist der gleiche Faktor A bzw. das gleiche Treatment nun in eine zweifaktorielle Varianzanalyse eingebunden, würde seine Varianzaufklärung mit $\hat{\eta}_p^2$ praktisch immer höher ausfallen als das entsprechende $\hat{\eta}^2$ in der einfaktoriellen Varianzanalyse. $\hat{\eta}_p^2$ gibt an, welchen Anteil der Faktor A an einer um den Haupteffekt B und die Interaktion $A \times B$ reduzierten Merkmalsvarianz hat. Dies wird im Folgenden begründet.

Für die Quadratsummenzerlegung in einer (orthogonalen) zweifaktoriellen Varianzanalyse erhält man (z. B. Bortz & Schuster, 2010, Kap. 14.1)

$$\mathrm{QS}_{\mathrm{tot}} = \mathrm{QS}_A + \mathrm{QS}_B + \mathrm{QS}_{A\times B} + \mathrm{QS}_{\mathrm{Fehler}}\,. \qquad (14.51)$$

Sind die Faktoren A und B **„organismische“ bzw. personengebundene Variablen** (z. B. Alter, Ausbildung etc., ▶ Kap. 8) und keine experimentellen Variablen (z. B. Ausprägungen eines mehrfach gestuften Treatments, das von den Forschenden gestaltet wird), entspricht die Merkmalsvarianz dem Quotienten $\mathrm{QS}_{\mathrm{tot}}/\mathrm{df}_{\mathrm{tot}}$. Würde man nun $\hat{\eta}_p^2$ nach Gl. 14.49 berechnen, hätte man ein Maß dafür, wie viel Varianz der Faktor A von der um den Faktor B und die Interaktion $A \times B$ reduzierten Merkmalsvarianz erklärt. Bezogen auf die Quadratsummen erhält man für den Nenner in Gl. 14.49

$$\begin{aligned}\mathrm{QS}_{\mathrm{Effekt}} + \mathrm{QS}_{\mathrm{prüf}} &= \mathrm{QS}_A + \mathrm{QS}_{\mathrm{Fehler}} \\ &= \mathrm{QS}_{\mathrm{tot}} - \mathrm{QS}_B - \mathrm{QS}_{A\times B}\,.\end{aligned} \qquad (14.52)$$

Diese Quadratsumme (bzw. Varianz) ist schwer vorstellbar und für **metaanalytische Zwecke gänzlich ungeeignet**.

Nehmen wir einmal an, es soll geprüft werden, ob Depressivität geschlechtsabhängig ist. Man fasst Geschlecht vereinfachend als binäre Variable auf (zur Operationalisierung von Geschlecht ▶ Abschn. 8.5.1) und vergleicht eine Frauen- und eine Männerstichprobe (Faktor A) bzgl. ihrer Depressivität (abhängige Variable) und berechnet nach Gl. 14.36 $\hat{\eta}^2$. Man erhält also einen Wert dafür, wie viel Varianz des Merkmals „Depressivität“ durch das Merkmal „Geschlecht“ erklärt wird.

Eine andere Untersuchung überprüft die gleiche Fragestellung, kontrolliert aber neben dem Faktor A (Geschlecht) einen weiteren Faktor B (z. B. Alter in 3 Stufen). Wenn man für Faktor A nun $\hat{\eta}_p^2$ nach Gl. 14.49 berechnet, würde die zweite Untersuchung voraussichtlich feststellen, dass der Geschlechtsfaktor einen höheren Varianzanteil erklärt als in der ersten Untersuchung. Aber einen Varianzanteil wovon?

Es ist nicht die Merkmalsvarianz, sondern die bzgl. Alter und der Interaktion Geschlecht × Alter bereinigte bzw. reduzierte Merkmalsvarianz, die mit der natürlichen Variabilität des Merkmals Depressivität wenig zu

tun hat. Wenn nun in weiteren, mehrfaktoriellen Plänen mit Depressivität als abhängige Variable neben dem Geschlecht jeweils untersuchungsspezifisch andere Merkmale kontrolliert werden, müsste man antizipieren, dass die jeweiligen $\hat{\eta}^2_p$-Werte für den Faktor Geschlecht in keiner Weise vergleichbar wären und damit einer kumulativen Depressionsforschung entgegenstünden.

Bemühungen, durch systematische Variation möglichst vieler unabhängiger Variablen (Faktoren) die Fehlervarianz zu reduzieren (und damit die Teststärke der Untersuchung zu erhöhen), können zwar zur Erkundung der Frage nach den Determinanten von Depressivität grundlegend von Bedeutung sein; sie sind jedoch für vergleichende Analysen der Bedeutung eines einzelnen Faktors anhand des jeweiligen $\hat{\eta}^2_p$-Wertes ungeeignet, solange die metaanalytisch einbezogenen Untersuchungen jeweils verschiedene unabhängige Variablen kontrollieren. Für diese Untersuchungen sollte man besser einen **generalisierten bzw. Gesamt-$\hat{\eta}^2$-Wert** ($\hat{\eta}^2_{gesamt}$; „generalized eta squared") angeben, der die $\hat{\eta}^2$-Werte der einzelnen Faktoren und Interaktionen zusammenfasst (in einem zweifaktoriellen Plan also $\hat{\eta}^2_{gesamt} = \hat{\eta}^2_A + \hat{\eta}^2_B + \hat{\eta}^2_{A\times B}$). Auf der Basis dieser Werte könnte man herausfinden, welche von verschiedenen „rivalisierenden Modellen" das fragliche Phänomen (hier Depressivität) am besten erklärt.

Die Sachlage ändert sich, wenn durch eine Behandlung (Treatment) in die natürliche Variabilität eines Merkmals eingegriffen wird und wir es mit **experimentellen unabhängigen Variablen** zu tun haben. Bei einer randomisierten Experimental-/Kontrollgruppenuntersuchung sollte – wie bereits erwähnt – ein an der Kontrollgruppenstreuung standardisierter Einzelvergleich berechnet werden bzw. – wenn man die Wirksamkeit mehrerer abgestufter Treatments global charakterisieren will, ein „normales" $\hat{\eta}^2$ nach Gl. 14.36.

Werden nun zusätzlich in einem mehrfaktoriellen Plan organismische Variablen kontrolliert (z. B. das Geschlecht oder das Alter der Probanden), so würde man den Behandlungseffekt überschätzen, wenn hierfür ein partielles $\hat{\eta}^2$ über Gl. 14.49 berechnet wird. Das Ausmaß der Überschätzung nimmt mit der Anzahl kontrollierter organismischer Variablen zu. Vergleichende Analysen verschiedener Behandlungen werden erschwert und machen nur Sinn, wenn jeweils identische organismische Variablen kontrolliert werden.

Hat man z. B. – wie in Abb. 14.4 – einen zweifaktoriellen Plan mit dem Faktor A (Experimental- vs. Kontrollgruppe) und Faktor B (weiblich vs. männlich) realisiert, empfiehlt sich folgendes Vorgehen:

Einzelvergleiche zu Haupteffekt A sollten an der über die Gruppen II und IV geschätzten Merkmalsstreuung standardisiert werden. Zusätzlich sollten die **bedingten**

Faktor B	Faktor A: Experimentalgruppe (a_1)	Faktor A: Kontrollgruppe (a_2)
♀ (b_1)	I	II
♂ (b_2)	III	IV

Abb. 14.4 Standardisierungsvarianten varianzanalytischer Effekte für einen zweifaktoriellen Plan (Erläuterungen s. Text)

Haupteffekte (bedingten Einzelvergleiche) charakterisiert werden (zu bedingten Haupteffekten und bedingten Einzelvergleichen Bortz & Schuster, 2010, Abschn. 13.1).

Für $A|b_1$, also den Vergleich Experimental- vs. Kontrollgruppe unter der Bedingung b_1 (weiblich), wäre die über Gruppe II (weibliche Kontrollgruppe) geschätzte Merkmalsstreuung für die Standardisierung adäquat und für $A|b_2$ die Streuung in der Gruppe IV.

Für Faktor B (die organismische Variable Geschlecht) sollten nur bedingte Einzelvergleiche dargestellt werden, denn der unbedingte Haupteffekt B aggregiert über die Experimental- und Kontrollgruppenbedingung, was für weiterführende Metaanalysen wenig Sinn ergibt. Sinnvoll ist demgegenüber der bedingte Effekt $B|a_1$, der charakterisiert, ob Frauen und Männer unterschiedlich auf die Behandlung reagieren. Er wäre an einer Streuung in den Gruppen I und III zu standardisieren. Der Effekt $B|a_2$ wiederum wäre einschlägig für Metaanalysen zum „normalen" Geschlechtseffekt und sollte dementsprechend an Gruppe II und IV standardisiert werden.

Schließlich sei in diesem Zusammenhang der klassische **zweifaktorielle Messwiederholungsplan** mit Experimental- und Kontrollgruppe erwähnt (Abb. 12.48). Der hier vorrangig interessierende Effekt ist der Interaktionseffekt, der als **Nettoeffekt** gem. Tab. 12.23 gemessen werden kann. Er wäre (bei randomisierten Gruppen) an einer Streuungsschätzung zu standardisieren, die auf den Pretestwerten in der Experimentalgruppe und der Kontrollgruppe basiert.

Allgemein sollte bei der Frage nach der „richtigen" Standardisierung von Effekten folgende Leitlinie beachtet werden:

Für die *Standardisierung varianzanalytischer Effekte* (bedingter oder unbedingter Einzelvergleiche) zur Berechnung standardisierter Effektgrößenmaße sollten *Streuungsschätzungen* verwendet werden, die die natürliche Variabilität des untersuchten Merkmals bestmöglich abbilden.

Bezogen auf die Varianzaufklärung $\hat{\eta}^2$ besagt diese Leitlinie, dass QS_{tot} im Nenner von Gl. 14.36 die Effektquadrat-

summe angemessen standardisiert, wenn QS_{tot}/df_{tot} die Merkmalsvarianz schätzt. Dies ist in **mehrfaktoriellen Plänen mit ausschließlich organismischen unabhängigen Variablen** meist der Fall. Wird durch Behandlungen, Instruktionen oder „Eingriffe" anderer Art die natürliche Variabilität eines Merkmals verändert, sollte die QS_{tot} durch eine Quadratsumme ersetzt werden, die der natürlichen Merkmalsvariabilität gut entspricht.

Dies ist auch zu beachten, wenn in einer Untersuchung Variablen **konstant** gehalten werden (z. B. nur männliche Gymnasiasten in einer Untersuchung über Maßnahmen zur Steigerung der Kreativität). Hier müssen gegebenenfalls externe Quellen zur Schätzung der **gesamten** (nicht der durch Konstanthaltung eingeschränkten) Merkmalsstreuung herangezogen werden, um die entsprechenden η^2-Werte zu schätzen. **Varianzaufklärungen, die sich auf die eingeschränkte Merkmalsvarianz beziehen, sind für metaanalytische Zwecke nur bedingt geeignet.**

Weitere Informationen zu dieser Thematik findet man bei Gillett (2003) bzw. Olejnek und Algina (2003).

Bezüglich der Bestimmung von Konfidenzintervallen für $\hat{\eta}^2$-Werte wird auf ▶ Abschn. 14.4.7 verwiesen. Für die Konfidenzintervallbestimmung für Einzelvergleiche in faktoriellen Plänen ist die derzeitig verfügbare Software noch nicht genügend ausgereift (Kline, 2004, S. 230).

Interaktionen Die A-priori-Bestimmung einer Effektgröße für Interaktionen setzt relativ genaue Vorkenntnisse über den Untersuchungsgegenstand voraus. Es ist erforderlich, dass man bereits vor Durchführung der Untersuchung die Größenordnung der zu erwartenden Mittelwerte $\overline{AB}_{ij}$ für alle Faktorstufenkombinationen angeben kann. Hierbei hilft eine grafische Darstellung der Interaktion (Abb. 12.38), in der jede Abweichung von der Parallelität der Mittelwertverläufe die Interaktionsvarianz erhöht. Die Größe der Haupteffekte spielt hierbei keine Rolle.

Ist das Muster der erwarteten Interaktion festgelegt, gestaltet sich die Bestimmung der Effektgröße $E_{A\times B}$ für die Interaktion relativ einfach. Zunächst ermitteln wir nach folgender Gleichung diejenigen Zellenmittelwerte μ'_{ij}, die nach der H_0 zu erwarten wären:

$$\mu'_{ij} = \mu_{i.} + \mu_{.j} - \mu_{..} \qquad (14.53)$$

wobei

$$\mu_{i.} = \frac{\sum_{j=1}^{q}\mu_{ij}}{q}$$

$$\mu_{.j} = \frac{\sum_{i=1}^{p}\mu_{ij}}{p}$$

und

$$\mu_{..} = \frac{\sum_{i=1}^{p}\mu_{i.}}{p} = \frac{\sum_{j=1}^{q}\mu_{.j}}{q}$$

(für gleich große Stichproben).

μ_{ij} sind die gem. der H_1 geschätzten Mittelwerte. Die Effektgröße $E_{A\times B}$ resultiert nach folgender Gleichung:

$$E_{A\times B} = \frac{1}{\sigma}\cdot\sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\left(\mu_{ij}-\mu'_{ij}\right)^2}{p\cdot q}}. \qquad (14.54)$$

(Kontrolle: $\sum_{i=1}^{p}\sum_{j=1}^{q}(\mu'_{ij}-\mu_{ij}) = 0.$) σ ist hierbei die Merkmalsstreuung, auf deren Schätzung wir bereits eingingen. Die in Tab. 14.5 genannte Klassifikation der varianzanalytischen Effekte gilt auch für Interaktionen.

Nach Durchführung der Untersuchung wird die Effektgröße $E_{A\times B}$ aufgrund der Daten analog zu Gl. 14.54 geschätzt. (Die μ'_{ij}-Parameter sind durch die entsprechenden Mittelwerte $\overline{AB}_{ij}$ und $\overline{AB}'_{ij}$ zu ersetzen.)

Zur Schätzung der Merkmalsstreuung ($\hat{\sigma}$) gelten die entsprechenden Ausführungen zu den Haupteffekten analog: Es werden diejenigen Daten zur Schätzung von σ herangezogen, die die „unverfälschte" Variabilität des Merkmals am besten widerspiegeln. Dies werden (bei quasi-experimentellen Untersuchungen) in der Regel die aus der QS_{tot} errechnete Merkmalsstreuung oder bei experimentellen Untersuchungen Streuungsschätzungen auf der Basis der Kontrollgruppe sein. Über Gl. 14.33 sollte $\hat{E}_{A\times B}$ in ein $\hat{\eta}^2$ transformiert werden, für das mit der SAS-Syntax (▶ Anhang zu diesem Buch) ein Konfidenzintervall konstruiert wird.

Interaktionseinzelvergleiche sollten dargestellt werden, wenn diese für das Ergebnis der Untersuchung besonders typisch und mit anderen Untersuchungsergebnissen gut vergleichbar sind (zur Konstruktion von Interaktionseinzelvergleichen Bortz & Schuster, 2010, S. 260, bzw. genauer Abelson & Prentice, 1997; die rechnerische Durchführung wird in ▶ Abschn. 14.4.7 demonstriert). Allerdings ist hier anzumerken, dass Interaktionseffekte nur schwer replizierbar sind, was deren metaanalytische Integration erschwert. Zur Standardisierung von Interaktionseinzelvergleichen gelten die Ausführungen zu (bedingten oder unbedingten) Einzelvergleichen der Haupteffekte analog.

■ Effektgrößen für dreifaktorielle Pläne

Mühelos lassen sich Gl. 14.28 bis 14.35 auch für Effektgrößenbestimmungen in dreifaktoriellen Varianzanalysen (mit $p\times q\times r$ Stufen) einsetzen. In den Bestimmungsgleichungen für die Effektgrößen der Haupteffekte (Gl. 14.29 bis 14.32) ersetzen wir p durch die Anzahl der Faktorstufen des jeweiligen Haupteffektes.

Für Interaktionen 1. Ordnung in einer dreifaktoriellen Varianzanalyse gilt die oben beschriebene Vorgehensweise. Will man – was selten vorkommt – eine Effektgröße für eine Interaktion 2. Ordnung bestimmen, fertigt man sinnvollerweise zunächst eine grafische Darstellung des gemäß der H_1 erwarteten Interaktionsmusters an (Abb. 12.41). Die gemäß der H_0 erwarteten Zellenmittelwerte bestimmt man nach folgender Gleichung:

$$\mu'_{ijk} = \mu_{ij.} + \mu_{i.k} + \mu_{.jk} - \mu_{i..} - \mu_{.j.} - \mu_{..k} + \mu_{...} \quad (14.55)$$

μ_{ijk} sind die gem. H_1 erwarteten Mittelwerte.

In Analogie zu Gl. 14.54 resultiert als Effektgröße

$$E_{A\times B\times C} = \frac{1}{\sigma} \cdot \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}\left(\mu'_{ijk} - \mu_{ijk}\right)^2}{p \cdot q \cdot r}} \quad . \quad (14.56)$$

Die bisherigen Ausführungen gelten für mehrfaktorielle Pläne, deren Faktoren **feste** Stufenauswahlen aufweisen (**„fixed-factors model“**, Modell I). Enthält ein mehrfaktorieller Plan einen oder mehrere Faktoren mit **zufälligen** Stufenauswahlen (**„random-factors model“**, Modell II), ändern sich dadurch die Prüfvarianzen (Bortz & Schuster, 2010, Kap. 14 und 16). Dies ist auch der Fall, wenn in einer Varianzanalyse gemischte Effekte, getestet werden (**„mixed-factors model“**, Modell III), also mindestens ein fester Faktor und mindestens ein zufälliger Faktor überprüft werden. Statt der Streuung innerhalb der Populationen in den Gl. 14.28, 14.54 und 14.56) – in varianzanalytischer Terminologie: Fehlerstreuung – verwenden wir allgemein eine Schätzung derjenigen Streuung, an deren Quadrat der zu prüfende Effekt getestet wird. Eine Gegenüberstellung der zu verwendeten Prüfvarianzen in den unterschiedlichen Modellen findet man bei Leonhart (2009).

Diese Vorgehensweise ist für die Planung des Stichprobenumfanges sinnvoll, wenn man eine Vorstellung von der Größenordnung der zu verwendenden Prüfvarianz hat – was in der Praxis sehr selten vorkommt. Für die Ex-post-Darstellung von Effektgrößen ist dieser Ansatz deshalb problematisch, weil die Prüfvarianzen von der Art und Anzahl der Faktoren mit zufälliger Stufenauswahl abhängen und damit metaanalytische Vergleiche erheblich erschwert werden. Dies gilt nach unserer Meinung auch für die entsprechenden Ausführungen bei Kline (2004, S. 232f.).

▪ Effektgrößen für Pläne mit Messwiederholungen

In der zweifaktoriellen Varianzanalyse mit Messwiederholungen (oder mit abhängigen Stichproben) werden der Messwiederholungsfaktor (z. B. Faktor B mit q Stufen) und die Interaktion $A \times B$ an der Interaktionsvarianz $B \times$ Vpn und der Gruppierungsfaktor (Faktor A mit p Stufen) an der Varianz innerhalb der Stichproben getestet (Bortz & Schuster, 2010, Abschn. 18.2). Dies sind gleichzeitig die Varianzen, die für die Bestimmung der Effektgrößen eines Haupteffekts nach den Gl. 14.28ff. bzw. für die Bestimmung der Effektgröße einer Interaktion nach Gl. 14.54 zu schätzen sind. (Vergleiche hierzu die Ausführungen zur einfaktoriellen Varianzanalyse mit und ohne Messwiederholungen.) Überwiegend interessiert in einer zweifaktoriellen Varianzanalyse mit Messwiederholungen jedoch die Interaktion, weil diese über gruppenspezifische Veränderungen informiert (z. B. Experimentalgruppe vs. Kontrollgruppe, ▶ Abschn. 12.5.3).

Eine Effektgröße für die Interaktion kann nach Gl. 14.54 bestimmt werden. Man beachte jedoch, dass σ in dieser Gleichung die auf individuellen Veränderungen basierende Streuung ($\sigma_{B\times \mathrm{Vpn}}$) in der Regel überschätzt. Eine günstigere Schätzung erhält man über Gl. 14.42, wenn man die Größenordnung für $\overline{\rho}$ (hier: durchschnittliche Korrelation zwischen den Messzeitpunkten, gemittelt über die Gruppen des Faktors A) kennt.

Auch hier muss angemerkt werden, dass diese Empfehlung nur Sinn ergibt, wenn es um die Planung des Stichprobenumfanges für eine Untersuchung geht. Für die Ex-post-Darstellung empirischer Effektgrößen kann diese Vorgehensweise nicht empfohlen werden. Auch hier sollte zur Erleichterung von **Metaanalysen** die bereits genannte Leitlinie beachtet werden, d. h., es ist eine Streuungsschätzung $\hat{\sigma}$ zu verwenden, die der natürlichen Merkmalsvariabilität am nächsten kommt.

Wie bereits erwähnt, interessiert bei einer Pretest/Posttest-Untersuchung mit Experimental- und Kontrollgruppe v. a. der in ▶ Abschn. 12.5.4 eingeführte „**Nettoeffekt**“, der dem Interaktionseffekt entspricht. Auch dieser Effekt sollte an einer der „natürlichen“ Merkmalsvariabilität entsprechenden Streuung standardisiert werden. Wir werden dieses Vorgehen in ▶ Abschn. 14.4.7 an einem Beispiel erläutern. Für die Berechnung von **Konfidenzintervallen** standardisierter Nettoeffekte steht unseres Wissens derzeit keine ausgereifte Software zur Verfügung.

14.3 Quantitative Studien mit optimalen Stichprobenumfängen planen

Bereits im Zuge der Planung einer empirisch-quantitativen hypothesenprüfenden Studie sollte man sich über die avisierte statistische Auswertung Gedanken machen und durch eine A-priori-Teststärkeanalyse (▶ Abschn. 14.1.2) dafür Sorge tragen, dass die später durchzuführenden statistischen Signifikanztests mit ausreichender Teststärke

(in der Regel mindestens $1 - \beta = .80$) ausgestattet sind. Dies wird erreicht, indem man vorab den optimalen Stichprobenumfang für den jeweiligen Signifikanztest bestimmt, der sich errechnen lässt, wenn man neben der Teststärke noch das Signifikanzniveau (meist $\alpha = .01$ bzw. $\alpha = .05$) sowie die erwartete bzw. interessierende Populationseffektgröße festlegt. Sollen, was der Regelfall ist, im Rahmen einer Studie mehrere Signifikanztests durchgeführt werden, so richtet sich der optimale Stichprobenumfang der Studie nach demjenigen **Signifikanztest, der den vergleichsweise größten optimalen Stichprobenumfang** benötigt.

Im Folgenden geben wir zunächst noch einige allgemeine Hinweise zum Vorgehen bei der Bestimmung des optimalen Stichprobenumfanges. Anschließend präsentieren wir eine Tabelle, aus der überblicksartig die optimalen Stichprobenumfänge der wichtigsten Signifikanztests ablesbar sind. Das Arbeiten mit derartigen Tabellen geht auf Cohen (1988) zurück. Es ist heute noch praktikabel, wird aber zunehmend durch den Einsatz von Computerprogrammen (z. B. G*Power) abgelöst.

14.3.1 Vorgehen bei der Festlegung des optimalen Stichprobenumfanges

Die Bestimmung des optimalen Stichprobenumfanges einer geplanten Studie, nachdem Teststärke, Signifikanzniveau und erwartete bzw. interessierende Populationseffektgröße festgelegt wurden, wird auch als **A-priori-Teststärkeanalyse** bezeichnet (▶ Abschn. 14.1.2).

Generell gilt, dass sich (bei sonst konstanten Einflussgrößen) der **optimale Stichprobenumfang verkleinert**,

- wenn die interessierende Populationseffektgröße zunimmt,
- wenn man die geforderte Teststärke verringert (z. B. $1 - \beta = .75$ statt $1 - \beta = .80$) und
- wenn man das Signifikanzniveau heraufsetzt (z. B. $\alpha = .10$ statt $\alpha = .05$).

Relativ unproblematisch bei einer A-priori-Teststärkeanalyse ist normalerweise die Festlegung von Teststärke und Signifikanzniveau, da sich hier konventionell $\boldsymbol{1 - \beta = .80}$ und $\boldsymbol{\alpha = .05}$ (bzw. $\alpha = .01$) eingebürgert haben. Anspruchsvoller ist es, die erwartete oder interessierende Effektgröße, die durch den geplanten Signifikanztest aufdeckbar sein soll, realistisch abzuschätzen.

In der Grundlagenforschung sind theoretische Modelle sowie vorliegende empirische Studien und ggf. auch eine eigene Vorstudie heranzuziehen, um die **erwartbare Effektgröße** vorab realistisch einschätzen zu können. Sei es, dass Vorläuferstudien direkt die Effektgröße zu entnehmen ist, sei es, dass einzelne Bestimmungsstücke des Effektgrößenmaßes bekannt sind und daraus dann die erwartete Effektgröße – anhand der Formeln in ▶ Abschn. 14.2.3 – abgeleitet werden kann. Dabei ist jedoch zu beachten, dass uns inhaltlich stets die Populationseffektgröße interessiert, die auf der Basis von Stichprobenuntersuchungen (insbesondere wenn es sich um kleine und nicht-probabilistische Stichproben handelt) nur sehr ungenau abgeschätzt werden kann. In der Evaluationsforschung (▶ Kap. 18) werden die interessierenden Effekte oft anhand praktischer Kriterien festgelegt, etwa wenn Maßnahmen nur dann als lohnend eingeschätzt werden können, wenn sie einen Mindesteffekt z. B. mittlerer Stärke erreichen (etwa beim Lernerfolg oder Therapiefortschritt). Wenn eine mittlere oder große Effektstärke angesetzt wird, resultiert im Rahmen der A-priori-Teststärkeanalyse ein geringerer optimaler Stichprobenumfang, der die empirische Arbeit erleichtert. Allerdings besteht die Gefahr, dass bei einer Überschätzung der Effektgröße und daraus resultierender Festlegung eines zu geringen Stichprobenumfanges dann Studien geplant werden, die letztlich doch keine ausreichende Teststärke aufweisen.

In der sozialwissenschaftlichen Forschungspraxis haben wir es meist mit komplexen Sachverhalten zu tun, die durch viele Einflussfaktoren determiniert sind. Werden in einer Studie einzelne Variablen herausgegriffen und analysiert, so kann das Phänomen dementsprechend nur in bestimmten Ausschnitten bzw. partiell erklärt werden. Dementsprechend fallen die **betrachteten Effektgrößen** oft nur mittel oder sogar **gering bis sehr gering** aus. Unter diesen Umständen ergeben sich laut A-priori-Teststärkeanalyse notgedrungen **sehr große optimale Stichprobenumfänge** (etwa im oberen dreistelligen oder gar im vierstelligen Bereich), um Signifikanztests mit hoher Teststärke zu gewährleisten. Aus forschungsökonomischen Gründen ist es jedoch oft gar nicht möglich, derart umfangreiche Samples zu erheben. Sei es, dass gar nicht genügend Untersuchungspersonen bzw. Untersuchungseinheiten vorhanden bzw. rekrutierbar sind, sei es, dass Zeit- und Personalaufwand für die Durchführung entsprechend vieler Versuchsdurchläufe die vorhandenen Ressourcen übersteigen.

Bedenkenswert bei der Planung einer hypothesenprüfenden quantitativen Studie sind zur **Sicherung ausreichender Teststärke** deswegen **neben dem Hochsetzen des Stichprobenumfanges** noch weitere Maßnahmen, die auf eine Steigerung der gemessenen Effektgröße sowie des Signifikanzniveaus hinauslaufen (Ellis, 2010, S. 81f.; Lipsey, 1997; Maxwell, 2000; Shadish, Cook, & Campbell, 2002):

1. Der in einer Studie nachweisbare Effekt wird v. a. dadurch vergrößert, dass die Fehlervarianz reduziert wird. Hier spielt die **Reliabilität der verwendeten**

Messinstrumente eine große Rolle: Alle Maßnahmen zur Erhöhung der Reliabilität der verwendeten Messinstrumente steigern die Teststärke. Um hohe Teststärke zu gewährleisten, sollte untersuchungsplanerisch bereits bei der Auswahl der untersuchten abhängigen Variablen mitberücksichtigt werden, ob hierfür hochmessgenaue Instrumente zur Verfügung stehen.

2. Ein nachweisbarer Effekt vergrößert sich zudem, wenn die Daten durch **Erhebung von Kontrollvariablen** um Störeinflüsse bereinigt werden können. Auch eine **Eliminierung von Störvariablen**, indem nur eine bestimmte merkmalsähnliche Teilgruppe untersucht wird, lässt den interessierenden Effekt deutlicher zutage treten und steigert mit der Effektgröße auch die Teststärke. **Messwiederholungsdesigns** reduzieren ebenfalls die Fehlervarianz, steigern somit (unter sonst gleichen Bedingungen) die Effektgröße und reduzieren folglich den notwendigen Mindeststichprobenumfang. (Maßnahmen zur Kontrolle und Eliminierung von Störvariablen können jedoch die externe Validität von Studien sowie ihre Vergleich- und Integrierbarkeit mit anderen Studien im Rahmen von Metaanalysen mindern und sollten deswegen mit untersuchungsplanerischem Augenmaß zum Einsatz kommen.)
3. Die Effektgröße kann in manchen Studien auch sinnvoll dadurch erhöht werden, dass man die **Dosierung der unabhängigen Variablen** („intervention intensity“) hochsetzt. Will man etwa unmittelbare Effekte der Mediennutzung nachweisen, so dürften kurze Versuchsdurchgänge (z. B. 10 Minuten Computerspiel) geringere Effekte nach sich ziehen als längere Versuchsdurchgänge (z. B. 45 Minuten Computerspiel). Ebenso können die Merkmale des Materials (z. B. Grad der Gewalthaltigkeit eines Computerspiels) unterschiedlich abgestuft werden. Hier sind neben theoretisch-inhaltlichen wiederum forschungspraktische und forschungsethische Entscheidungskriterien zu berücksichtigen, wenn es darum geht, die Gestaltung und Dosierung der unabhängigen Variable(n) mit Blick auf die erwartete Effektgröße festzulegen. Auch Aspekte der externen Validität sind zu beachten, denn extrem dosierte Interventionen mögen zwar das Teststärkeproblem beheben, aber auf Kosten der Generalisierbarkeit, Vergleichbarkeit und praktischen Anwendbarkeit der Befunde gehen.
4. Die **Wahl des verwendeten Signifikanztests** hat Einfluss auf die Teststärke:
   - Einseitige Tests sind teststärker als zweiseitige Tests.
   - Parametrische Tests sind (sofern ihre Voraussetzungen erfüllt sind) in der Regel teststärker als nonparametrische Tests.
   - Tests mit gleichgroßen Untersuchungsgruppen sind oft teststärker als Tests mit ungleichen Gruppengrößen (dies gilt z. B. für t-Test und Varianzanalyse).
5. Durch eine **Liberalisierung des Signifikanzniveaus** z. B. von 5 % auf 10 % erhöht sich die Teststärke.

Der Optimierung der genannten Sekundärparameter sollte im Rahmen der Untersuchungsplanung mit Blick auf die Erhöhung von Effektgröße und Teststärke mehr Aufmerksamkeit geschenkt werden, da entsprechende Maßnahmen – mit dem angesprochenen Augenmaß umgesetzt – häufig kostengünstiger realisierbar sind als extrem große Stichprobenumfänge im Bereich mehrerer Tausend Untersuchungspersonen (Erdfelder et al., 2010).

Weichen der per A-priori-Teststärkeanalyse ermittelte optimale Stichprobenumfang und der mit den vorhandenen Ressourcen voraussichtlich realisierbare Stichprobenumfang stark voneinander ab, so ist der Sinn der geplanten Studie zu hinterfragen und möglicherweise das zu bearbeitende Forschungsproblem (▶ Kap. 5) einzugrenzen oder zu verändern.

Die Durchführung einer A-priori-Teststärkeanalyse nach Festlegung der Eingangsparameter kann mithilfe entsprechender **Software** erfolgen, etwa mit dem bereits vorgestellten, kostenlosen Programm G*Power (▶ Hinweise zur Durchführung von Teststärkeanalysen). Darüber hinaus stehen auch **Tabellenwerke** zur Verfügung, denen für häufig auftretende Parameterkonstellationen die optimalen Stichprobenumfänge zu entnehmen sind. Beide Vorgehensweisen sollten zu identischen Ergebnissen führen.

14.3.2 Optimale Stichprobenumfänge der wichtigsten Signifikanztests

In ◘ Tab. 14.11 sind die optimalen Stichprobenumfänge für wichtige Signifikanztests für $\alpha = .01$ bzw. $\alpha = .05$ und kleine, mittlere sowie große Effekte enthalten (zur **Effektgrößenklassifikation** ▶ Abschn. 14.2.2). Alle Stichprobenumfänge basieren auf einer Teststärke von $1 - \beta = .80$. Sie wurden den entsprechenden Tabellen von Cohen (1988, 1992) entnommen. Wenn zwei abweichende Werte angegeben sind, bezieht sich die erste Zahl auf Cohen (1988) und die zweite auf Cohen (1992).

Abweichend von Cohen (1992) gelten die zu den Ziffern 1–5 genannten Stichprobenumfänge für **einseitige Tests**. Wie in ▶ Abschn. 12.5.1 ausgeführt, sind wir der Auffassung, dass es bei einer zu prüfenden Theorie oder einer zu evaluierenden Maßnahme möglich sein müsste, die Richtung ihrer Wirkung in Form **gerichteter Hypothesen** vorzugeben, so dass einseitige Tests gerechtfertigt

Tabelle 14.11 Optimale Stichprobenumfänge für verschiedene Signifikanztests nach Effektgrößen ($1 - \beta = .80$; Nr. 1–5 einseitige Tests; Nr. 6–7 zweiseitige Tests; mod. nach Cohen, 1988, S. 54, 101, 134, 167, 205, 253f., 258f., 381f., 384f. Republished by permission of Taylor & Francis Ltd, from Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Permission conveyed through Copyright Clearance Center, Inc.; 1992, S. 158. Copyright © 1992 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA; abweichende Werte nach Cohen (1992) sind nach dem Schrägstrich angegeben; Erläuterungen s. Text)

Test	$\alpha = .01$; Effektgrößen			$\alpha = .05$; Effektgrößen		
	Klein	Mittel	Groß	Klein	Mittel	Groß
1. t-Test	503/ 586	82/ 95	33/ 38	310/ 393	50/ 64	20/ 26
2. Bivariate Korrelation	1 000/ 1 163	108/ 125	36/ 41	617/ 783	68/ 85	22/ 28
3. Korrelationsdifferenz	2 010/ 2 339	226/ 263	83/ 96	1 240/ 1 573	140/ 177	52/ 66
4. Differenz eines Anteilswertes von .50	1 001/ 1 165	109/ 127	37/ 44	616/ 783	67/ 85	23/ 30
5. Differenz zweier Anteilswerte	502/ 584	80/ 93	31/ 36	309/ 392	49/ 63	19/ 25
6. χ^2-Test						
df = 1	1 168	130	47/ 38	785	87	31/ 26
df = 2	1 388	154	56	964	107	39
df = 3	1 546	172	62	1 090	121	44
df = 4	1 648/ 1 675	183/ 186	66/ 67	1 194	133	48
df = 5	1 787	199	71	1 283/ 1 293	143	51
df = 6	1 887	210	75	1 362	151	54
7. Einfaktorielle Varianzanalyse						
df = 1	586	95	38	393	64	26
df = 2	464	76	30	322	52	21
df = 3	388	63	25	274	45	18
df = 4	336	55	22	240	39	16
df = 5	299	49	20	215	35	14
df = 6	271	44	18	195	32	13
8. Multiple Korrelation (Ausführungen siehe Text)						

sind. Die Stichprobenumfänge für Ziffer 6–7 gelten für **zweiseitige Tests**. Optimale Stichproben für andere Konfigurationen sind dem Standardwerk von Cohen (1988) zu entnehmen.

t-Test

Will man z. B. einen in der Population gültigen, großen Effekt ($\delta = 0.80$) über die Mittelwerte zweier unabhängiger Stichproben auf dem 5 %igen Signifikanzniveau statistisch absichern, so benötigt man bei einseitigem Test pro Stichprobe $n = 20$ Untersuchungsobjekte.

Ungleich große Stichproben Es empfiehlt sich, den ermittelten Gesamtstichprobenumfang (im Beispiel $n = 40$) auf die beiden Stichproben gleich zu verteilen, da sonst der t-Test an Teststärke verliert (Kraemer & Thiemann, 1987). Sollten die Untersuchungsumstände eine ungleiche Verteilung erfordern, ist wie folgt vorzugehen: Man entnimmt zunächst Tab. 14.11 den optimalen Stichprobenumfang (im Beispiel $n = 20$). Wenn nun $n_A = 15$ für die Stichprobe A vorgesehen ist, errechnet man den Stichprobenumfang n_B wie folgt:

$$n_B = \frac{n_A \cdot n}{2 \cdot n_A - n} = \frac{15 \cdot 20}{30 - 20} = 30 \,. \qquad (14.57)$$

Mit $n_B = 30$ und $n_A = 15$ hat der t-Test die gleiche Teststärke wie mit gleich großen Stichproben ($n_A = n_B = 20$), d. h., mit diesen Stichproben entscheidet der t-Test bei einem großen Effekt und $\alpha = .05$ mit 80 %iger Wahrscheinlichkeit zugunsten von H_1 (hier: gerichtete spezifische Hypothese, die einen großen Mittelwertunter-

Tabelle 14.12 Optimale Stichprobenumfänge für den Vergleich von zwei Mittelwerten aus abhängigen/korrelierten Stichproben bei unterschiedlichen Korrelationen nach Effektgrößen ($1 - \beta = .80$; einseitiger Test)

Korrelation	$\alpha = .01$			$\alpha = .05$		
	Klein	Mittel	Groß	Klein	Mittel	Groß
$r = .20$	403	66	26	248	41	16
$r = .40$	302	49	20	187	31	13
$r = .60$	202	33	14	125	21	9
$r = .80$	101	17	7	63	11	5

schied postuliert), falls diese zutrifft. Man beachte, dass Gl. 14.57 $2 \times n_A > n$ voraussetzt.

Abhängige Stichproben Als standardisiertes Effektgrößenmaß für die Mittelwertdifferenz bei abhängigen bzw. korrelierten Mittelwerten (auf der Basis von Messwertpaaren) wurden in ▶ Abschn. 14.2.3 zwei Varianten eingeführt, die auch bei der Bestimmung optimaler Stichprobenumfänge zu unterscheiden sind.

Variante a: Für die Ermittlung optimaler Stichprobenumfänge für den Vergleich zweier abhängiger Stichproben kann man ebenfalls die Cohen-Tabellen heranziehen. Allerdings sind diese für zwei unabhängige Stichproben ausgelegt, so dass eine Korrektur des hypothetisch vorgegebenen Effektes erforderlich wird. Sind 2 Stichproben voneinander unabhängig, wird die Korrelation $\rho = 0$, d. h., man erhält für Gl. 14.9) $\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2}$ bzw. für homogene Varianzen $\sigma_D = \sqrt{2} \cdot \sigma$. Die Varianz der Differenzen σ_D^2 ist also zweimal so groß wie die Merkmalsvarianz σ^2 (genauer hierzu Bortz, 2005, Anhang B, Gl. B37). Die Vergleichbarkeit von δ (Tab. 14.5) und δ' (Gl. 14.11) wird hergestellt, wenn man δ' in (Gl. 14.11) mit dem Faktor $\sqrt{2}$ multipliziert, um so ein zu δ äquivalentes Effektmaß ($\delta_{äquiv}$) zu erhalten:

$$\delta_{äquiv} = \delta' \cdot \sqrt{2} = \frac{\mu_D}{\sigma \cdot \sqrt{1-\rho}} = \frac{\delta}{\sqrt{1-\rho}}. \qquad (14.58)$$

Mit größer werdender Korrelation vergrößert sich auch $\delta_{äquiv}$ mit der Folge, dass zur Absicherung eines δ'-Wertes bei größeren (positiven) Korrelationen erheblich kleinere Stichproben erforderlich sind als zur Absicherung eines entsprechenden δ-Wertes. Wie sich einige ausgewählte Korrelationskoeffizienten auf die optimalen Stichprobenumfänge auswirken, zeigt Tab. 14.12.

In Tab. 14.12 werden die Anzahl der Untersuchungsobjekte (mit zweimaliger Messung) bzw. die Anzahl der benötigten Paare von Untersuchungsobjekten genannt. Die Tabelle gilt für die Überprüfung gerichteter Hypothesen (einseitige Tests).

Beispiel: Es soll ein mittlerer Effekt ($\delta = 0.50$) mit $\alpha = .05$ abgesichert werden. Vergleichbaren Untersuchungen ist zu entnehmen, dass die Messwerte zum Untersuchungszeitpunkt t_1 zu $\rho = .60$ mit den Messwerten zum Untersuchungszeitpunkt t_2 korrelieren. Aus Tab. 14.12 entnehmen wir, dass für diese Untersuchung 21 zweimal zu untersuchende Untersuchungsteilnehmer (bzw. 21 Paare von Untersuchungsteilnehmern) ausreichend sind. Diesen Wert ermitteln wir wie folgt:

Gl. 14.58 führt zu

$$\delta_{äquiv} = \frac{0.5}{\sqrt{1-0.6}} = 0.79.$$

Den optimalen Stichprobenumfang errechnen wir nach folgender Gleichung (Cohen, 1988, S. 53):

$$n_{opt} = \frac{n_{0.10}}{100 \cdot \delta_{äquiv}^2} + 1 \qquad (14.59)$$

mit $n_{0.10} = 1\,237$ für $\alpha = .05$ und $n_{0.10} = 2\,009$ für $\alpha = .01$.

Für das Beispiel erhält man den in Tab. 14.12 genannten Wert

$$n_{opt} = \frac{1\,237}{100 \cdot 0.79^2} + 1 = 20.8 \approx 21.$$

Unter **Variante b** wird μ_D nicht an der Streuung der Differenzen (σ_D), sondern an der Merkmalsstreuung σ standardisiert. Es gelten deshalb die Regeln für unabhängige Stichproben bzw. die in Tab. 14.11 genannten optimalen Stichprobenumfänge. Im oben genannten Beispiel wären gem. Tab. 14.11 unter den gleichen Bedingungen ($\alpha = .05$, $1 - \beta = .80$, $\delta = 0.50$, einseitiger Test) 2 Stichproben mit jeweils 50 Untersuchungsteilnehmern erforderlich.

Bivariate Korrelation

Will man z. B. eine Populationskorrelation, die einem mittleren Effekt entspricht ($\rho = .30$), mit $\alpha = .05$ statistisch absichern (einseitiger Test), so sollte als optimaler Stichprobenumfang $n = 64$ realisiert werden.

Korrelationsdifferenz

Soll z. B. ein mittlerer Unterschied zweier Korrelationen ρ_A und ρ_B ($Q = 0.3$) aus unabhängigen Stichproben mit $\alpha = .01$ abgesichert werden, benötigt man aus den Populationen A und B jeweils eine Stichprobe mit dem Umfang $n = 226$. Der mittleren Effektgröße $Q = 0.3$ entsprechen z. B. die Korrelationspaare .00–.29; .20–.46; .40–.62; .60–.76; .80–.885 oder .90–.945 (Tab. 9 im ▶ Webangebot zu diesem Buch).

Ungleich große Stichproben Wenn es die Untersuchungsumstände erforderlich machen, dass die Stichproben aus den Populationen A und B nicht gleich groß sein können, geht man wie folgt vor: Man entnimmt zunächst Tab. 14.11 den optimalen Stichprobenumfang n und legt n_A fest. Der Wert für n_B ergibt sich dann nach folgender Gleichung:

$$n_B = \frac{n_A \cdot (n+3) - 6 \cdot n}{2 \cdot n_A - n - 3} . \qquad (14.60)$$

Der Test mit diesen Stichprobenumfängen hat die gleiche Teststärke wie der entsprechende Test mit $n_A = n_B = n$.

Beispiel (für $n = 226$ und $n_A = 150$):

$$n_B = \frac{150 \cdot (226+3) - 6 \cdot 226}{2 \cdot 150 - 226 - 3} = 465$$

(n_A ist so zu wählen, dass $2 \times n_A > (n+3)$ ist).

Differenz eines Anteilswertes von .50

Zur statistischen Absicherung einer kleinen Abweichung eines Anteilswertes π von .50 ($G = 0.55 - 0.50 = 0.05$) werden (für $\alpha = .01$) $n = 1\,001$ Untersuchungsobjekte benötigt, um eine ausreichende Teststärke von mindestens 80 % sicherzustellen.

Differenz zweier Anteilswerte

Wird erwartet, dass die Differenz zweier Anteilswerte π_A und π_B in zwei unabhängigen Populationen klein ist ($H = 0.2$), benötigt man für $\alpha = .05$ pro Stichprobe $n_A = n_B = 309$ Untersuchungsobjekte. Der Effektgröße $H = 0.2$ entsprechen z. B. die folgenden Anteilsdifferenzen: .05–.10; .20–.29; .40–.50; .60–.70; .80–.87 oder .90–.95 (Tab. 10 ▶ Webangebot zu diesem Buch).

Ungleich große Stichproben Muss die Planung von ungleich großen Stichproben aus den Populationen A und B ausgehen, wählt man n gem. Tab. 14.11, legt n_A fest und errechnet n_B nach folgender Gleichung:

$$n_B = \frac{n \cdot n_A}{2 \cdot n_A - n} . \qquad (14.61)$$

Beispiel (mit $n = 309$ und $n_A = 200$):

$$n_B = \frac{309 \cdot 200}{2 \cdot 200 - 309} = 679.$$

Man beachte, dass $2n_A > n$ ist.

χ^2-Test

Ein χ^2-Test über eine rxc-Kontingenztafel hat $(r-1) \times (c-1)\times$ Freiheitsgrade (df). Erwartet man z. B. für eine 3×4-Tafel eine große Kontingenz der geprüften Merkmale ($W = .50$ gem. Tab. 14.5), ergäben sich für df = 6 und für $\alpha = .05$ ein optimaler Gesamtstichprobenumfang von $n = 54$. Ein Goodness-of-Fit-Test auf Gleichverteilung mit k Kategorien hat $k - 1$ Freiheitsgrade.

Varianzanalysen

Einfaktorielle Varianzanalyse

Eine einfaktorielle Varianzanalyse über p Gruppen hat $p - 1$ Zählerfreiheitsgrade (df). Erwartet man z. B., dass sich vier Gruppen (df = 3) insgesamt mittelmäßig unterscheiden ($E = f = 0.25$), benötigt man für eine statistische Absicherung der Unterschiede mit $\alpha = .05$ und $1-\beta = .80$ pro Gruppe 45 oder als Gesamtstichprobe $4 \times 45 = 180$ Untersuchungsobjekte. Die 180 Untersuchungsobjekte können – falls erforderlich – auch auf ungleich große Stichproben verteilt werden.

Die Planung des Stichprobenumfanges sollte sich auf alle p Gruppen der einfaktoriellen Varianzanalyse beziehen, auch wenn vorrangig nur ausgewählte Gruppen für spezielle Einzelvergleiche interessieren. Man beachte, dass die Varianzanalyse mit df = 1 dem t-Test für unabhängige Stichproben entspricht. Die varianzanalytischen Angaben in Tab. 14.11 gelten für den zweiseitigen und die t-Test-Angaben für den einseitigen Test.

Einfaktorielle Varianzanalyse mit Messwiederholungen

Durch die mehrfache Untersuchung derselben Stichprobe (oder durch den Einsatz von p „Matched Samples“) lässt sich der optimale Stichprobenumfang erheblich reduzieren. Hierfür benötigt man allerdings gem. Gl. 14.43 eine Schätzung von σ_{res}. Liegen keine vergleichbaren Untersuchungen vor, kann man σ_{res} über Gl. 14.42 unter Verwendung der durchschnittlichen Korrelation $\overline{\rho}$ zwischen den Messwertreihen schätzen.

Die Planung des optimalen Stichprobenumfanges für eine Varianzanalyse mit Messwiederholungen bereitet ohne Zuhilfenahme vergleichbarer Untersuchungen einige Probleme. Hat man weder eine plausible Schätzung der zu erwartenden Residualvarianz noch eine Vorstellung über die durchschnittliche Korrelation der p Messwertreihen, ist man immer auf der sicheren Seite, wenn man von $\overline{\rho} = 0$ ausgeht und damit die optimalen Stichprobenumfänge der Varianzanalyse mit unabhängigen Stichproben einsetzt. Geht man jedoch von der für viele Fragestellungen vorsichtigen Annahme aus, dass die **Messwertreihen im Durchschnitt etwa zu $\overline{\rho} = .50$** korrelieren, ergeben sich die in Tab. 14.13 genannten optimalen Stichprobenumfänge (mit df = $p - 1$).

Um mittlere Veränderungen bei dreimaliger Untersuchung (df = 2) auf einem α-Niveau von .05 statistisch mit einer Teststärke von $1 - \beta = .80$ abzusichern, sollte eine Stichprobe mit $n = 27$ Untersuchungsobjekten dreifach untersucht werden. (Ohne Messwiederholungen wären – gem. Tab. 14.11 – immerhin $3 \times 52 = 156$ Untersuchungsobjekte erforderlich.)

Tabelle 14.13 Optimale Stichprobenumfänge der einfaktoriellen Varianzanalyse mit Messwiederholungen und $\overline{\rho} = .50$ nach Effektgrößen ($1 - \beta = .80$)

Freiheitsgrade	$\alpha = .01$			$\alpha = .05$		
	Klein	Mittel	Groß	Klein	Mittel	Groß
df = 1	293	49	20	197	33	14
df = 2	232	39	16	162	27	11
df = 3	195	33	14	138	23	10
df = 4	169	29	12	121	20	9
df = 5	150	26	11	108	18	8
df = 6	136	23	10	99	17	7

Mehrfaktorielle Varianzanalyse

Die Auswertung faktorieller Pläne erfolgt mit mehrfaktoriellen Varianzanalysen. Dieses Verfahren überprüft die in einer Untersuchung interessierenden Haupteffekte und Interaktionen. Erwartet man keine Interaktionen, werden die optimalen Stichproben zur Absicherung der Haupteffekte nach den Regeln für einfaktorielle Pläne bestimmt. Wenn hierbei in Abhängigkeit von den Haupteffekten unterschiedliche Gesamtstichprobenumfänge resultieren, entscheidet man sich im Regelfall für die größere Gesamtstichprobe, wodurch sich die Teststärke für Haupteffekte erhöht, für deren Absicherung eigentlich ein kleinerer Gesamtstichprobenumfang ausreichen würde. Entscheidet man sich für eine kleinere Stichprobe, sind Teststärkeeinbußen für diejenigen Effekte hinzunehmen, deren Absicherung größere Stichproben erforderlich machen.

Typischerweise ist man jedoch bei mehrfaktoriellen Plänen an Interaktionen interessiert und sollte deshalb die Festlegung des Stichprobenumfanges vom **erwarteten Interaktionseffekt** abhängig machen. Ausgehend von den in Tab. 14.11 unter Ziffer 7 für unterschiedliche Zählerfreiheitsgrade (df) genannten optimalen Stichprobenumfängen (n) errechnet sich der optimale Stichprobenumfang für eine Zelle des mehrfaktoriellen Planes wie folgt:

$$n_{\text{Zelle}} = \frac{(n-1)\cdot(\text{df}+1)}{\text{Anzahl der Zellen}} + 1\,. \qquad (14.62)$$

In einem dreifaktoriellen Plan mit p Stufen für Faktor A, q Stufen für Faktor B und r Stufen für Faktor C erhält man $p \times q \times r$ Zellen. Will man z. B. in einem $2 \times 3 \times 3$-Plan für die $A \times B$-Interaktion einen mittleren Effekt ($E = 0.25$) auf dem $\alpha = .05$-Niveau absichern, resultieren für Gl. 14.62

$$\text{df} = (2-1)\cdot(3-1) = 2$$
$$(= \text{Freiheitsgrade der fraglichen Interaktion})\,,$$

$n = 52$ (gem. Tab. 14.11 für df $= 2$, $\alpha = .05$ und einen mittleren Effekt). Anzahl der Zellen $= 2 \cdot 3 \cdot 3 = 18$ und damit

$$n_{\text{Zelle}} = \frac{(52-1)\cdot(2+1)}{18} + 1 = 9.5 \approx 10.$$

Man benötigt also pro Zelle 10 Untersuchungsobjekte bzw. eine Gesamtstichprobe von $n = 18 \cdot 10 = 180$ (diese und die folgenden Ausführungen gehen von gleich großen Stichproben pro Zelle aus).

Für die Absicherung eines mittleren Effektes ($E = 0.25$) für die Interaktion 2. Ordnung erhält man entsprechend (mit $\alpha = .05$):

$$\text{df} = (2-1)\cdot(3-1)\cdot(3-1) = 4\,,$$
$$n = 39\,,$$

Anzahl der Zellen $= 2 \cdot 3 \cdot 3 = 18$,

$$n_{\text{Zelle}} = \frac{(39-1)\cdot(4+1)}{18} + 1 = 11.5 \approx 12.$$

Als optimale Gesamtstichprobe wäre hier also $n = 18 \cdot 12 = 216$ anzusetzen.

Die optimalen Stichprobenumfänge pro Zelle für einige ausgewählte Versuchspläne enthält Tab. 14.14 Die Stichprobenumfänge orientieren sich jeweils an der **höchsten Interaktion**, also bei zweifaktoriellen $A \times B$-Plänen an der $A \times B$-Interaktion und bei dreifaktoriellen $A \times B \times C$-Plänen an der Interaktion 2. Ordnung ($A \times B \times C$).

Beispiel: In einem 3×4- (oder 4×3-)Plan soll für die $A \times B$-Interaktion eine mittlere Effektgröße mit $\alpha = .05$ und $1 - \beta = .80$ abgesichert werden. Hierfür sollten pro Zelle 19 bzw. insgesamt $12 \cdot 19 = 228$ Untersuchungsobjekte vorgesehen werden. Im 3×4-Plan würde der Haupteffekt A auf $4 \cdot 19 = 76$ Objekten pro A-Stufe und der Haupteffekt B auf $3 \cdot 19 = 57$ Objekten pro B-Stufe beruhen, d. h., der Stichprobenumfang wäre für den Haupteffekt A (df $= 2$) und auch für den Haupteffekt B (df $= 3$) ausreichend, um jeweils einen mittleren Effekt mit $\alpha = .05$ abzusichern ($n_{\text{opt}(A)} = 52$; $n_{\text{opt}(B)} = 45$ gem. Tab. 14.11).

Man beachte, dass die hier dargestellte Planung eines optimalen Stichprobenumfanges voraussetzt, dass mit der Varianzanalyse nur ***eine* spezifische Effekthypothese** geprüft werden soll. Die Stichprobenumfänge verändern sich (bei gleichem Signifikanzniveau und gleicher Teststärke) beträchtlich, wenn man z. B. nur daran interessiert

Tabelle 14.14 Optimale Stichprobenumfänge für einige mehrfaktorielle Versuchspläne nach Effektgrößen ($1 - \beta = .80$)

Versuchsplan	$\alpha = .01$			$\alpha = .05$		
	Klein	Mittel	Groß	Klein	Mittel	Groß
2 × 2	294	48	20	197	33	14
2 × 3	233	39	16	162	27	11
3 × 3	187	32	13	134	22	9
3 × 4	159	26	11	114	19	8
4 × 4	136	23	10	99	17	7
2 × 2 × 2	147	25	10	99	17	7
2 × 2 × 3	117	20	8	81	14	6
2 × 3 × 3	94	16	7	67	12	5
3 × 3 × 3	77	13	6	57	10	4
4 × 4 × 4	80	14	6	58	10	5

Tabelle 14.15 Optimale Stichprobenumfänge für einige mehrfaktorielle Messwiederholungspläne mit $\overline{\rho} = .50$ nach Effektgrößen ($1 - \beta = .80$)

Versuchsplan	$\alpha = .01$			$\alpha = .05$		
	Klein	Mittel	Groß	Klein	Mittel	Groß
2 × 2	147	25	11	99	17	8
2 × 3	117	20	9	82	14	6
3 × 3	94	17	7	68	12	6
3 × 4	80	14	6	59	11	5
4 × 4	68	13	6	50	9	5
2 × 2 × 2	74	13	6	50	9	5
2 × 2 × 3	59	11	5	42	8	4
2 × 3 × 3	48	9	4	35	7	4
3 × 3 × 3	39	8	4	29	6	3
2 × 3 × 4	41	8	4	30	7	3

ist, dass irgendein beliebiger Effekt signifikant wird oder wenn man begründen kann, dass alle Effekte signifikant werden müssten. Die hiermit verbundene Abhängigkeit der Teststärke von der Anzahl simultan durchgeführter Signifikanztests **(multiples Testen)** wird bei Maxwell (2004) untersucht.

Mehrfaktorielle Varianzanalyse mit Messwiederholungen

Hat man einen mehrfaktoriellen Untersuchungsplan mit Messwiederholungen, ergeben sich – wie bei einfaktoriellen Plänen mit Messwiederholungen – in Abhängigkeit von $\overline{\rho}$ gegenüber Tab. 14.14 Stichprobenersparnisse. Gehen wir erneut von $\overline{\rho} = .50$ aus, werden die unter Ziffer 7 in Tab. 14.5 genannten Effektgrößen zunächst durch $\sqrt{1 - 0.5}$ dividiert, um für die so korrigierten Effektgrößen die optimalen Stichprobenumfänge festzulegen. Die Resultate haben wir bereits in Tab. 14.13 kennen gelernt; die optimalen Stichprobenumfänge für mehrfaktorielle Pläne ergeben sich hieraus über Gl. 14.62.

Die Ergebnisse für einige ausgewählte Pläne fasst Tab. 14.15 zusammen. Bei $A \times B$-Plänen ist A der Gruppierungsfaktor und ***B* der Messwiederholungsfaktor**, und bei $A \times B \times C$-Plänen sind A und B die Gruppierungsfaktoren mit ***C* als Messwiederholungsfaktor**. Die Stichprobenumfänge orientieren sich erneut jeweils an der höchsten Interaktion (Interaktion 1. Ordnung bei zweifaktoriellen und Interaktion 2. Ordnung bei dreifaktoriellen Plänen) und gelten für jede Stufe des Gruppierungsfaktors oder – bei dreifaktoriellen Plänen – für jede Kombination der gruppenbildenden Faktorenstufen.

Beispiel: Für einen 2 × 2-Plan (z. B. Pre-/Posttest-Plan mit Experimental- und Kontrollgruppe) würde man 17 Untersuchungsteilnehmer für die Kontrollgruppe und 17 Untersuchungsteilnehmer für die Experimentalgruppe benötigen, wenn ein mittlerer Interaktionseffekt mit $\alpha = .05$ abgesichert werden soll. Da der Gruppierungsfaktor A von der Messwiederholung nicht „profitiert“ (er wird an der Streuung innerhalb der Gruppen bzw. an $\hat{\sigma}^2_{\text{in } S}$ getestet), reicht dieser Stichprobenumfang nur aus, um einen „sehr“ großen Effekt bzgl. Faktor A abzusichern ($17 < 26 = n_{\text{opt}(A)}$ für einen großen Effekt gemäß Tab. 14.11). Der Messwiederholungsfaktor B hingegen basiert pro Stufe auf $2 \times 17 = 34$ Untersuchungsteilnehmern, was zur Absicherung eines mittleren Effektes ausreicht (Tab. 14.13).

Für einen 2 × 2 × 3-Plan (z. B. Kontrollgruppe vs. Experimentalgruppe als Faktor A und männlich vs. weiblich als Faktor B mit 3 Messungen pro $A \times B$-Kombination) benötigt man 4 Stichproben à 59 Untersuchungsteilnehmer, wenn ein kleiner Effekt für die Interaktion 2. Ordnung auf dem $\alpha = .01$-Niveau abgesichert werden soll. Die Haupteffekte A und B basieren damit pro Stufe jeweils auf $2 \times 59 = 118$ Untersuchungsteilnehmern, was zur Absicherung mittlerer Effekte ausreicht ($n_{\text{opt}(A)} = n_{\text{opt}(B)} = 95$ für $df = 1$, $\alpha = .01$ und mittlerem Effekt gem. Tab. 14.11 für einfaktorielle Pläne). Für die $A \times B$-Interaktion hat man pro Gruppe 59 Un-

Tabelle 14.16 Optimale Stichprobenumfänge und multiple Korrelationen (in Klammern) in Abhängigkeit von der durchschnittlichen Multikollinearität ($\overline{\rho}_{xx}$) und der durchschnittlichen Validität ($\overline{\rho}_{xy}$) (5 Prädiktorvariablen, $1-\beta = .80, \alpha = .05$). (Mod. nach Maxwell, 2000, Tab. 1 und 2. Copyright © 2000 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

		$\overline{\rho}_{xy}$		
		.20	.30	.40
$\overline{\rho}_{xx}$	.30	1 070 (.30)	419 (.45)	191 (.60)
	.40	1 731 (.28)	692 (.42)	328 (.55)
	.50	2 752 (.26)	1 117 (.39)	544 (.52)

Tabelle 14.17 Optimale Stichprobenumfänge für die multiple Korrelation in Abhängigkeit von der Anzahl der Prädiktoren ($\overline{\rho}_{xx} = \overline{\rho}_{xy} = .30; 1-\beta = .80; \alpha = .05$). (Mod. nach Maxwell, 2000, Tab. 5. Copyright © 2000 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

Anzahl der Prädiktoren	Optimaler Stichprobenumfang
2	141
3	218
4	311
5	419
6	543
7	682
8	838
9	1 009
10	1 196

tersuchungsteilnehmer. Da diese Interaktion einer $A \times B$-Interaktion in einem 2×2-Plan ohne Messwiederholungen entspricht, entnimmt man Tab. 14.14 $n_{\text{opt}(A\times B)} = 48$ für einen mittleren Effekt und $\alpha = .01$. Der Stichprobenumfang $n = 59$ ist also zur Absicherung eines mittleren bis kleinen Effektes ausreichend.

Für die $A \times C$- (bzw. $B \times C$-)Interaktion stehen pro Faktorstufenkombination $2 \times 59 = 118$ Messungen zur Verfügung. Dieser Wert ist mit dem optimalen Stichprobenumfang für eine Interaktion in 2 × 3-Plänen mit Messwiederholungen zu vergleichen. Wir entnehmen Tab. 14.15 $n_{\text{opt}} = 117$ für die Absicherung eines kleinen Interaktionseffektes mit $\alpha = .01$, d. h., mit 59 Untersuchungsteilnehmern pro Gruppe können ebenfalls kleine Effekte für die $A \times C$- und die $B \times C$-Interaktion mit $\alpha = .01$ abgesichert werden.

Man beachte, dass diese Stichprobenumfänge für $\overline{\rho} = .50$ gelten. Muss man mit einer geringeren Durchschnittskorrelation rechnen, sind größere Stichprobenumfänge anzusetzen. Wenn man ungeachtet der Höhe von $\overline{\rho}$ sichergehen will, dass mindestens mit einer Teststärke von .80 geprüft wird, sind die in Tab. 14.14 genannten Stichprobenumfänge einzusetzen.

Multiple Korrelation

Die optimalen Stichprobenumfänge für die multiple Korrelation hängen von α und $1-\beta$ ab sowie vom Populationseffekt K^2. Der Populationseffekt ist dabei sowohl von der Höhe der Interkorrelationen der Prädiktorvariablen **(Multikollinearität)** als auch von den Korrelationen der Prädiktorvariablen mit der Kriteriumsvariablen **(Validitäten)** abhängt. Vernünftige Vorabschätzungen des zu erwartenden Populationsparameters K^2 setzen also voraus, dass zumindest die Größenordnung von $\frac{p\cdot(p+1)}{2}$ bivariaten Korrelationen bekannt ist (p = Anzahl der Prädiktoren).

Wie stark der optimale Stichprobenumfang von der Größe dieser Parameter abhängt, verdeutlicht eindrucksvoll Tab. 14.16.

Die optimalen Stichprobenumfänge schwanken also zwischen 191 ($\rho_{xx} = .30$; $\rho_{xy} = .40$) und 2 752 ($\rho_{xx} = .50$; $\rho_{xy} = .20$). Der Tabelle ist zu entnehmen, dass die Stichproben mit größer werdender Validität und mit sinkender Multikollinearität kleiner werden.

Wenn man annehmen kann, dass alle bivariaten Korrelationen im Durchschnitt ($\overline{\rho}_{xx}$ und $\overline{\rho}_{xy}$) Werte von .30 aufweisen (mittlerer Korrelationseffekt gem. Tab. 14.5), sind die in Tab. 14.17 genannten Stichprobenumfänge in Abhängigkeit von der Anzahl der Prädiktorvariablen optimal.

Aus Tab. 14.16 und Tab. 14.17 lassen sich die folgenden allgemeinen Richtlinien für die Planung von multiplen Korrelationsstudien ableiten. Der **optimale Stichprobenumfang für die multiple Korrelation wird kleiner**, wenn

- die Prädiktorvariablen möglichst hoch mit der Kriteriumsvariablen korrelieren (hohe Validitäten),
- die Prädiktorvariablen-Interkorrelationen möglichst niedrig sind (geringe Multikollinearität),
- die Anzahl der Prädiktorvariablen möglichst klein ist.

Konfidenzintervalle sollten auch für β-Gewichte im Rahmen der **multiplen Regressionsrechnung** bestimmt werden. In Verbindung mit optimalen Stichprobenumfängen kann hierbei die Frage interessieren, wie groß der Stichprobenumfang sein muss, damit die Breite eines Konfi-

denzintervalls einem vorgegebenen Wert entspricht. Wie dieser Stichprobenumfang berechnet wird, beschreiben Kelley und Maxwell (2003).

14.4 Beispiele für die Planung optimaler Stichprobenumfänge und das Berichten von Effektgrößen

Hinsichtlich der inferenzstatistischen Auswertung hypothesenprüfender quantitativer Studien halten wir folgende Schritte für jeden einzelnen zu untersuchenden Effekt bzw. jede einzelne Forschungshypothese für unabdingbar:

- **In der Planungsphase:**
  - Festlegung von Signifikanzniveau (α) und Teststärke ($1-\beta$) (Empfehlung: $\alpha = .05$ und $1-\beta = .80$);
  - Festlegung einer Forschungs-/Alternativhypothese (Empfehlung: gerichtete H_1) und Angabe der erwarteten Effektgröße nach Maßgabe von Tab. 14.5 (Empfehlung: Im Zweifelsfall gehen wir bei mangelnden konkreten Vorkenntnissen zur fraglichen Effektgröße in den Sozialwissenschaften von kleinen bis mittleren Effekten aus, große Effekte sind selten);
  - Festlegung der zur H_1 passenden zu prüfenden Nullhypothese: traditionelle Nil-Nullhypothese (H_{00}), die gar keinen Effekt in der Population postuliert, oder Minimum-Effekt-Nullhypothese, die von einem vernachlässigbar kleinen Effekt in der Population bis maximal 1 % oder maximal 5 % Varianzaufklärung ausgeht (H_{01} bzw. H_{05}, ▶ Kap. 15; konventionell wird meist mit der H_{00} gearbeitet);
  - Festlegung des optimalen Stichprobenumfanges (n_{opt}), der mindestens notwendig ist, um bei festgelegter Effektgröße mindestens eine Teststärke von .80 zu erzielen (▶ Abschn. 14.3).
- **Für die Ergebnisdarstellung:**
  - Darstellung der Resultate des Signifikanztests (Teststatistik, Irrtumswahrscheinlichkeit) für jede Hypothese;
  - jeweils auch Darstellung der ermittelten unstandardisierten bzw. standardisierten Effektgröße ergänzend zur Signifikanzaussage in Form einer Punktschätzung sowie möglichst auch einer Intervallschätzung mittels Konfidenzintervall (Empfehlung: 95 %iges Konfidenzintervall der Effektgröße). Art und Umfang der Darstellung der Effektgrößen sollten möglichst so gewählt werden, dass die Studie alle für eine Metaanalyse notwendigen Angaben enthält (zur Metaanalyse ▶ Kap. 16);
  - die Angabe von Konfidenzintervallen für die standardisierten Effektgrößenmaße wird zwar vom APA Publication Manual inzwischen gefordert (APA, 2009), hat sich in der Forschungsliteratur bislang aber noch nicht stark etabliert (Fritz et al., 2012). Dies liegt auch daran, dass Softwaretools zur Bestimmung von Konfidenzintervallen für standardisierte Effektgrößen bislang nur begrenzt zur Verfügung stehen (▶ Hinweise zur Durchführung von Teststärkeanalysen).

Mit der Darstellung des Konfidenzintervalls erübrigt sich eigentlich die Darstellung des Signifikanztestergebnisses, denn wenn sich – bei Prüfung der Nil-Nullhypothese H_{00} – der Wert Null im Konfidenzintervall befindet, kann die Nullhypothese nicht abgelehnt werden (▶ Abschn. 12.5.1). Bei einem signifikanten Ergebnis hingegen befindet sich der Wert Null außerhalb des Konfidenzintervalls. Dies gilt analog für die Prüfung von Minimum-Effekt-Nullhypothesen H_{01} und H_{05}, wenn man die Grenzen des Konfidenzintervalls (z. B. für ρ oder δ) gem. Gl. 15.1 in η^2-Werte transformiert. Befindet sich $\eta^2 = .01$ (.05) nicht im Konfidenzintervall, kann die H_{01} (H_{05}) abgelehnt werden (zu Minimum-Effektgrößen-Tests ▶ Kap. 15).

Wir empfehlen jedoch, auf die Darstellung des Signifikanztestergebnisses nicht zu verzichten, auch wenn zusätzlich das Effektgrößen-Konfidenzintervall berichtet wird. Überlegungen zum Signifikanztest sind essenziell für die Untersuchungsplanung, weil in dieser Phase (mit Festlegung von α, $1-\beta$, der Effektgröße und der Art der Nullhypothese) der optimale Stichprobenumfang festgelegt wird (oder werden sollte). Zwar könnte man den optimalen Stichprobenumfang auch über die maximal tolerierbare Breite des Konfidenzintervalls festlegen (Kelley & Maxwell, 2003), denn mit größer werdendem Stichprobenumfang verringert sich die Konfidenzintervallbreite. Diese Denkweise ist jedoch bislang unüblich.

Im Folgenden sollen Planung und Auswertung von (fiktiven) Untersuchungen aus der Evaluationsforschung (▶ Kap. 18) demonstriert werden. Da wir nicht davon ausgehen, dass sich die Forschungspraxis „von heute auf morgen“ ändern wird, basieren die Untersuchungsplanungen auf der Überprüfung der traditionellen Nullhypothese (H_{00}). Dennoch wird in jedem Falle überprüft, ob die jeweilige Untersuchung auch geeignet ist, die Minimum-Effekt-Nullhypothese H_{01} (ggf. sogar die H_{05}) zu verwerfen. Hierbei muss man jedoch in Rechnung stellen, dass die Teststärke mit Stichprobenumfängen, die für die Prüfung von H_{00} optimal sind (mit $1-\beta = .80$), für die Überprüfung von H_{01} suboptimal bzw. mit einer Teststärke $1-\beta < .80$ verbunden sind. Erneut folgen wir der in Tab. 14.5 vorgegebenen Gliederung der Signifikanztests (ausführliche Informationen zur Planung hypothesenprüfender Untersuchungen findet man auch bei Hager, 2004)

14.4.1 Vergleich von zwei Mittelwerten

t-Test für unabhängige Stichproben

Eine Schulpsychologin möchte die Effektivität eines neu auf dem Markt erschienenen Lernspiels für den Englischunterricht evaluieren. Hierzu will sie eine herkömmlich unterrichtete und eine nach der neuen Methode unterrichtete Schülerstichprobe vergleichen. Die Planung sieht vor, zwei gleich große Stichproben (Experimental- und Kontrollgruppe) durch Randomisierung zusammenzustellen. Den Lernerfolg operationalisiert die Evaluatorin durch die Fehleranzahl in einem Testdiktat.

Da man bislang noch keine Erfahrungen mit der neuen Methode gemacht hat, entscheidet sich die Evaluatorin für einen mittleren Effekt ($\delta = 0.5$ gem. Tab. 14.5), d. h., sie erwartet, dass die durchschnittliche Leistung der mit dem Lernspiel lernenden Schüler um mindestens eine halbe Streuungseinheit der Fehlerzahlen unter dem Durchschnittswert der Kontrollgruppe liegt. Da die neue Methode nach erfolgreicher Evaluation der zuständigen Schulbehörde empfohlen werden soll, ist die Evaluatorin vorsichtig und will nur mit einer Wahrscheinlichkeit von höchstens $\alpha = .01$ fälschlicherweise für die Überlegenheit der neuen Methode plädieren. Eine Teststärke von $1 - \beta = .80$ erscheint ihr angemessen.

Ausgerüstet mit diesen Informationen entnimmt sie Tab. 14.11, dass pro Gruppe 82 Schüler untersucht werden sollen. Mit **G*Power** wird derselbe optimale Stichprobenumfang von $n = 82$ pro Gruppe bestimmt (Abb. 14.5). Dazu wird zunächst die Testfamilie (t-Tests), dann der konkrete Test (Mittelwertvergleich bei unabhängigen Stichproben) und schließlich die Art der Teststärkeanalyse (A priori) ausgewählt, bevor die Inputparameter (einseitiger Test; Effektgröße $d = 0.5$; Signifikanzniveau $\alpha = .01$; Teststärke $1 - \beta = .80$; Verhältnis der Gruppengrößen $N1/N2 = 1$, d. h. gleichgroße Gruppen) eingegeben werden.

Die statistische Auswertung der Untersuchung führt zu $\overline{x}_{Exp} = 14.50$ und $\overline{x}_{Kon} = 18.00$ mit $\hat{\sigma}_{Exp} = 7.00$ sowie $\hat{\sigma}_{Kon} = 6.00$. Über Gl. 14.3 wird die Merkmalsstreuung mit $\sigma = 6.52$ geschätzt. Dieser Wert führt über $\hat{\delta} = \frac{(\overline{x}_{Kon} - \overline{x}_{Exp})}{\hat{\sigma}}$ zu einer Effektgrößenschätzung von $\hat{\delta} = 0.54$.

Konfidenzintervall Das 95 %ige **Konfidenzintervall** der standardisierten Populationseffektgröße hat nach den Ausführungen in ▶ Abschn. 14.2.2 die Grenzen $\delta_u = 0.22$ und $\delta_o = 0.85$: $0.22 \leq \delta \leq 0.85$

Das Konfidenzintervall kann mittels des Wilson-Online-Rechners (▶ Hinweise zur Durchführung von Teststärkeanalysen) bestimmt werden: Zunächst muss dazu mittels G*Power der t-Wert berechnet werden. Hierbei werden die gleichen Einstellungen wie zur Berechnung

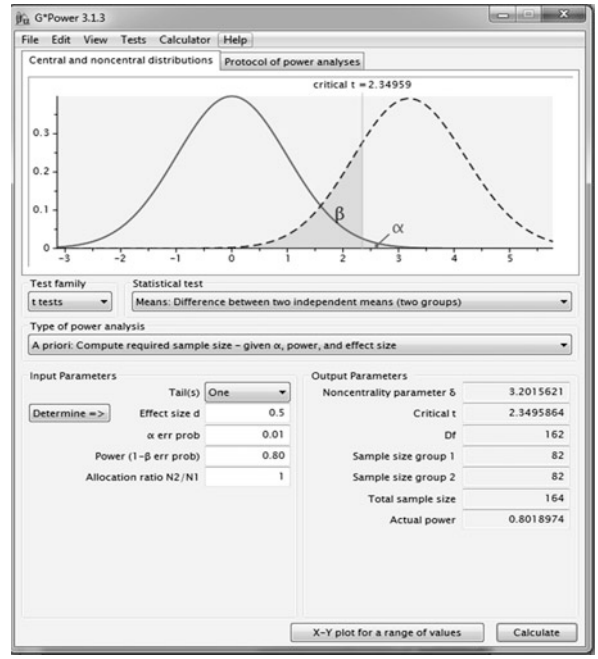


Abb. 14.5 Durchführung einer A-priori-Teststärkeanalyse beim t-Test für unabhängige Stichproben mit G*Power (▶ Hinweise zur Durchführung von Teststärkeanalysen). (http://www.gpower.hhu.de. Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. Mit freundlicher Genehmigung)

von n_{opt} gewählt, jedoch unter der Art der Teststärkeanalyse „Post hoc" eingegeben und das Feld „Determine" angeklickt. Darauf öffnet sich das linke Dialogfenster, in dem die Mittelwerte und Standardabweichungen der beiden Gruppen eingegeben werden. Durch einen Klick auf das Feld „Calculate and transfer to main window" werden die Werte berechnet und ins Hauptfenster übertragen. Nach Gl. 14.6 entspricht der Nonzentralitätsparameter δ dem nichtzentralen t-Wert ($t = 3.438$, Abb. 14.6).

Dieser Wert wird in den Online-Rechner eingegeben, ebenso wie der Gesamtstichprobenumfang. Durch das Anklicken des Feldes „Calculate" werden der geschätzte δ-Wert (im Dialogfeld bezeichnet als „d") sowie die Grenzen des Konfidenzintervalls angegeben (Abb. 14.7).

Gemäß t-Test ist die gefundene Differenz der Mittelwerte auf dem $\alpha = .01$-Niveau signifikant, d. h., die Evaluatorin kann der Schulbehörde guten Gewissens die neue Methode empfehlen. Allerdings ist es – entgegen der ursprünglichen Annahme – durchaus möglich, dass der „wahre" Effekt ein kleiner Effekt ist ($\delta = 0.2 \approx 0.22$) und kein mittlerer ($\delta = 0.5$).

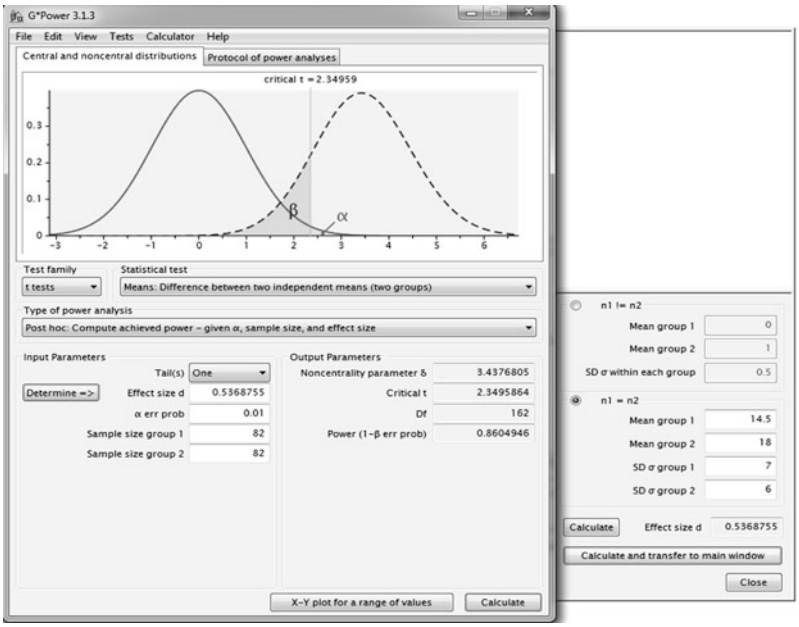


Abb. 14.6 Bestimmung des nichtzentralen *t*-Werts beim *t*-Test für unabhängige Stichproben mit G*Power (► Hinweise zur Durchführung von Teststärkeanalysen). (http://www.gpower.hhu.de. Faul et al., 2007, Faul et al., 2009. Mit freundlicher Genehmigung)

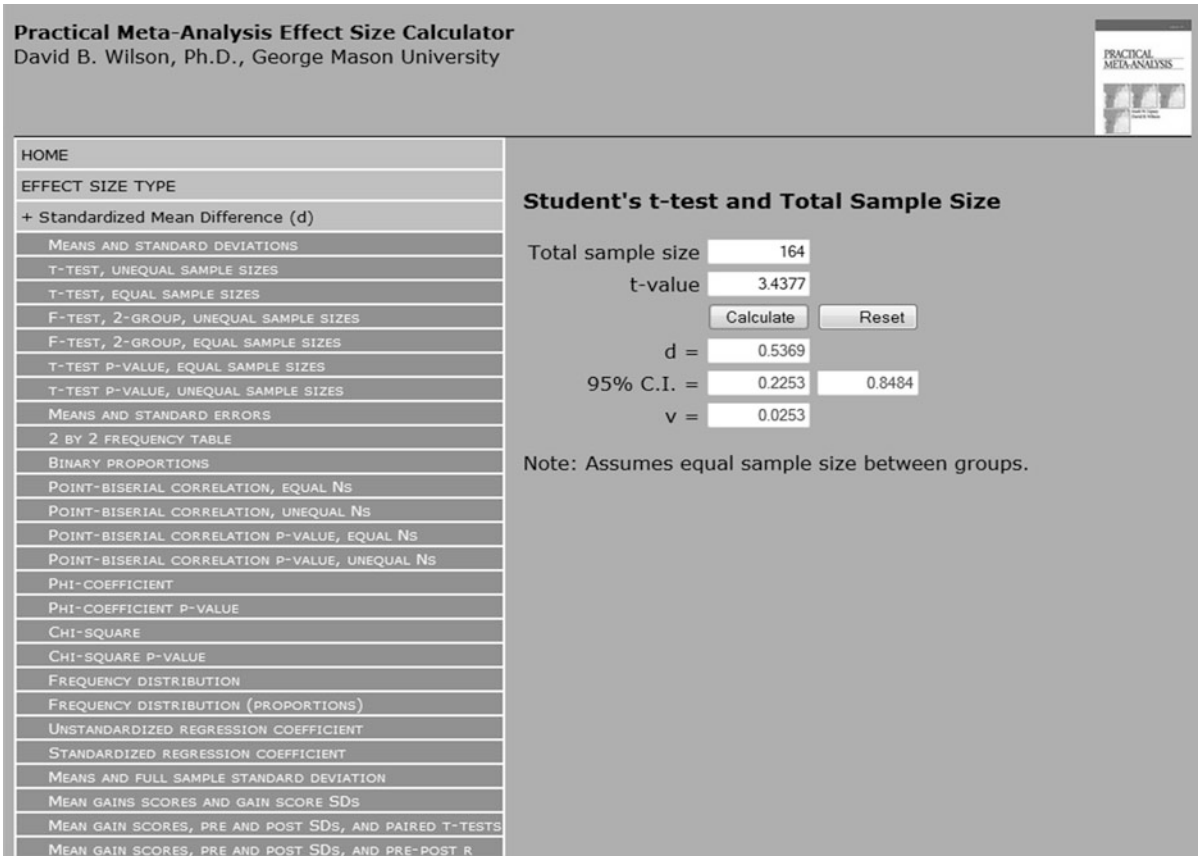


Abb. 14.7 Berechnung des Konfidenzintervalls für die Populationseffektgröße δ bzw. *d* mit dem Wilson-Online-Rechner (► Hinweise zur Durchführung von Teststärkeanalysen). (Mit freundlicher Genehmigung von David B. Wilson, http://campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php)

Prüfung von H_{01} Kann mit dem erzielten Ergebnis auch die Minimum-Effekt-Nullhypothese H_{01} verworfen werden? Wir transformieren $\hat{\delta} = 0.54$ über Gl. 6 in Tab. 15.4 in ▶ Kap. 15 in ein F-Äquivalent:

$$F_{(1.162)} = \frac{0.54^2 \cdot 162}{4} = 11.81 .$$

Tabelle 11 im ▶ Webangebot zu diesem Buch entnehmen wir (für $df_N = 150$ ohne Interpolation), dass die H_{01} für $\alpha = .05$ zu verwerfen wäre ($F_{crit} = 8.61$), aber nicht für $\alpha = .01$ ($F_{crit} = 13.04$). Diese Angaben gelten für den zweiseitigen Test.

t-Test für abhängige Stichproben

Nach einer schweren Flutkatastrophe registrieren Vertreter der evangelischen Kirche eine Zunahme der Gottesdienstbesuche. Sie beauftragen ein demoskopisches Institut zu überprüfen, ob diese Veränderung durch Zufall zu erklären sei oder ob die Flutkatastrophe eine verstärkte Hinwendung zur Kirche bewirkt haben könnte.

Nach einem ersten Kontaktgespräch mit den Auftraggebern schlägt der Evaluator vor, von einem kleinen Effekt auszugehen. Man einigt sich ferner auf $\alpha = .05$ und $1 - \beta = .80$. Aus Statistiken über die Frequenzen sonntäglicher Kirchenbesuche errechnet der Evaluator, dass die Kirchenbesuche in einer kleinen Zufallsauswahl von Gemeinden in einem 8-wöchigen Intervall von Sonntag zu Sonntag im Durchschnitt zu $\overline{r} = .65$ korrelieren. Mit diesen Angaben entnimmt er Tab. 14.12, dass eine Zufallsstichprobe von ca. $n = 125$ von der Sturmflut betroffener Gemeinden für die Untersuchung ausreichen müsste (die Tabelle weist n_{opt} für $r = .60$ aus, da die Korrelation jedoch mit $r = .65$ etwas höher liegt, ist n_{opt} etwas geringer anzusetzen). Den genauen Wert errechnet der Evaluator über Gl. 14.58 und Gl. 14.59:

$$\hat{\delta}_{äquiv} = \frac{0.2}{\sqrt{1 - .65}} = 0.3381 ,$$

$$n_{opt} = \frac{1.237}{100 \cdot 0.3381^2} + 1 = 109.2 \approx 109 .$$

Zu vergleichen ist pro Gemeinde die Anzahl der Gottesdienstbesuchenden an 4 Sonntagen vor der Katastrophe mit der entsprechenden Anzahl danach. Berechnet man den optimalen Stichprobenumfang mit G*Power, ermittelt man $n = 110$ (test family: t-tests, statistical test; means: difference between two dependent means [matched pairs], type of power analysis: a priori; tails: one; effect size $dz = 0.239$; $\alpha = .05$; Teststärke $1 - \beta = .80$). Die dabei von G*Power genutzte Effektgröße $dz = 0.239$ wird berechnet, indem man unter „determine“ → „from group parameters“ die Korrelation .65 zwischen den Gruppen eingibt sowie ein beliebiges Muster von Gruppenmittelwerten und Gruppenstreuungen, das der vorgegebenen Effektgröße von $d = 0.20$ entspricht (z. B. Mean 1 = 5, Mean 2 = 4, SD1 = 5, SD2 = 5).

Nach Abschluss der Untersuchung werden $\overline{x}_{vor} = 166$, $\overline{x}_{nach} = 180$ und $\hat{\sigma}_D = 35$ errechnet, was nach Gl. 14.8) zu $\hat{\delta}' = \frac{14}{35} = 0.40$ führt. Der t-Test für abhängige Stichproben bzw. Beobachtungspaare (Bortz & Schuster, 2010, S. 124ff.) ergibt mit df = 124 einen signifikanten t-Wert ($t = 4.47$).

Konfidenzintervall Nach den Ausführungen in ▶ Abschn. 14.2.3 ermittelt man für Variante a auf der Basis der Streuung der Differenzwerte ein Konfidenzintervall für die standardisierte Populationseffektgröße von $0.22 \leq \delta' \leq 0.58$.

Der „wahre“ Effekt δ' befindet sich nach dieser Berechnung also mit hoher Konfidenz (95 %) in einem Intervall oberhalb des ursprünglich angenommenen kleinen Effektes (0.20).

Für eventuelle spätere Metaanalysen wird auch noch das Konfidenzintervall für Variante b (Cohens d) auf der Basis der Merkmalsstreuung errechnet. Als Schätzung der Merkmalsstreuung ermittelt man 41.8 (über Gl. 14.10 mit $r = .65$), d. h., man errechnet

$$\hat{\delta} = \frac{14}{41.8} = 0.33 .$$

Für das approximative Konfidenzintervall dieser Effektgröße benötigen wir den Standardfehler von $\hat{\delta}$ nach Gl. 14.13):

$$\hat{\sigma}_{\hat{\delta}} = \sqrt{\frac{0.33^2}{2 \cdot (125 - 1)} + \frac{2 \cdot (1 - 0.65)}{125}}$$
$$= \sqrt{0.00044 + 0.0056} = 0.078 .$$

Damit erhält man über Gl. 14.14 folgendes 95 %ige Konfidenzintervall für die standardisierte Populationseffektgröße:

$$KI_\delta = 0.33 \pm 1.96 \cdot 0.078 = 0.33 \pm 0.15 \quad \text{bzw.}$$
$$KI(95\,\%) = 0.18 \leq \delta \leq 0.48 .$$

Der wahre Effekt kann nach dieser Berechnung mit einer Konfidenz von 95 % also auch dem vorgegebenen kleinen Effekt ($\delta = 0.20$) entsprechen. Dass die erste Berechnungsform gemäß Konfidenzintervall einen etwas stärkeren Effekt in der Population schätzt, liegt an der hohen Korrelation ($r = .65$) der beiden abhängigen Messwertreihen (Vorher-Nachher-Messung pro Kirche), die sich in einer geringeren Differenzwert-Streuung im Vergleich zur Merkmalsstreuung widerspiegelt. Beide Berechnungsformen kommen bei $r = .50$ zu identischen Resultaten.

Bei der Ergebnispräsentation weist der Evaluator schließlich zu Recht darauf hin, dass die interne Validität der Untersuchung nicht überschätzt werden dürfe, da auf die parallele Untersuchung einer Kontrollgruppe (Gemeinden aus Gebieten, die nicht von der Flutkatastrophe betroffen waren) verzichtet wurde.

Prüfung von H_{01} Über Gl. 7 in ◘ Tab. 15.4 in ▶ Kap. 15 errechnet man

$$F_{(1,124)} = \frac{0.33^2 \cdot 124}{4 \cdot \sqrt{1 - 0.65}} = 5.71 \; .$$

Dieser Wert ist gem. Tab. 11 im ▶ Webangebot zu diesem Buch kleiner als der kritische Wert für $\alpha = .05$ ($F_{crit} = 7.76$ für $df_N = 120$), d. h., die H_{01} kann nicht verworfen werden.

14.4.2 Bivariate Korrelation

Die zahnärztliche Kassenvereinigung will in Erfahrung bringen, ob es sich lohnt, unter Schulkindern eine Aufklärungsbroschüre über Mundhygiene zu verteilen. Da diese Schrift vermutlich primär die Einstellung der Kinder zur Mundhygiene verändert, ist man daran interessiert, in einer Pilotstudie die Einstellungen bzgl. Zahnpflege und Mundhygiene mit der tatsächlich für die Zahnpflege aufgewendeten Zeit in Beziehung zu setzen. Der Zusammenhang soll mit dem Signifikanztest für eine Produkt-Moment-Korrelation statistisch überprüft werden. Der Einsatz der Broschüre – so wird argumentiert – sei nur dann sinnvoll, wenn zwischen den Einstellungen und dem tatsächlichen Verhalten ein statistisch bedeutsamer Zusammenhang besteht. Hinsichtlich der Altersgruppe will man sich auf 12- bis 14-jährige Kinder (6.–8. Klassenstufe) konzentrieren, dem Lebensalter, in dem alle Milchzähne durch das bleibende Gebiss ersetzt sind und der Übergang von einer Junior- zu einer Erwachsenenzahnpasta empfohlen wird.

Bezüglich der Höhe der Korrelation ist man anspruchslos, denn bereits geringfügige Verbesserungen in der Mundhygiene, mit denen bei einer geringen Korrelation zu rechnen ist, können hochgerechnet auf die Population aller 12- bis 14-jährigen Schulkinder den gesamten aktuellen und zukünftigen Behandlungsaufwand und die damit verbundenen Kosten für die Krankenkassen erheblich verringern (die exakte Kosten-Nutzen-Bilanz der Maßnahme wäre mit einer ökonomischen Evaluation zu bestimmten, ▶ Abschn. 18.2.1). Deswegen hält man bereits einen kleinen Effekt ($\rho = .10$) für praktisch bedeutsam. Mit $\alpha = .01$ und $1 - \beta = .80$ wird gem. ◘ Tab. 14.11 geplant, den Zusammenhang von Einstellung und Verhalten an einer Stichprobe von $n = 1\,000$ Schulkindern der 6.–8. Klassenstufe unterschiedlicher Schultypen zu überprüfen. Mittels G*Power wird $n_{opt} = 997$ berechnet (test family: t-tests; statistical test: correlation: point biserial model, type of power analysis: A priori; tail: one, effect size: $\rho = .10$; α: .01; $1 - \beta$: .80).

Die Untersuchung führt zu der signifikanten Korrelation von $r = .48$, also einer Korrelation, die nahezu einem großen Effekt entspricht. Möglicherweise hätte man in der Planungsphase mehr Wert auf die Recherche vergleichbarer Untersuchungen legen sollen. Hätte sich hierbei herausgestellt, dass Korrelationen in dieser Größenordnung zu erwarten sind, wäre eine erhebliche Einsparung möglich gewesen, denn statt der untersuchten 1 000 Kinder wären dann gem. ◘ Tab. 14.11 ca. 40 Schülerinnen und Schüler für einen Signifikanznachweis ausreichend gewesen.

Konfidenzintervall Nach den Ausführungen in ▶ Abschn. 14.2.3 ermittelt man für die Korrelation folgendes 95 %ige Konfidenzintervall:

$$Z\,(r = .48) = 0.523,$$
$$\sigma_Z = \sqrt{\frac{1}{998 - 3}} = 0.032,$$
$$KI_Z = 0.523 \pm 1.96 \cdot 0.032 = 0.523 \pm 0.062 \; .$$

Die Grenzen für KI_Z heißen also 0.461 und 0.585. Man erhält über Tab. 9 ▶ Webangebot zu diesem Buch:

$$KI(95\,\%) = .43 \leq \rho \leq .53 \; .$$

Der Wilson-Online-Rechner (▶ Hinweise zur Durchführung von Teststärkeanalysen) berechnet für $r = .48$ und $n = 1\,000$ dasselbe Konfidenzintervall.

Angesichts der unerwartet hohen Korrelation beschließt man, die Broschüre herzustellen und unter Schulkindern zu verteilen.

Prüfung von H_{01} Man errechnet über Gl. 2 in ◘ Tab. 15.4 in ▶ Kap. 15

$$F_{(1,996)} = \frac{0.48^2 \cdot 996}{\left(1 - 0.48^2\right)} = 298.18 \; .$$

Nach Tab. 11 im ▶ Webangebot zu diesem Buch kann mit diesem Wert nicht nur die Minimum-Effekt-Nullhypothese H_{01} ($F_{crit} = 30.44$), sondern sogar die H_{05} mit $\alpha = .01$ verworfen werden ($F_{crit} = 92.43$).

14.4.3 Korrelationsdifferenz

Die Personalchefin einer großen Werbeagentur hat einen branchenspezifischen Kreativitätstest für die Bewerberauswahl entwickelt, der allerdings noch wenig tauglich ist, weil seine Testhalbierungsreliabilität (▶ Abschn. 10.4.1) nur $r_{tt(A)} = .54$ beträgt. Die Geschäftsleitung erwartet eine Revision der Testskala und fordert, dass die Endversion mindestens eine Reliabilität von $r_{tt(B)} = .80$ aufweist. Nach Überarbeitung des Tests überlegt die Personalchefin, an wie vielen Testpersonen sie die Reliabilität des revidierten Tests überprüfen soll. Der Reliabilitätszuwachs soll mit $\alpha = .05$ bei einer Teststärke von 80 % abgesichert werden.

Die Fishers-Z-Werte der Korrelationen lauten nach Tab. 9 im ▶ Webangebot zu diesem Buch $Z_A(r = .54) = 0.604$ und $Z_B(r = .80) = 1.099$, d. h., nach ◘ Tab. 14.5 resultiert eine Effektgröße von $Q = 1.099 - 0.604 = 0.495$ (das Vorzeichen von Q ist hier unerheblich). Dieser Wert entspricht nahezu exakt einem großen Effekt ($Q = 0.50$), für dessen Absicherung nach ◘ Tab. 14.11 pro Stichprobe (A und B) $n = 52$ Probanden benötigt werden. (Mittels G*Power wird $n_{opt} = 53$ ermittelt; test family: z tests, statistical test: correlations, two independent Pearson's r; type of power analysis: A priori; tails: one; effect size q: 0.50; $\alpha = .05$; $1 - \beta = .80$; allocation ratio: 1).

Nun hat die Personalchefin jedoch die Reliabilität der ersten Version ihres Tests nur für $n_A = 40$ Probanden ermittelt. Um den Korrelationsunterschied dennoch mit $\alpha = .05$ und $1 - \beta = .80$ absichern zu können, ist es erforderlich, für die zweite Stichprobe n_B mehr als 52 Probanden vorzusehen. Nach Gl. 14.60 wird ermittelt:

$$n_B = \frac{40 \cdot (52 + 3) - 6 \cdot 52}{2 \cdot 40 - 52 - 3} = 76 .$$

Nach Abschluss der Studie errechnet die Personalchefin eine Reliabilität von $r_{tt(B)} = .72$. Die Erhöhung der Reliabilität $r_{tt(A)} = .54$ auf $r_{tt(B)} = .72$ erweist sich als nicht signifikant. Da dieser Erhöhung ein mittlerer Effekt entspricht ($Z_A = 0.604$; $Z_B = 0.908$; $Q = Z_B - Z_A = 0.304 \approx 0.30$), wären – bei gleicher Verteilung – $n_A = n_B = 140$ Testpersonen erforderlich gewesen, um den Reliabilitätszugewinn mit $\alpha = .05$ statistisch absichern zu können. Da jedoch $n_A = 40$ für die erste Testform bereits festliegt, wird probeweise über Gl. 14.60 errechnet, an wie vielen Probanden die zweite Testform hätte geprüft werden müssen, um einen mittleren Effekt mit $\alpha = .05$ und $1 - \beta = .80$ abzusichern. Hierbei stellt sich leider heraus, dass dieser Stichprobenumfang nicht existiert (der Nenner in Gl. 14.60 wird negativ).

Es ist somit empfehlenswert, den Test inhaltlich zu verbessern und seine Reliabilität erneut zu prüfen, zumal die Geschäftsleitung ohnehin nur an einem Messinstrument mit $r_{tt} \geq .80$ interessiert ist.

Konfidenzintervall Das 95 %ige Konfidenzintervall für $q = 0.304$ ergibt sich wie folgt:

Für $n_A = 40$ und $n_B = 76$ berechnet sich nach Gl. 14.17

$$\sigma_q = \sqrt{\frac{1}{40 - 3} + \frac{1}{76 - 3}} = 0.202 ,$$

so dass nach Gl. 14.18

$$KI_Q = 0.304 \pm 1.96 \cdot 0.202 = 0.304 \pm 0.400 .$$

Dieses Intervall befindet sich zwischen den Grenzen $Z_u = -0.096$ und $Z_o = 0.704$. Transformiert in Korrelationsäquivalente (Tab. 9 im ▶ Webangebot zu diesem Buch) erhält man

$$KI(95\,\%) = -0.096 \leq Q \leq 0.61 .$$

Das Intervall umschließt eine Korrelationsdifferenz von 0 und bestätigt damit den bereits erwähnten Tatbestand, dass die Reliabilitätserhöhung nicht signifikant ist.

Prüfung von H_{01} Da in der Untersuchung nicht einmal die Nil-Nullhypothese H_{00} verworfen werden konnte, kann die Minimum-Effekt-Nullhypothese H_{01} erst recht nicht verworfen werden. Eine Gleichung für die Bestimmung eines F-Äquivalentes des Signifikanztests für Korrelationsdifferenzen ist in ◘ Tab. 15.4 in ▶ Kap. 15 nicht enthalten. Eine entsprechende Gleichung ergibt sich jedoch aus dem Signifikanztest für Korrelationsdifferenzen (Bortz & Schuster, 2010, S. 166, Gl. 10.17) und einer Transformation des resultierenden z-Wertes in einen F-Wert nach der Beziehung $z^2 = F_{(1,\infty)}$ (Bortz & Schuster, 2010, S. 77).

14.4.4 Differenz eines Anteilswertes von .50

Ein pharmazeutischer Konzern hat ein „sanftes" blutzuckersenkendes Mittel entwickelt, dessen Wirksamkeit in einem Feldversuch evaluiert werden soll. Der Konzernleitung ist sehr daran gelegen, dass die Studie eine „signifikante Wirkung" des Medikaments nachweist, weil diese Qualifikation für den späteren Verkaufserfolg von großer Bedeutung sei.

Die biometrische Abteilung plant, in einem Großversuch Proben des Medikaments über Arztpraxen an Diabetes-Patienten (Typ IIA) verteilen zu lassen. Die Patienten erhalten außerdem 2 Harnteststreifen, mit denen der

Zuckergehalt vor und nach Medikamenteneinnahme geprüft werden soll. Die Instruktion für die Patienten weist u. a. darauf hin, dass man anhand der Einfärbung der Teststreifen den Zuckergehalt feststellen kann. Die Patienten werden gebeten, auf einem vorgefertigten Kontrollzettel zu markieren, ob der Zuckergehalt abgenommen (−) bzw. zugenommen hat (+) oder ob keine Veränderung der Einfärbung festzustellen ist (0). Für die Mitwirkung an der Untersuchung erhalten die Patienten ein Aufwandsentschädigung von 20 €.

Der Großversuch soll mit einer repräsentativen Stichprobe von $n = 25\,000$ Patienten (realisiert als Klumpenstichprobe aus der Population bundesdeutscher Arztpraxen) durchgeführt werden. Die Nullhypothese (das Medikament hat keine Wirkung bzw. positive und negative Veränderungen sind mit $\pi = .50$ zufällig bzw. gleich wahrscheinlich) soll einseitig mit $\alpha = .01$ getestet werden.

Bei der Auswertung der Daten weist man Patienten der (0)-Kategorie (keine Veränderung) zu gleichen Teilen der (−)-Kategorie und der (+)-Kategorie zu. Der Signifikanztest bestätigt die Alternativhypothese: Der einseitige Test ist auf dem 1 %-igen Signifikanzniveau ($\alpha = .01$) signifikant.

Der Ergebnisbericht wird von der Konzernleitung aufmerksam studiert. Hierbei stellt man mit Entsetzen fest, dass sich in der (−)-Kategorie, also der Kategorie mit Blutzuckerabnahme, lediglich 51 % der Patienten (einschließlich der Hälfte der Patienten aus der (0)-Kategorie) befinden. Dies entspricht einer Effektgröße von $g = 0.51 - 0.50 = 0.01$. Man beschließt, auf eine Veröffentlichung dieses schwachen, klinisch bedeutungslosen Ergebnisses zu verzichten (zu Folgen derartiger Publikationsentscheidungen siehe den „Publication Bias", ▶ Abschn. 16.1.3.) und die Arbeit am Projekt einzustellen, zumal finanzielle Überlegungen deutlich gemacht hatten, dass eine Fortführung des Projektes nur sinnvoll ist, wenn der medikamentöse Effekt um mindestens 5 % über der Zufallserwartung von 50 % liegt.

Konfidenzintervall Unter Bezugnahme auf ▶ Abschn. 14.2.3 errechnen wir

$$\sigma_P = \sqrt{\frac{0.51 \cdot (1 - 0.51)}{25\,000}} = 0.0032$$

und (für das 99 %ige Konfidenzintervall)

$$KI_\pi = 0.51 \pm 2.58 \cdot 0.0032 = 0.51 \pm 0.0083 .$$

Das Konfidenzintervall hat damit die Grenzen

$$KI(99\,\%) = .5017 \leq \pi \leq .5183 .$$

Wegen des sehr großen Stichprobenumfanges ($n = 25\,000$) ist das Konfidenzintervall sehr eng. Der Wert $\pi = .50$ befindet sich nicht in diesem Intervall, d. h., die Abweichung .51 von .50 ist – wie bereits gesagt – für $\alpha = .01$ signifikant.

In dieser Untersuchung wurden offenbar statistische Signifikanz und praktische Bedeutsamkeit verwechselt. Der Planungsfehler, der der biometrischen Abteilung anzulasten ist, besteht in dem Versäumnis, die Geschäftsleitung nach einem praktisch bzw. klinisch bedeutsamen Mindesteffekt gefragt zu haben. Hätte man gewusst, dass die Geschäftsleitung einen Mindesteffekt von $G = 0.05$ (kleiner Effekt gem. Tab. 14.5 erwartet, wäre gemäß Tab. 14.11 ein Stichprobenumfang von $n = 1\,001$ Patienten bzw. gemäß G*Power ein Stichprobenumfang von $n = 1\,007$ ausreichend gewesen (test family: exact; statistical test: proportion: difference from constant, one sample case; type of power analysis: a priori; tail: one; effect size g: .05; $\alpha = .01$; $1 - \beta = .80$; constant proportion: 0.5).

Diese Untersuchung hätte aller Voraussicht nach zwar zu keinem statistisch signifikanten Ergebnis geführt; der Firma wären jedoch erhebliche Kosten erspart geblieben. Da sowohl klinische Bedeutungslosigkeit des Präparats als auch ein nicht-signifikanter Effekt Gründe sind, das Projekt einzustellen, hätte man diese Entscheidung besser auf der Basis des weniger aufwändigen und weniger teuren Samples treffen sollen.

Prüfung von H_{01} Über den Vorzeichentest bzw. – asymptotisch – über den Zweifelder-χ^2-Test (Bortz & Lienert, 2008, Abschn. 3.3.1) ermittelt man $\chi^2 = 10$ bzw. – nach Gl. 5 in Tab. 15.4 in ▶ Kap. 15 – $F_{(1,\infty)} = 10$. Dieser Wert ist gem. Tab. 11 (▶ Webangebot zu diesem Buch) sehr viel kleiner als der kritische Wert für $\alpha = .05$ ($F_{crit} = 135.8$), d. h., die Minimum-Effekt-Nullhypothese H_{01} kann nicht verworfen werden.

14.4.5 Differenz zweier Anteilswerte

Ein wenig populärer Politiker steht vor einem wichtigen Fernsehauftritt. Er möchte überprüfen lassen, ob dieser Fernsehauftritt dazu beitragen wird, seine Popularität in der Bevölkerung zu verbessern. Das mit dieser Aufgabe beauftragte Meinungsforschungsinstitut weiß aus älteren Untersuchungen, dass nur ca. 20 % der Bevölkerung diesen Politiker sympathisch finden (Skala: unsympathisch, neutral, sympathisch, keine Meinung). In Vorgesprächen mit dem Politiker stellt sich nun heraus, dass er nicht daran interessiert ist, eine zu vernachlässigende Sympathiesteigerung nachgewiesen zu bekommen. Das Ganze

sei erst dann interessant für ihn, wenn sein Sympathiewert nach dem Fernsehauftritt auf mindestens 30 % steigt.

Das Meinungsforschungsinstitut plant die Befragung einer repräsentativen Stichprobe A vor dem Fernsehauftritt und einer weiteren Stichprobe B danach. Man rechnet damit, dass über den Fernsehauftritt auch in den Printmedien berichtet wird und legt deshalb keinen Wert darauf, dass die Stichprobe nur aus der Fernsehbevölkerung bzw. aus den Nutzern der fraglichen Sendung gezogen wird.

Zur Klärung der Frage, wie viele Personen pro Stichprobe befragt werden sollen, ist zunächst die Effektgröße H zu bestimmen (Tab. 14.5). Ausgehend von $\pi_A = 0.20$ und $\pi_B = 0.30$ ergibt sich nach Tab. 10 (▶ Webangebot zu diesem Buch) $\varphi(A) = 0.9273$ und $\varphi(B) = 1.1593$ und damit $H = 1.1593 - 0.9273 = 0.23$. Dieser Wert entspricht ungefähr einem kleinen Effekt ($H = .20$). Eine Entscheidung zugunsten von H_1: $\pi_B - \pi_A \geq 0.30 - 0.20 = 0.10$ will man mit einem Signifikanzniveau von $\alpha = .01$ absichern. Bei Gültigkeit von H_1 sollte der Test mit einer Wahrscheinlichkeit von 80 % ($1-\beta = .80$) zugunsten von H_1 entscheiden. Nach Tab. 14.11 sind pro Stichprobe ca. 500 Personen zu befragen. Da die Kosten für die Untersuchung (Befragung von 2×500 Personen plus Auswertung) akzeptiert werden, gibt der Politiker (bzw. seine Partei) die Untersuchung in Auftrag. (Mit G*Power wird ein optimaler Stichprobenumfang von $n = 502$ pro Gruppe bestimmt; Test Family: z-tests; statistical test: Proportions: difference between two independent proportions; type of power analysis: a priori; $h = .20$; $p_1 = .30$. α: .01; $1 - \beta$: .80. allocation ratio: 1; options: computation: use arcsin transform und input: use Cohen's effect size inde xh).

Die Auswertung der Befragungen führt zu $P_A = .18$ und $P_B = .25$. Dieser Unterschied ist bei einseitigem Test und $\alpha = .01$ signifikant. Als Effektgröße resultiert $h = 1.0472 - 0.8763 = .17$, d. h., der angestrebte kleine Effekt wurde nicht ganz erreicht.

Konfidenzintervalle Gefragt wird anhand des Stichprobenergebnisses $P_A - P_B = .18 - .25 = -.07$, das als Punktschätzer für den „wahren" Populationsparameter $\pi_A - \pi_B$ genutzt werden kann, auch nach dem Konfidenzintervall, das die Schätzgenauigkeit angibt. Wir errechnen das Konfidenzintervall gem. Gl. 14.21

$$\sigma_{(P_A - P_B)} = \sqrt{\frac{0.18 \cdot (1 - 0.18)}{500} + \frac{0.25 \cdot (1 - 0.25)}{500}} = 0.0259$$

und über Gl. 14.22

$$KI_{(\pi_A - \pi_B)} = -0.07 \pm 2.58 \cdot 0.0259 = -0.07 \pm 0.0668$$

Tabelle 14.18 Beispiel für eine Kontingenztafelanalyse

	Film A	Film B	Film C	Summe
Experimentalgruppe	7	30	17	54
Kontrollgruppe	26	12	16	54
Gesamt	33	42	33	108

bzw.

$$KI(99\,\%) = -0.1368 \leq \pi_A - \pi_B \leq -0.0032 .$$

Durch den Fernsehauftritt hat sich also der „wahre" Anteil der Sympathisanten um 0.3–13.6 % erhöht ($\alpha = 0.01$).

Prüfung von H_{01} Über den Vierfelder-χ^2-Test (Bortz & Schuster, 2010, S. 137) ergibt sich $\chi^2 = 7.26$ bzw. nach Gl. 5 in Tab. 15.4 in ▶ Kap. 15 $F_{(1,\infty)} = 7.26$. Aus Tab. 11 (▶ Webangebot zu diesem Buch) entnehmen wir für die Überprüfung der Minimum-Effekt-Nullhypothese H_{01} auf dem 5 %-Niveau $F_{crit} = 135.8 > 7.26$, d. h., die H_{01} kann nicht verworfen werden.

14.4.6 χ^2-Test

Eine Fernsehanstalt will den Einfluss kurzer Inhaltsangaben über Fernsehfilme überprüfen, die in Fernsehzeitschriften abgedruckt werden. In einer experimentellen Untersuchung soll eine Gruppe von Personen nach dem Lesen der Inhaltsangaben von 3 Filmen (inkl. Angaben über die Hauptdarsteller) entscheiden, welchen Film sie sich ansehen würden, falls die 3 Filme im Fernsehen parallel angeboten werden (Experimentalgruppe). Eine zweite Gruppe trifft ihre Entscheidung nur aufgrund des Titels und der Hauptdarsteller der Filme (Kontrollgruppe). Die Nullhypothese („Die Inhaltsangaben haben keinen Einfluss auf die Programmpräferenzen") soll über einen 3×2-χ^2-Test mit $\alpha = .05$ und $1 - \beta = .80$ geprüft werden. Man entscheidet sich für einen mittleren Effekt ($W = .30$) und benötigt damit für die Untersuchung gem. Tab. 14.11 (df = 2) eine Gesamtstichprobe von $n = 107$ bzw. (um gleich große Gruppen bilden zu können) 108 Testpersonen. G*Power berechnet $n_{opt} = 108$ (test family χ^2; statistical test goodness of fit-test: contingency tables, type of power analysis: a priori; $w = 0.30$; α: .05; $1 - \beta$: .80).

Per Zufall werden 54 Personen der Kontrollbedingung und 54 Personen der Experimentalbedingung zugeordnet. Die Ergebnisse der Untersuchung zeigt Tab. 14.18.

Wir berechnen einen χ^2-Wert von $\chi^2 = 18.68$, der mit df = 2 statistisch signifikant ist ($\alpha = .01$). (Zur Be-

rechnung von χ^2-Werten Bortz & Schuster, 2010, S. 140, Beispiel 9.1) Über Gl. 6 in Tab. 14.5 wird W durch $w = \sqrt{\frac{18.68}{108}} = .42$ geschätzt.

Konfidenzintervall Im Weiteren wollen wir annehmen, dass v. a. Unterschiede bzgl. der Film-A-Präferenzen interessieren. Der Anteil derjenigen, die Film A präferieren, beträgt in der Experimentalgruppe $P_A = 7/54 = .13$ und in der Kontrollgruppe $P_B = 26/54 = .48$. Es resultiert also eine Differenz von $.48 - 0.13 = .35$ mit folgendem Konfidenzintervall:

Über Gl. 14.21 errechnen wir als Standardfehler

$$\sigma_{(P_A - P_B)} = \sqrt{\frac{.13 \cdot (1 - .13)}{54} + \frac{.48 \cdot (1 - .48)}{54}} = 0.082 .$$

Das 95 %ige Konfidenzintervall für die Prozentwertdifferenz in der Population lautet gem. Gl. 14.22

$$KI_{(\pi_A - \pi_B)} = 0.35 \pm 1.96 \cdot 0.082 = 0.35 \pm 0.16 \quad \text{bzw.}$$
$$KI(95\,\%) = .19 \leq \pi_A - \pi_B \leq .51 .$$

Der „wahre" Unterschied derjenigen, die in der Experimentalgruppe bzw. in der Kontrollgruppe Film A präferieren, liegt also zwischen 19 % und 51 % ($\alpha = .05$).

Prüfung von H_{01} Wir transformieren $\chi^2_{(2)} = 18.68$ nach Gl. 5 in Tab. 15.4 in ▶ Kap. 15 in ein F-Äquivalent: $F_{(2,\infty)} = 18.68/2 = 9.34$. Wegen $F_{crit} = 68.43 > 9.34$ ($\alpha = .05$) gem. Tab. 11 (▶ Webangebot zu diesem Buch) kann die Minimum-Effekt-Nullhypothese H_{01} nicht verworfen werden.

14.4.7 Varianzanalysen

Einfaktorielle Varianzanalyse

Im Amt für Soziales einer Großstadt interessiert man sich für die Frage, durch welche Kanäle Personen, die Anspruch auf Transferleistungen haben (Arbeitslosengeld II/„Hartz IV", Wohngeld, Elterngeld, Kindergeld, Ausbildungshilfen/„BAföG" etc.), über die ihnen zustehende Hilfe informiert werden. Vor allem will man wissen, wie viel Zeit von der ersten Information bis zur tatsächlichen Entgegennahme der Leistung vergeht. Die folgenden Informationskanäle sollen vergleichend evaluiert werden:

- a_1) Bekannte, Freunde, Verwandte,
- a_2) öffentliche Medien,
- a_3) Beratungsstellen der Leistungsträger.

Man plant, aus dem bereits geförderten Personenkreis 3 Stichproben zu ziehen, deren Mitglieder retrospektiv neben dem Informationskanal angeben sollen, wie viel Zeit in Tagen (abhängige Variable) vom ersten Bekanntwerden der Hilfsmöglichkeit bis zur Entgegennahme der konkreten Leistung verging. Da vergleichbare Daten nicht bekannt sind, entscheidet man sich einfachheitshalber für einen optimalen Stichprobenumfang, der eine mittlere Effektgröße ($E = 0.25$; Tab. 14.5) mit $\alpha = .05$ und $1 - \beta = .80$ absichert. Die Nullhypothese (die Informationsquellen unterscheiden sich nicht in Bezug auf die abhängige Variable) soll mit einer einfaktoriellen Varianzanalyse ($df = 2$) überprüft werden. Gemäß Tab. 14.11 benötigt man pro Informationsquelle $n = 52$ Leistungsempfänger. Falls gleich große Stichproben nicht zu realisieren sind, soll darauf geachtet werden, dass sich eine Gesamtstichprobe von $n = 3 \times 52 = 156$ ergibt. (Mittels G*Power wird ein $n_{opt} = 159$ ermittelt; Test family: F Tests, Statistical test: ANOVA: Fixed effects, omnibus, one-way; Type of power analysis: A priori; Effect size $F = 0.25$, $\alpha = .05$, $1 - \beta$: .80. number of groups: 3).

Nach der Datenerhebung ermittelt man die folgenden Durchschnittswerte und Stichprobenumfänge:

- a_1: 13 Tage ($n = 62$),
- a_2: 18 Tage ($n = 58$),
- a_3: 16 Tage ($n = 36$).

Die Unterschiede sind statistisch signifikant, d. h., die H_0 ist abzulehnen. Zur Ex-post-Bestimmung der in der Untersuchung erzielten Effektgröße wird zunächst ein Schätzwert für σ_μ bestimmt. Ausgehend von einem Gesamtmittelwert von

$$\frac{62 \cdot 13 + 58 \cdot 18 + 36 \cdot 16}{156} = 15.55$$

erhält man nach Gl. 14.35:

$$\hat{\sigma}^2_\mu = \frac{62 \cdot (13 - 15.55)^2 + 58 \cdot (18 - 15.55)^2}{156} + \frac{136 \cdot (16 - 15.55)^2}{156} = 4.99 \quad \text{bzw.} \quad \hat{\sigma}_\mu = 2.20 .$$

Für die Merkmalsstreuung (Fehlerstreuung) schätzt man aus den (hier nicht wiedergegebenen) Einzeldaten $\hat{\sigma} = 8.80$, so dass sich $e = 2.20/8.80 = 0.25$ ergibt. Der geplante mittlere Effekt ist auch faktisch eingetreten.

Für η^2 errechnen wir über Gl. 14.33 folgenden Schätzwert:

$$\hat{\eta}^2 = \frac{0.25^2}{1 + 0.25^2} = .06 .$$

Setzen wir in Gl. 14.36 ein, erhält man den gleichen Wert:

$$\hat{\eta}^2 = \frac{758.59}{12\,606.91} = .06$$

(zur Berechnung von $QS_{treat} = 758.59$ ohne Einzelmessungen auf der Basis von Aggregatwerten Bortz & Schuster, 2010, S. 213; QS_{Fehler} ergibt sich zu $\hat{\sigma}^2_{Fehler} \times df_{Fehler} = 8.80^2 \times 153 = 11\,848.32$ und damit $QS_{tot} = QS_{treat} + QS_{Fehler} = 758.59 + 11\,848.32 = 12\,606.91$).

Konfidenzintervall Für die Berechnung des Konfidenzintervalls von η^2 benötigen wir neben den Freiheitsgraden (im Beispiel $df_Z = 2$, $df_N = 153$) den F-Wert als Nichtzentralitätsparameter. Wir errechnen $F = 379.30/8.8^2 = 4.90$. Mit diesen Eingangsparametern erhält man mit der SAS-Syntax (▶ Anhang zu diesem Buch) folgendes 95 %ige Konfidenzintervall:

$$KI(95\,\%) = .0049 \leq \eta^2 \leq .138 .$$

Die „wahre" Varianzaufklärung ($\alpha = .05$) liegt also zwischen 0.49 % und 13.8 %.

Für weitergehende Interpretationen sollen ungerichtete Einzelvergleichshypothesen getestet werden. Vor allem interessiert, ob sich die Beratungsstellen der Leistungsträger von den beiden anderen Informationskanälen unterscheiden. Der erste Vergleich hat also die c-Koeffizienten $(1, 0, -1)$ und der zweite Vergleich die c-Koeffizienten $(0, 1, -1)$. Diese beiden Einzelvergleiche sind nicht orthogonal.

Nach Gl. 14.37 errechnet man für den **ersten Einzelvergleich**

$$\hat{\psi}_1 = (1) \cdot 13 + (0) \cdot 18 + (-1) \cdot 16 = -3 .$$

Damit erhält man über Gl. 14.39 folgende Quadratsumme:

$$QS_{\hat{\psi}_1} = \frac{(-3)^2}{\frac{1}{62} + \frac{0}{58} + \frac{1}{36}} = 204.98 .$$

Die Merkmalsstreuung wurde über die Fehlervarianz geschätzt, d. h., man erhält $\hat{\sigma}_{Fehler} = 8.8$ bzw. $\hat{\sigma}^2_{Fehler} = 77.44$. Der F-Bruch nach Gl. 14.38 hat folgenden Wert:

$$F = \frac{204.98}{77.44} = 2.65 .$$

Dieser Wert ist mit $df_N = 1$ und $df_Z = (62 + 58 + 36) - 3 = 153$ nicht signifikant.

Der standardisierte Einzelvergleich hat nach Gl. 14.40 folgenden Wert

$$\hat{\delta}_{\hat{\psi}_1} = \frac{-3}{8.80} = -0.34 .$$

Das Konfidenzintervall berechnen wir nach den Ausführungen in ▶ Abschn. 14.2.3 z. B. entweder mit der SAS-Syntax im Webabgebot zu diesem Buch oder mit dem Rechner **Noncentral Distribution Calculator** (NDC, verfügbar unter www.statpower.net/software). Die Eingangsparameter lauten:

$$t = \sqrt{2.65} = 1.63, \quad df = 153,$$
$$n_1 = 62, \quad n_2 = 58, \quad n_3 = 36,$$
$$c_1 = 1, \quad c_2 = 0, \quad c_3 = -1 .$$

Das Programm errechnet (für $\alpha = .05$)

$$NZt_u = -0.34108 \quad \text{und} \quad NZt_o = 3.59580 \quad \text{bzw.}$$
$$KI(95\,\%) = -0.07 \leq \delta_{\psi_1} \leq 0.75 .$$

Für den **zweiten Einzelvergleich** erhält man: $\hat{\psi}_2 = 2$; $QS_{\hat{\psi}_2} = 88.85$, $F = 1.15$ (n. s.) und $\hat{\delta}_{\hat{\psi}_2} = 0.23$. Die Eingangsparameter für die SAS-Syntax (▶ Anhang zu diesem Buch) oder den o. g. Rechner NDC sind

$$t = \sqrt{1.15} = 1.07 \quad \text{und} \quad df = 153$$
$$\text{mit } c_1 = 0, c_2 = 1, c_3 = -1 .$$

Die Stichprobenumfänge entsprechen den oben genannten Werten. Man erhält folgendes Ergebnis:

$$NZt_u = -0.89536 \quad NZt_o = 3.03188 \quad \text{bzw.}$$
$$KI(95\,\%) = -0.19 \leq \delta_{\psi_2} \leq 0.64 .$$

Prüfung von H_{01} Die Überprüfung von H_{01} gestaltet sich in diesem Beispiel besonders einfach, da die empirischen F-Werte bereits vorliegen. Für den Haupteffekt haben wir $F_{(2.153)} = 4.90$ ermittelt. Wegen $F_{crit(0.5)} = 5.01 > 4.90$ (Tab. 11, ▶ Webangebot zu diesem Buch) kann die auf den Haupteffekt bezogene Minimum-Effekt-Nullhypothese H_{01} nicht verworfen werden. Die F-Werte der beiden Einzelvergleiche ($F_{\hat{\psi}_1} = 2.65$; $F_{\hat{\psi}_2} = 1.15$) sind nicht groß genug, um die Nil-Nullhypothese H_{00} verwerfen zu können. Dementsprechend kann in beiden Fällen auch die H_{01} nicht verworfen werden.

Einfaktorielle Varianzanalyse mit Messwiederholungen

Es soll ein neues Mittel gegen Nikotinabhängigkeit geprüft werden. Geplant ist ein Untersuchungszeitraum von 16 Wochen, der in vier vierwöchige Phasen unterteilt wird. Um sich den Aufwand einer Kontrollgruppe zu ersparen, werden nur Raucher in die Untersuchung einbezogen, deren Zigarettenkonsum seit mehreren Jahren stabil ist. Man beabsichtigt, einen ABAB-Plan (▶ Abschn. 12.5.5) einzusetzen, bei dem das Rauchverhalten in der ersten und dritten Phase ohne und in der

zweiten und vierten Phase mit Medikamenten registriert wird (abhängige Variable: durchschnittlicher Tageskonsum pro Woche). Die Auswertung soll mit einer einfaktoriellen Varianzanalyse mit Messwiederholungen erfolgen ($df = 3$).

Die Planung des Stichprobenumfanges geht von einem kleinen Effekt ($E = 0.1$), $\alpha = .05$ und $1 - \beta = .80$ aus. Außerdem ist man davon überzeugt, dass die durchschnittliche Korrelation zwischen dem viermal gemessenen Rauchverhalten keinesfalls unter $\overline{\rho} = .50$ liegt. Für diese Konstellation entnimmt man Tab. 14.13 einen Stichprobenumfang von $n = 138$. G*Power (Test family: F-tests, statistical test: ANOVA, repeated measures, within factors; type of power analysis: a priori, effect size $f = 0.1$, $\alpha = .05$; $1 - \beta = .80$; number of groups: 2; number of measurements: 4; correlation among rep measures: 0.5; nonsphericity correlation $\Lambda = 1$) ermittelt den gleichen optimalen Stichprobenumfang.

Nach Abschluss der Untersuchung liegen folgende Durchschnittswerte vor:

- 1. Phase: 24 Zigaretten,
- 2. Phase: 20 Zigaretten,
- 3. Phase: 22 Zigaretten,
- 4. Phase: 18 Zigaretten.

Das Ergebnis der Varianzanalyse ist nicht signifikant. Als ex post bestimmte Effektgröße erhält man mit $\hat{\sigma}_\mu = 2.58$ und $\hat{\sigma} = 24.80$ (geschätzt durch $\hat{\sigma}_{res}$) $e = 2.80/24.80 = .10$. Der Effekt entspricht also dem für praktisch bedeutsam erachteten Mindesteffekt von $E = .10$. Allerdings beträgt die durchschnittliche Korrelation der 4 Messwertreihen nur $\overline{r} = .46$. Eine höhere Korrelation hätte $\hat{\sigma}$ ($= \hat{\sigma}_{res}$) stärker reduziert, was mit einem größeren Effekt und einem vermutlich signifikanten Ergebnis verbunden wäre.

Zusätzlich interessiert vorrangig ein Vergleich der A-Phasen mit den B-Phasen, also ein Vergleich der Phasen mit und ohne Medikamente. Für diesen Einzelvergleich benötigen wir für Gl. 14.37 die c-Koeffizienten $1/2$, $-1/2$, $1/2$ und $-1/2$ und errechnen

$$\hat{\psi} = \frac{1}{2} \cdot 24 + \left(-\frac{1}{2}\right) \cdot 20 + \frac{1}{2} \cdot 22 + \left(-\frac{1}{2}\right) \cdot 18 = 4 \, .$$

Als Quadratsumme (bzw. wegen $df = 1$ als Varianz) dieses Einzelvergleiches errechnet man nach Gl. 14.39

$$QS_\psi = \hat{\sigma}^2_\psi = \frac{4^2}{4 \cdot \frac{0.25}{138}} = 2.208 \, .$$

Der F-Test ergibt nach Gl. 14.45

$$F = \frac{2.208}{24.80^2} = 3.59 \text{ (n. s.)}$$

($df_Z = 1$; $df_N = (4 - 1) \cdot (138 - 1) = 411$).

Da es fraglich ist, ob die Zirkularitätsvoraussetzung erfüllt ist, prüfen wir Einzelvergleiche auch über Gl. 14.46. Die hierfür benötigte Streuung der Differenzen berechnet sich wie folgt: Über Gl. 14.42 ergibt sich nach Umstellen

$$\hat{\sigma} = \frac{\hat{\sigma}_{Res}}{\sqrt{1 - \overline{r}}} = \frac{24.8}{\sqrt{1 - 0.46}} = 33.75 \, .$$

Mit diesem Wert erhält man über Gl. 14.10

$$\hat{\sigma}_D = 33.75 \cdot \sqrt{2 \cdot (1 - 0.46)} = 35.07$$

(wegen $\overline{r} < .50$ ergibt sich in diesem Beispiel $\hat{\sigma} < \hat{\sigma}_D$).

Als t-Wert (mit $df = 138 - 1 = 137$) resultiert nach Gl. 14.46

$$t = \frac{4}{\frac{35.07}{\sqrt{138}}} = 1.34 \text{ (n. s.)} \, .$$

Der an σ standardisierte Einzelvergleich lautet

$$\hat{\delta}_{\hat{\psi}} = \frac{4}{33.75} = 0.12 \, .$$

Die Medikamente haben also insgesamt eine Zigarettenreduktion von nur 12 % der Merkmalsstreuung bewirkt.

Konfidenzintervall Das asymptotische 95 %ige Konfidenzintervall für diesen Effekt errechnet sich über Gl. 14.50 zu

$$KI_{\delta_\psi} = 0.12 \pm \frac{1.97 \cdot \frac{35.07}{\sqrt{138}}}{33.75} = 0.12 \pm 0.17 \quad \text{bzw.}$$
$$KI(95\,\%) = -0.05 \le \delta_\psi \le 0.29 \, .$$

Sogar eine Unterlegenheit der medikamentösen Phasen gegenüber den Kontrollphasen wäre also mit $\hat{\delta}_{\hat{\psi}} = 0.12$ vereinbar ($\alpha = .05$). Bezogen auf die Anzahl geraucher Zigaretten errechnet man als Konfidenzintervall

$$KI_\psi = 4 \pm 1.97 \cdot \frac{35.07}{\sqrt{138}} = 4 \pm 5.88 \quad \text{bzw. gerundet}$$
$$KI(95\,\%) = -2 \le \psi \le 10 \, .$$

Die Therapie muss also als gescheitert gelten.

Der Vollständigkeit halber berechnen wir auch noch $\hat{\delta}'_{\hat{\psi}}$ nach Gl. 14.48

$$\hat{\delta}'_{\hat{\psi}} = \frac{4}{35.07} = 0.11$$

sowie $\hat{\eta}^2_p$ für das Treatment und den geprüften Einzelvergleich (Gl. 14.36). Auf die Berechnung der hierfür benötigten QS_{treat} wird verzichtet; sie ergibt sich aus $\hat{\sigma}^2_\mu =$

2.58. QS_{res} erhält man über $QS_{res} = 24.8^2 \cdot (4-1) \cdot (138-1) = 252\,781.44$. Festzuhalten bleibt:

$$\hat{\eta}^2 = \frac{126\,960}{379\,741.44} = .33,$$
$$\hat{\eta}^2_{\hat{\psi}} = \frac{2\,208}{379\,741.44} = .006\,.$$

Prüfung von H_{01} Da weder die Einzelvergleiche noch der Haupteffekt signifikant sind, erübrigt sich eine Überprüfung von H_{01}.

Zweifaktorielle Varianzanalyse

Eine Regierung plant, die gesetzlichen Fördermaßnahmen zum Mutterschutz einzuschränken. Zuvor will man durch eine Befragung mögliche Reaktionen auf diese Gesetzesänderung erkunden, denn man befürchtet, dass dieses Vorhaben der Regierungspartei (a_1) wichtige Wählerstimmen kosten könnte. Es interessiert die Einstellung zur Gesetzesänderung (abhängige Variable) in Abhängigkeit vom Geschlecht der Befragten (Faktor B: männlich/weiblich) und von ihren Parteipräferenzen (Faktor A: die Parteien a_1, a_2 und a_3). Man vermutet, dass zwischen dem Geschlecht und den Parteipräferenzen in Bezug auf die Einstellung eine Interaktion besteht: Die Anhänger der Partei a_1 befürworten die Änderung stärker als die Anhängerinnen, während die Anhänger der Partei a_2 die Änderung stärker ablehnen als die Anhängerinnen (H_1). Für Angehörige der Partei a_3 werden keine geschlechtsspezifischen Unterschiede vorhergesagt. Die abhängige Variable soll mit einem Einstellungsfragebogen erhoben werden, der zu Einstellungswerten zwischen −5 (starke Ablehnung) und +5 (starke Befürwortung) führen kann.

In Vorgesprächen mit einer Evaluationsagentur stellt sich natürlich die Frage nach den Kosten für die Studie. Diese – so die Agentur – hingen v. a. von der einzusetzenden Stichprobe ab, deren Größe nicht beliebig sei, sondern sehr genau kalkuliert werden könne. Dies setze allerdings voraus, dass die Auftraggeber eine Vorstellung davon haben, welche parteispezifischen Effekte der Gesetzesänderung für praktisch bedeutsam gehalten werden. In diesem Fall könne der Stichprobenumfang so festgelegt werden, dass bei Gültigkeit von H_1 genau dieser Effekt und keine kleineren, unbedeutenden Effekte statistisch signifikant werden können.

Da die Auftraggeber nur sehr vage Vorstellungen davon haben, wie Wählerinnen und Wähler der 3 Parteien auf die Gesetzesänderung reagieren würden, präsentiert die Evaluationsagentur einige vorbereitete Folien (nach Art von Abb. 14.8), die einen kleinen, einen mittleren und einen großen Interaktionseffekt veranschaulichen.

Bei der Vorbereitung dieser Folien ging die Evaluationsagentur davon aus, dass in der Befragung der gesamte Range der Einstellungsskala von −5 bis +5 genutzt wird. Unter der Annahme normalverteilter Einstellungen schätzt sie die Merkmalsstreuung unter Bezug auf Abb. 12.18 mit $\hat{\sigma} = 1.60$. Man einigt sich auf einen mittleren Effekt ($E = .25$). Da der Vorschlag, von $\alpha = .01$ und $1-\beta = .80$ auszugehen, akzeptiert wird, legt die Agentur eine Kostenkalkulation vor, die von 39 Personen pro Faktorstufenkombination bzw. einer Gesamtstichprobe von $6 \times 39 = 234$ Personen ausgeht (Tab. 14.14 für einen 2×3-Versuchsplan). Mittels G*Power wird ein optimaler Stichprobenumfang $n_{opt} = 227$ berechnet (test family: F-tests; statistical test: ANOVA, fixed effects, special, main effects and interactions; type of power analysis: a priori; effect size f: .25; $\alpha = .01$; $1-\beta = .80$; numerator df: 2; number of groups: 6).

Nach Abschluss der genehmigten Untersuchung resultieren die in Tab. 14.19 zusammengefassten empirischen Zellenmittelwerte $\overline{AB}_{ij}$. Der mit der Interaktion verbundene Effekt beträgt nach Gl. 14.54) $e_{A\times B} = .19$. (Die Populationsparameter sind wegen der Effektgrößenschätzung durch Stichprobenmittelwerte zu ersetzen.)

Die zur Berechnung dieses Wertes benötigten Mittelwerte $\overline{AB}'_{ij}$ (Zellenmittelwerte ohne Interaktion, sprich: „A B quer Strich“) ergeben sich nach Gl. 14.53 zu den in Tab. 14.20 dargestellten Werten.

Der Effekt ist kleiner als erwartet und damit praktisch zu vernachlässigen. Da die Planung von einem optimalen Stichprobenumfang für $f_{A\times B} = .25$ ausging, ist der Effekt auch statistisch unbedeutend:

$$F_{(2;228;1\,\%)} \approx 4.65 > 2.85 = F_{emp}\,.$$

Tabelle 14.19 Empirische Zellenmittelwerte $\overline{AB}_{ij}$ des Beispiels für eine zweifaktorielle Varianzanalyse

	Partei			
	a_1	a_2	a_3	$\overline{B}_j$
b_1 männlich	1.00	–3.00	–1.50	–1.17
b_2 weiblich	0.50	–2.00	–1.50	–1.00
$\overline{A}_i$	0.75	–2.50	–1.50	$\overline{G} = -1.08$

Tabelle 14.20 Theoretische Zellenmittelwerte $\overline{AB}_{ij}$ des Beispiels bei additiver Wirkung der Haupteffekte

	Partei			
	a_1	a_2	a_3	$\overline{B}_j$
b_1 männlich	0.66	–2.59	–1.59	–1.17
b_2 weiblich	0.83	–2.42	–1.42	–1.00
$\overline{A}_i$	0.75	–2.50	–1.50	$\overline{G} = -1.08$

Konfidenzintervall Über Gl. 14.33 schätzen wir $\hat{\eta}^2_{A\times B} = \frac{0.19^2}{(1+0.19^2)} = .035$. Hierfür ermitteln wir über die SAS-Syntax (▶ Anhang zu diesem Buch) folgendes Konfidenzintervall ($\alpha = .05$):

$$.000 \leq \eta^2_{A\times B} \leq .092 .$$

(Die SAS-Syntax berechnet nur die obere Grenze. Die untere Grenze kann nicht berechnet werden, weil es keine nichtzentrale F-Verteilung mit den genannten Freiheitsgraden gibt, von der $F = 2.85$ die oberen 2.5 % der Fläche abschneidet. Wir setzen deshalb die untere Grenze auf den Wert Null).

Zu Demonstrationszwecken soll noch ein an der Merkmalsstreuung standardisierter Interaktionseinzelvergleich bestimmt werden. Wir fragen, ob Einstellungsunterschiede zwischen der Regierungspartei (a_1) und den beiden übrigen Parteien (a_2 und a_3) bei Männern genauso groß sind wie bei Frauen (H_0). Nach den Ausführungen bei Bortz und Schuster (2010, S. 261ff., 308, 311) berechnen wir zunächst **bedingte Einzelvergleiche bzw. einfache Treatmentkontraste** „Regierungspartei vs. übrige Parteien“ für die Gruppe der Männer ($\hat{\psi}_{(A|b_1)}$) und für die Gruppe der Frauen ($\hat{\psi}_{(A|b_2)}$):

$$\begin{aligned}
&\hat{\psi}_{(A|b_1)} \\
&= 1 \cdot 1.00 + \left(-\frac{1}{2}\right) \cdot (-3.00) + \left(-\frac{1}{2}\right) \cdot (-1.50) \\
&= 3.25 , \\
&\hat{\psi}_{(A|b_2)} \\
&= 1 \cdot 0.50 + \left(-\frac{1}{2}\right) \cdot (-2.00) + \left(-\frac{1}{2}\right) \cdot (-1.50) \\
&= 2.25 .
\end{aligned}$$

Diese beiden Einzelvergleiche werden nun gegeneinander kontrastiert, d. h. zu dem o. g. **Interaktionseinzelvergleich** ($\hat{\psi}_{(A\times B)}$) kombiniert:

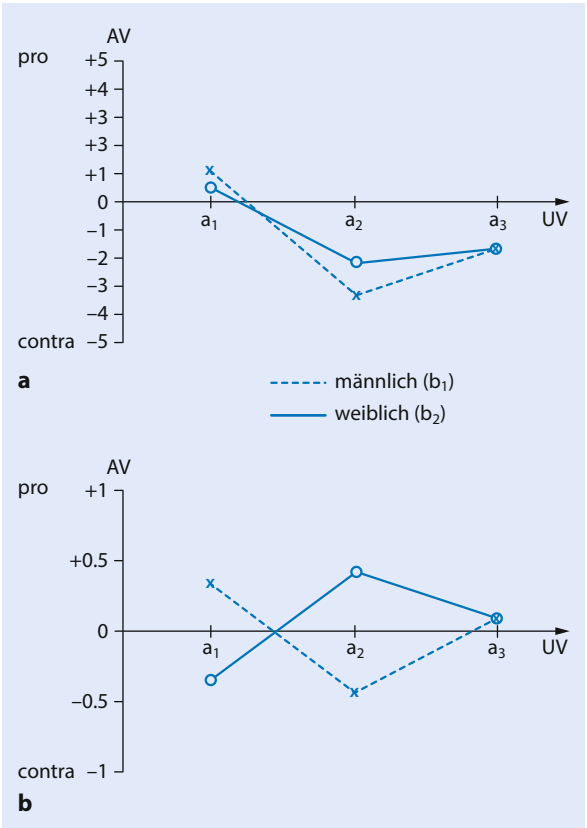
$$\hat{\psi}_{(A\times B)} = (1) \cdot 3.25 + (-1) \cdot 2.25 = 1.00 .$$

Der an der Merkmalsstreuung ($\hat{\sigma} = 1.6$) standardisierte Interaktionseinzelvergleich ergibt sich also zu

$$\hat{\delta}_{\hat{\psi}_{A\times B}} = \frac{1}{1.6} = 0.63 .$$

Der Interaktionseinzelvergleich ist nicht signifikant.

In ◘ Abb. 14.8 wird das Interaktionsmuster grafisch veranschaulicht (**Interaktionsdiagramm**). In ◘ Abb. 14.8a sind die empirischen Zellenmittelwerte abgetragen, deren Größe auch von den Haupteffekten beeinflusst ist. Ein treffenderes Bild von der Interaktion vermittelt



◘ **Abb. 14.8** Beispiel einer spezifischen Alternativhypothese für eine Interaktion. **a** Mit Haupteffekten, **b** ohne Haupteffekte; *AV* abhängige Variable, *UV* unabhängige Variable

◘ Abb. 14.8b, bei der die Haupteffekte aus den Zellenmittelwerten „herausgerechnet“ sind ($\overline{AB}_{ij} - \overline{AB}'_{ij}$).

Der Haupteffekt A (Parteienunterschiede) ist statistisch signifikant und mit einem sehr großen Effekt von $e_A = 1.36/1.6 = .85$ verbunden. Wenn die für die Stichprobe errechneten Mittelwerte tatsächlich den Populationsverhältnissen entsprechen, hätten zur Absicherung dieses Effektes sehr viel kleinere Stichprobenumfänge ausgereicht ($n < 16$ gem. ◘ Tab. 14.14).

Der Haupteffekt B (Geschlechtsunterschied) ist statistisch nicht signifikant. Der ihm zugeordnete Effekt ist mit $e_B = 0.085/1.6 = .05$ sehr klein. Hätte man sich dafür interessiert, einen derart kleinen Effekt statistisch abzusichern, wären pro Faktorstufenkombination Stichproben mit $n > 233$ erforderlich gewesen.

Zusammenfassend stellt die Evaluationsagentur fest, dass die geplante Gesetzesänderung zwar von der Anhängerschaft der Parteien a_2 und a_3 abgelehnt, von den Anhängerinnen und Anhängern der Regierungspartei (a_1) jedoch eher positiv aufgenommen wird.

Prüfung von H_{01} Die auf den Haupteffekt B und die Interaktion $A \times B$ bezogenen Nil-Nullhypothesen (H_{00})

Tabelle 14.21 Mittelwerte des Beispiels (Hausmüll in kg) für eine zweifaktorielle Varianzanalyse mit Messwiederholung

		Faktor *B*	
		Vorher	Nachher
Faktor *A*	Experimentalgruppe	45	37
	Kontrollgruppe	47	45

konnten nicht verworfen werden. Es erübrigt sich damit eine Überprüfung der entsprechenden Minimum-Effekt-Nullhypothesen (H_{01}). Für Haupteffekt A wurde eine Effektgröße von $e = .85$ geschätzt. Wir ermitteln hierfür über Gl. 14.33 $\hat{\eta}^2_A = .42$ und analog zu Gl. 15.4 in ▶ Kap. 15 ein F-Äquivalent von $F_{(2.228)} = 82.56$. Aus Tab. 11 (▶ Webangebot zu diesem Buch) ist zu entnehmen, dass mit diesem Wert nicht nur die H_{01}, sondern auch die H_{05} für $\alpha = .01$ zu verwerfen ist. Für $df_Z = 2$ und $df_N = 300$ lesen wir $F_{crit} = 9.18$ (H_{01}) und $F_{crit} = 20.94$ (H_{05}) ab.

Zweifaktorielle Varianzanalyse mit Messwiederholungen

Die Stadtreinigung beabsichtigt, eine Aufklärungsbroschüre über die Notwendigkeit der Mülltrennung zu evaluieren. Insbesondere ist ihr daran gelegen, die Akzeptanz der öffentlich aufgestellten Papier- und Glascontainer zu erhöhen. Man wählt als Untersuchungsdesign einen Pretest-Posttest- bzw. Prä-Post-Plan mit randomisierter Experimental- und Kontrollgruppe (▶ Abschn. 12.5.4). In der Experimentalgruppe soll die Broschüre über die Hauspostkästen verteilt werden; die Kontrollgruppe erhält keine diesbzgl. Informationen. Das durchschnittliche, an 4 Entleerungstagen gemessene Gewicht (in kg) des Mülls in den Hausmülltonnen – gemessen vor und nach der Maßnahme – dient als abhängige Variable.

Schon ein geringer Effekt macht eine Verteilung der Broschüre an die Gesamtbevölkerung rentabel. Für die Kalkulation der hierfür optimalen Stichprobenumfänge für Experimental- und Kontrollgruppe benötigt man Angaben über die Korrelation der Müllmengen bei den wöchentlichen Entleerungen. Experten schätzen, dass diese Korrelation (in der Kontrollgruppe) nicht unter $\rho = .50$ liegen dürfte. Man entnimmt deshalb (für $\alpha = .01$ und $1 - \beta = .80$) Tab. 14.15, dass für einen 2×2-Messwiederholungsplan $n_{Exp} = n_{Kon} = 147$ Haushalte optimal wären. Die Untersuchung führt zu den in Tab. 14.21 zusammengefassten Abfallgewichten.

Die Interaktion (Gruppe × Messzeitpunkt) ist statistisch nicht signifikant ($\alpha = .01$). Ihr entspricht ein Effekt von $e_{A\times B} = .09$ (für $\hat{\sigma}_\mu = 1.50$ und einer ex post ermittelten Fehlerstreuung von $\hat{\sigma}_{B\times Vpn} = 16.7$).

Für metaanalytische Zwecke sollte zusätzlich ein **standardisierter Nettoeffekt** berechnet werden. Nach den Angaben in Tab. 12.23 „Schema zur Ermittlung eines Treatmenteffektes“ ermitteln wir $E = 45 - 37 = 8$ und $K = 47 - 45 = 2$, d. h., wir erhalten einen Nettoeffekt von $NE = 8 - 2 = 6$.

Die Streuung des Merkmals „Durchschnittliches Abfallgewicht in Hausmülltonnen“ wäre in diesem Beispiel auf der Basis der Pretestwerte in Experimental- und Kontrollgruppe zu schätzen. Auch die Posttestmessungen in der Kontrollgruppe könnten mit einbezogen werden, wenn – was zu erwarten ist – die Pretestmessung die Posttestmessung in dieser Gruppe nicht beeinträchtigt.

Wenn wir von einer Streuung von $\hat{\sigma} = 18$ ausgehen, resultiert ein standardisierter Nettoeffekt von $\hat{\delta}_{\hat{\psi}_{NE}} = \frac{6}{18} = 0.33$. Verglichen mit ähnlichen Untersuchungen erweist sich dieser Effekt als sehr klein (auf die Berechnung eines Konfidenzintervalls wird verzichtet, weil u. E. hierfür bislang keine ausgereifte Software existiert).

Dieser Effekt ist der Stadtreinigung zu klein. Man beschließt deshalb mit einer ansprechender gestalteten Informationsbroschüre einen neuen Evaluationsversuch.

Prüfung von H_{01} Für den hier primär interessierenden Interaktionseffekt konnte die Nil-Nullhypothese H_{00} nicht verworfen werden, was bedeutet, dass auch die H_{01} nicht zu verwerfen ist.

14.4.8 Multiple Korrelation

Die Betriebspsychologin eines großen Werkes möchte eine Testbatterie zur Vorhersage von Arbeitszufriedenheit (Kriteriumsvariable Y) zusammenstellen. Sie beabsichtigt, die folgenden 5 Prädiktorvariablen einzusetzen:

- X_1: Entlohnung,
- X_2: Möglichkeiten zur flexiblen Arbeitszeitgestaltung,
- X_3: Abwechslungsreichtum am Arbeitsplatz,
- X_4: Betriebsklima,
- X_5: Beeinträchtigungen durch Lärm, Staub, Hitze etc.

In der Literatur wird über eine ähnliche Untersuchung berichtet, die bei $p = 7$ vergleichbaren Prädiktoren zu einer multiplen Korrelation von $R = .50$ führte. Dieser Wert entspricht einer standardisierten Effektgröße von $k^2 = .50^2/(1-.50^2) = .33$, die in Tab. 14.5 als nahezu „großer Effekt“ klassifiziert wird.

Nach den Ausführungen in ▶ Abschn. 14.2.3 wird davon abgeraten, den optimalen Stichprobenumfang nur von der Größe des Effekts K^2 und der Anzahl der Prädiktorvariablen abhängig zu machen. Wichtiger sind die Validitäten und die Multikollinearitätsstruktur. Aufgrund

vergleichbarer Untersuchungen vermutet die Betriebspsychologin, dass die Validitäten im Durchschnitt bei $\overline{\rho}_{xy} = .40$ liegen könnten, bei einer durchschnittlichen Multikollinearität von $\overline{\rho}_{xx} = .30$. Mit diesen Angaben entnimmt sie Tab. 14.16 $n_{opt} = 191$. Sollten ihre Vermutungen zutreffen, wäre mit einer multiplen Korrelation von .60 zu rechnen.

Nach einer Befragung von 191 Beschäftigten errechnet die Betriebspsychologin $R = .49$. Offensichtlich hat sie die Validitäten überschätzt und/oder die Multikollinearität unterschätzt. Für $R = .49$ errechnet man als Effekt $k^2 = .32$. Ferner entnimmt man Tab. 12 (▶ Webangebot zu diesem Buch) für $R^2 = .24$, $p = 5$ und $n = 200$ ($\approx$191) eine untere Grenze des 95 %igen Konfidenzintervalls von .139. Der Signifikanztest für multiple Korrelationen (F-Test gem. Gl. 3 in Tab. 15.4 in ▶ Kap. 15) weist diese Korrelation als signifikant aus ($F_{emp} = F_{(5,185)} = 11.69 > F_{crit} = 3.11$; $\alpha = .01$). Ein akzeptables Modell der Determinanten von Arbeitszufriedenheit setzt dennoch Replikationen von multiplen Korrelationsstudien mit identischen Prädiktoren voraus, die metaanalytisch integriert werden könnten.

Prüfung von H_{01} Wie Tab. 11 (▶ Webangebot zu diesem Buch) zeigt, können mit $F_{emp} = 11.69$ und $F_{crit} = 4.18$ auch die Minimum-Effekt-Nullhypothese H_{01} (▶ Kap. 15) und mit $F_{crit} = 7.51$ sogar die H_{05} verworfen werden ($\alpha = .01$, für $df_N = 200$).

Lernquiz

1. Was trifft zu? Ein Signifikanztest wird umso eher signifikant,
   a) je größer/kleiner der Effekt,
   b) je größer/kleiner der Stichprobenumfang,
   c) je größer/kleiner das Signifikanzniveau,
   d) je größer/kleiner die Teststärke.
2. Was versteht man unter der „Effektgröße“?
3. Was ist mit der „Power“ im Kontext von Signifikanztests gemeint und wovon hängt sie ab?
4. Anhand von Urlauberbefragungen hat ein Tourismusunternehmen 30 zufällig ausgewählte Ferienorte am Meer mit einem Punktwert versehen, der die Zufriedenheit mit dem Urlaubsort ausdrückt. Dann wurde untersucht, welche Bedeutung folgende Faktoren für die Bewertung des Urlaubsortes haben: Tageshöchsttemperatur, Wassertemperatur, Seegang, Wasserverschmutzung, Verschmutzung des Strandes, Anzahl der Besucher am Strand sowie Anzahl der aktiven Wassersportler im Umkreis von 1 km vor dem Strand. Interessanterweise konnte kein einziger signifikanter Effekt nachgewiesen werden. Woran könnte das liegen?
5. Welches Problem tritt auf, wenn eine empirische Studie „underpowered“ ist?
6. Grenzen Sie Vorgehensweise und Zielsetzung der A-priori- und der Post-hoc-Teststärkeanalyse voneinander ab!
7. Warum gibt es viele verschiedene Effektgrößenmaße?
8. Welche Typen von Effektgrößenmaßen werden unterschieden?
9. Warum sollten die Ergebnisse einer quantitativen empirischen Studie immer auch in Form von Effektgrößenmaßen berichtet werden?
10. Welche Typen von Teststärkeanalysen werden differenziert?

Übungsaufgaben

1. Sie möchten die Hypothese überprüfen, dass es einen Zusammenhang zwischen dem Geschlecht einer in einer Werbeanzeige dargestellten Person und ihrer Funktion als „Dekorationsobjekt“ neben dem beworbenen Produkt gibt. Sie ziehen dazu eine Stichprobe von Werbeanzeigen, in denen neben dem beworbenen Produkt auch Personen dargestellt werden. Sie kodieren jeweils das Geschlecht der Person (weiblich/männlich) sowie deren Funktion in Abhängigkeit davon, ob die Person in inhaltlichem Bezug zum Produkt dargestellt wird oder nur als schmückendes Beiwerk fungiert (kein Dekorationsobjekt/Dekorationsobjekt). Sie gehen anhand des Forschungsstandes von einem mittleren Effekt in der Richtung aus, dass Frauen eher als Dekorationsobjekte auftauchen als Männer und möchten das Ergebnis auf einem α-Niveau von 5 % absichern. Bestimmen Sie den optimalen Stichprobenumfang mittels G*Power und anhand der Tabellen in diesem Kapitel!
2. In einer Studie stellen Sie fest, dass die Händigkeit (0 = Rechtshändigkeit, 1 = Linkshändigkeit) mit dem Ergebnis eines Kreativitätstests zu $r_{p\,bis} = .20$ korreliert. Sie wollen Ihr Ergebnis mit anderen Studien vergleichen, die das d-Maß von Cohen als Effektgrößenmaß angeben. Berechnen Sie aus Ihrem Ergebnis das d-Maß!
3. Eine neue Therapiemaßnahme wirbt mit „garantierten Besserungsraten“ von 67 % (neueste Untersuchung) und 59 % (Untersuchung 3 Jahre zuvor). Berechnen Sie die Effektgröße für die Veränderung. Handelt es sich gemäß Cohen-Klassifikation um einen kleinen, mittleren oder großen Effekt? Was ist bei der Einordnung der Effektgröße hinsichtlich klinischer Bedeutsamkeit zu beachten?
4. Wie viele Personen müssen Sie untersuchen, um folgende Hypothesen bei einer Teststärke von

80 % auf dem $\alpha = 5\,\%$-Niveau abzusichern? Nennen Sie jeweils auch den indizierten Signifikanztest!

a) Der Erfolg einer betrieblichen Weiterbildungsmaßnahme (gemessen anhand eines Leistungstestes am Ende der Maßnahme) hängt davon ab, wie lange die Maßnahme dauert (1 Tag, 2 Tage, 3 Tage oder eine Woche), ob und wie oft die Maßnahme mit denselben Teilnehmenden wiederholt und vertieft wird (keinmal, einmal, zweimal) und ob die Schulung im Betrieb oder in einem Tagungshotel durchgeführt wird. Erwartet wird ein mittlerer Interaktionseffekt zweiter Ordnung.

b) 50 % aller Fahrschüler/innen fallen bei der ersten Führerscheinprüfung durch (fiktive Angabe). Anhand einer Zufallsauswahl ihrer Kundinnen und Kunden will eine Fahrschule nachweisen, dass die von ihr angebotene Fahrausbildung sehr viel besser auf die Prüfung vorbereitet. Wie groß sollte die Stichprobe sein? ($\alpha = .05$; $1 - \beta = .80$)

Weiterführende Literatur

- Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences.* New York: Psychology Press.
- Bausell, R. B. & Li, Y.-F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences.* Cambridge, UK: Cambridge University Press.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale: Erlbaum.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Ellis, P. D. (2010). *The essential guide to effect sizes: An introduction to statistical power, meta-analysis and the interpretation of research results*. United Kingdom: Cambridge University Press.
- Fritz, C., Morris, P., & Richler, J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology, 141*(1), 2–18.
- Grissom, R. J. & Kim, J. J. (2011). *Effect sizes for research: A broad practical approach* (2nd ed.). New York, NY: Psychology Press.
- Murphy, K. R., Myors, B., & Wolach, A. (2008). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (3rd ed.). Mahwah, NJ: Erlbaum.

Literatur

Abelson, R. P. & Prentice, D. A. (1997). Contrast tests of interaction hypothesis. *Psychological Methods, 2*(4), 315–328.

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York: Psychology Press.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent group case. *Psychological Methods, 10*(3), 317–328.

APA, American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington DC: American Psychological Association.

APA, American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington DC: American Psychological Association.

Bausell, B. R. & Li, Y.–F. (2003) *Power analysis for experimental research: a practical guide for the biological, medical and social sciences*. Cambridge: Cambridge University Press.

Bird, K. D. (2002). Confidence intervalls for effect sizes in analysis of variance. *Educational Psychological Measurement, 62*(2), 197–226.

Bortz, J. (2005). *Statistik* (6. Aufl.). Berlin: Springer.

Bortz, J. & Lienert, G. A. (2003). *Kurzgefasste Statistik für die klinische Forschung* (2. Aufl.). Berlin: Springer.

Bortz, J. & Lienert, G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung* (3. Aufl.). Heidelberg: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Berlin: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.

Brizendine, L. (2006). *The female brain*. New York: Broadway Books.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics – Simulation and Computation 39*(4), 860–864.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Research, 65*(3), 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49*(12), 997–1003.

Cramer, E. M. & Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika, 44*(1), 43–54.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. & Finch, S. (2001). A primer on the understanding, Use, and calculation of confidence intervals that are based on central and noncentral distribution. *Educational Psychological Measurement, 61*(4), 532–574.

Dunlop, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measure designs. *Psychological Methods, 1*(2), 170–177.

Ellis, P. D. (2010). *The essential guide to effect sizes: an introduction to statistical power, meta-analysis and the interpretation of research results*. Cambridge Cambridge University Press.

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des beta-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle. *Zeitschrift für Sozialpsychologie, 15*(1), 18–32.

Erdfelder, E., Faul, F., Buchner, A., & Cüpper, L. (2010). Effektgröße und Teststärke. In H. Holling & B. Schmitz (Hrsg.), *Handbuch der Psychologischen Methoden und Evaluation* (S. 358–369). Göttingen: Hogrefe.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 246–259). New York: Sage.

Fowler, R. L. (1987). A general method for comparing effect-magnitudes in ANOVA designs. *Educational and Psychological Measurement, 47*, 361–367.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18.

Gatsonis, C. & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin, 106*(3), 519–524.

Gigerenzer, G. (1993). The superego, the ego, and the ID in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioural sciences. Methodological Issues* (pp. 311–339). Hillsdale: Erlbaum.

Gillett, R. (1994). The average power criterion for sample size estimation. *Statistician, 43*, 389–394.

Gillett, R. (2003). The metric comparability of meta-analytic effectsize estimators from factorial designs. *Psychological Methods, 8*(4), 419–433.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Grissom, R. J. & Kim, J. J. (2011). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). Milton Park: Routledge.

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data. A primer on methods and issues. *Psychological Methods, 3*(3), 339–353.

Hager, W. (2004). *Testplanung zur statistischen Prüfung psychologischer Hypothesen*. Göttingen: Hogrefe.

Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association, 288*(3), 358–362.

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt College Publishers.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*(2), 490–499.

Held, U. (2010). Was ist eine „Odds Ratio" und wann wird sie verwendet? *Schweiz Med Forum, 10*(37), 634–645.

Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19–24.

Hsu, L. M. (2004). Biases of success rate differences shown in binomial effect size displays. *Psychological Methods, 9*(2), 183–197.

James, D. & Drakich, J. (1993). Understanding gender differences in amount of talk. In D. Tannen (Ed.), *Gender and conversational interaction* (pp. 281–312). Oxford: Oxford University Press.

Janosky, J. E. (2002). The ethics of underpowered clinical trials. *Journal of the American Medical Association, 288*(17), 2118.

Kelley, K. & Maxwell, S. E. (2003). Sample Size for multiple regression. Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8*(3), 305–321.

Keren, G. & Lewis, C. (1979). Partial omega squared for ANOVA designs. *Educational and Psychological Measurement, 39*, 119–128.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746–759.

Kline, R. B. (2004). *Beyond significance testing*. Washington: American Psychological Association.

Kraemer, H. C. & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Thousand Oaks: Sage.

Kshirsagar, A. M. (1972). *Multivariate analysis*. New York: Dekker.

Küchler, M. (1980). The analysis of nonmetric data. *Sociological Methods and Research, 8*(4), 369–388.

Leonhart, R. (2009). *Statistik – Einstieg und Vertiefung* (2. Aufl.). Bern: Huber.

Lilford, R., & Stevens, A. J. (2002). Underpowered studies. *British Journal of Surgery, 89*(2), 129–131.

Lipsey, M. W. (1997). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods* (pp. 39–68). Thousand Oaks: Sage.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*(4), 434–458.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.

Mehl, M. R., Vazire, S., Ramirez–Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science, 317*(5834), 82.

Mendoza, J. L. & Stafford, K. L. (2001). Confidence Intervals, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational Psychological Measurement, 61*(4), 650–667.

Murphy, K. R., Myors, B., & Wolach, A. (2008). *Statistical power analysis: 2*. Milton Park: Routledge.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301.

Olejnik, S. & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–477.

Peng, C.-Y. J., Long, H., & Abaci, S. (2012). Power analysis software for educational researchers. *Journal of Experimental Education, 80*(2), 113–136.

Reedera, H. M. (1996). A critical look at gender difference in communication research. *Communication Studies, 47*(4), 318–330.

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 85–94). Thousand Oaks: Sage.

Rosenthal, R. & Rubin, D. B. (1982). A simple, general purpose display of magnitudes of experimental effect. *Journal of Educational Psychology, 74*(2), 166–169.

Sachs, L. (2002). *Statistische Auswertungsmethoden* (10. Aufl.). Berlin: Springer.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309–316.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Smithson, M. J. (2003). *Confidence intervals*. Thousand Oaks: Sage.

Steiger, J. H. (2004). Beyond the F-Test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164–182.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837–847.

Thompson, B. (2002). „Statistical", „practical", and „clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*(1), 64–71.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur Psychologischen Methodenlehre*. Göttingen: Hogrefe.

Wilkinson, L. & Inference, T. T. F. o. S. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*. New York: Mc–Graw Hill.

Minimum-Effektgrößen-Tests

Nicola Döring, Jürgen Bortz

15.1 Das Good-Enough-Prinzip – 869

15.2 Überprüfung von Minimum-Effekt-Nullhypothesen – 870

15.3 Zur Frage der „Bestätigung“ von Nullhypothesen – 885

Literatur – 892

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_15, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Wissen, was das Good-Enough-Prinzip besagt.
- Untersuchungen unter Berücksichtigung von Minimum-Effekt-Nullhypothesen planen können.
- Minimum-Effekt-Nullhypothesen prüfen können.
- Prinzipien von Nullhypothesen als „Wunschhypothesen" verstehen.
- Nullhypothesen als Wunschhypothesen testen können.

Der **klassische Signifikanztest** läuft darauf hinaus, die bedingte Wahrscheinlichkeit empirischer Ergebnisse bei Gültigkeit der Nullhypothese H_0 zu bestimmen (► Abschn. 12.5.1). Wir sprechen von einem statistisch signifikanten Ergebnis, wenn das gefundene Ergebnis einer Ergebnisklasse angehört, die bei Gültigkeit von H_0 nur mit einer sehr geringen Wahrscheinlichkeit auftritt. Konkret darf diese Irrtumswahrscheinlichkeit $p(D|H_0)$ das Signifikanzniveau von $\alpha = .05$ ($\alpha = .01$) nicht überschreiten. Wenn die Daten also nicht gut zum H_0-Modell passen, entscheidet man sich für die Annahme der Alternativ- bzw. Forschungshypothese H_1, zu der alle mit der Nullhypothese nicht erfassten Populationsverhältnisse zählen.

In dieser Logik des klassischen Signifikanztests („null hypothesis significance test", NHST) liegt – so wurde in ► Abschn. 12.5.1 argumentiert – ein Nachteil des Verfahrens. Behauptet die klassische Nullhypothese (sog. **Nil-Nullhypothese**; „nil hypothesis": Cohen 1994), es existiere keinerlei Effekt in der Population (also z. B. überhaupt kein Zusammenhang oder keinerlei Unterschied), so geben auch die kleinsten Effekte Anlass zur Entscheidung für die Alternativhypothese, wenn sie sich als statistisch signifikant erweisen. Wie bereits in ► Abschn. 14.1 dargestellt, hängt die statistische Signifikanz eines Effekts u. a. vom Umfang der untersuchten Stichprobe ab. Die Nullhypothese als theoretische Aussage, die auf die Realität praktisch niemals exakt zutrifft, ist also gewissermaßen chancenlos. Setzte der Untersuchungsaufwand der Wahl des Stichprobenumfanges keine Grenzen, wäre wohl jede H_0 zu verwerfen, weil in der Realität eher minimale Effekte als exakte Nulleffekte auftreten. Dementsprechend würde bei großem Stichprobenumfang beim klassischen Signifikanztest jeder noch so kleine, praktisch unbedeutende Effekt statistisch signifikant.

Statistische Signifikanz kann deshalb nicht allein als Gradmesser des Aussagegehaltes hypothesenprüfender Untersuchungen angesehen werden. Neben die wichtige Forderung, an Stichproben gewonnene Ergebnisse gegen den Zufall abzusichern, tritt eine weitere: Diese besagt, dass relevante statistische Ergebnisse für Populationsverhältnisse sprechen müssen, die in einer für die Praxis nicht zu vernachlässigenden Weise von den in der H_0 behaupteten Populationsverhältnissen abweichen – oder kurz: signifikante Ergebnisse müssen auch **praktisch bedeutsam** sein. Diese Forderung sowie Alternativen zum klassischen Signifikanztest werden in der sozialwissenschaftlichen Forschungsgemeinschaft bereits seit mehreren Jahren ausführlich diskutiert (für einen Überblick Harlow, Mulaik, & Steiger, 2009; Kline, 2005; Morrison & Henkel, 2006).

Die Kritik gipfelte in dem Vorwurf, der Signifikanztest sei dafür verantwortlich zu machen, dass sich z. B. die Psychologie nicht zu einer **kumulativen Wissenschaft** entwickeln konnte (Kline, 2004, S. 90). Mit kumulativer Wissenschaft ist gemeint, dass mit der Zeit gut gesicherte und bedeutsame wissenschaftliche Befunde angesammelt (kumuliert) werden und damit der Wissensbestand in einem Fach kontinuierlich wächst. Kritiker des klassischen Signifikanztests befürchten, dass die Fixierung allein auf statistisch signifikante Ergebnisse (im Sinne einer Zurückweisung der Nil-Nullhypothese) letztlich zu viele unbedeutende Befunde mit sich bringt. Die Erkenntnislage werde damit unklar und unsicher. Auch wenn wir diese harsche Kritik im Wesentlichen nachvollziehen können, zeigt doch die Verbreitung des klassischen Signifikanztests in der empirischen Forschung, dass ein völliger Verzicht auf diese Testmethode zum momentanen Zeitpunkt wohl kaum durchzusetzen ist. Unsere Leitlinie besteht darin, Ergänzungen und Alternativen zum traditionellen Signifikanztest aufzuzeigen.

Der wichtigste Hinweis besteht darin, beim **Einsatz des klassischen Signifikanztests** neben der statistischen Signifikanz immer auch die **Effektgröße** (bei einem signifikanten Ergebnis) sowie die **Teststärke** (vor allem bei einem nicht-signifikanten Ergebnis) zu betrachten und in die Interpretation einzubeziehen (► Kap. 14). Darüber hinaus ist es empfehlenswert, ergänzend zu den Teststatistiken (z. B. empirischer t-Wert und Irrtumswahrscheinlichkeit beim Vergleich zweier Mittelwerte) häufiger die **Konfidenzintervalle für die Populationseffektgrößen** zu berichten (z. B. die Konferenzintervalle der beiden verglichenen Mittelwerte; Colegrave & Ruxton, 2003; Johnson, 1999, 2005). Bei einem nicht-signifikanten Ergebnis umschließt das Konfidenzintervall den Wert Null. Die Breite und Lage des Konfidenzintervalls – also des Effektgrößenbereiches, der durch die Daten gut gesichert ist – lässt eine bessere Interpretation des Populationseffektes zu als der alleinige Blick auf die Signifikanz (bzw. die Irrtums- bzw. α-Fehler-Wahrscheinlichkeit) oder – bei nicht-signifikanten Befunden – auf die Teststärke (bzw. die β-Fehler-Wahrscheinlichkeit).

Einen Schritt weiter gehen die in diesem Kapitel behandelten **Minimum-Effektgrößen-Tests**. Bei diesen wird anstelle der üblichen Nil-Nullhypothese, die überhaupt keinen Effekt in der Population postuliert und somit unrealistisch ist, eine Nullhypothese zugrunde gelegt, die

einen **minimalen Effekt** (z. B. 1 % Varianzaufklärung oder 5 % Varianzaufklärung) zulässt. Wird ein Minimum-Effektgrößen-Test statistisch signifikant, so schlussfolgert man, dass der Effekt in der Population mit hoher Wahrscheinlichkeit größer als der Minimaleffekt ist.

Bei der Hypothesenprüfung sollte nicht nur die statistische Signifikanz, sondern auch die theoretische und/oder praktive Bedeutsamkeit des empirischen Ergebnisses betrachtet und inhaltlich interpretiert werden. Dafür gibt es drei Strategien:

1. **Verwendung des klassischen Signifikanztests und zusätzliche Diskussion von Effektgröße und Teststärke des Ergebnisses.**
2. **Anstelle des klassischen Signifikanztests werden die Konfidenzintervalle der interessierenden Populationsparameter berichtet, so dass automatisch die Aufmerksamkeit auf der Effektgröße liegt.**
3. **Anstelle des klassischen Signifikanztests werden Minimum-Effektgrößen-Tests durchgeführt.**

Das vorliegende Kapitel beschreibt zunächst das **Good-Enough-Prinzip**, das den Minimum-Effektgrößen-Tests zugrunde liegt (► Abschn. 15.1). Anschließend wird die **Durchführung von Minimum-Effektgrößen-Tests** detailliert erläutert (► Abschn. 15.2). Im Prinzip laufen sie darauf hinaus, die empirische Testgröße des traditionellen Signifikanztests jeweils mit einem modifizierten kritischen Wert zu vergleichen, der aus einer **nichtzentralen Prüfverteilung** („non-central distribution“) stammt. Während eine zentrale Prüfverteilung (z. B. zentrale t-, χ^2-oder F-Verteilung), wie sie im traditionellen Signifikanztest verwendet wird, auf einer Normalverteilung mit dem Mittelwert Null (für den Nulleffekt gemäß H_0) basiert, werden die entsprechenden nichtzentralen Verteilungen (z. B. nichtzentrale t-; χ^2- oder F-Verteilung), die bei Minimum-Effektgrößen-Tests genutzt werden, für unterschiedliche Mittelwerte (d. h. für unterschiedliche Minimumeffekte) konstruiert. Eine nichtzentrale Prüfverteilung wird durch ihren Nichtzentralitätsparameter gekennzeichnet. Kritische Werte für nichtzentrale Prüfverteilungen sind teilweise austabelliert (Tab. 11 im ► Webangebot zu diesem Buch), teilweise lassen sie sich mit entsprechenden Statistik-Tools generieren.

Neben den Minimum-Effektgrößen-Tests stellen **Strukturgleichungsmodelle** eine weitere Alternative zum klassischen Signifikanztest dar. Mithilfe von Strukturgleichungsmodellen (► Kap. 17) lassen sich anstelle von Einzelhypothesen komplexe Hypothesensysteme prüfen, indem man berechnet, wie gut die empirischen Daten zum postulierten Mess- und Strukturmodell passen. Eine gute Passung (Fit) der Daten an das Modell ist hier wünschenswert. Somit ist die Nullhypothese – dergemäß die Daten nicht vom Modell abweichen – in diesem Fall die Wunschhypothese. Auch bei klassischen Signifikanztests sowie bei Minimum-Effektgrößen-Tests ist aus inhaltlichen Erwägungen zuweilen die Nullhypothese (dergemäß es keinen oder nur einen minimalen Effekt in der Population gibt) die von den Forschenden favorisierte „Wunschhypothese“, die es vorläufig zu belegen gilt. So kann die Information, dass zwei Variablen *nicht* zusammenhängen (z. B. Haschischkonsum und spätere Abhängigkeit von harten Drogen) oder dass sich bestimmte Gruppen **nicht** unterscheiden (z. B. Frauen und Männer in der Intelligenz) im Kontext theoretischer Erklärungen sozialwissenschaftlicher Phänomene durchaus aussagekräftig sein. In ► Abschn. 15.3 gehen wir auf die Frage der **„Bestätigung“ von Nullhypothesen** deswegen näher ein. Bei der Prüfung einer Wunsch-Nullhypothese ist die Nil-Nullhypothese zu nutzen und idealerweise mit einer **Minimum-Effektgrößen-Alternativhypothese** zu arbeiten.

Es gibt natürlich auch Kritik an Minimum-Effektgrößen-Tests und den damit verbundenen Post-hoc-Teststärkebestimmungen (Hoenig & Heisey, 2001). Als Alternative zum klassischen Signifikanztest haben sich Konfidenzintervalle für Effektgrößen (z. B. Colegrave & Ruxton, 2003; Johnson, 1999, 2005) bislang stärker etabliert als Minimum-Effektgrößen-Tests, obwohl ihre Durchführung relativ komfortabel möglich ist, wie wir im Folgenden demonstrieren werden.

15.1 Das Good-Enough-Prinzip

Wie bereits ausgeführt ist jede Nullhypothese letztlich chancenlos, wenn man nur genügend große Stichproben untersucht, weil sich dadurch die Teststärke vergrößert und am Ende auch winzige Populationseffekte statistisch signifikant werden. Das bedeutet, dass statistische Signifikanz beim klassischen Nullhypothesen-Signifikanztest letztlich nur eine Frage des Stichprobenumfanges ist. Zudem ist die Nil-Nullhypothese in der Regel eine reine Fiktion: Eine Hypothese, die behauptet, es gäbe überhaupt keinen (nicht einmal den kleinsten) Zusammenhang, überhaupt keinen Unterschied oder keinerlei (auch nur minimale) Maßnahmenwirkungen, ist eigentlich von vorne herein falsch. Mit der Folge, dass die Ablehnung einer H_0 immer richtig ist, es also auch keinen α-Fehler gibt (► Abschn. 12.5.1). Angesichts dieser aus den beiden genannten Gründen wenig ermutigenden Situation stellt sich die Frage, ob bzw. wie man im Rahmen statistischer Hypothesenprüfung eine Theorie-falsifizierende Instanz festlegen könnte, die im Sinne einer strengen

Theorieprüfung in der Logik des Kritischen Rationalismus (▶ Abschn. 2.2) die potenzielle Falsifizierbarkeit von Theorien und daraus abgeleiteten Hypothesen ermöglicht.

Das von Serlin und Lapsley (1993) vorgeschlagene **Good-Enough-Prinzip** („good-enough principle“) geht von der Vorstellung aus, dass die Nullhypothese im traditionellen Signifikanztest, die als **Punkthypothese exakt einen Nulleffekt** postuliert, eine reine Fiktion darstellt, die zwar theoretisch postuliert werden kann, in den allermeisten Fällen aber kein praktisches Pendant hat. Wann ist der Unterschied zwischen realen Populationen wirklich exakt Null? Ist es vorstellbar, dass zwischen zwei Merkmalen überhaupt kein Zusammenhang besteht? Gibt es Maßnahmen, die keinerlei Wirkung zeigen? All diese Fragen sind letztlich zu verneinen. Eine empirische Zurückweisung der – ohnehin unrealistischen – Nil-Nullhypothese wird somit als Indikator für die Gültigkeit der Alternativhypothese als „nicht gut genug“ befunden. Stattdessen soll gemäß dem Good-Enough-Prinzip eine **Nullhypothese als Bereichshypothese** bzw. zusammengesetzte Hypothese („range null hypotheses“, „good-enough belts“) aufgestellt werden, die alle vernachlässigbar kleinen Populationseffekte umfasst. Wenn diese Minimum-Effekt-Nullhypothese anhand der Daten zurückzuweisen ist, dann ist das ein Indikator, der „gut genug“ ist, um eine Alternativhypothese anzunehmen, die einen substanziellen bzw. praktisch bedeutsamen Populationseffekt postuliert. Kurz gesagt:

Das Good-Enough-Prinzip transformiert die Nullhypothese von einer Punkthypothese in eine Bereichshypothese. Während die traditionelle Nil-Nullhypothese – bezogen auf einen ungerichteten Mittelwertvergleich – formuliert wird als

$$H_0 : \mu_1 - \mu_2 = 0 \text{ bzw. } H_0 : \mu_1 = \mu_2$$

formulieren wir nach dem Good-Enough-Prinzip eine Minimum-Effekt-Nullhypothese („minimum effect null hypothesis“) als Bereichshypothese:

$$H_0 : |\mu_1 - \mu_2| < \Delta_K \text{ bzw. } H_0 : \mu_1 < \mu_2 + \Delta_K$$

wobei Δ_K den kleinsten Effekt symbolisiert, der gerade noch als „gut genug“ akzeptiert werden kann. Jede Differenz bzw. jeder Effekt, der kleiner ist als Δ_K, bedeutet Falsifikation der Alternativhypothese und Beibehaltung der Nullhypothese.

Good-Enough-Prinzip – Das Good-Enough-Prinzip („good enough principle“) besagt, dass für die Annahme einer Alternativhypothese festgelegt werden sollte, welche Effektgrößen hinsichtlich praktischer Bedeutsamkeit „gut genug“ sind. Dass ein Effekt vom Wert Null abweicht, wird als nicht gut genug angesehen. Stattdessen sollte ein ganzer Bereich unbedeutsam kleiner Effekte definiert und die Alternativhypothese nur dann akzeptiert werden, wenn der empirische Effekt außerhalb des Bereichs vernachlässigbar kleiner Effekte liegt.

Minimum-Effekt-Nullhypothese – Eine Minimum-Effekt-Nullhypothese („minimum effect null hypotheses“) postuliert im Sinne des Good-Enough-Prinzips als Bereichshypothese einen Bereich vernachlässigbar kleiner Populationseffekte.

Minimum-Effekt-Nullhypothesen sind keineswegs bei großen Stichproben chancenlos, denn diese Nullhypothesen sind keine reine Fiktion. Sie können tatsächlich richtig sein, womit auch das α-Fehler-Konzept wieder sinnvoll ist: Wenn sich der „wahre“ Effekt nicht im „Schutzgürtel“ des Good-Enough-Bereichs befindet, kann es niemals zu einem signifikanten Ergebnis kommen, egal wie groß die Stichproben sind.

Mit dieser Modifikation des traditionellen Signifikanztests – Nil-Nullhypothesen werden durch Minimum-Effekt-Nullhypothesen ersetzt – sind wir besser in der Lage, Theorien „qualifiziert“ zu falsifizieren. Obwohl diese Modifikation in der Methodenliteratur seit mehreren Jahrzehnten intensiv und durchaus positiv diskutiert wird (Denis, 2003; Fowler, 1985; Greenwald, 1975; Murphy, 1990; Murphy & Myors, 1999; Murphy et al., 2009; Rindskopf, 2009; Serlin & Lapsley, 1993), hat sie sich in der empirischen Forschungspraxis bisher nicht durchgesetzt (Cortina & Dunlap, 1997; Kline, 2005). Dies liegt vermutlich u. a. darin begründet, dass die einschlägigen Statistik-Softwarepakete diese Modifikation von klassischen Signifikanztests noch nicht integriert haben. Wie wir jedoch im folgenden ▶ Abschn. 15.2 sehen werden, ist das Good-Enough-Prinzip ohne besondere Mühe praktisch umsetzbar.

15.2 Überprüfung von Minimum-Effekt-Nullhypothesen

Wir haben schon festgehalten, dass Minimum-Effekt-Nullhypothesen einen Bereich vernachlässigbar kleiner Effekte umfassen. Doch wie ist dieser **Bereich vernachlässigbar geringer Effektgrößen** festzulegen? Generell müssen Effektgrößen immer im Kontext des jeweiligen Forschungsgegenstandes betrachtet und interpretiert werden. Denn ein Populationszusammenhang von $\rho = +.20$ kann beispielsweise in einem Forschungsgebiet theoretisch und praktisch sehr bedeutsam, gleichzeitig in einem anderen vernachlässigbar gering sein – das hängt vom vorliegenden Forschungsstand und von den Schlussfolgerungen ab, die aus den Daten gezogen werden sollen. Trotz der Notwendigkeit der **kontextspezifischen Interpretation von Effektgrößen** haben sich grobe Richtwerte eingebürgert, etwa die in Anlehnung an den Statistiker Cohen etablierte Einteilung von Effekten in die Kategori-

en kleiner, mittlerer oder großer Effekt (► Abschn. 14.2). Das Arbeiten mit dieser **allgemeinen Effektgrößen-Klassifikation** ist pragmatisch oft hilfreich (und allemal besser als ein Ignorieren der Effektgrößen-Problematik), muss aber immer wieder zugunsten inhaltsspezifischer Einordnungen von Effektgrößen hinterfragt und relativiert werden.

Dasselbe Spannungsverhältnis gilt auch für die Festlegung von Minimum-Effekt-Nullhypothesen. Idealerweise würde man hier die Grenze zwischen vernachlässigbar geringen und praktisch bedeutsamen Effektgrößen jeweils gegenstands- bzw. studienspezifisch festlegen. Pragmatisch operiert man aber oft mit einer allgemeinen Grobeinteilung, die sich auf die **Varianzaufklärung** („percentage of variance" PV) bezieht (Murphy et al., 2009). Als Effektgrößenmaß für die Varianzaufklärung hat sich der aus der Varianzanalyse oder Korrelationsrechnung bekannte **η^2-Koeffizient** (Eta-Quadrat-Koeffizient) etabliert. Der η^2-Koeffizient lässt sich auch als Determinationskoeffizient auffassen, d. h. als quadrierter Korrelationskoeffizient ρ (rho) in der Population (► Abschn. 14.2).

Mit Blick auf die Varianzaufklärung lässt sich über die traditionelle **Nil-Nullhypothese** H_0 sagen, die einen exakten Null-Effekt postuliert, dass sie von **0 % Varianzaufklärung** ausgeht (H_0: $\eta^2 = 0$). Dies lässt sich symbolisieren als $\boldsymbol{H_{00}}$.

Will man nun einen Bereich vernachlässigbar geringer Varianzaufklärungen für eine Minimum-Effekt-Nullhypothese definieren, so wird häufig eine **Varianzaufklärung zwischen 0 % und 1 %** (H_0: $\eta^2 \leq .01$) als vernachlässigbar gering eingestuft. Dies lässt sich symbolisieren als $\boldsymbol{H_{01}}$. Wird die H_{01} abgelehnt und die Alternativhypothese H_1 angenommen, bedeutet dies, dass der geprüfte Effekt in dem Sinne nicht zu vernachlässigen ist, dass die Varianzaufklärung größer als 1 % bzw. der Korrelationskoeffizient größer als .10 ausfällt.

Nun gibt es auch Forschungsbereiche, in denen man üblicherweise mit recht beachtlichen Effekten rechnet (z. B. Trainingseffekte in schul- oder entwicklungspsychologischen Untersuchungen). Wenn dies der Fall ist, sollte der zu vernachlässigende Minimaleffekt auf eine **Varianzaufklärung zwischen 0 % und 5 %** heraufgesetzt werden (H_0: $\eta^2 \leq .05$). Dies lässt sich symbolisieren als $\boldsymbol{H_{05}}$. Wird die H_{05} abgelehnt und die Alternativhypothese H_1 angenommen, kann man davon ausgehen, dass der geprüfte Effekt in der Population mit einer Varianzaufklärung von mehr als 5 % bzw. einer Korrelation über .22 verbunden ist.

Effektgrößen in Minimum-Effekt-Nullhypothesen – Bei Minimum-Effekt-Nullhypothesen wird als Effektgröße die Varianzaufklärung herangezogen. Dabei wird der Bereich vernachlässigbar kleiner Effektgrößen traditionell entweder auf 0 %–1 % Varianzaufklärung (H_{01}: $\eta^2 \leq .01$) oder auf 0 %–5 % Varianzaufklärung (H_{05}: $\eta^2 \leq .05$) festgelegt.

Man mag nun vielleicht einwenden, dass auch ein Minimaleffekt von 5 % noch vergleichsweise klein wirkt – denn er geht mit immerhin 95 % unaufgeklärter Varianz einher. In den Sozialwissenschaften, in denen viele Phänomene durch eine Fülle von Einflussfaktoren zustande kommen, von denen eine einzelne Studie stets nur einen kleinen Ausschnitt beleuchtet, sind jedoch Varianzaufklärungen im geringen einstelligen Bereich durchaus typisch. Ein Testen gegen Minimaleffekte über 5 % Varianzaufklärung ist somit nur sinnvoll, wenn theoretisch begründet alle kleineren Effekte bedeutungslos sind.

Man beachte, dass H_{01} und H_{05} als Minimum-Effekt-Nullhypothesen durchaus zutreffen können, dass also das α-Fehler-Konzept bei dieser Art von Nullhypothesen sinnvoller ist als bei der eigentlich immer falschen Nil-Nullhypothese. Dies bedeutet auch, dass Minimum-Effekt-Nullhypothesen keineswegs – anders als Nil-Nullhypothesen – bei genügend großen Stichprobenumfängen immer abgelehnt werden.

Im Folgenden wenden wir uns der Frage zu, wie Minimum-Effekt-Nullhypothesen getestet werden und welche Teststärke mit Tests dieser Art verbunden ist.

15.2.1 Signifikanzschranken und Teststärkeanalysen

Zentral für die Durchführung von Minimum-Effektgrößen-Tests in der Tradition von Murphy und Myors (1998, 2004; Murphy et al., 2009) ist die sog. **„One Stop F Table"** (Alles-auf-einen-Blick-Tabelle), mit deren Hilfe Signifikanzüberprüfungen und Teststärkeanalysen denkbar einfach (sozusagen bei einem „einzigen Stop" bzw. auf „einen Blick") durchzuführen sind. Die Tabelle enthält **kritische F-Werte**, die – ähnlich wie F-Tabellen in Statistikbüchern (z. B. Tab. E bei Bortz & Schuster, 2010) – unterschiedlichen Perzentilen von F-Verteilungsfunktionen mit variablen Zählerfreiheitsgraden (df_Z) und Nennerfreiheitsgraden (df_N) entsprechen. In der **Alles-auf-einen-Blick-Tabelle** (Tab. 11 im ► Webangebot zu diesem Buch) sind jedem df_Z/df_N-Paar 12 F-Werte zugeordnet mit folgender Bedeutung:

- 6 Werte entsprechen den Signifikanzschranken (Signifikanzniveau von 5 %: $\alpha = .05$ und von 1 %: $\alpha = .01$) zur Prüfung von H_{00}, H_{01} und H_{05} (Nullhypothesen, die von einem Mindesteffekt von 0 %, 1 % oder 5 % Varianzaufklärung ausgehen).
- 3 Werte stellen F-Äquivalente dar, die zu erreichen sind, wenn eine Untersuchung zur Überprüfung von H_{00}, H_{01} oder H_{05} mit einer Teststärke von $1 - \beta = 50$ % ausgestattet sein soll.
- 3 Werte stellen F-Äquivalente dar, die zu erreichen sind, wenn eine Untersuchung zur Überprüfung von

H_{00}, H_{01} oder H_{05} mit einer Teststärke von $1 - \beta = 80\,\%$ ausgestattet sein soll.

Wir werden diese 12 Werte weiter unten an einem Beispiel genauer erklären.

F-Verteilungen bzw. F-Tests werden v. a. im Rahmen der Varianzanalyse bzw. des Allgemeinen Linearen Modells (ALM) eingesetzt. Dementsprechend handelt es sich bei dem gleich präsentierten einführenden Beispiel um eine Varianzanalyse. Dass diese Tabelle jedoch nicht nur für varianzanalytische Auswertungen von Wert ist, sondern für die wichtigsten in der Inferenzstatistik verwendeten Signifikanztests einsetzbar sind (indem man ihre Kennwerte in F-Äquivalente umrechnet), werden wir in ▶ Abschn. 15.2.3 erläutern.

Ein Beispiel (in Anlehnung an Murphy & Myors, 1998, S. 41ff.) soll die Handhabung der Alles-auf-einen-Blick-Tabelle verdeutlichen: Mit einer einfaktoriellen Varianzanalyse werden vier verschiedene Methoden (unabhängige Variable = 4 Treatmentstufen) verglichen ($p = 4$). Eine Stichprobe von $N = 54$ Probanden wird zufällig den 4 Treatmentstufen zugewiesen, d. h., man erhält in diesem Beispiel ungleich große Stichproben wie z. B. $n_1 = n_2 = 13$ und $n_3 = n_4 = 14$. Man formuliert eine **spezifische Alternativhypothese**, die behauptet, dass die unabhängige Variable mindestens 15 % ($\eta^2 = .15$) der Varianz der abhängigen Variablen erklärt. Dieser Wert entspricht nach Tab. 14.5 einem mittleren bis großen Effekt.

Die Varianzanalyse führt zu $F_{emp} = 3.50$ mit 3 Zählerfreiheitsgraden ($df_Z = p - 1 = 3$) und 50 Nennerfreiheitsgraden ($df_N = N - p = 50$). Über Gl. 14.36 schätzt man eine Varianzaufklärung von 17.4 % ($\hat{\eta}^2 = .174$). Dieser Wert lässt sich auch direkt aus dem F-Wert ableiten. Es gilt

$$\hat{\eta}^2 = \frac{df_Z \cdot F}{df_Z \cdot F + df_N}, \qquad (15.1)$$

im Beispiel: $\hat{\eta}^2 = \frac{3 \cdot 3.50}{3 \cdot 3.50 + 50} = 0.174.$

(Die Gleichung gilt auch für den Populationsparameter η^2, ▶ Abschn. 14.2.3)

Prüfung von H_{00}

In Tab. 15.1 werden noch einmal die 12 Werte gezeigt, die in der Alles-auf-einen-Blick-Tabelle, Tab. 11 (▶ Webangebot zu diesem Buch) unter $df_Z = 3$ und $df_N = 50$ genannt sind. Die Werte sind hier durchnummeriert, worauf bei den folgenden Erklärungen Bezug genommen wird.

Der erste Wert stellt die Signifikanzschranke ($\alpha = .05$) für die Überprüfung der traditionellen Nil-Nullhypothese (H_{00}) dar. Man findet diesen Wert in jeder F-Tabelle von

Tabelle 15.1 Auszug aus der Alles-auf-einen-Blick-Tabelle, Tab. 11 im ▶ Webangebot zu diesem Buch für $df_Z = 3$ und $df_N = 50$

nil $\alpha = .05$	2.79	(1)
nil $\alpha = .01$	4.20	(2)
pow .50	1.99	(3)
pow .80	3.88	(4)
1 % $\alpha = .05$	3.24	(5)
1 % $\alpha = .01$	4.85	(6)
pow .50	2.46	(7)
pow .80	4.48	(8)
5 % $\alpha = .05$	4.84	(9)
5 % $\alpha = .01$	6.98	(10)
pow .50	4.08	(11)
pow .80	6.55	(12)

Statistikbüchern wie z. B. bei Bortz & Schuster (2010, Tab. E). Er lautet für $df_Z = 3$ und $df_N = 50$ dementsprechend $F_{crit(.05)} = 2.79$. Wegen $F_{emp} > F_{crit(.05)}$ ($3.50 > 2.79$) ist der F-Wert für $\alpha = .05$ signifikant, d. h., die H_{00} wird mit einer Irrtumswahrscheinlichkeit $p < .05$ verworfen.

Der zweite Wert (4.20) ist analog zum ersten Wert zu interpretieren für $\alpha = .01$. Da $F_{emp} = 3.50$ kleiner ist als $F_{crit(.01)} = 4.20$. kann die H_{00} für $\alpha = .01$ bzw. auf dem 1 %-Signifikanzniveau nicht verworfen werden.

Bevor wir uns der dritten und vierten Zahl zuwenden, sind einige erläuternde Vorbemerkungen angebracht. Wir wollen einmal annehmen, dass eine Untersuchung mit einer **Teststärke** („power") von $1 - \beta = .50$ durchgeführt wird. Nach den Ausführungen von ▶ Abschn. 14.1 besagt dieser Wert, dass der Signifikanztest mit einer Wahrscheinlichkeit von 50 % zu einem signifikanten Ergebnis führt, wenn die H_1 gilt. Die Chance für ein signifikantes Ergebnis entspricht also der Chance für z. B. „Kopf" beim Münzwurf, d. h., hier wird Wissenschaft zu einem reinen Glücksspiel.

Wenn man nun zusätzlich in Rechnung stellt, dass viele wissenschaftliche Zeitschriften bevorzugt Artikel mit statistisch signifikantem Ergebnis publizieren – oft ohne dass dabei die Teststärke der Untersuchungen ausreichend kontrolliert wird – kommt man leicht zu der in ▶ Abschn. 14.1 zitierten Behauptung, dass die Psychologie (und sicherlich auch andere Human- und Sozialwissenschaften) keine kumulative Wissenschaft sei: Studien, die **„under-**

Das Problem mangelnder Teststärke

Auf Studien mit einer Teststärke („power“) von 50 % oder sogar weniger sollte verzichtet werden. Denn hier ist bei einem statistisch nicht-signifikanten Ergebnis unklar, ob tatsächlich in der Population kein Effekt vorliegt oder der Test nur eine zu geringe Teststärke hatte (▶ Abschn. 14.1), um den vorhandenen Effekt nachzuweisen. Studien mit mangelnder Teststärke („underpowered“) sollten durch entsprechende Untersuchungsplanung (Orientierung an optimalen Stichprobenumfängen, ▶ Abschn. 14.3) möglichst im Vorfeld vermieden werden. Wird die mangelnde Teststärke erst im Nachhinein festgestellt, so ist diese Schwäche der Studie ausdrücklich zu benennen. Eine solche Studie sollte in der Regel unveröffentlicht bleiben, da ihr Ergebnis nicht eindeutig interpretierbar ist und sie somit keinen Beitrag zum kumulativen Erkenntnisgewinn im Forschungsfeld liefert.

powered“ sind (d. h. mit Teststärken von $1-\beta \leq .50$ operieren), produzieren zufällige und damit möglicherweise auch widersprüchliche Ergebnisse (▶ Das Problem mangelnder Teststärke).

Nach diesen Vorbemerkungen wenden wir uns dem dritten Wert zu. Es handelt sich um ein ***F*-Äquivalent**, das einem Populationseffekt entspricht, bei dem die Untersuchung (mit $df_Z = 3$, $df_N = 50$, $\alpha = .05$) eine Teststärke von 50 % hat ($1-\beta = .50$). Für das Beispiel ist dies der Wert $F_{.50} = 1.99$. Die Varianzaufklärung (η^2), die diesem F-Wert entspricht, errechnet man über Gl. 15.1). Für das Beispiel ergibt sich

$$\eta^2 = \frac{3 \cdot 1.99}{50 + 3 \cdot 1.99} = .107.$$

Die traditionelle Nullhypothese (H_{00}) wäre (bei gegebenem N und $\alpha = .05$) mit einer Wahrscheinlichkeit von 50 % zu verwerfen, wenn die unabhängige Variable ca. 11 % der Varianz der abhängigen Variable erklären würde. Die Untersuchungsplanung ging von 15 % Varianzaufklärung aus, d. h., die Untersuchung hat eine Teststärke über 50 %.

Der vierte Wert stellt ein F-Äquivalent für einen Effekt dar, bei dem der Signifikanztest (bei gegebenem N und $\alpha = .05$) eine Teststärke von 80 % ($1-\beta = .80$) haben würde. Wir entnehmen hierfür Tab. 15.1 den Wert $F_{.80} = 3.88$. Über Gl. 15.1 resultiert hierfür $\eta^2 = .189$ bzw. eine Varianzaufklärung von ca. 19 %, die über der angenommenen Varianzaufklärung von 15 % liegt. Die Untersuchung hat also eine Teststärke, die zwischen 50 % und 80 % liegt.

Eine genauere Schätzung der Teststärke lässt sich durch einfache lineare Interpolation ermitteln. Hierfür nennen Murphy und Myors (1998, S. 46) folgende Formel:

$$(1-\beta)_{\text{int pol}} = .50 + \left(\frac{F_{\text{Hyp}} - F_{.50}}{F_{.80} - F_{.50}} \cdot .30\right). \qquad (15.2)$$

$1-\beta_{\text{int pol}}$ ist die interpolierte Teststärke und $F_{.50}$ sowie $F_{.80}$ sind die in Tab. 15.1 genannten Werte (Ziffer 3 und 4). Den F-Wert, der dem hypothetisch vorgegebenen η^2 entspricht, bezeichnen wir als F_{Hyp}. F_{Hyp} ergibt sich nach folgender Gleichung:

$$F_{\text{Hyp}} = \frac{\eta^2 \cdot df_N}{(1-\eta^2) \cdot df_Z}, \qquad (15.3)$$

im Beispiel:

$$F_{\text{Hyp}} = \frac{.15 \cdot 50}{(1-.15) \cdot 3} = 2.94.$$

Eingesetzt in Gl. 15.2 erhält man

$$(1-\beta)_{\text{int pol}} = .50 + \left(\frac{2.94 - 1.99}{3.88 - 1.99}\right) \cdot .30 = .65.$$

Die Untersuchung hat also eine Teststärke von 65 %.

Statt die Werte $F_{.50}$ und $F_{.80}$ über Gl. 15.1 in η^2-Werte zu transformieren, um so herauszufinden, dass der hypothetisch angenommene η^2-Wert ($\eta^2 = .15$) zwischen diesen beiden η^2-Werten liegt ($.107 < .15 < .189$), kann man alternativ $\eta^2 = .15$ direkt über Gl. 15.3) in einen F_{Hyp}-Wert transformieren. Ist $F_{\text{Hyp}} \leq F_{.50}$, hat die Untersuchung eine zu geringe Teststärke. Für $F_{\text{Hyp}} \geq F_{.80}$ ist die Teststärke zufriedenstellend und für $F_{.50} < F_{\text{Hyp}} < F_{.80}$ erhält man eine (für praktische Zwecke ausreichend genaue) Teststärkenschätzung über Gl. 15.2.

Man beachte, dass Gl. 15.2 nur für die Schätzung von Teststärken zwischen .50 und .80 geeignet ist. Teststärken außerhalb dieses Bereiches können über Gl. 15.2 nicht ermittelt werden. Dies ist in der Regel auch nicht erforderlich, denn Untersuchungen mit einer Teststärke unter .50 sollten – wie gesagt – nicht durchgeführt und schon gar nicht veröffentlicht werden. Genauere Angaben für diesen Bereich erübrigen sich also.

Eine **Teststärke von mindestens .80** wird mittlerweile von der Scientific Community als ausreichend akzeptiert. Hat man also eine Untersuchung so geplant, dass eine Teststärke von mindestens 80 % gewährleistet ist, sind weitere Korrekturen nicht erforderlich. Man erreicht diese Teststärke mit den in Tab. 14.11 genannten **optimalen Stichprobenumfängen** (für $\alpha = .05$ bzw. .01 und kleine, mittlere und große Effekte). Will man erfahren, mit welchem Effekt, Stichprobenumfang und α-Fehler Teststärken über 80 % erzielt werden, seien die ausführlichen Tabellen bei Cohen (1988) empfohlen sowie das **kostenlose Software-Tool G*Power** von Erdfelder, Faul und Buchner (1996).

Nach diesen Ausführungen wollen wir fragen, was angesichts einer zu geringen Teststärke (im Beispiel 65 %) zu tun ist.

In der **Planungsphase** hätte dieses Ergebnis Maßnahmen zur Erhöhung der Teststärke durch die Wahl eines **optimalen Stichprobenumfanges** veranlassen müssen. Hierbei ist zunächst zu prüfen, ob größere Stichproben untersucht werden können. Wie groß die Stichprobe sein muss, um mit $\alpha = .05$ und einer angenommenen Effektgröße von $\eta^2 = .15$ eine Teststärke von $1 - \beta = .80$ zu erzielen, entnimmt man einfachheitshalber Tab. 15.2 (eine erste Orientierung hierfür bietet auch Tab. 14.11 für kleine, mittlere und große Effekte; genauere Werte für variable Effektgrößen und α-Fehler enthalten die **Sample Size Tables** von Cohen, 1988).

Wir entnehmen Tab. 15.2 für $df_Z = 3$ und $\eta^2 = .15$ den Wert $df_N = 65$. Der optimale Stichprobenumfang N_{opt} ergibt sich (wegen $df_N = N - p$) zu $df_N + p$, d. h., wir erhalten $N_{opt} = 65 + 4 = 69$. Somit hätten 15 zusätzliche Untersuchungspersonen (und somit $n = 69$ statt $n = 54$) bereits ausgereicht, um die Teststärke von 65 % auf 80 % zu steigern.

Eine weitere Maßnahme zur Erhöhung der Teststärke besteht – wie in ▶ Abschn. 14.3.1 bereits erwähnt – darin, das **maximal tolerierbare α-Fehler-Niveau** („type I error") von z. B. $\alpha = .05$ auf $\alpha = .10$ zu erhöhen (Murphy & Myors, 1998, S. 15, 80f.). Dies ist angesichts der Tatsache, dass die Nil-Nullhypothese (H_{00}) praktisch immer falsch ist (die H_{00} also unter diesen Umständen niemals fälschlicherweise, sondern nur korrekterweise verworfen werden kann), eine durchaus akzeptable Maßnahme. Wie sich die Teststärke durch Vergrößern des α-Fehler-Niveaus erhöht, ist den **Sample Size Tables** (Cohen, 1988) zu entnehmen.

Eine dritte Maßnahme, um die Teststärke zu erhöhen, besteht – neben der Steigerung des Stichprobenumfangs und der Lockerung des Signifikanzniveaus – in der **Vergrößerung des angenommenen Effektes**. Im Beispiel hat die unabhängige Variable 17.3 % der abhängigen Variablen erklärt. Entspräche dieser Schätzwert dem wahren η^2, könnte man für diesen Wert (bzw. für sein F-Äquivalent von $F = 3.50$) über Gl. 15.2 die Teststärke schätzen. Man erhält mit $(1-\beta)_{\text{int pol}} = .74$ einen günstigeren Wert als die oben ermittelte Teststärke von 65 % (zur Problematik dieser Bestimmung von „**observed power**" s. Hoenig & Heisey, 2001).

An dieser Stelle wollen wir noch einmal darauf zurückkommen, dass hypothesenprüfende Untersuchungen mit einer Teststärke von 50 % oder weniger in der Regel nicht veröffentlicht werden sollten. Falls sich diese Regel durchsetzt, könnte es da nicht zu nachträglichen Effektstärkemanipulationen kommen?

Angenommen, eine Untersuchung mit einem signifikanten Ergebnis soll zur Publikation eingereicht werden, hat aber – ermittelt über Tab. 15.1 (oder Alles-auf-einen-Blick-Tabelle, Tab. 11 im ▶ Webangebot zu diesem Buch) – leider nur eine Teststärke unter 50 %. Die Untersuchung würde also nicht publiziert werden. Die Untersuchungsplanung möge von einem mittleren Effekt ausgegangen sein, d. h. von $\eta^2 = .10$. Könnte man nun nicht im Nachhinein den **angenommenen** Effekt vergrößern, um damit – bei unverändertem N und α – die Teststärke zu erhöhen? Oder noch deutlicher formuliert: Könnte man den angenommenen Effekt nicht so weit erhöhen, bis eine akzeptable Teststärke erreicht ist?

Diese Manipulationsmöglichkeit ist fraglos gegeben. Um ihr zu entgegnen, müsste eine neue Herausgeberpolitik nicht nur eine Angabe zur Teststärke verlangen, sondern zusätzlich auf der Information bestehen, auf welcher Grundlage (z. B. frühere Studien) von welchem Effekt die Teststärkeberechnung (bei gegebenem N und α) ausging. Wenn sich hierbei herausstellt, dass die Untersuchungsplanung mit einem unrealistisch großen Effekt gerechnet hat, müsste dies ebenfalls eine Ablehnung (oder zumindest eine gründliche Revision) des Artikels nach sich ziehen. Dies zu entscheiden, setzt natürlich entsprechend sensibilisierte Gutachterinnen und Gutachter im Rahmen des Peer-Review-Prozesses voraus (zu Problemen der Post-hoc-Teststärkeanalyse ▶ Abschn. 14.1.1).

Untersuchungen in einem Wissensbereich ohne etablierte Forschungstradition sollten jedoch wegen ihres potenziellen Innovationspotenzials nicht nach dieser Regel bewertet werden. Hier kann die erwartete Effektgröße und damit auch die notwendige Teststärke im Vorfeld nicht genau eingegrenzt werden. Empirische Studien zu bislang unerforschten Gegenständen werden jedoch in der Regel ohnehin eher explorativ und nicht hypothesenprüfend angelegt und ausgewertet (zur explorativen Datenanalyse ▶ Abschn. 12.3).

Prüfung von H_{01}

Nachdem im vergangenen Abschnitt die ersten 4 Ziffern der Tab. 15.1 ausführlich erläutert wurden, sind die nächsten 4 Ziffern schnell erklärt. Vom Prinzip her haben sie die gleiche Bedeutung wie die ersten 4 Ziffern mit der Besonderheit, dass jetzt nicht die traditionelle Nil-Nullhypothese getestet wird, sondern eine Minimum-Effekt-Nullhypothese mit höchstens 1 % Varianzaufklärung (H_{01}).

Auf die Erörterung technischer Aspekte wird hier verzichtet. Die H_{00} wird über die **zentrale F-Verteilung** geprüft und die H_{01} (sowie die H_{05} und alle Teststärkeangaben) über sog. **nichtzentrale F-Verteilungen**. Hin-

Tabelle 15.2 Optimale Stichprobenumfänge (df_N) bei der Überprüfung der traditionellen Nil-Nullhypothese (H_{00}) in Abhängigkeit von der Effektgröße (η^2) und df_Z für $1 - \beta = .80$ und $\alpha = .05$; Erläuterungen s. Text. (Mod. nach Murphy & Myors, 1998, Tab. 3.1. Republished by permission of Taylor & Francis Ltd, from Murphy, K. R., & Myors, B. (1998). Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests. Permission conveyed through Copyright Clearance Center, Inc.)

	df_Z																
η^2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
.01	775	952	1 072	1 165	1 260	1 331	1 394	1 451	1 504	1 580	1 670	1 825	1 992	2 302	2 565	3 027	4 016
.02	385	473	533	579	627	662	694	722	762	787	832	909	993	1 176	1 313	1 513	2 010
.03	255	313	353	384	416	439	460	479	505	522	552	603	660	782	874	1 008	1 341
.04	190	233	263	286	310	328	343	358	377	390	413	451	494	585	654	774	1 031
.05	151	186	209	228	247	261	273	285	300	310	329	359	402	466	522	618	825
.06	125	154	173	189	204	216	227	236	249	257	273	298	333	388	434	514	687
.07	106	131	148	161	174	184	193	204	212	220	233	255	285	331	371	440	601
.08	92	114	128	140	152	160	168	178	185	191	203	222	248	289	324	384	525
.09	81	100	113	124	134	142	149	157	164	169	179	196	220	256	287	341	466
.10	73	90	101	110	120	127	133	141	146	152	161	176	197	230	258	312	419
.11	66	81	91	101	108	115	120	127	132	137	148	159	178	208	238	283	388
.12	60	74	83	92	99	104	110	116	121	125	135	145	163	190	218	259	355
.13	55	68	76	84	90	96	101	106	111	115	124	133	150	178	200	238	327
.14	50	62	70	78	83	88	94	98	102	106	114	123	138	165	185	220	302
.15	47	58	65	72	77	82	87	91	95	98	106	115	129	153	172	205	286
.16	43	54	61	67	72	76	81	85	88	92	99	107	120	143	161	192	268
.17	40	50	57	63	68	72	76	80	83	86	93	101	112	134	151	183	251
.18	38	47	53	59	63	67	71	75	78	81	87	96	106	126	142	172	236
.19	36	44	50	55	59	63	67	70	73	77	82	90	101	119	136	163	227
.20	34	42	47	52	56	60	64	67	69	73	77	85	96	112	129	154	214
.22	30	37	42	47	51	54	57	60	62	65	70	76	86	102	116	139	194
.24	27	34	39	42	46	49	52	54	57	59	63	69	78	93	105	128	178
.26	25	31	35	38	42	44	47	49	52	54	58	63	71	85	96	117	163
.28	22	28	32	35	38	41	43	45	48	49	53	58	65	78	90	107	152
.30	21	26	30	32	35	37	40	42	44	45	49	53	61	72	83	100	142
.32	19	24	27	30	33	35	37	39	40	42	45	50	56	68	76	93	131
.34	18	22	25	28	30	32	34	36	38	39	42	46	52	63	72	87	123

weise zur Mathematik nichtzentraler Verteilungen findet man z. B. bei Cumming und Finch (2001, inkl. Software), Johnson und Kotz (1970), Kendall und Stuart (1973) oder auch Murphy und Myors (1998, Appendix A).

Die fünfte Ziffer ist mit 3.24 kleiner als der empirische F-Wert ($F_{emp} = 3.50 > 3.24$), d. h., auch die H_{01} kann auf dem $\alpha = .05$-Niveau verworfen werden. Mit anderen Worten: die Behauptung, das Treatment erklärt höchstens

1 % der Varianz der abhängigen Variablen (H_{01}), wird mit $\alpha = .05$ verworfen. Stattdessen wird die Alternativhypothese, dass das Treatment mehr als 1 % Varianzaufklärung leistet, angenommen.

Die sechste Zahl entspricht dem kritischen F-Wert für die Überprüfung von H_{01} auf dem $\alpha = .01$-Niveau. Der Wert ist größer als der empirische F-Wert ($F_{emp} = 3.50 < 4.85$), d. h., die H_{01} kann – ebenso wie bei der Nil-Nullhypothese – für $\alpha = .01$ nicht verworfen werden.

Den nächsten beiden Zahlen ist zu entnehmen, wie groß der F-Wert sein müsste, damit der F-Test (für $\alpha = .05$) eine Teststärke von 50 % (Ziffer 7) bzw. eine Teststärke von 80 % hätte (Ziffer 8). Dies sind die Werte $F_{.50} = 2.46$ und $F_{.80} = 4.48$. Diese Werte sind zu vergleichen mit dem oben bereits berechneten Wert $F_{Hyp} = 2.94$, der der angenommenen Varianzaufklärung ($\eta^2 = .15$) entspricht. Wir registrieren $2.46 < 2.94 < 4.48$, d. h., die Teststärke zur Überprüfung von H_{01} liegt zwischen 50 % und 80 %. Den genauen Wert errechnen wir wieder über Gl. 15.2:

$$(1-\beta)_{\text{int pol}} = .50 + \left(\frac{2.94-2.46}{4.48-2.46}\right) \cdot .30 = .57.$$

Wir stellen also fest, dass die Teststärke des F-Tests zur Überprüfung der H_{01} mit 57 % geringer ist als die Teststärke zur Überprüfung von H_{00} (mit 65 %).

Ein Signifikanztest zur Prüfung von H_{01} (Minimum-Effektgrößen-Nullhypothese, die eine Varianzaufklärung von bis zu 1 % als unbedeutend klassifiziert) hat bei sonst gleichen Bedingungen eine geringere Teststärke als der entsprechende Signifikanztest zur Prüfung von H_{00}.

Wie groß hätte die Stichprobe sein müssen, um die $\boldsymbol{H_{01}}$ mit einer Teststärke von 80 % (für $\boldsymbol{\alpha = .05}$) verwerfen zu können? Eine Antwort gibt Tab. 15.3.
Wir entnehmen dieser Tabelle für $df_Z = 3$ und $\eta^2 = .15$ den Wert $df_N = 80$ und erhalten damit $N_{opt} = 80 + 4 = 84$. Eine Stichprobe von $N_{opt} = 84$ wäre also genügend groß gewesen, um die H_{01} für $\alpha = .05$ und einem angenommenen Effekt von $\eta^2 = .15$ mit einer Teststärke von 80 % verwerfen zu können. Die gewählte Stichprobe von $N = 54$ war demgegenüber suboptimal und lieferte eine zu geringe Teststärke von nur 57 %.

Prüfung von H_{05}

Die Bedeutung der letzten 4 Ziffern in Tab. 15.1 liegt nach den bisherigen Ausführungen auf der Hand. Die Werte 4.84 (9. Wert) und 6.98 (10. Wert) sind die kritischen Signifikanzschranken ($\alpha = .05$ und $\alpha = .01$) zur Überprüfung der H_{05}. Der empirische F-Wert ($F_{emp} = 3.50$) ist kleiner als diese kritischen Werte, d. h., die H_{05} kann auf dem $\alpha = .05$-Niveau nicht verworfen werden (und damit auch nicht für $\alpha = .01$). Mit anderen Worten: Die Behauptung, die „wahre" Treatmentwirkung sei zu vernachlässigen, weil sie höchstens eine Varianzaufklärung von 5 % erzielt (H_{05}), kann für $\alpha = .05$ nicht verworfen werden. Warum dies trotz des relativ hohen $\hat{\eta}^2$-Wertes = 17.4 % der Fall ist, verdeutlichen die folgenden Teststärkeüberlegungen:

Nach Ziffer 11 aus Tab. 15.1 hätte der F-Test zur Überprüfung der H_{05} eine Teststärke von 50 %, wenn der wahre Effekt einem F-Äquivalent von 4.08 entspräche (für $\alpha = .05$). Da $F_{Hyp} = 2.94$ (▶ Abschn. 15.2.1 „Prüfung von H_{01}") kleiner ist als $F_{.50} = 4.08$, hat die Untersuchung zur Überprüfung von H_{05} eine Teststärke unter 50 %.

Um herauszufinden, wie groß η^2 sein müsste, damit der Test von H_{05} eine Teststärke von 50 % aufweist, setzen wir die entsprechenden Werte in Gl. 15.1 ein:

$$\eta^2 = \frac{3 \cdot 4.08}{50 + 3 \cdot 4.08} = .197.$$

Erst wenn das Treatment ca. 20 % Varianz erklärt, hätte – bei sonst gleichen Bedingungen – der Signifikanztest zur Prüfung von H_{05} eine Teststärke von 50 %. Die Untersuchung (mit $N = 54$) ist also – bei einem angenommenen $\eta^2 = .15$ – deutlich „underpowered", wenn die H_{05} geprüft werden soll. Die genaue Teststärke für $\eta^2 = .15$, $N = 54$ und $\alpha = .05$ kann über Gl. 15.2 nicht ermittelt werden, da $(1-\beta)_{\text{int pol}} < .50$ nicht im vorgesehenen Interpolationsbereich der Gl. 15.2 liegt (.50 bis .80).

Damit erübrigt sich eine Interpretation der letzten Ziffer in Tab. 15.1 (Ziffer 12). Wenn nicht einmal eine Teststärke von 50 % erzielt wird, dann schon gar nicht eine von 80 % ($6.55 > 2.94 = F_{Hyp}$). Über Gl. 15.1 ermitteln wir, dass der wahre Effekt einer Varianzaufklärung von ca. 28 % ($\eta^2 = .281$) entsprechen müsste, um die H_{05} – bei sonst gleicher Untersuchungsanlage – verwerfen zu können.

Auf eine Tabelle der optimalen Stichprobenumfänge zur Prüfung von H_{05} wird verzichtet, da diese Nullhypothese (bzw. die komplementäre Alternativhypothese) in der praktischen Forschung nur selten begründet werden kann: 5 % Varianzaufklärung ist in vielen sozialwissenschaftlichen Forschungsfeldern kein vernachlässigbar geringer, sondern ein inhaltlich bedeutsamer Effekt. Eine Übersicht von „typischen" Effektgrößen verschiedener Forschungsgebiete findet man bei Lipsey und Wilson (1993; zit. nach Murphy & Myors, 1998, Tab. 1.2).

15.2.2 Hinweise zur Untersuchungsplanung

Im Folgenden fassen wir zusammen, welche Fragen in der Planungsphase einer hypothesenprüfenden Untersuchung beantwortet werden sollten.

Tabelle 15.3 Optimale Stichprobenumfänge (df_N) bei der Überprüfung der Minimum-Effekt-Nullhypothese (H_{01}) in Abhängigkeit von der Effektgröße (η^2) und df_Z für $1 - \beta = .80$ und $\alpha = .05$; Erläuterungen s. Text. (Mod. nach Murphy & Myors, 1998, Tab. 3.3. Republished by permission of Taylor & Francis Ltd, from Murphy, K. R., & Myors, B. (1998). Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests. Permission conveyed through Copyright Clearance Center, Inc.)

	df_Z															
η^2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
.02	3 225	3 242	3 301	3 266	3 334	3 349	3 364	3 429	3 442	3 454	3 479	3 570	3 621	3 900	4 042	4 379	5 260
.03	1 058	1 086	1 104	1 122	1 139	1 176	1 185	1 199	1 212	1 254	1 271	1 303	1 377	1 518	1 615	1 833	2 299
.04	573	590	607	623	650	658	670	683	694	716	736	779	836	920	993	1 151	1 458
.05	373	389	405	422	434	445	457	472	483	492	509	541	586	652	728	833	1 075
.06	269	285	299	313	323	334	343	357	365	373	387	414	450	506	568	654	854
.07	208	223	235	246	255	267	275	283	290	297	315	338	362	419	460	533	718
.08	166	180	192	202	211	219	226	236	243	249	265	279	307	357	393	457	606
.09	139	151	161	170	180	187	193	199	208	214	224	241	266	303	343	400	532
.10	117	130	139	147	154	162	168	174	179	187	196	212	234	268	297	355	473
.11	101	113	122	129	136	143	149	154	159	166	174	189	205	240	266	318	426
.12	89	99	108	115	121	127	133	138	142	147	157	170	185	217	241	289	388
.13	80	89	97	104	109	114	121	125	129	133	142	152	168	197	220	264	355
.14	72	80	87	94	99	104	110	114	118	121	130	139	154	181	202	243	327
.15	65	73	80	86	91	95	101	105	108	112	120	128	142	168	187	225	303
.16	59	67	73	79	84	88	93	97	100	103	111	119	132	156	174	209	283
.17	54	61	68	73	77	81	85	90	93	96	103	110	123	145	162	195	269
.18	49	57	63	68	72	76	79	83	86	89	96	103	115	136	152	183	252
.19	45	53	58	63	67	71	74	78	81	84	90	97	108	127	144	172	238
.20	42	49	55	59	63	67	69	73	76	79	84	91	101	120	137	162	224
.22	37	43	48	52	56	59	62	65	68	70	75	81	91	107	123	146	204
.24	32	38	43	47	50	53	56	59	61	63	68	74	83	97	111	134	185
.26	29	34	38	42	45	48	51	53	55	57	61	67	75	90	101	122	169
.28	26	31	35	38	41	43	46	48	50	53	56	61	69	82	92	111	156
.30	24	28	32	35	37	40	42	44	46	48	51	56	63	75	86	103	144
.32	21	26	29	32	34	36	39	40	42	44	47	52	58	69	79	96	135
.34	20	24	27	30	32	34	36	37	39	41	44	48	54	64	73	89	125

- Welche Nullhypothese soll geprüft werden: die traditionelle „Nil-Nullhypothese“ (H_{00}) oder eine Minimum-Effekt-Nullhypothese (H_{01} bzw. sogar H_{05})?
  - Es wird empfohlen, vorzugsweise die H_{01} (1 % Varianzaufklärung sind zu vernachlässigen) zu prüfen. Diese Nullhypothese wird nicht zwangsläufig mit wachsendem Stichprobenumfang verworfen. Außerdem kann man davon ausgehen, dass 1 % Varianzaufklärung für die meisten Forschungsfragen eine „Quantité négligeable“ ist (bzw. eine Varianzaufklärung, die in der Tat zu vernachlässigen ist).

- Die Planung einer Untersuchung auf der Basis von H_{01} führt zu einer Teststärke, die zwangsläufig kleiner ist als die Teststärke des Signifikanztests von H_{00}. Der optimale Stichprobenumfang für die Überprüfung von H_{01} sichert also in jedem Falle auch eine ausreichende Teststärke ($1-\beta > .80$) für den Test von H_{00}.
- Die H_{05} sollte nur in begründeten Ausnahmefällen die zu überprüfende Nullhypothese sein. Zu behaupten, dass 5 % Varianzaufklärung zu vernachlässigen sei, bedeutet, dass man eine Korrelation von $\rho \approx .22$ für bedeutungslos hält. Dies ist jedoch nach der Cohen-Effektgrößen-Klassifikation bereits ein kleiner ($\rho = .10$) bis mittlerer ($\rho = .30$) Effekt (Tab. 14.5). Die Prüfung von H_{05} mit $1-\beta = .80$ und $\alpha = .05$ erfordert im Übrigen sehr große Effekte und/oder Stichprobenumfänge.

- Wie groß ist der wahre Populationseffekt η^2?
  - Falls hierüber keine zuverlässigen Angaben zur Verfügung stehen, sollte man von einem kleinen bis mittleren Effekt ($.01 < \eta^2 < .10$) ausgehen. Bezüglich der Kalkulation des optimalen Stichprobenumfanges ist man damit – falls der „wahre" Effekt tatsächlich größer ist – immer auf der „sicheren Seite".
- Welches Signifikanzniveau sollte gewählt werden?
  - Es wird $\alpha = .05$ empfohlen. Bei diesem Signifikanzniveau hat der Signifikanztest eine höhere Teststärke als bei $\alpha = .01$. Das 1 %ige Signifikanzniveau verschärft die Überprüfung von H_{00} in der Regel unnötigerweise, da die H_{00} ohnehin meistens falsch ist.
- Wie groß muss der Gesamtstichprobenumfang N_{opt} sein, um dem Signifikanztest für gegebenes η^2 und α eine Teststärke von mindestens 80 % zu verleihen?
  - Eine erste Orientierung hierfür liefert Tab. 14.11, wenn die H_{00} gegen einen kleinen, mittleren oder großen Effekt getestet werden soll. Entspricht der angenommene Effekt keinem dieser Werte, erhält man eine genauere Schätzung des optimalen Stichprobenumfanges über Tab. 15.2 (für Tests der H_{00}) bzw. Tab. 15.3 (für Tests der H_{01}).

In den vergangenen Abschnitten wurden Teststärkeanalysen und die Überprüfung von Minimum-Effekt-Nullhypothesen am Beispiel der Varianzanalyse bzw. für die F-Statistik erläutert. Wie jedoch ist zu verfahren, wenn man Hypothesen mit anderen Signifikanztests (t-Test, χ^2-Test etc.) überprüfen will (bzw. überprüft hat)? Diese Frage soll im folgenden Abschnitt beantwortet werden.

15.2.3 Transformation statistischer Test- und Kennwerte in die F-Statistik

In Lehrbüchern zum **Allgemeinen Linearen Modell** (ALM; z. B. Bortz & Schuster, 2010, Kap. 22 bzw. die dort zitierte Literatur) wird gezeigt, dass die meisten statistischen Verfahren Spezialfälle der multiplen Regression bzw. Korrelation sind, die ihrerseits über die F-Statistik auf Signifikanz geprüft werden. Diese Zusammenhänge wollen wir uns im Folgenden zu Nutze machen, indem wir die wichtigsten statistischen Prüfgrößen und einige statistische Kennwerte in ***F*-Äquivalente** transformieren. Dies hat den großen Vorteil, dass die Alles-auf-einen-Blick-Tabelle (Tab. 11 im ▶ Webangebot zu diesem Buch) für praktisch alle wichtigen statistischen Verfahren genutzt werden kann: Signifikanzschranken für H_{00}, H_{01} und H_{05} sowie Teststärkeangaben müssen also nicht testspezifisch entwickelt werden, sondern können einheitlich dieser Tabelle entnommen werden.

In Tab. 15.4 wird gezeigt, welche Teststatistiken wie transformiert werden (siehe hierzu auch Bortz & Schuster, 2010, Abschn. 5.5.4). Die Durchführung der unter den Ziffern 1–5 genannten statistischen Tests kann z. B. Bortz & Schuster (2010) entnommen werden.

Für die F-Äquivalente können Nennerfreiheitsgrade (df_N) resultieren, die in der Alles-auf-einen-Blick-Tabelle nicht aufgeführt sind. In diesem Fall ermittelt man den gesuchten F-Wert approximativ über eine einfache lineare Interpolation:

$$F_{\text{int pol}} = F_k + \frac{df_{N_{\text{int}}} - df_{N_k}}{df_{N_g} - df_{N_k}} \cdot (F_g - F_k). \qquad (15.4)$$

F_k ist der kleinere F-Wert des Intervalls, in dem sich der gesuchte F-Wert befindet und F_g der größere. df_{N_k} und df_{N_g} sind die Nennerfreiheitsgrade dieser beiden F-Werte; $df_{N_{\text{int}}}$ kennzeichnet die Nennerfreiheitsgrade des gesuchten F-Wertes.

Beispiel: Gesucht wird der kritische F-Wert (Signifikanzschranke zur Prüfung von H_{00} mit $\alpha = .05$) für 2 Zählerfreiheitsgrade und $df_{N_{\text{int}}} = 38$. Wir entnehmen der Alles-auf-einen-Blick-Tabelle, Tab. 11 im ▶ Webangebot zu diesem Buch, $F_g = 3.32$ ($df_{N_g} = 30$) und $F_k = 3.23$ ($df_{N_k} = 40$). Man erhält also folgende interpolierte Signifikanzschranke

$$F_{\text{int pol}} = 3.23 + \frac{38-40}{30-40} \cdot (3.32 - 3.23) = 3.25.$$

Die H_{00} wäre also für $F_{(2,38)} \geq 3.25$ mit $\alpha = .05$ abzulehnen.

Tabelle 15.4 Transformation der wichtigsten Teststatistiken in ein F-Wert-Äquivalent; Erläuterungen siehe Text. (Mod. nach Murphy & Myors, 1998, Tab. 2.1. Republished by permission of Taylor & Francis Ltd, from Murphy, K. R., & Myors, B. (1998). Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests. Permission conveyed through Copyright Clearance Center, Inc.)

Teststatistik	F-Äquivalent	df_Z	df_N
1. t-Test (t)	$F_{(df_Z, df_N)} = t^2_{(df_N)}$	1	$N-2$
2. Bivariate Korrelation (r)	$F_{(df_Z, df_N)} = \frac{r^2 \cdot df_N}{(1-r^2)}$	1	$N-2$
3. Multiple Korrelation (R)	$F_{(df_Z, df_N)} = \frac{R^2 \cdot df_N}{(1-R^2) \cdot df_Z}$	p	$N-p-1$
4. Hierarchische Regression ($R^2_v - R^2_r$)	$F_{(df_Z, df_N)} = \frac{(R^2_v - R^2_r) \cdot df_N}{(1-R^2_v) \cdot df_Z}$	k	$N-p-1$
5. χ^2-Test (χ^2)	$F_{(df_Z, df_N)} = \frac{\chi^2}{df_Z}$	df_{χ^2}	∞
6. Standardisierte Mittelwerts-Differenz ($\hat{\delta}$)	$F_{(df_Z, df_N)} = \frac{\hat{\delta}^2 \cdot df_N}{4}$	1	$N-2$
7. Standardisierte Mittelwerts-Differenz bei abhängigen Stichproben ($\hat{\delta}$)	$F_{(df_Z, df_N)} = \frac{\hat{\delta}^2 \cdot df_N}{4 \cdot \sqrt{1-r_{AB}}}$	1	$N-1$
8. Zwei- und mehrfaktorielle Varianzanalysen	Ein F-Äquivalent wird nicht benötigt, da empirische F-Werte vorliegen.		
9. Zweifaktorielle Varianzanalyse mit Messwiederholungen	Ein F-Äquivalent wird nicht benötigt, da empirische F-Werte vorliegen.		

Anwendungen

Im Folgenden sollen die 9 Transformationsregeln (Tab. 15.4) im Verbund mit der Alles-auf-einen-Blick-Tabelle, Tab. 11 im ▶ Webangebot zu diesem Buch, an einfachen Beispielen erläutert werden. Es könnte sich hierbei um Ergebnisse empirischer Untersuchungen handeln, die wir ex post analysieren, um gegebenenfalls auf Planungsfehler (typischerweise: der gewählte Stichprobenumfang ist kleiner als der optimale Stichprobenumfang, der ausreichende Teststärke garantieren würde) aufmerksam zu machen.

1. t-Test

F-Werte (besser: F-Verteilungen) sind durch Zählerfreiheitsgrade (df_Z) und Nennerfreiheitsgrade (df_N) bestimmt. Ein t-Wert mit $N-2$ Freiheitsgraden wird nach der unter „F-Äquivalent“ genannten Gleichung in einen $F_{(1,N-2)}$-Wert transformiert. Für $t_{(100)} = 1.80$ z. B. erhält man $F_{(1,100)} = 1.80^2 = 3.24$.

Der t-Wert möge für eine gesamte Stichprobe $N = n_1 + n_2 = 102$ berechnet worden sein (beachte: „N“ steht hier sowohl für den Gesamtstichprobenumfang als auch als Abkürzung für „Nenner“). Aus der ▶ Alles-auf-einen-Blick-Tabelle entnehmen wir, dass mit diesem Ergebnis die H_{00} für $\alpha = .05$ nicht verworfen werden kann ($F_{emp} = 3.24 < F_{crit(.05)} = 3.93$). Damit können auch die H_{01} und die H_{05} nicht verworfen werden, denn Signifikanztests zur Prüfung von H_{01} und H_{05} haben eine geringere Teststärke als der entsprechende Test zur Prüfung der H_{00}. Nach Gl. 15.1 entspricht der empirische F-Wert einer Varianzaufklärung von ca. 3 %.

Nun wollen wir annehmen, dass die Untersuchungsplanung von einer „wahren“ Varianzaufklärung von 9 % ($\eta^2 = .09$), d. h. von einem mittleren Effekt ausging (Tab. 14.5). Wir transformieren diesen Wert über Gl. 15.3 in einen F-Wert und erhalten $F_{Hyp} = 9.89$. Dieser Wert ist nun zu vergleichen mit demjenigen F-Äquivalent, bei dem die Untersuchung eine Teststärke von 80 % hätte (mit $df_N = 100$ und $\alpha = .05$). Aus der Alles-auf-einen-Blick-Tabelle, Tab. 11 im ▶ Webangebot zu diesem Buch, entnehmen wir hierfür $F_{.80} = 7.95$. Der F_{Hyp}-Wert ist größer ($9.89 > 7.95$), d. h., die Untersuchung hat für $\eta^2 = .09$ mit $1-\beta > .80$ eine akzeptable Teststärke. Aufgrund der aus dem Untersuchungsergebnis geschätzten Varianzaufklärung von 3 % ($\hat{\eta}^2 = .03$) und wegen des nicht-signifikanten Ergebnisses können wir vermuten, dass der wahre Effekt kleiner ist als der angenommene Effekt von 9 %.

Wir wollen nun überprüfen, wie groß die Teststärke der Untersuchung wäre, wenn der wahre Effekt dem geschätzten Effekt entsprechen würde ($\eta^2 = .03$). Wir transformieren diesen Effekt in einen F_{Hyp}-Wert und erhalten nach Gl. 15.3 $F_{Hyp} = 3.09$. Dieser Wert ist kleiner als das F-Äquivalent, bei dem die Untersuchung eine Teststärke von 50 % hat ($3.09 < 3.85$ gem. der Alles-auf-einen-Blick-Tabelle, Tab. 11 im ▶ Webangebot zu diesem Buch), d. h., die Untersuchung wäre mit $N = 102$ deutlich „underpowered“ (für $\alpha = .05$). Aus Tab. 15.2 entnehmen wir, dass für $\eta^2 = .03$ ein Stichprobenumfang

von $N = 255 + 2 = 257$ „optimal“ wäre ($1 - \beta = .80$; $\alpha = .05$).

Wollte man die H_{01} mit einer Teststärke von 80 % verwerfen ($\alpha = .05$), wäre sogar ein Stichprobenumfang von $N = 1\,058 + 2 = 1\,060$ erforderlich (Tab. 15.3).

2. Bivariate Korrelation

Zwei Merkmale x und y korrelieren in einer Stichprobe mit $N = 62$ Untersuchungspersonen zu $r_{xy} = .48$. Das F-Äquivalent hierzu lautet (Tab. 15.4):

$$F_{(1,60)} = \frac{.48^2 \cdot 60}{(1 - .48^2)} = 17.96.$$

Mit diesem Wert könnte nicht nur die H_{00} ($F_{crit(.01)} = 7.07 < 17.96$), sondern auch die H_{01} ($F_{crit(.01)} = 10.35 < 17.96$) auf dem $\alpha = 1$ %-Niveau verworfen werden. Auch die H_{05} wäre – allerdings nur auf dem 5 %-Niveau – zu verwerfen ($F_{crit(.05)} = 12.49 < 17.96$).

Die Ex-post-Teststärkeanalyse ergibt für einen angenommenen großen Effekt ($\eta^2 = .25$ gem. Tab. 14.5), dass alle 3 Nullhypothesen (H_{00}, H_{01} und H_{05}) mit einer Teststärke über 80 % ($\alpha = .05$, $df_Z = 1$; $df_N = 60$) geprüft wurden: $F_{Hyp} = 20.0$ (gemäß Gl. 15.3). Die F-Äquivalente für $1 - \beta = .80$ und $\alpha = .05$ lauten gem. Tab. 11 (▶ Webangebot zu diesem Buch) 8.06 (für H_{00}), 11.13 (für H_{01}) und 19.10 (für H_{05}).

3. Multiple Korrelation

In einer Stichprobe mit $N = 72$ Untersuchungspersonen besteht zwischen $p = 4$ Prädiktorvariablen und einer Kriteriumsvariablen eine multiple Korrelation von $R = .53$. Man errechnet als F-Äquivalent gem. Tab. 15.4

$$F_{(4,67)} = \frac{.53^2 \cdot 67}{(1 - .53^2) \cdot 4} = 6.54.$$

Mit diesem Wert sind alle 3 Nullhypothesen für $\alpha = .01$ abzulehnen. Dies ist daran zu erkennen, dass F_{emp} größer ist als der entsprechende kritische Wert für $df_Z = 4$ und $df_N = 60$ (als untere Grenzen des kritischen Intervalls, in dem sich $F_{crit(4,67)}$ befindet). Die Berechnung der genauen Werte über Gl. 15.4 wollen wir für den kritischen Wert zur Überprüfung von H_{05} verdeutlichen. Man erhält

$$F_{int\,pol} = 5.90 + \frac{67 - 60}{70 - 60} \cdot (6.13 - 5.90) = 6.06.$$

Der empirische Wert ist größer ($F_{emp} = 6.54 > 6.06$), d. h. – wie bereits bemerkt –, sogar die H_{05} ist für $\alpha = .01$ zu verwerfen.

Für die Ex-post-Analyse der Teststärke wollen wir davon ausgehen, dass im fraglichen Forschungsgebiet Varianzaufklärungen (mit vergleichbaren Prädiktoren) von ca. 15 % ($\eta^2 = .15$ bzw. $R = \sqrt{.15} = .39$) typisch sind. Geprüft werden sollte – wie bislang in diesem Fachgebiet üblich – die H_{00} mit $\alpha = .05$. Wie groß ist die Teststärke für $N = 72$?

Wir berechnen zunächst den F_{Hyp}-Wert nach Gl. 15.3:

$$F_{Hyp} = \frac{.15 \cdot 67}{(1 - .15) \cdot 4} = 2.96.$$

Da $df_N = 67$ in Tab. 11 (▶ Webangebot zu diesem Buch) nicht aufgeführt ist, müssen wir das F-Äquivalent für $1 - \beta = .80$ ($\alpha = .05$) per Interpolation gem. Gl. 15.4 ermitteln. Der Wert befindet sich im Bereich $F_k = 3.16$ ($df_Z = 4$, $df_N = 70$) bis $F_g = 3.20$ ($df_Z = 4$, $df_N = 60$). Wir erhalten also

$$F_{int\,pol} = 3.16 + \frac{67 - 70}{60 - 70} \cdot (3.20 - 3.16) = 3.17.$$

Der F_{Hyp}-Wert ist kleiner als dieser Wert ($F_{Hyp} = 2.96 < 3.17$), d. h. die Untersuchung hat eine Teststärke unter 80 %.

Das F-Äquivalent für eine Teststärke von 50 % liegt (wegen $df_N = 67$) zwischen 1.68 ($df_N = 70$) und 1.70 ($df_N = 60$). Wir übernehmen (ohne Interpolation) den kleineren Wert (1.68) und stellen fest, dass die Teststärke der Untersuchung über 50 % liegt ($F_{Hyp} = 2.96 > 1.68$). Die genaue Teststärke können wir über die Interpolationsformel (Gl. 15.2) errechnen

$$(1 - \beta)_{int\,pol} = .50 + \left(\frac{2.96 - 1.68}{3.16 - 1.68} \cdot .30\right) = .76.$$

Entspräche der „wahre“ Effekt einer Varianzaufklärung von 15 % ($\eta^2 = .15$), so hätte die Untersuchung (mit $N = 72$ und $\alpha = .05$) eine Teststärke von ca. 76 %. Aus Tab. 15.2 entnehmen wir, dass eine geringfügige Vergrößerung der Stichprobe die Teststärke auf 80 % erhöht hätte: Wir lesen für $\eta^2 = .15$ und $df_Z = 4$ den Wert $df_N = 72$ ab, d. h., der optimale Stichprobenumfang hat den Wert $N = 72 + 4 + 1 = 77$. Es wären also nur fünf zusätzliche Untersuchungsobjekte notwendig gewesen, um eine ausreichende Teststärke zu erzielen.

4. Hierarchische Regression

In einer Stichprobe ($N = 135$) korrelieren 5 Prädiktorvariablen mit einem Kriterium zu $R_r = .32$. Durch das Hinzufügen von $k = 2$ weiteren Prädiktoren erhöht sich die multiple Korrelation auf $R_v = .40$. Gefragt wird, ob der Zuwachs signifikant ist bzw. wie das F-Äquivalent für diesen Zuwachs lautet.

In diesem Beispiel steht $R_v = .40$ für das **vollständige Modell** mit $p = 5 + 2 = 7$ Prädiktoren und $R_r = .32$ für das **reduzierte Modell** mit $p - k = 5$ Prädiktoren. Als

F-Äquivalent errechnen wir (Tab. 15.4):

$$F_{(2,127)} = \frac{(.40^2 - .32^2) \cdot 127}{(1 - .40^2) \cdot 2} = 4.35.$$

Diese Gleichung wird bei Bortz und Schuster (2010, S. 358) im Kontext der schrittweisen Regressionstechnik diskutiert. Man verwendet sie auch, um eine multiple Semipartialkorrelation auf Signifikanz zu testen.

Aus Tab. 11 (▶ Webangebot zu diesem Buch) ist zu entnehmen, dass sich die multiple Korrelation durch das Hinzufügen von $k = 2$ Prädiktorvariablen signifikant erhöht hat, d. h., die H_{00} wird verworfen ($\alpha = .05$). Den kritischen Wert ($F_{\text{crit}(.05)} = 3.07 < F_{\text{emp}} = 4.35$) entnehmen wir dem Werteblock für $df_Z = 2$ und $df_N = 120$. Der nächst höhere Wert (für $df_N = 150$) unterscheidet sich nur marginal ($F_{\text{crit}(.05)} = 3.06$), so dass wir auf eine Interpolation verzichten.

Für die Analyse der Teststärke wenden wir uns zunächst dem Modell mit 5 Prädiktoren zu (mit $R_r = 32$). Für ein angenommenes $\eta^2 = .10$ (mittlerer Effekt) ergibt sich $F_{\text{Hyp}} = 2.87$ (über Gl. 15.3 mit $df_Z = 5$ und $df_N = 129$). Um mit der vorliegenden Untersuchung (mit $\alpha = .05$) eine Teststärke von 80 % zu erzielen, müsste gem. Tab. 11 (▶ Webangebot zu diesem Buch) der Wert $F_{.80} \approx 2.65$ erreicht oder überschritten werden. Dies ist der Fall, d. h., die H_{00} wird in dieser Untersuchung mit einer Teststärke über 80 % verworfen.

Durch das Hinzunehmen von $k = 2$ weiteren Prädiktoren erhalten wir $p = 7$ und $R_v = .40$. Wir gehen davon aus, dass man erwartet hat, mit dieser Modellerweiterung ca. 5 % Varianz mehr erklären zu können ($\Delta\eta^2 = .05$; kleiner bis mittlerer Effekt). Für diesen Effekt errechnen wir $F_{\text{Hyp}} = 3.34$ (Gl. 15.3 mit $df_Z = 2$ und $df_N = 127$). Die F-Äquivalente, die für eine Teststärke von 80 % bzw. 50 % erreicht werden müssen, lauten gem. Tab. 11 (▶ Webangebot zu diesem Buch) (für $df_N = 120$) $F_{.80} = 4.91$ bzw. $F_{.50} = 2.51$. Die Teststärke liegt also zwischen 50 % und 80 %. Wir interpolieren über Gl. 15.2 und erhalten

$$(1 - \beta)_{\text{int pol}} = .50 + \left(\frac{3.34 - 2.51}{4.91 - 2.51}\right) \cdot .30 = .60.$$

Die Wahrscheinlichkeit, die H_{00} (keine zusätzliche Varianzaufklärung) für $\alpha = .05$ und $N = 135$ verwerfen zu können, wenn sich die erklärte Varianz tatsächlich um 5 % erhöht ($\Delta\eta^2 = .05$), beträgt also nur ca. 60 %. Man hat in dieser Untersuchung, in der der Zuwachs an Varianzaufklärung signifikant ist, großes Glück gehabt. Wenn gefordert wird, bei Gültigkeit von H_1: $\Delta\eta^2 \geq .05$ die H_{00} mit einer Teststärke von mindestens von 80 % zu verwerfen, hätte man nach Tab. 15.2 mindestens $186 + 7 + 1 = 194$ Untersuchungsteilnehmer einsetzen müssen ($\alpha = .05$).

Die H_{01} ($F_{\text{crit}(.05)} = 4.74$) und damit auch die H_{05} ($F_{\text{crit}(.05)} = 9.64$) bzgl. des Zuwachses an Varianzaufklärung können nicht verworfen werden. Die Teststärke für die Überprüfung von H_{01} liegt unter 50 % ($F_{.50} = 4.04 > 3.4$). Somit ist das Ergebnis nicht eindeutig interpretierbar, sondern als Zufallsergebnis zu werten.

■ 5. χ^2-Test

Die χ^2-Analyse einer 3×4-Kontingenztafel mit $N = 100$ führte zu $\chi^2 = 12.00$. Mit $df_{\chi^2} = (3 - 1) \times (4 - 1) = 6$ errechnet man (Tab. 15.4):

$$F_{(6,\infty)} = \frac{12.00}{6} = 2.00.$$

Für weiterführende Analysen mit Tab. 11 im ▶ Webangebot zu diesem Buch („Alles auf einen Blick“) wird $df_N \to \infty$ hinreichend gut durch $df_N = 10\,000$ approximiert (letzte Zeile in Tab. 11, ▶ Webangebot zu diesem Buch).

Für df_{χ^2} in Spalte df_Z der Tab. 15.4 sind die Freiheitsgrade des verwendeten χ^2-Tests einzusetzen. Informationen hierzu findet man z. B. bei Bortz & Schuster (2010).

Aus Tab. 11 (▶ Webangebot zu diesem Buch) entnehmen wir für $df_Z = 6$ und $df_N \to \infty$ den kritischen Wert $F_{\text{crit}(.05)} = 2.10$. Die H_{00} kann also für $\alpha = .05$ nicht verworfen werden.

Für die Ex-post-Teststärkeanalyse gehen wir davon aus, dass für den Populationsparameter ein kleiner bis mittlerer Effekt ($W = .20$ gem. Tab. 14.5) angenommen wurde. Mit $N = 100$ entspricht dieser Effekt dem Wert $\chi^2 = N \times W^2 = 100 \times .20^2 = 4$ (Cohen, 1988, S. 216f.). Die Transformation dieses Wertes in ein F-Äquivalent (Tab. 15.4, Ziffer 5) führt zu

$$F_{\text{Hyp}} = \frac{4}{6} = .67.$$

Dieser Wert ist kleiner als der Wert $F_{.80} = 2.23$ ($df_Z = 6$, $df_N = 10\,000$), d. h., die Teststärke der Untersuchung (zur Prüfung von H_{00} mit $\alpha = .05$) liegt unter 80 %. Er ist zudem auch kleiner als $F_{.50} = 1.20$, d. h., die Teststärke liegt unter 50 % und ist damit nicht akzeptabel. Cohen (1988, Tab. 7.4) ist zu entnehmen, dass für diese Untersuchung ($W = 0.2$, $\alpha = .05$, $1 - \beta = .80$) ein Stichprobenumfang von $N = 341$ optimal gewesen wäre (Tab. 15.2 und Tab. 15.3 sind für die Bestimmung optimaler Stichprobenumfänge hier nicht geeignet, da diese Tabellen von Varianzaufklärungen ausgehen, die für χ^2-Analysen – mit nominalen/ordinalen Daten – nicht definiert sind).

Die H_{01} und H_{05} können folgerichtig auch nicht abgelehnt werden. Die Teststärke für die Überprüfung dieser Nullhypothesen liegt deutlich unter 50 %, so dass das Ergebnis nicht sinnvoll interpretierbar ist.

6. Standardisierte Mittelwertdifferenz

Wenn als Effektgröße ein $\hat{\delta}$-Wert für eine standardisierte Mittelwertdifferenz vorliegt (Tab. 14.5, 1. Zeile), kann diese nach der in Tab. 15.4 unter Punkt 6 genannten Gleichung in ein F-Äquivalent transformiert werden. Für $\hat{\delta} = 1$ resultiert (für $N = 80$):

$$F_{(1,78)} = \frac{1^2 \cdot 78}{4} = 19.5.$$

Tab. 11 (▶ Webangebot zu diesem Buch) entnehmen wir, dass mit diesem Wert sowohl die H_{00} ($F_{emp} = 19.5 > F_{crit(.01)} \approx 6.96$) als auch die H_{01} für ($F_{emp} = 19.5 > F_{crit(.01)} \approx 10.98$) für $\alpha = .01$ verworfen werden können. Auch die H_{05} wäre – allerdings nur für $\alpha = .05$ – zu verwerfen ($F_{emp} = 19.5 > F_{crit(.05)} \approx 14.39$). Die Nullhypothese, nach der die Varianzaufklärung höchstens 5 % beträgt (und damit im fraglichen Untersuchungskontext zu vernachlässigen ist), wird abgelehnt.

Für die Überprüfung der H_{01} mit $\alpha = .05$ hat diese Untersuchung bei einem angenommenen Effekt von $\delta = 0.80$ eine Teststärke von über 80 % ($F_{Hyp} = .80^2 \times 78/4 = 12.48 > F_{.80} \approx 11.95$).

Für das bereits genannte Beispiel (mit $\delta = 0.5$) ergibt sich $F_{(1,38)} = 0.50^2 \times 38/4 = 2.38$. Dieser Wert ist kleiner als $F_{.50} \approx 3.97$, d. h., die Teststärke der Untersuchung liegt unter 50 %.

7. Standardisierte Mittelwertdifferenz bei abhängigen Stichproben

Eine Stichprobe von 91 Personen wird vor und nach einer Behandlung untersucht. Die standardisierte Differenz von Pre- und Posttestmittelwert beträgt $\hat{\delta} = 0.50$ bei einer Korrelation von Pre- und Posttestwerten von $r = .60$. Wir ermitteln nach Ziffer 7 von Tab. 15.4

$$F_{(1,90)} = \frac{.50^2 \cdot 90}{4 \cdot \sqrt{1 - .60}} = 8.89.$$

Dieser Wert ist größer als der kritische F-Wert zur Überprüfung der H_{00} mit $\alpha = .01$ ($F_{crit(.01)} = 6.92$), aber auch größer als der kritische F-Wert zur Überprüfung der H_{01} mit $\alpha = .05$ ($F_{crit(.05)} = 6.97$). Die Minimum-Effekt-Nullhypothese, nach der die Behandlung eine Veränderung von höchstens 1 % der Merkmalsvarianz bewirkt, kann mit $\alpha = .05$ verworfen werden.

Die Untersuchung möge auf der Annahme basieren, dass mit der Behandlung eine Veränderung von $\delta = 0.40$ einhergeht bei einer Pre-Posttest-Korrelation von $p = .50$. Man ermittelt hierfür

$$F_{Hyp} = \frac{.04^2 \cdot 90}{4 \cdot \sqrt{1 - .50}} = 5.09.$$

Nach Tab. 11 (▶ Webangebot zu diesem Buch) hat die Untersuchung für die Überprüfung von H_{00} eine Teststärke zwischen 50 % und 80 % ($F_{.50} = 3.86 < 5.09 < F_{.80} = 7.97$). Die interpolierte Teststärke ergibt sich nach Gl. 15.2 zu

$$(1 - \beta)_{int\,pol} = .50 + \frac{5.09 - 3.86}{7.97 - 3.86} \cdot .30 = .59.$$

Die Untersuchung hat also eine Teststärke von ca. 60 %. Dass die Untersuchung dennoch zu einem signifikanten Ergebnis führte, ist v. a. darauf zurückzuführen, dass der wahre Effekt vermutlich größer ist als $\delta = 0.40$ und dass die Korrelation möglicherweise höher ist als $\rho = .50$. Verwenden wir die Stichprobenergebnisse ($\hat{\delta} = 0.50$ und $r = .60$) als Schätzungen für δ und ρ, ergibt sich (für $N = 91$ und $\alpha = .05$) eine Teststärke von über 80 % ($F_{Hyp} = 8.89 > F_{.80} = 7.97$). Im Nachhinein kann also nur konstatiert werden: Glück gehabt – die Effekte waren größer als erwartet und konnten somit trotz geringer Teststärke (aufgrund zu kleiner Stichprobe) dennoch signifikant werden.

Hätte man auf der Basis von $\delta = 0.40$ und $\rho = .50$ geplant, wäre zur Überprüfung von H_{00} (für $1-\beta = .80$ und $\alpha = .05$) ein größerer Stichprobenumfang erforderlich gewesen. Wir transformieren $F_{Hyp} = 5.09$ über Gl. 15.1 in $\eta^2 = .05$ und entnehmen Tab. 15.2 $N_{opt} = 151+1 = 152$.

Die Teststärke zur Überprüfung von H_{01} liegt unter 50 % ($F_{.50} = 6.86 > 5.09$).

8. Zwei- und mehrfaktorielle Varianzanalysen

Es wurde eine zweifaktorielle Varianzanalyse mit $p = 2$, $q = 3$ und $n = 20$ durchgeführt. Es folgt eine Ex-post-Analyse der Ergebnisse (zur Terminologie und rechnerischen Durchführung mehrfaktorieller Varianzanalysen s. Bortz & Schuster, 2010, Kap. 14–16). Hier und im Folgenden gehen wir von Faktoren mit festen Effekten und orthogonalen Versuchsplänen (d. h. von gleichgroßen Stichproben pro Faktorstufenkombination) aus.

Für den Haupteffekt A möge sich $F_A = 11.6$ ergeben haben mit $df_A = 1$ und $df_{Fehler} = 114$. Aus Tab. 11 (▶ Webangebot zu diesem Buch) entnehmen wir für die Überprüfung von H_{00} den Wert $F_{crit(.01)} \approx 6.85$ und für die Überprüfung von H_{01} den Wert $F_{crit(.05)} \approx 7.76$. (Wir verwenden hier einfachheitshalber die kritischen Werte für $df_N = 120$, die sich nur geringfügig von den entsprechenden Werten für $df_{Fehler} = 114$ unterscheiden.) Die H_{01} wird für $\alpha = .05$ verworfen.

Die Teststärkeanalyse möge von einem Populationseffekt $\eta_A^2 = .10$ ausgehen (mittlerer Effekt). Hierfür ergibt sich nach Gl. 15.3

$$F_{Hyp} = \frac{0.10 \cdot 114}{0.90} = 12.67.$$

Wir entnehmen Tab. 11 (▶ Webangebot zu diesem Buch) für die Überprüfung von H_{00} ($\alpha = .05$), $F_{.80} = 7.93$ und für H_{01} $F_{.80} = 13.10$. Der F-Test des Haupteffekts A hat also unter den genannten Randbedingungen nur für die Überprüfung von H_{00} eine Teststärke über 80 %.

Für den Haupteffekt B ergibt sich $F_B = 2.80$ mit $df_B = 2$ und $df_{Fehler} = 114$. Dieser Wert ist gem. Tab. 11 (▶ Webangebot zu diesem Buch) nicht signifikant (H_{00}: $F_{crit(.05)} = 3.07 > 2.80$).

Die Teststärkeanalyse geht von einem kleinen bis mittleren Effekt aus ($\eta^2_B = .05$). Dieser Wert entspricht einem F-Äquivalent von

$$F_{Hyp} = \frac{.05 \cdot 114}{.95 \cdot 2} = 3.00.$$

Für eine Teststärke von 80 % wäre ein F-Äquivalent von $F_{.80} = 4.91$ erforderlich und für eine Teststärke von 50 % $F_{.50} = 2.51$. $F_{Hyp} = 3.00$ befindet sich zwischen diesen Werten, so dass wir nach Gl. 15.2 interpolieren:

$$(1-\beta)_{int\,pol} = .50 + \left(\frac{3.00 - 2.51}{4.91 - 2.51}\right) \cdot .30 = .56.$$

Die Untersuchung hat also zur Überprüfung der H_{00} des Haupteffekts B mit 56 % eine sehr geringe Teststärke ($\alpha = .05$).

Der F-Test des Interaktionseffektes führt zu $F_{A\times B} = 17.12$ mit $df_{A\times B} = 2$ und $df_{Fehler} = 114$. Mit diesem F-Wert kann sogar die H_{05} für $\alpha = .01$ verworfen werden (H_{05}: $F_{crit(.01)} = 13.05 < 17.12$). Die Minimum-Effekt-Nullhypothese, nach der der Interaktionseffekt höchstens 5 % der Gesamtvarianz erklärt, wird also mit $\alpha = .01$ abgelehnt.

Für die Interaktion hat man einen starken Effekt erwartet ($\eta^2_{A\times B} = .25$). Hierfür erhält man

$$F_{Hyp} = \frac{.25 \cdot 114}{.75 \cdot 2} = 19.00.$$

Dieser Wert ist größer als $F_{.80} = 13.02$ zur Prüfung von H_{05} (mit $\alpha = .05$), d. h., die Untersuchung hatte zur Absicherung eines starken Effektes auch gegen die Minimum-Effekt-Nullhypothese H_{05} eine ausreichende Teststärke von über 80 %.

Das Beispiel zeigt, dass die Teststärke für die beiden Haupteffekttests und für den Interaktionstest sehr unterschiedlich ausfallen. Dies liegt zum einen daran, dass für die 3 Effekte unterschiedliche Varianzaufklärungen angenommen wurden ($\eta^2_A = .10$; $\eta^2_B = .05$ und $\eta^2_{A\times B} = .25$). Zum anderen basieren die verglichenen Mittelwerte auf unterschiedlich großen Gesamtstichproben ($n_A = 60$; $n_B = 40$; $n_{A\times B} = 20$). Für die Untersuchungsplanung ergibt sich hieraus die Empfehlung, den Gesamtstichprobenumfang für eine mehrfaktorielle Varianzanalyse so zu bestimmen, dass jeder Effekt mit einer ausreichenden Teststärke geprüft werden kann.

Re-Analysen von **drei-** oder **mehrfaktoriellen Varianzanalysen** sind analog zum hier vorgeführten Beispiel vorzunehmen:

- Über Tab. 11 (▶ Webangebot zu diesem Buch) wird entschieden, ob die H_{00} bzw. sogar eine Minimum-Effekt-Nullhypothese (H_{01} oder H_{05}) abgelehnt werden kann.
- Für die Teststärkeanalysen gem. Tab. 11 (▶ Webangebot zu diesem Buch) werden Populationsparameter einer spezifischen Alternativhypothese benötigt. Diese für alle Haupteffekte und Interaktionen festzulegen, dürfte allerdings nicht unproblematisch sein. Es sollte darauf geachtet werden, dass zumindest der „wichtigste" Effekt mit ausreichender Teststärke geprüft wird.

9. Zweifaktorielle Varianzanalyse mit Messwiederholungen

Für die Ex-post-Analyse einer zweifaktoriellen Varianzanalyse mit Messwiederholungen wählen wir das bei Bortz und Schuster (2010, S. 290) genannte Beispiel (Faktor A: 3 verschiedene Kreativitätstrainings, Faktor B: 3 aufeinander folgende Kreativitätsmessungen; $n = 5$).

Es wird $F_A = 3.78$ errechnet mit $df_A = 2$ und $df_{inS} (= df_N) = 12$. Dieser Wert ist nicht signifikant (H_{00}: $F_{crit(.05)} = 3.89$). Ausgehend von $\eta^2_A = .10$ ergibt sich nach Gl. 15.3

$$F_{Hyp} = \frac{.10 \cdot 12}{.90 \cdot 2} = 0.67.$$

Wir entnehmen Tab. 11 (▶ Webangebot zu diesem Buch) für H_{00} $F_{.50} = 3.17$ ($\alpha = .05$), d. h., die Untersuchung ist massiv „underpowered" ($1 - \beta < .50$).

Für Faktor B resultiert $F_B = 44.03$ mit $df_B = 2$ und $df_{B\times Vpn} (= df_N) = 24$. Mit diesem Wert kann sogar die H_{05} auf dem $\alpha = .01$-Niveau verworfen werden ($F_{crit(.01)} = 8.43$).

Mit einem angenommenen $\eta^2_B = .25$ erhält man über Gl. 15.3

$$F_{Hyp} = \frac{.25 \cdot 24}{.75 \cdot 2} = 4.00.$$

Für diesen Wert hat der F-Test zur Überprüfung der H_{00} eine Teststärke zwischen 50 % ($F_{.50} = 2.76$) und 80 % ($F_{.80} = 5.44$). Wir interpolieren und erhalten nach Gl. 15.2

$$(1-\beta)_{int\,pol} = .50 + \left(\frac{4.00 - 2.76}{5.44 - 2.76}\right) \cdot .30 = .64.$$

Auch diese Teststärke lässt zu wünschen übrig (mit $\alpha = .05$).

Für die Interaktion ergibt sich $F_{A\times B} = 2.71$ mit $df_{A\times B} = 4$ und $df_{B\times Vpn} (= df_N) = 24$. Dieser Wert ist nicht signifikant ($F_{crit(.05)} = 2.78$ zur Prüfung von H_{00}).

Für $\eta^2_{A\times B} = .05$ hat der F-Test eine Teststärke deutlich unter 50 % (H_{00}: $F_{.50} = 1.94$). F_{Hyp} ist kleiner als dieser Wert:

$$F_{Hyp} = \frac{.05 \cdot 24}{.95 \cdot 4} = 0.32.$$

Zusammenfassend ist also festzustellen, dass das Beispiel den Teststärkeansprüchen einer realistischen Untersuchung in keiner Weise genügt. Dies war aber auch nicht intendiert, denn das Beispiel sollte „lediglich" den Rechengang einer zweifaktoriellen Varianzanalyse mit Messwiederholungen überschaubar demonstrieren.

Die Teststärke einer zweifaktoriellen Varianzanalyse mit Messwiederholungen erhöht sich, wenn die Voraussetzungen dieser Analyse verletzt sind (**Zirkularitäts-/Sphärizitätsannahme**, dergemäß die Varianzen der Differenzen zwischen je 2 Faktorstufen homogen sein müssen; Bortz & Schuster, 2010, S. 301). Dies gilt zumindest für den Messwiederholungsfaktor B und die Interaktion A × B (der Gruppierungsfaktor A „profitiert" nicht von der Messwiederholung).

Verletzungen dieser Voraussetzung erfordern über die sog. ε-Korrektur (**Epsilon-Korrektur**) eine Verringerung der Freiheitsgrade. Will man auf die ε-Korrektur verzichten, kann man stattdessen konservative F-Tests durchführen, deren Freiheitsgrade bei Bortz (2005, Tab. 9.25) aufgeführt sind. Wir wollen die in diesem Falle erforderlichen Modifikationen am oben genannten Beispiel verdeutlichen (obwohl die Zirkularitätsannahme in diesem Beispiel nicht verletzt ist).

- Für Faktor A ändert sich nichts.
- Faktor B wird mit $df_Z = 1$ und $df_N = 12$ konservativ getestet. Die H_{05} kann auch mit diesen Freiheitsgraden für $\alpha = .01$ verworfen werden ($F_{crit(.05)} = 14.07 < F_{emp} = 44.03$). Wir lassen $\eta^2_B = .25$ unverändert und errechnen erneut

$$F_{Hyp} = \frac{.25 \cdot 12}{.75} = 4.$$

  Als F-Äquivalent für $(1 - \beta) = .50$ entnehmen wir Tab. 11 (▶ Webangebot zu diesem Buch) $F_{.50} = 4.52$, d. h., der F-Test zur Überprüfung von H_{00} ($\alpha = .05$) hat mit den korrigierten Freiheitsgraden eine Teststärke unter 50 % (ohne Freiheitsgradkorrektur: 64 %).

Ähnliches gilt für die Interaktion A × B, die konservativ mit $df_Z = 2$ und $df_N = 12$ getestet wird. Der Wert $F_{A\times B} = 2.71$ ist mit den korrigierten Freiheitsgraden erst recht nicht signifikant, weil sich die kritische Signifikanzschranke erhöht ($F_{crit(.05)} = 3.89$ für H_{00}). Die Teststärke sinkt um ein Weiteres.

$$F_{Hyp} = \frac{.05 \cdot 12}{.95 \cdot 2} = 0.32$$

ist zwar unverändert, aber das F-Äquivalent für eine Teststärke von 50 % ($\alpha = .05$) wird größer ($F_{.50} = 3.17$ für H_{00}).

Zusammenfassend ist festzustellen, dass Verletzungen der Zirkularitätsannahme mit Teststärkezugewinnen einhergehen. Versäumt man es, derartige Verletzungen durch eine ε-Korrektur (bzw. durch konservative F-Tests) zu kompensieren, haben die „normalen" F-Tests (ohne Freiheitsgradkorrektur) eine zu hohe, nicht zu rechtfertigende Teststärke, die zu progressiven Testentscheidungen führt (hierzu auch Bortz & Schuster, 2010, S. 301).

Kurzanleitung zur Alles-auf-einen-Blick-Tabelle

Die Tab. 11 im ▶ Webangebot zu diesem Buch wird v. a. zur **Ex-post-Analyse** von Untersuchungen eingesetzt. Sie erübrigt sich für Untersuchungsplanungen, wenn man mit den in ◘ Tab. 14.11 genannten „optimalen" Stichprobenumfängen operiert und nur an der Überprüfung der H_{00} interessiert ist. Will man Untersuchungsergebnisse im Nachhinein analysieren, ist so vorzugehen:

- Das Testergebnis wird über eine passende Transformationsgleichung der ◘ Tab. 15.4 in einen F_{emp}-Wert überführt. (Dieser Schritt erübrigt sich natürlich, wenn das Testergebnis – wie z. B. bei der Varianzanalyse – bereits ein F-Wert ist)
- Über Tab. 11 (▶ Webangebot zu diesem Buch) wird entschieden, ob die traditionelle Nil-Nullhypothese (H_{00}) oder eine der beiden Minimum-Effekt-Nullhypothesen (H_{01} bzw. seltener auch H_{05}) für $\alpha = .05$ oder $\alpha = .01$ verworfen werden kann. Den hierfür „zuständigen" Werteblock findet man über df_Z und df_N des F_{emp}-Wertes.
- Es muss auf der Basis theoretischer und/oder empirischer Vorkenntnisse ein plausibler Wert für die in der Population vermutlich gültige Varianzaufklärung (η^2 für eine spezifische H_1) vorgegeben werden. Im Zweifelsfalle wählt man einen kleinen bis mittleren Effekt (◘ Tab. 14.5). Der entsprechende Wert wird über Gl. 15.3) in einen F_{Hyp}-Wert transformiert.
- Aus Tab. 11 (▶ Webangebot zu diesem Buch) wird entnommen, wie groß F_{Hyp} sein müsste, um für den in der Untersuchung durchgeführten Signifikanztest (zur Prüfung von H_{00}, H_{01} oder H_{05}) eine Teststärke von mindestens 50 % ($F_{.50}$) oder 80 % ($F_{.80}$) sicherzustellen (für $\alpha = .05$).
- Sollte $F_{Hyp} < F_{.50}$ sein, sind für Replikationen erhebliche Designänderungen (v. a. eine größere Stichprobe) erforderlich. Für $F_{Hyp} \geq F_{.80}$ hat die Untersuchung eine ausreichende Teststärke, d. h., Designänderungen sind nicht nötig. Für $F_{.50} < F_{Hyp} < F_{.80}$ wird die Teststärke über Gl. 15.2 interpoliert. Auch in diesem Falle sind Designänderungen empfehlenswert.

Werteblöcke für Freiheitsgrade, die in Tab. 11 (▶ Webangebot zu diesem Buch) nicht aufgeführt sind, kann man durch Interpolation über Gl. 15.4 bestimmen. Oftmals lohnt sich diese Interpolation jedoch nicht, wenn die Grenzen des Interpolationsbereiches sehr eng beieinander liegen.

Man beachte, dass für Planungszwecke oder auch für Re-Analysen vorliegender Untersuchungsergebnisse in der Regel die Größenordnung des erforderlichen Stichprobenumfanges (gem. Tab. 15.2 bzw. Tab. 15.3) oder der Teststärkeangaben vollkommen ausreichend ist.

15.3 Zur Frage der „Bestätigung" von Nullhypothesen

Üblicherweise überprüft man mit empirischen Untersuchungen Forschungshypothesen, die sich in eine statistische Alternativhypothese umsetzen lassen und das **Vorhandensein von Effekten in der Population** – d. h. von Gruppenunterschieden, Variablenzusammenhängen oder Veränderungen über die Zeit – postulieren. Die Forschungshypothese gilt als bestätigt, wenn die komplementär zur Alternativhypothese formulierte Nullhypothese (H_{00}, H_{01} oder H_{05}), die entsprechende Effekte negiert, mit einer akzeptablen Irrtumswahrscheinlichkeit (Unterschreitung des Signifikanzniveaus von $\alpha = .05$ oder $\alpha = .01$) abgelehnt werden kann.

Gelegentlich ist jedoch die traditionelle Nullhypothese (H_{00}) die Forschungs- bzw. „Wunschhypothese". So mag es z. B. theoriebasiert von Interesse sein nachzuweisen, dass bestimmte Alltagsvorstellungen über Gruppenunterschiede (z. B. zwischen Altersgruppen, zwischen Geschlechtsgruppen, zwischen ethnischen Gruppen) auf Vorurteilen basieren und empirisch gar nicht nachweisbar sind.

Wunsch-Nullhypothese – Eine Wunsch-Nullhypothese liegt vor, wenn man theoriebasiert mithilfe einer empirischen Studie belegen möchte, dass hinsichtlich der betrachteten Variablen kein Effekt (kein Unterschied, kein Zusammenhang, keine Veränderung) in der Population vorliegt.

Damit taucht die Frage auf, ob bzw. wie man eine Nullhypothese „bestätigen" kann. Hierzu muss konstatiert werden, dass leider immer wieder irrtümlicherweise behauptet wird, ein nicht-signifikantes Ergebnis sei ein Beleg für die Gültigkeit einer Nullhypothese. Diese Auffassung ist falsch (Bortz & Schuster, 2010, S. 106): Ist ein **Untersuchungsergebnis nicht signifikant**, so weiß man zunächst nicht,

- ob tatsächlich der von der H_1 postulierte Effekt in der Population zu verneinen und somit die H_0, anzunehmen ist, dergemäß in der Population kein Effekt oder nur ein minimaler Effekt vorliegt, oder
- ob die Studie aufgrund zu kleiner Stichprobe und damit zu geringer Teststärke gar nicht in der Lage gewesen ist, den vorhandenen Populationseffekt aufzudecken. Das nicht-signifikante Ergebnis wäre dann Ausdruck dafür, dass die Studie überhaupt keine Aussage über das aus H_1 und H_0 bestehende Hypothesenpaar treffen kann.

Was ist also zu tun, wenn man z. B. zeigen will, dass 2 Merkmale **nicht** korrelieren oder dass die Differenz zweier Populationsmittelwerte **Null** oder vernachlässigbar klein ist? In der empirischen Forschung wird diese Problematik unter dem Stichwort **Äquivalenztests** („equivalence testing") behandelt (zusammenfassend Klemmert, 2004). Hier soll ein Vorschlag aufgegriffen werden, den Cohen (1988, S. 16f.) skizziert hat.

Ähnlich wie Kevin Murphy und Brett Myors (2004) bei der Begründung von Minimum-Effektgrößen-Tests ausführen, argumentiert Jacob Cohen (1988), dass echte Nulleffekte unrealistisch seien, und dass die meisten Nullhypothesen als „bestätigt" angesehen werden können, wenn der fragliche Effekt zu vernachlässigen bzw. trivial sei. Wenn man nun als **Alternativhypothese** einen Minimaleffekt postuliert, der nahezu Null bzw. unbedeutend ist, besagt ein nicht-signifikantes Ergebnis, dass der Populationseffekt vermutlich nicht größer ist als dieser Minimaleffekt.

Diese Interpretation setzt allerdings voraus, dass man den Signifikanztest bei einer Wunsch-H_{00} mit einer **hohen Teststärke** ausgestattet hat. Dies geschieht in der Regel durch den Einsatz **großer Stichproben**. Gewährleistet die eingesetzte Stichprobe z. B. eine Teststärke von $1 - \beta = .95$, riskiert man mit einer nur geringen β-Fehler-Wahrscheinlichkeit von $\beta = .05$ eine fälschliche Annahme von H_{00} bzw. die irrtümliche Ablehnung von H_1.

Eine weitere Determinante der Teststärke ist das α-Fehler-Niveau. Denn α-(„type 1 error"-) und β-Fehler-(„type 2 error"-)Wahrscheinlichkeit verhalten sich gegenläufig. Wir plädieren dafür, $\alpha = .10$ zu setzen, wenn H_{00} die Wunschhypothese ist. Bei diesem **gelockerten α-Fehler-Niveau** hat der Signifikanztest eine höhere Teststärke als für $\alpha = .05$ (oder gar $\alpha = .01$), d. h., das β-Fehler-Risiko ist bei diesem α-Fehler-Niveau kleiner als bei den konventionellen Signifikanzschranken. In manchen Lehrbüchern (so z. B. auch bei Bortz & Schuster, 2010, S. 146) wird sogar für $\alpha = .25$ (bzw. $\alpha = .20$) plädiert. Bei diesem α-Fehler-Niveau ist die Teststärke noch höher als bei $\alpha = .10$. d. h., das Risiko eines β-Fehlers wird um ein weiteres gesenkt. Allerdings sind bei derart hohem α-Fehler-Niveau nicht-signifikante Ergebnisse bei großen Stichproben ziemlich unwahrscheinlich, was die Chance für die „Bestätigung" einer Wunsch-H_{00} mindert.

Tabelle 15.5 Optimale Stichprobenumfänge (Gesamtstichprobenumfang *N*) für die „Bestätigung" von Nullhypothesen ($\alpha = .10$; zweiseitig). (Mod. nach Cohen, 1988. Republished by permission of Taylor & Francis Ltd, from Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Permission conveyed through Copyright Clearance Center, Inc.)

	β-Fehler-Wahrscheinlichkeit						
Effektgröße	**.01**	**.05**	**.10**	**.20**	**.30**	**.40**	**.50**
a) Kleine Effekte							
$\delta = 0.20$	1 578	1 084	858	620	472	362	272
$\rho = .10$	1 570	1 078	854	617	470	361	272
$Q = 0.10$	3 157	2 167	1 716	1 240	944	724	544
$G = .05$	1 568	1 077	853	616	469	360	271
$H = 0.20$	1 578	1 082	856	618	470	360	270
$W = .10$ bei							
df = 1	1 577	1 082	856	618	470	360	270
df = 2	1 856	1 302	1 046	771	597	465	356
df = 3	2 051	1 457	1 180	880	688	541	418
df = 4	2 209	1 583	1 288	968	763	604	469
df = 5	2 344	1 691	1 382	1 045	827	658	514
df = 6	2 465	1 787	1 465	1 113	884	706	553
$E = .10$ bei							
df = 1	1 578	1 084	858	620	472	360	272
df = 2	1 857	1 305	1 047	774	606	465	357
df = 3	2 052	1 460	1 184	884	692	541	420
df = 4	2 215	1 585	1 290	965	770	604	475
df = 5	2 352	1 698	1 386	1 044	834	658	516
df = 6	2 471	1 792	1 470	1 113	889	706	560

Tests zur Bestätigung einer Wunsch-H_{00} sollten **zweiseitig** durchgeführt werden. Zwar haben zweiseitige Tests gegenüber einseitigen Tests eine geringere Teststärke und damit ein höheres β-Fehler-Risiko; für eine Wunsch-H_{00} in Bezug auf Korrelationen (oder auch Differenzen) dürfte es jedoch meist unerheblich sein, ob eine unbedeutende Korrelation (Differenz) positiv oder negativ ausfällt. Unter diesen Rahmenbedingungen haben wir aus den Sample Size Tables von Jacob Cohen (1988) Tab. 15.5 zusammengestellt.

Es handelt sich um optimale Stichprobenumfänge (Gesamtstichprobenumfang N) für Tests zur „Bestätigung" von H_{00} für $\alpha = .10$ (zweiseitig). Die Effektgrößen in der ersten Spalte haben wir bereits in Tab. 14.5 definiert. In Tab. 15.5 gehen wir unter a) davon aus, dass die hier genannten Effekte (kleine Effekte gem. Tab. 14.5) praktisch unbedeutend sind.

Die β-Fehler-Wahrscheinlichkeit wurde bis zu einem Wert von $\beta = .50 = 50\,\%$ variiert, um zu verdeutlichen, dass Entscheidungen zugunsten von H_{00} auch dann noch reine 50:50-Glücksspiele sind, wenn die untersuchte Gesamtstichprobe (N) schon recht beachtlich ist. Legt man für die fälschliche Ablehnung von H_1 zugunsten von H_{00} die gleichen Maßstäbe an wie für die fälschliche Ablehnung von H_{00} zugunsten von H_1, wäre $\beta = \alpha = .05$ (oder gar .01) zu setzen. Die dann erforderlichen Stichprobenumfänge sind „imposant" und für viele Forschungsfragen (mit Wunsch-H_{00}) wohl unrealistisch.

Tabelle 15.5 (Fortsetzung)

	β-Fehler-Wahrscheinlichkeit						
Effektgröße	**.01**	**.05**	**.10**	**.20**	**.30**	**.40**	**.50**
b) Kleine bis mittlere Effekte							
$\delta = 0.30$	702	482	382	276	210	162	122
$\rho = .20$	387	266	211	153	117	91	69
$Q = 0.20$	792	544	431	312	238	183	138
$G = .10$	385	265	210	152	116	90	68
$H = 0.30$	700	480	380	274	210	160	120
$W = .20$ bei							
df = 1	394	271	214	155	118	90	68
df = 2	464	326	261	193	149	116	89
df = 3	513	364	295	220	172	135	104
df = 4	552	396	322	242	191	151	117
df = 5	586	423	345	261	207	164	128
df = 6	616	447	366	278	221	176	138
$E = .15$ bei							
df = 1	702	482	382	276	210	162	122
df = 2	828	582	468	345	267	210	159
df = 3	916	652	528	396	308	248	188
df = 4	985	705	575	435	340	277	215
df = 5	1 050	756	618	462	366	300	234
df = 6	1 099	798	658	497	399	325	252

Sehr viel günstiger stellt sich die Situation dar, wenn als „Nulleffekte“ auch kleine bis mittlere Effekte akzeptiert werden können, wie in Tab. 15.5b. Hier könnte z. B. die H_{00}: $\rho = 0$ bei einem nicht-signifikanten Ergebnis mit einer β-Fehler-Wahrscheinlichkeit von 5 % als „bestätigt“ gelten, wenn eine Stichprobe von $N = 266$ untersucht wird.

Bei der statistischen Überprüfung einer Wunsch-Nullhypothese sind folgende Hinweise zu beachten:

- Geprüft wird als H_0 die klassische Nil-Nullhypothese H_{00} (keine Minimum-Effektgrößen-Nullhypothese).
- Idealerweise wird die H_1 als spezifische Alternativhypothese formuliert, die eine Mindesteffektgröße postuliert.
- Das β-Fehler-Niveau wird konventionell nicht auf 5 % (wie beim α-Fehler-Niveau), sondern meist auf 20 % festgelegt.
- Die Festlegung des β-Fehler-Niveaus auf 20 % impliziert eine Teststärke von mindestens $1 - \beta = 80\,\%$.
- Da sich α- und β-Fehler-Wahrscheinlichkeit gegenläufig verhalten, wird bei einer Wunsch-Nullhypothese gegen ein liberalisiertes α-Niveau von 10 % (statt üblicherweise 5 %) getestet.
- Bei einer Wunsch-Nullhypothese erfolgt ein zweiseitiger Test.
- Der optimale Stichprobenumfang für die Prüfung einer Wunsch-Nullhypothese ist entsprechenden Tabellen zu entnehmen oder mit Software-Tools wie

G*Power zu bestimmen und abhängig von a) dem β-Fehler-Niveau (in der Regel 20 %), b) dem α-Fehler-Niveau (in der Regeln 10 %, zweiseitig) sowie c) der Effektgröße des jeweiligen Tests (kleiner, mittlerer oder großer Effekt).

Bevor wir die Handhabung von Tab. 15.5 an Beispielen veranschaulichen, soll noch kurz begründet werden, warum wir davon ausgehen, dass im Normalfall die traditionelle Nullhypothese (H_{00}) als Wunschhypothese zur „Bestätigung" ansteht und nicht Minimum-Effekt-Nullhypothesen (H_{01} oder H_{05}). Zunächst einmal gilt auch für H_{01} und H_{05}, dass ein nicht-signifikantes Ergebnis diese Nullhypothesen **nicht** bestätigt. Selbst bei hoher Teststärke ($1-\beta$) bedeutet ein nicht-signifikantes Ergebnis lediglich, dass die H_1 mit einer niedrigen β-Fehler-Wahrscheinlichkeit fälschlicherweise abgelehnt wird. Auch bei sehr großen Stichproben (bzw. hoher Teststärke) kann es passieren, dass eine spezifische H_1 (z. B. $\eta^2_{Hyp} = .15$) wegen eines nicht-signifikanten Ergebnisses als nicht bestätigt gilt, obwohl sich der wahre Parameter hiervon nur geringfügig unterscheidet (z. B. $\eta^2 = .13$). Hieraus zu folgern, die Nullhypothese sei „bestätigt", wäre zweifellos ein Fehler.

Kurzum: Bei einem nicht-signifikanten Ergebnis muss der „wahre" Parameter keineswegs den Parametern der Nullhypothesen (H_{00}, H_{01}, H_{05}) entsprechen.

Man kommt dem H_{00}-Parameter jedoch näher, wenn mit der H_1 ein unbedeutender Effekt spezifiziert wird. Nicht-signifikant heißt dann, der „wahre" Effekt ist mit hoher Wahrscheinlichkeit praktisch unbedeutend bzw. „nahezu" Null.

Wenn nun auch noch der H_{01}-(H_{05}-)Parameter einen kleinen Effekt behauptet, könnte es sein, dass die H_{05}-(H_{01}-)Parameter und der H_1-Parameter identisch oder nahezu identisch sind, was inhaltlich natürlich keinen Sinn ergibt. Eine hohe Teststärke (und damit bei einem nicht-signifikanten Ergebnis eine geringe β-Fehler-Wahrscheinlichkeit für die fälschliche Ablehnung von H_1) würde zudem riesige, völlig unrealistische Stichprobenumfänge erfordern.

15.3.1 Beispiele für Wunsch-Nullhypothesen

Standardisierte Differenz (δ)

Es soll die Wunsch-Nullhypothese überprüft werden, dass sich die Intelligenz weiblicher und männlicher Geschwister nicht unterscheidet. Da die „Bestätigung" dieser H_{00} grundlagenwissenschaftlich sehr bedeutend wäre (sie spräche gegen den Einfluss von Umweltfaktoren auf die Intelligenzleistung), will man die „strengere" H_1: $\delta = 0.20$ nur dann zugunsten von H_{00} verwerfen, wenn die β-Fehler-Wahrscheinlichkeit höchstens 1 % beträgt. Für diese Untersuchung sind $N = 1\,578$ Personen (oder $1\,578/2 = 789$ Geschwisterpaare) erforderlich.

Wenn man nun bei einem nicht-signifikanten Ergebnis (via t-Test mit $\alpha = .10$) die H_1 ablehnt, ist das Risiko einer falschen Ablehnung mit $\beta = .01$ sehr gering. Oder anders formuliert: Es ist sehr wahrscheinlich, dass die H_1 korrekt abgelehnt wird. Der wahre Parameter ist mit hoher Konfidenz kleiner als 0.20, d. h., er befindet sich in einem Bereich, der „praktisch" einer Nulldifferenz entspricht (genauere Angaben entnimmt man dem Konfidenzintervall für das Stichprobenergebnis $\hat{\delta}$, das wegen des nicht-signifikanten Ergebnisses $\delta = 0$ umschließt; zur Berechnung ▶ Anhang in diesem Buch (SAS-Syntax). Diesen Sachverhalt wollen wir im Folgenden kurz als „Bestätigung" der H_{00} bezeichnen. Tatsächlich bestätigt ein empirisches Signifikanztestergebnis nicht die Hypothese, sondern wird zum Anlass genommen, sich für oder gegen eine Hypothese zu entscheiden (▶ Abschn. 12.5.1).

Korrelation (ρ)

Im Rahmen einer Konstruktvalidierung soll gezeigt werden, dass die Merkmale „Glauben an Verschwörungstheorien" und „Paranoia" nicht korreliert sind (Neumann, 2005). Hierbei werden Korrelationen von $\rho < .20$ für unbedeutend gehalten. Aus Tab. 15.5b ist zu entnehmen, dass bei einer Stichprobe von $N = 266$ Untersuchungspersonen die H_1: $\rho = .20$ mit einer β-Fehler-Wahrscheinlichkeit von $\beta = .05$ abgelehnt werden und die H_{00} als „bestätigt" gelten kann, wenn die Korrelation für $\alpha = .10$ nicht signifikant ist.

Korrelationsdifferenz (Q)

Es soll gezeigt werden, dass die Merkmale „Intelligenz" und „Schulnote" bei Schülern und Schülerinnen aus der Unterschicht genauso hoch korrelieren wie bei solchen aus der Oberschicht. Korrelationsdifferenzen von $Q < 0.20$ werden für unbedeutend gehalten (zur Bedeutung von Q-Werten ▶ Abschn. 14.2.3; zur Überprüfung des Unterschiedes zweier Korrelationen s. Bortz & Schuster, 2010, Gl. 10.17). Bei einem nicht-signifikanten Unterschied ($\alpha = .10$) könnte mit $\beta = .05$ die H_1: $Q = 0.20$ zugunsten von H_{00} abgelehnt werden, wenn jeweils $544/2 = 272$ Schüler/innen der Unterschicht und der Oberschicht untersucht werden (Tab. 15.5b).

Differenz $\pi - .50$ (G)

Mit einem Vorzeichentest (Bortz & Lienert, 2008, Kap. 3.3.1) soll gezeigt werden, dass ein kostengünstiges, neu entwickeltes B-Präparat zu einem bewährten, aber teureren A-Präparat äquivalent ist. Man plant eine Untersuchung mit „Matched Samples" und will die beiden Präparate zufällig den Paarlingen (d. h. jeweils zwei vergleichbaren Patienten) zuordnen. Gemäß H_{00} wird erwartet, dass bei 50 % aller Patientenpaare das A-Präparat

und bei den restlichen 50 % das B-Präparat wirksamer ist. Betrachtet man die Vorzeichen der Wirkdifferenzen bei allen A/B-Paarlingen, müssten gemäß H_{00} 50 % der Vorzeichen positiv und 50 % negativ sein. Man geht davon aus, das bestenfalls eine Abweichung von $G = .05$ tolerierbar ist (H_1: $G = .05$). Eine falsche Ablehnung von H_1 sollte möglichst vermieden werden, d. h., man will maximal eine β-Fehler-Wahrscheinlichkeit von 1 % tolerieren.

Aus Tab. 15.5a ist zu entnehmen, dass eine Stichprobe von 1 568 Differenzen (d. h. im Beispiel 1 568 Patientenpaare bzw. 3 136 Patienten) benötigt wird. Unter diesen Rahmenbedingungen könnten die beiden Präparate als äquivalent gelten, wenn die Abweichung des P-Wertes (z. B. Anteil der +/−-Differenzen von .50) nicht signifikant ist ($\alpha = .10$). Eine klinische Studie mit mehr als 3 000 Testpersonen ist jedoch als extrem aufwändig und kostspielig einzuschätzen.

Differenz zweier Anteilswerte πA und πB (*H*)

Es wird behauptet, dass sich zwei Unterrichtsmethoden A und B nicht nennenswert im Unterrichtserfolg unterscheiden. Zu vergleichen sind die Anteile P_A und P_B derjenigen Schüler/innen, die einen Abschlusstest nicht bestehen. Die Methoden sollen als vergleichbar gelten, wenn für H höchstens ein Wert von $H = 0.20$ resultiert (kleiner Effekt; zur Berechnung von H ▶ Abschn. 14.2.3). Die zu überprüfende H_1 lautet also H_1: $H = 0.20$. Man bildet zwei randomisierte Stichproben mit $N_A = N_B = 100$ und überprüft die H_0 mit einem Vierfelder-χ^2-Test. Das Ergebnis ist nicht signifikant. Kann aus diesem Ergebnis geschlossen werden, die beiden Methoden seien vergleichbar in Bezug auf ihren Erfolg, dass also die H_0 angenommen werden kann?

Wie oben dargelegt, wäre diese Schlussfolgerung falsch. Aus Tab. 15.5a ist zu entnehmen, dass eine Entscheidung zugunsten von H_{00} für $H = 0.20$ und $N = N_A + N_B = 200$ mit einer β-Fehler-Wahrscheinlichkeit versehen ist, die über 50 % liegt ($N = 270$ wäre für $\beta = .50$ erforderlich). Auch wenn man sich mit $\beta = .10$ zufrieden geben würde, hätte man immer noch $N = 856$ Schüler prüfen müssen ($\alpha = .10$). Hier wird erneut deutlich, dass eine „Bestätigung" von Nullhypothesen ein forschungsökonomisch besonders aufwändiges Unterfangen ist.

χ^2-Test (*W*)

Es soll überprüft werden, ob ein Merkmal normalverteilt ist. Die Überprüfung erfolgt mit dem χ^2-Anpassungstest auf Normalverteilung (Goodness-of-Fit-Test; Bortz & Schuster, 2010. S. 144ff.) und dient der Überprüfung der für die meisten parametrischen Verfahren (t-Test, F-Test etc.) wichtigen Normalverteilungsvoraussetzung. Die Wunschhypothese entspricht also der H_{00}.

Da die meisten parametrischen Verfahren robust auf Voraussetzungsverletzungen reagieren (zumal bei großen Stichproben), werden moderate Abweichungen der Merkmalsverteilung von einer Normalverteilung akzeptiert. Man legt deshalb den H_1-Parameter auf $W = .20$ fest. Wenn die Messwerte in $k = 9$ Kategorien eingeteilt werden, hat der Anpassungs-χ^2-Wert $9 - 3 = 6$ Freiheitsgrade. Falls der χ^2-Wert nicht signifikant sein sollte ($\alpha = .10$), will man sich bei Ablehnung von H_1 zugunsten von H_{00} bzw. mit der Behauptung, das Merkmal sei normalverteilt, sozusagen nur mit einer Wahrscheinlichkeit von 5 % irren ($\beta = .05$). Für diese Konstellation entnehmen wir Tab. 15.5a eine optimale Stichprobe von $N = 447$.

Varianzanalyse (*E*)

Die in Tab. 15.5 (a und b) unter E genannten optimalen Stichprobenumfänge sind einzusetzen, wenn man zeigen will, dass ein p-stufiger **Haupteffekt** (in einer einfaktoriellen oder mehrfaktoriellen Varianzanalyse) zu vernachlässigen ist. Der Stichprobenumfang n für die einzelnen Faktorstufen ergibt sich über $n = N/(\mathrm{df} + 1) = N/p$.

Für die „Bestätigung" einer **Interaktions**-Wunsch-Nullhypothese wird Gl. 14.62 benötigt. Das folgende Beispiel demonstriert das Vorgehen: Es wird behauptet, dass männliche und weibliche Patienten (Faktor A; $p = 2$) **nicht** geschlechtsspezifisch auf $q = 3$ unterschiedliche Dosierungen eines Antidepressivums (Faktor B) reagieren bzw. dass die A×B-Interaktion zu vernachlässigen sei. Da die abhängige Variable „Depressivität" mit einem nur mäßig reliablen Fragebogen gemessen wird, begnügt man sich mit $E = .15$ (H_1: $E = .15$). Dies entspricht nach Gl. 14.33 einer Varianzaufklärung von $\eta^2 = .02$. (Zur schlechten Kompatibilität der Cohen-Klassifikation von η^2 und E Tab. 14.5 bzw. ▶ Abschn. 14.2.3. Beispiel: In Tab. 14.5 sind $\eta^2 = .25$ und $E = .40$ jeweils große Effekte. Überführt man jedoch E in η^2 über Gl. 14.33, resultiert $\eta^2 = .14$, was eher einem mittleren als großen Effekt entspricht.)

Das β-Fehler-Niveau wird auf 5 % festgelegt. Für Gl. 14.62 benötigen wir die Freiheitsgrade der Interaktion: $\mathrm{df}_{A\times B} = (p - 1) \cdot (q - 1) = 2$; aus Tab. 15.5b) entnehmen wir $N = 582$ für $\mathrm{df} = 2$ und $\beta = .05$. Dies wäre der Gesamtstichprobenumfang für eine einfaktorielle Varianzanalyse mit $p = \mathrm{df} + 1 = 3$ Gruppen, d. h., man erhält $n = 194$. Mit $2 \times 3 = 6$ Zellen errechnet man über Gl. 14.62

$$n_{\mathrm{Zelle}} = \frac{(194 - 1) \cdot (2 + 1)}{6} + 1 = 97.5 \approx 98.$$

Pro Faktorstufenkombination wären also 98 Patienten zu untersuchen (orthogonales Design, gleichgroße Stichproben vorausgesetzt) bzw. insgesamt $6 \times 98 = 588$ Patienten. Sollte der Interaktionseffekt für $\alpha = .10$ nicht signifikant sein, kann die H_{00} als „bestätigt" gelten.

15.3.2 Alternative Prüfung von Wunsch-Nullhypothesen

Ein anderer Weg zur „Bestätigung" von Nullhypothesen wird bei Serlin und Lapsley (1993, S. 219f.) beschrieben. Wenn man zeigen will, dass ein Effekt trivial ist, sollte – so die Autoren – das logische Komplement zu diesem trivialen Effekt als Nullhypothese postuliert werden. Bezeichnen wir einen trivialen Effekt – Cohens Notation folgend (1988, S. 16) – mit i, wäre also (z. B. bezogen auf Korrelationen) folgendes Hypothesenpaar zu prüfen:

$$H_0 : \rho \geq i; \qquad H_1 : \rho < i.$$

Führt ein entsprechender Signifikanztest (basierend auf der nichtzentralen t-Verteilung) zu einem „signifikanten Ergebnis", wäre die H_0 abzulehnen und die H_1 könnte angenommen werden, d. h., man hätte gezeigt, dass die Behauptung, der Effekt sei trivial, mit einer Irrtumswahrscheinlichkeit von $\alpha \leq .05$ (.01) richtig ist (Klemmert, 2004).

Serlin und Lapsley (1993, S. 220) weisen darauf hin, dass dieser Ansatz dem Ansatz von Cohen, den wir hier übernommen haben, entspricht. Es macht im Ergebnis keinen Unterschied, ob man die oben genannte H_0 (z. B. mit $\alpha = .05$) ablehnt oder ob man – im Cohen-Ansatz – die H_1: $\rho \geq i$ mit $\beta = .05$ ablehnt.

Eine Technik zur Überprüfung der **Äquivalenz zweier Mittelwerte** wurde von Tryon (2001) vorgeschlagen. Dieser Vorschlag lässt sich wie folgt zusammenfassen: Man definiert einen Äquivalenzbereich Δ für Parameterdifferenzen, die man vernachlässigen kann. Dann werden für die beiden Mittelwerte Konfidenzintervalle bestimmt, die einander überlappen müssen. Wenn die Differenzen zwischen der oberen Konfidenzintervallgrenze des größeren Mittelwertes und der unteren Grenze des kleineren Mittelwertes kleiner ist als Δ, wird von einer Äquivalenz der Mittelwerte ausgegangen.

15.3.3 Modellanpassungstests und Wunsch-Nullhypothesen

Wenn wir uns hier mit der Frage befassen, wie man eine Nullhypothese „bestätigen" kann, darf eine kurze Anmerkung zu den für viele Verfahren essenziellen Modellanpassungstests nicht fehlen. Zu diesen Verfahren zählen z. B.

- Strukturgleichungsmodelle („structural equation modeling", SEM, ▶ Kap. 17),
- Modellanpassungen im Rahmen der Item-Response-Theorie (IRT; z. B. Rasch-Modell; ▶ Abschn. 10.4.4),
- log-lineare oder logistische Modelle (▶ Abschn. 12.5.2),
- Zeitreihenanalyse (ARIMA-Modelle; ▶ Abschn. 12.5.4),
- Test zur Voraussetzungsüberprüfung für statistische Verfahren (z. B. Tests auf Normalverteilung wie der Goodness-of-Fit-χ^2-Test, Varianzhomogenitätstests).

Generell sind hier Verfahren angesprochen, bei denen die Güte der Anpassung empirischer Daten an ein theoretisches Modell getestet werden soll. Bei all diesen Tests ist die Nullhypothese die Wunschhypothese: Es soll gezeigt werden, dass Modell und Daten übereinstimmen bzw. dass mögliche Abweichungen der Daten vom Modell zu vernachlässigen sind. Zu fragen ist, wie dieser Nachweis geführt werden kann.

Es ist zu beachten, dass allein ein nicht-signifikantes Ergebnis kein Nachweis einer guten Modellanpassung bzw. kein Indiz für die Bestätigung der Nullhypothese ist. Denn bei Entscheidungen zugunsten von H_{00} geht es nicht um große α-Fehler-Wahrscheinlichkeiten (d. h. um nicht-signifikante Ergebnisse mit $P > \alpha$), sondern um eine möglichst **geringe β-Fehler-Wahrscheinlichkeit**. Diese ist – wie oben ausgeführt – nur durch eine **hohe Teststärke** $(1-\beta)$ zu erzielen, was wiederum große bzw. sehr **große Stichproben** erforderlich macht.

Doch mit welcher Teststärke soll die H_{00} verworfen werden, um im Falle eines nicht-signifikanten Ergebnisses mit genügend kleiner β-Fehler-Wahrscheinlichkeit eine erfolgreiche Modellanpassung konstatieren zu können? Wie soll verfahrensspezifisch entschieden werden, gegen welche H_1 zu testen ist? Was ist gemäß dieser H_1 ein zu vernachlässigender Effekt bzw. eine vernachlässigbare Abweichung vom Modell? Diese Fragen sind nicht leicht zu beantworten. Allein die Vielzahl verschiedener Modelltests für Strukturgleichungsmodelle – die zuweilen als **„Dschungel" verschiedener Fit-Indizes** bezeichnet wird (Nachtigall, Kroehne, Funke, & Steyer, 2003) – spricht nicht gerade für Eindeutigkeit und Transparenz der Entscheidungsregeln bei der Überprüfung von Strukturgleichungsmodellen. Auch das „Ausprobieren" vieler Modelle, um schließlich ein Modell mit einem „günstigen Fit" zu finden („Post Hockery" und „Fitishism" nach Nachtigall et al., 2003) bietet keine Gewähr für eine Modellannahmestrategie mit niedriger β-Fehler-Wahrscheinlichkeit. Diese Probleme sprechen indessen nicht pauschal gegen Modellierungsverfahren, sondern für eine genaue Kenntnis der Logik dieser Verfahrensgruppe sowie der verschiedenen Fit-Indizes und ihrer Bedeutungen

(zu Modellanpassungstests für Strukturgleichungsmodelle ▶ Kap. 17).

Lernquiz

1. Worin unterscheidet sich die traditionelle Nullhypothese des statistischen Signifikanztests von einer Minimum-Effekt-Nullhypothese?
2. Was stimmt? Die Teststärke eines Signifikanztests ist umso höher,
   a) je größer der Stichprobenumfang,
   b) je liberaler das Signifikanzniveau,
   c) je größer der Populationseffekt und/oder
   d) je geringer die β-Fehler-Wahrscheinlichkeit.
3. Empfiehlt es sich, in Zukunft nur noch H_{05}-Hypothesen zu überprüfen um sicherzustellen, dass keine praktisch zu vernachlässigenden Effekte postuliert werden? Begründen Sie!
4. Welche Aussage trifft jeweils zu?
   a) Eine H_{05} erklärt eine bivariate Populationskorrelation von $\geq .15$ für unbedeutend klein.
   b) Bei einem traditionellen Signifikanztest kann eine multiple Korrelation von $R = .007$ in der Stichprobe statistisch signifikant werden.
   c) Beim Test einer H_{05} kann eine bivariate Korrelation von $r = .20$ nicht signifikant werden.
   d) Wenn ein Test der H_{01} signifikant wird, ist von mangelhafter Teststärke auszugehen.
   e) Wenn ein Test gegen die H_{05} signifikant geworden ist, weiß man, dass auch ein Test gegen H_{01} und H_{00} statistisch signifikant ausfallen würde.
5. Begründen Sie, warum Untersuchungen mit einer Teststärke unter 50 % nicht durchgeführt bzw. nicht publiziert werden sollten. Welche Teststärke ist zu fordern?
6. Was versteht man unter dem optimalen Stichprobenumfang und wann und wie wird er festgelegt?
7. Was lässt sich über die Teststärken von Signifikanztests aussagen, bei denen eine Nil-Nullhypothese, eine 1 %-Minimum-Effekt-Nullhypothese oder eine 5 %-Minimum-Effekt-Nullhypothese verwendet wird?
8. Was ist zu beachten, wenn bei einem Signifikanztest die Nullhypothese (Nil-Nullhypothese oder Minimum-Effekt-Nullhypothese) die von den Forschenden inhaltlich begründete und zu prüfende Wunschhypothese ist?
9. Eine Grundsatzkritik am klassischen Nil-Nullhypothesen-Signifikanztest lautet, er sei informationslos, weil letztlich jede empirische Studie ein signifikantes Ergebnis erzeugen könne, wenn nur eine entsprechend große Stichprobe gewählt wird. Wie ist mit dieser Kritik bei der statistischen Hypothesenprüfung umzugehen?
10. Was versteht man im Kontext von Minimum-Effekt-Tests unter einem F-Äquivalent?

Übungsaufgaben

1. Zwei in ihren Englisch-Vorkenntnissen vergleichbare Gruppen von Studierenden lernen Englischvokabeln a) mit einem Vokabelheft ($n_1 = 52$) oder b) einem handygestützten Vokabeltrainer ($n_2 = 52$). Die Studie basiert auf der Forschungshypothese, dass der Handy-Vokabeltrainer zu einem größeren Lernerfolg führt, wobei man von einem mittleren Effekt ($\eta^2 = .08$) ausgeht. Prüfen Sie diese Hypothese gegen die H_{00} und die H_{01} auf einem α-Niveau von .05 und berücksichtigen Sie dabei die Teststärke und Effektgröße. Der t-Test für unabhängige Stichproben möge einen empirischen t-Wert von $t_{(100)} = 1.9$ ergeben haben.
2. Laden Sie folgenden Artikel herunter:
   Staiano, A.E., & Calvert, S.L. (2011). Wii tennis play for low-income African American adolescents' energy expenditure. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace,* 5(1), article 4. http://cyberpsychology.eu/view.php?cisloclanku=2011060801&article=4
   Überprüfen Sie für das von den Autorinnen ermittelte Ergebnis, dass Mädchen mehr Spaß daran hatten, Wii-Tennis zu spielen als Jungen, die H_{00}, H_{01} und die H_{05} und berücksichtigen Sie dabei die Teststärke und Effektgröße.
3. In einer sportpsychologischen Studie mit 10-jährigen Kindern zeigte sich eine bivariate Korrelation von $r_{xy} = .39$ zwischen sportlicher Aktivität und Selbstwertgefühl. Prüfen Sie diese Korrelation gegen die H_{00} und die H_{01} auf einem α-Niveau von .05 und berücksichtigen Sie Effektgröße und Teststärke.

Weiterführende Literatur

- Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler.* München: Pearson Studium.
- Denis, D. J. (2003). Alternatives to null hypothesis significance testing. *Theory and Science, 4*(1). Retrieved November 1, 2013, from http://theoryandscience.icaap.org/content/vol4.1/02_denis.html.

- Murphy, K. R., Myors, B., & Wolach, A. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (3rd ed.). New York, NY, US: Routledge/Taylor & Francis Group.
- Nachtigall, C. & Wirtz, M. (2009). *Wahrscheinlichkeitsrechnung und Inferenzstatistik. Statistische Methoden für Psychologen, 2. Teil* (5. Aufl.). Weinheim: Juventa.

Literatur

Bortz, J. (2005). *Statistik* (6. Aufl.). Berlin: Springer.

Bortz, J., & Lienert, G. (2008). *Kurzgefasste Statistik für die klinische Forschung* (3. Aufl.). Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human–und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. MahWah: Erlbaum.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49*(12), 997–1003.

Colegrave, N., & Ruxton, G. D. (2003). Confidence Intervals are a More Useful Complement to Nonsignificant Tests Than are Power Calculations. *Behavioral Ecology, 14*(3), 446–447.

Cortina, J. M., & Dunlap, W. P. (1997). On the Logic and Purpose of Significance Testing. *Psychological Methods, 2*(2), 161–172.

Cumming, G., & Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals That are Based on Central and Noncentral Distribution. *Educational Psychological Measurement, 61*(4), 5323–5574.

Denis, D. J. (2003). Alternatives to Null Hypothesis Significance Testing. *Theory & Science, 4*(1). Retrieved November 1, 2013, from http://theoryandscience.icaap.org/content/vol4.1/02_denis.html

Erdfelder, E., Faul, F., & Buchner, A. (1996). G Power: A general power analysis program. *Behaviour Research Methods, Instruments and Computers, 28*(1), 1–11.

Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Psychology, 70*(1), 215–218.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1–20.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (2009). *What if there were no significance tests?* New York: Psychology Press, Taylor & Francis.

Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19–24.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management, 63*(3), 763–772.

Johnson, D. H. (2005). What hypothesis tests are not: A response to colegrave and ruxton. *Behavioral Ecology, 16*(1), 323–324.

Johnson, N. L. & Kotz, S. (1970). *Continuous univariate distributions*. Boston: Houghton Mifflin.

Kendall, M. G. & Stuart, A. (1973). *The advanced theory of statistics* (2nd ed.). London: Griffin.

Klemmert, H. (2004). *Äquivalenz- und Effekttests in der psychologischen* Forschung. Frankfurt/Main: Lang.

Kline, R. B. (2004). *Beyond significance testing*. Washington: American Psychological Association.

Kline, R. B. (2005). *Beyond significance testing. Reforming data analysis methods in behavioral research* (2nd ed.). Washington, DC: American Psychological Association.

Lipsey, M. W. & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48*(12), 1181–1209.

Morrison, D. E. & Henkel, R. E. (Eds.). (2006). *The significance test controversy*. Chicago: Aldine.

Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist, 45*(3), 403–404.

Murphy, K. R. & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah: Erlbaum.

Murphy, K. R. & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology, 84*(2), 234–248.

Murphy, K. R. & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah: Erlbaum.

Murphy, K. R., Myors, B., & Wolach, A. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (3rd ed.). New York, NY, US: Routledge/Taylor & Francis Group.

Nachtigall, C., Kroehne, U., Funke, F., & Steyer, R. (2003). (Why) Should we use SEM? Pros and cons of structural equation modelling. *Methods of Psychological Research Online, 8*(2), 1–22.

Neumann, M. (2005). *Entwicklung einer Skala zur Erfassung des Glaubens an Verschwörungstheorien (GVT-Skala)*. Unveröffentlichte Diplomarbeit, TU Berlin, Berlin.

Rindskopf, D. M. (2009). Testing „small," not null, hypotheses: classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319–332). New York: Psychology Press, Taylor & Francis.

Serlin, R. C. & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues* (pp. 199–228). Hillsdale: Erlbaum.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals. An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*(4), 371–386.

Metaanalyse

Nicola Döring, Jürgen Bortz

16.1 Quantitative Metaanalyse – 895

16.2 Arbeitsschritte einer Metaanalyse – 904

16.3 Aggregierung von Primärstudien ohne gemeinsame Effektgrößenschätzung – 924

16.4 Beispiel: Eine Fixed-Effect-Metaanalyse zum Lehrererwartungs-Effekt – 927

16.5 Software für Metaanalysen – 937

Literatur – 940

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_16, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Ziele und Anwendungsmöglichkeiten der Metaanalyse kennen.
- Die Metaanalyse von narrativen Reviews abgrenzen können.
- Vorliegende Metaanalysen kritisch rezipieren können.
- Die Schritte bei der Durchführung einer quantitativen Metaanalyse erläutern können.
- Mögliche Fehler bei der Durchführung einer Metaanalyse sowie entsprechende Gegenmaßnahmen erklären können.

Eine Metaanalyse fasst im Sinne der Forschungssynthese die Ergebnisse einzelner empirischer Untersuchungen zusammen. Es handelt sich also um eine Analyse von Datenanalysen vorliegender Studien (deswegen „Meta-Analyse“): Die Untersuchungseinheiten einer Metaanalyse sind nicht einzelne Untersuchungspersonen bzw. -objekte, sondern jeweils ganze Primärstudien, die sich alle auf dieselbe Fragestellung oder Hypothese beziehen.

Man unterscheidet **quantitative Metaanalysen**, die die Ergebnisse Dutzender oder Hunderter quantitativer Studien zum gleichen Thema statistisch zu einem Gesamteffekt verrechnen und **qualitative Metaanalysen**, die die Ergebnisse qualitativer Studien zum gleichen Forschungsthema interpretativ zu einer Gesamtaussage aggregieren (Paterson et al., 2009; Sandelowski & Barroso, 2006; Timulak, 2009). Die quantitative Metaanalyse wird seit den 1970er-Jahren intensiv eingesetzt und erlebt in jüngster Zeit einen regelrechten Boom, v. a. in der Medizin, der Psychologie und der Erziehungswissenschaft: Die Zahl der in wissenschaftlichen Literaturdatenbanken nachgewiesenen quantitativen Metaanalysen bzw. methodischen Arbeiten über das Verfahren ist exponentiell gestiegen (Kontopantelis & Reeves, 2009). Demgegenüber ist die qualitative Metaanalyse bislang sehr viel weniger methodisch elaboriert und wird auch seltener eingesetzt. In dem Maße, in dem ideologische Abgrenzungen zwischen quantitativer und qualitativer Forschung abgebaut und verstärkt integrative Ansätze im Sinne von Mixed-Methods-Designs verfolgt werden (► Abschn. 2.4 und ► Abschn. 7.1), ist auch die Forschungssynthese zunehmend gefordert, bei der Bearbeitung einer Fragestellung einschlägige Erkenntnisse sowohl aus quantitativen als auch qualitativen Studien zu integrieren (für entsprechende Ansätze Petticrew & Roberts, 2006; Popay, 2006; Pope, Mays, & Popay, 2007). Man spricht hier auch von **„mixed methods research synthesis“** (Heyvaert, Maes, & Onghena, 2011).

Allgemein werden mit **Forschungssynthese** („research synthesis“) Methoden angesprochen, die darauf abzielen, den Forschungsstand zu einer bestimmten Fragestellung systematisch zu recherchieren, zu bewerten und zu bündeln (Cooper, 2009). In Zeiten weltweit wachsender Produktion und Zugänglichkeit wissenschaftlicher Ergebnisse – nicht zuletzt durch internetgestützte Datenbanken und Netzwerke – sind einzelne Forschende, Multiplikatoren sowie Entscheidungsträger/innen in der Praxis immer weniger in der Lage, sich angesichts der Informationsfülle ad hoc selbst einen Überblick über den Forschungsstand zu verschaffen. Insbesondere **systematische Reviews** bzw. **systematische Übersichtsarbeiten** („systematic reviews“), die nicht selten statistische Metaanalysen enthalten, sollen hier Abhilfe schaffen (Kunz, Khan, Kleijnen, & Antes, 2009; Littell, Corcoran, & Pillai, 2008). Vorhandene Ergebnisse nach wissenschaftlichen Kriterien zu aggregieren, ist für den wissenschaftlichen Erkenntnisfortschritt sowie die praktische Anwendbarkeit von Forschungsergebnissen (z. B. evidenzbasierte Medizin, evidenzbasierte Bildungs- und Sozialreformen) ebenso wichtig wie die Durchführung weiterer Primärstudien. Es ist zu beachten, dass die Bezeichnung „Forschungssynthese“ einen weiten Oberbegriff darstellt, unter den verschiedene Typen von Forschungssynthesen fallen (z. B. Übersichtsarbeiten zu Theorien, Methoden, Anwendungsfeldern), während mit der quantitativen Metaanalyse ein ganz **spezieller Typ der Forschungssynthese** gemeint ist, nämlich die rechnerische Zusammenfassung von statistischen Befunden (speziell Effektgrößen) aus Primärstudien (Cooper & Hedges, 1993, S. 5).

Angesichts der bislang noch geringen Verbreitung der qualitativen Metaanalyse (im Sinne der interpretativen Aggregierung empirischer Befunde qualitativer Studien) wird in der vorliegenden Auflage dieses Lehrbuchs auf deren Darstellung verzichtet (Sandelowski & Barroso, 2006; Zhao, 1991). Wenn im Folgenden von „Metaanalyse“ die Rede ist, ist somit immer der quantitative Ansatz der Ergebniszusammenfassung vergleichbarer Primärstudien gemeint. Die quantitative Metaanalyse wird zunächst hinsichtlich ihrer Merkmale und Probleme charakterisiert (► Abschn. 16.1), bevor wir die einzelnen Arbeitsschritte mitsamt der notwendigen Formeln darstellen (► Abschn. 16.2). Wie Primärstudien zu aggregieren sind, zu denen keine vollständigen Informationen über die Effektgrößen vorliegen und somit auch keine klassische Metaanalyse möglich ist, erläutert ► Abschn. 16.3. Schließlich wird die Durchführung einer Fixed-Effect-Metaanalyse an einem kleinen Beispiel zum Lehrererwartungseffekt praktisch demonstriert (► Abschn. 16.4). Hinweise zur Software-Unterstützung von Metaanalysen schließen das Kapitel ab (► Abschn. 16.5).

16.1 Quantitative Metaanalyse

Wie ist die quantitative Metaanalyse definiert, und welche Zielsetzungen verfolgt sie? Wie ist sie vom narrativen Review abzugrenzen? Welche Probleme können bei der Durchführung einer Metaanalyse auftreten, und wie ist ihnen zu begegnen? Die folgenden Abschnitte geben Antworten auf diese Fragen.

16.1.1 Definition und Zielsetzung von Metaanalysen

Unter der Bezeichnung **(quantitative) Metaanalyse** versteht man eine Gruppe von Verfahren, mit denen die empirischen Ergebnisse verschiedener quantitativer Untersuchungen mit gemeinsamer Thematik statistisch zusammengefasst werden, um so einen Überblick über den Forschungsstand zu einem **bestimmten statistischen Effekt** – z. B. Mittelwert, Mittelwertsdifferenz, Korrelation, Prozentwert, relatives Risiko oder Quotenverhältnis (Odds Ratio) – zu gewinnen (zu Effekten und Effektgrößen ▶ Abschn. 14.2). Die Metaanalyse läuft auf eine **Schätzung des Gesamteffekts** („summary effect") hinaus sowie auf das **Aufspüren von Moderatorvariablen**, die dafür verantwortlich sind, dass der Effekt unter bestimmten Umständen größer oder kleiner ausfällt. Dabei besteht die Herausforderung zunächst darin, alle einschlägigen Studien zu sammeln. Vor der Berechnung des Gesamteffekts sind die gefundenen Studien dann noch hinsichtlich ihrer Aussagekraft kritisch zu bewerten und entsprechend **unterschiedlich zu gewichten**: Experimentelle Studien haben eine höhere interne Validität als quasi- oder nicht-experimentelle Studien, weshalb z. B. manche Metaanalysen von vorne herein nur experimentelle Kontrollgruppenstudien berücksichtigen (z. B. bei der Wirksamkeitsprüfung von Medikamenten). Studien mit probabilistischen Stichproben verfügen im Unterschied zu Untersuchungen mit nicht-probabilistischen Samples über höhere externe Validität, so dass manche Metaanalysen Ergebnisse auf der Basis nicht-zufälliger Stichproben ausschließen (also mit dem Faktor Null gewichten). Studien mit größeren Stichprobenumfängen liefern statistisch präzisere Effektgrößenschätzungen (bzw. engere Konfidenzintervalle) als solche mit kleinen Stichproben – präzisere Studien gehen oft mit einem größeren Gewichtungsfaktor in die Berechnung des Gesamteffektes ein. Ebenso können methodische und inhaltliche Merkmale der Studien im Zuge der Moderatorvariablen-Analyse bedeutsam sein, etwa wenn sich nachweisen lässt, dass der interessierende Effekt bei bestimmten Populationen oder bestimmten Dosierungen der unabhängigen Variablen stärker oder schwächer ausfällt.

Hinsichtlich der Zielsetzung, vorliegende Studien eines Forschungsgebietes systematisch zusammenzufassen, rivalisiert die Metaanalyse („meta-analytic review") mit dem **narrativen Review** („narrative review", „classical literature review"), dessen primäre Aufgabe ebenfalls darin besteht, den aktuellen Forschungsstand zu einer bestimmten Thematik anhand vorliegender Literatur aufzuarbeiten und zu verdichten (Cooper & Hedges, 1993). Der Hauptunterschied zwischen narrativem Review einerseits und metaanalytischem Review andererseits besteht in der Auswertung der recherchierten Studien: Während die Metaanalyse auf eine statistische Analyse hinausläuft, bietet das narrative Review eine verbale Zusammenfassung und Bewertung des Forschungsstandes (▶ Abschn. 16.1.2).

Bei der Metaanalyse handelt es sich **nicht** – wie z. B. bei der Varianz- oder Regressionsanalyse – um eine einheitliche Familie statistischer Datenanalysemethoden, sondern vielmehr um ein bestimmtes Untersuchungsdesign bzw. einen **Studientyp**, bei dem unterschiedliche konzeptionelle und methodische Verfahren (z. B. Literaturrecherche; Einschätzung der Aussagekraft der Studien) – darunter spezielle statistische Methoden (z. B. Berechnung und Signifikanztest des Gesamteffekts) – zum Einsatz kommen. Deswegen kann es vorkommen, dass zwei Metaanalysen, die zur selben Fragestellung durchgeführt werden und sich auf dieselben Primärstudien beziehen, dennoch zu unterschiedlichen Ergebnissen führen (Steiner, Lane, Dobbins, Schnur, & McConnell, 1991). Dies kann z. B. daran liegen, dass unterschiedlich strenge Kriterien an die methodische Qualität der Primärstudien angelegt werden, um in die Metaanalyse aufgenommen zu werden. Es kann auch daran liegen, dass zur Vereinheitlichung und Zusammenfasssung der Befunde der Primärstudien unterschiedliche statistische Verfahren eingesetzt werden. Ein weiterer Grund sind unvollständige statistische Angaben in den Primärstudien, die dann in der Metaanalyse auf unterschiedlichen Wegen kompensiert werden müssen (zum Missing-Data-Problem ▶ Abschn. 16.1.3). Eine wesentliche Voraussetzung für aussagekräftige Metaanalysen sind somit Primärstudien, deren Ergebnisse vollständig zugänglich sind.

Insgesamt zeichnet sich mit wachsender Verbreitung der Metaanalyse eine zunehmende Standardisierung der Vorgehensweisen ab (zur Methodik der Metaanalyse s. Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper & Hedges, 1993; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001; Schulze, 2004).

Quantitative Metaanalyse – Die quantitative **Metaanalyse** („meta analysis", „meta-analytic review") ist ein spezieller Typ der Forschungssynthese. Sie fasst den Forschungsstand zu einer bestimmten Fragestellung zusammen, indem sie die statistischen Einzelergebnisse inhaltlich vergleichbarer, aber unabhängiger quantitativer Primärstudien integriert. Dazu werden unterschiedliche konzeptionelle und methodische Verfahren verwendet, darunter auch spezielle statistische Analysetechniken. Das Ergebnis der Metaanalyse basiert auf einer Gesamtstichprobe, die sich aus den Stichproben der einzelnen Primärstudien zusammensetzt.

Gesamteffekt – Bei der Metaanalyse wird aus den Effektgrößenmaßen der einzelnen Primärstudien ein gewichteter Mittelwert berechnet, der den **Gesamteffekt** in der Population schätzt. Im Zuge der Metaanalyse kann überprüft werden, ob ein fraglicher Effekt in der Population vorliegt (Signifikanztest für den Gesamteffekt) und wie groß er ist (Effektgrößenschätzung des Gesamteffekts als Punkt- und/oder Intervallschätzung). Häufig werden die Wirksamkeit einer Maßnahme oder die Auswirkung einer Verhaltensweise mittels Metaanalyse untersucht.

Moderatorvariablen – Neben der Schätzung des Gesamteffekts wird durch die Metaanalyse der Wirkprozess analysiert, indem man diejenigen **Moderatorvariablen** identifiziert, die dafür verantwortlich sind, dass der Effekt unter bestimmten Bedingungen stärker oder schwächer ausfällt (Moderatorvariablen- bzw. Subgruppen-Analyse).

Ein Ziel von Metaanalysen ist die statistische Aggregierung der Ergebnisse von Primäruntersuchungen zur selben Fragestellung im Hinblick auf einen ganz konkreten Effekt (z. B. einen Unterschied zwischen 2 Gruppen: Unterschiedshypothese; einen Zusammenhang zwischen 2 Variablen: Zusammenhangshypothese; eine Veränderung zwischen Vorher- und Nachher-Messung/en: Veränderungshypothese). Die Metaanalyse berechnet diesen **Gesamteffekt** und prüft ihn auf **Signifikanz**. Dabei kann die Metaanalyse einen „wahren" Populationseffekt mit höherer Wahrscheinlichkeit aufdecken als eine Einzelstudie. Denn eine Metaanalyse, die auf einer Gesamtstichprobe basiert, die sich aus allen unabhängigen Einzelstichproben der einbezogenen Primärstudien addiert, hat gegenüber den Primärstudien eine höhere Teststärke (Cohn & Becker, 2003) und somit eine höhere **statistische Validität** (zur Validitäts-Typologie in der Campbell-Tradition ▶ Abschn. 3.2.2). Auch die **Konstruktvalidität** ist höher, da in eine Metaanalyse unterschiedliche Operationalisierungen der interessierenden Konstrukte eingehen. Der **internen Validität** eines metaanalytischen Befundes kommt es zugute, dass die methodische Qualität der eingehenden Primärstudien genau geprüft und diskutiert wird und oft nur experimentelle Kontrollgruppenstudien berücksichtigt werden. Die **externe Validität** von Metaanalysen ist hoch, da anhand des Studienpools sowie der Merkmale der Studien mittels **Moderatorvariablenanalyse** relativ genau angegeben werden kann, auf welche Bedingungen (Personengruppen, Settings, Zeitpunkte) der Effekt gut oder nicht so gut übertragbar ist.

Wenn wir also die Bestimmung des Gesamteffektes (Berechnung der mittleren Effektgröße) sowie die Moderatorvariablen-Analyse (Erklärung der Varianz der Effektgrößen) als die beiden wesentlichen Ziele der Metaanalyse betrachten, stellt sich zunächst die Frage, wie der Gesamteffekt („summary effect") im Einzelnen zu bestimmen ist. Da er als **gewichteter Durchschnitt** berechnet wird, müssen für alle eingehenden Effektgrößen der Primärstudien **Gewichtungsfaktoren** bestimmt werden. Die Gewichtung der Studien hängt dabei von ihren Eigenschaften ab sowie von dem zugrunde gelegten **Effektmodell der Metaanalyse** (Hedges & Vevea, 1998):

- Das Modell fester Effekte (**„fixed-effect model"**) geht davon aus, dass alle Studien exakt denselben Populationseffekt (deswegen: „fixed effect" im Singular) erfassen und dass die Unterschiedlichkeit der Effektgrößen zwischen den Studien allein auf Stichprobenfehler („sampling error", „within-study variance") zurückgeht. (In der Literatur wird die Fixed-Effect-Metaanalyse teilweise jedoch auch im Plural als „Fixed-Effects"-Metaanalyse bezeichnet.) Eine Studie mit geringem Stichprobenumfang weist einen hohen Stichprobenfehler auf und geht im Fixed-Effect-Modell mit geringerem Gewichtungsfaktor in die Metaanalyse ein.
- Das Modell zufallsbedingter Effekte (**„random-effects model"**) basiert dagegen auf der theoretischen Annahme, dass sich in jeder Studie **eine andere Populationseffektgröße** (deswegen: „random effects" im Plural) ausdrückt, etwa weil sich Studien hinsichtlich der Merkmale der Studienteilnehmenden oder der Untersuchungsbedingungen unterscheiden („between-studies variance"). Das Random-Effects-Model postuliert also, dass studienspezifische Effekte gemessen werden, ohne dass die **studienspezifischen Varianzquellen** im Einzelnen genau bekannt sind oder sein müssen – sie gehen vielmehr als Zufallsfehler in das Modell ein. Im Random-Effects-Modell wird neben dem Stichprobenfehler (d. h. den Messwertunterschieden innerhalb der Studien: Präzision) also noch die Unterschiedlichkeit zwischen den Studien (Heterogenität) berücksichtigt: Eine Studie mit geringem Stichprobenumfang wird dementsprechend geringer gewichtet (analog dem Fixed-Effect-Modell). Gleichzeitig wird eine Studie, deren Effekt sich deutlich von den anderen Studien unterscheidet, stärker gewichtet, denn diese Studie bringt offenbar eine Information ein, die die anderen Studien nicht liefern. Insgesamt fallen die Studiengewichte durch die Berücksichtigung von zwei Varianzquellen im Random-Effects-Modell gleichmäßiger aus als im Fixed-Effect-Modell (Borenstein et al., 2009, S. 79).

Welches metaanalytische Effektmodell – und somit welche Berechnungsweise der Studiengewichte – bei einer selbst durchgeführten Metaanalyse zu wählen ist,

hängt in erster Linie von **theoretischen Überlegungen** ab (▶ Abschn. 16.2.4). Zudem existieren unterschiedliche metaanalytische Schulen mit abweichenden Berechnungsansätzen, etwa die homogenitätsbasierte Metaanalyse und die psychometrische Metaanalyse (▶ Abschn. 16.2).

Effektmodelle der Metaanalyse – Das Effektmodell der Metaanalyse gibt vor, wie die Gewichtungsfaktoren für die Ergebnisse der einzelnen Primärstudien-Effekte zu bestimmen sind, auf deren Basis der Gesamteffekt als gewichteter Mittelwert berechnet wird. Man unterscheidet Fixed-Effect-, Random-Effects- und Mixed-Effects-Modelle der Metaanalyse.

Fixed-Effect-Modell – Das Fixed-Effect-Modell (Modell fester Effekte) geht davon aus, dass die Primärstudien im Studienpool alle denselben Populationseffekt abbilden und Differenzen nur auf Stichprobenfehler zurückgehen. In die Gewichtungsfaktoren geht nur der Stichprobenumfang ein.

Random-Effects-Modell – Das Random-Effects-Modell (Modell zufallsbedingter Effekte) geht davon aus, dass die Primärstudien im Studienpool unterschiedliche Populationseffekte abbilden, dass also Differenzen nicht nur auf Stichprobenfehler, sondern auch auf inhaltliche Unterschiede der Studien zurückzuführen sind. In die Gewichtungsfaktoren geht neben dem Stichprobenumfang auch die Abweichung des Ergebnisses einer Primärstudie von den anderen Untersuchungen im Studienpool ein.

Mixed-Effects-Modell – Das Mixed-Effects-Modell kombiniert das Fixed-Effect- und das Random-Effects-Modell.

Eine wichtige Komponente der Metaanalyse neben der Schätzung des Gesamteffekts ist die **Moderatorvariablen-Analyse**. Sie hat das Ziel, **systematische Unterschiede** zwischen den Studien herauszuarbeiten. Hierbei macht man es sich zunutze, dass in dem in die Metaanalyse eingehenden Studienpool der interessierende Effekt meist unter verschiedenen Bedingungen sowie an verschiedenen Populationen untersucht wurde. Für die inhaltlich relevanten **Moderatorvariablen**, bei denen es sich z. B. um Merkmale des Designs (z. B. experimentelle vs. nicht-experimentelle Studie), der unabhängigen Variablen (z. B. unterschiedliche Dosierung eines Medikamentes), der abhängigen Variablen (z. B. Messung des subjektiven vs. des objektiven Lernerfolges) sowie der Studienteilnehmenden (z. B. ältere vs. jüngere Studienteilnehmende) oder auch der Studienfinanzierung (z. B. finanziert mit oder ohne Drittmittel von Interessensgruppen) handeln kann, wird geprüft, ob sie einen substanziellen Einfluss auf den Effekt haben und wenn ja, wie groß dieser ist. Handelt es sich bei der betrachteten Moderatorvariable um eine kategoriale Variable, so lassen sich die Studien gemäß den Ausprägungen der betreffenden Moderatorvariable in Subgruppen aufteilen. Unterschiede zwischen den Gruppen weisen dann auf die Wirkung der Moderatorvariable hin (z. B. wenn bei der Untersuchung der Wirkung eines Medikamentes industriegeförderte Studien systematisch geringere Risiken oder bessere Wirkungen zeigen als Studien ohne einen interessensgeleiteten Sponsor). Die Durchführung einer solchen **Subgruppen-Analyse** zum Nachweis des Einflusses kategorialer Moderatorvariablen wird in ▶ Abschn. 16.2.8 sowie in ▶ Abschn. 16.4.5 dargestellt. Sind mehrere und auch kontinuierliche Moderatorvariablen zu berücksichtigen, so wird die Moderatorvariablen-Analyse in der Regel mittels **Meta-Regression** durchgeführt (kategoriale Variablen können dabei als Dummy-kodierte dichotome Variablen eingehen; Viechtbauer, 2007). Die Meta-Regression ist das zu bevorzugende Verfahren, da Moderatorvariablen aller Skalenniveaus integrierbar sind und die Interkorrelationen aller Moderatoren mitberücksichtigt werden.

Inhaltlich geht es in Metaanalysen häufig um die **Wirksamkeit einer Maßnahme** (z. B. psychotherapeutische Behandlung, Medikamentengabe, Lehrmethode), weshalb Metanalysen in der Therapieforschung (▶ Geschichte der Metaanalyse) und in der Medizin, aber auch in der Erziehungswissenschaft verbreitet sind. Zudem beziehen sich Metaanalysen auch auf die **Auswirkung bestimmter Verhaltensweisen** (z. B. Metaanalyse zu Wirkungen der Nutzung gewalthaltiger Computerspiele: Anderson et al., 2010). Metaanalysen sind nicht nur zur Beschreibung des Forschungsstandes in einem begrenzten Forschungsfeld, sondern häufig auch für politische Entscheidungen und öffentliche Meinungsbildung wichtig, d. h. die Methode fördert den **Wissenstransfer** zwischen Forschung und Praxis. So spielen Metaanalysen z. B. auch bei der Vorbereitung größerer **Evaluationsprojekte** eine wichtige Rolle: Denn Evaluationsstudien sollen sich oft nicht mit dem Nachweis einer irgendwie gearteten Maßnahmenwirkung begnügen, sondern eine Effektgröße vorgeben, die mindestens zu erreichen ist, damit die Maßnahme als praktisch oder klinisch bedeutsam einzuordnen ist bzw. damit der Nutzen die Kosten übertrifft (zur Evaluationsforschung ▶ Kap. 18). Derartige Angaben werden erheblich erleichtert, wenn man auf bereits vorhandene bzw. selbst durchgeführte Metaanalysen zurückgreifen kann.

Der Boom der Metaanalyse hat dazu geführt dass mittlerweile **Meta-Metaanalysen** möglich sind, also systematische Zusammenfassungen vorliegender Metaanalysen. Viel beachtet wurde die Meta-Metaanalyse des neuseeländischen Erziehungswissenschaftlers John Hattie (2008), der über 800 Metaanalysen (basierend auf mehr als 50 000 Primärstudien) zum schulischen Lernerfolg zusammenfasste. Er leitete daraus sein Modell für guten Unterricht nach dem Prinzip „Lernen sichtbar machen" ab. Die Studie wird in der breiten Öffentlichkeit teilweise

Geschichte der Metaanalyse

Die Metaanalyse wurde in der Psychologie entwickelt und ist eng mit der Forschung zur **Wirksamkeit von Psychotherapie** verknüpft: 1952 veröffentliche der deutsch-britische Psychologe **Hans Jürgen Eysenck** einen Artikel, der auf der Basis von 24 Primärstudien die Heilerfolge der Psychoanalyse mit anderen Therapieansätzen sowie einer unbehandelten Kontrollgruppe verglich (Eysenck, 1952). Er kam zu dem Ergebnis, dass bei schwer neurotischen Patienten bei 72 % nach zwei Jahren eine Spontanheilung ohne Behandlung feststellbar war, während von den psychoanalytisch behandelten Patienten nur 44 % nach zwei Jahren als stark gebessert oder geheilt einzustufen waren (5 Studien, $N = 760$, Tab. 16.1).
Andere Therapieformen als die Psychoanalyse brachten es auf einen Heilerfolg von 64 % (19 Studien, $N = 7293$). Eysenck rief – auch auf der Basis seiner anderen Beiträge zur Psychotherapieforschung – dazu auf, die subjektive Überzeugung von der **Wirksamkeit der Psychoanalyse zu hinterfragen** und stattdessen mehr empirische Forschung in diesem Bereich zu betreiben. Sein Befund, dass Psychoanalyse nicht nur nicht hilft, sondern womöglich sogar den Selbstheilungsprozess behindert und somit schadet, wurde in der Scientific Community äußerst kontrovers diskutiert.
In den folgenden Jahren boomte die Psychotherapieforschung. Anfang der 1970er-Jahre lagen dann Hunderte entsprechender Primärstudien vor, die teils positive, teils negative, teils keine Effekte psychotherapeutischer Behandlungen zeigten. Somit bestand die Notwendigkeit einer systematischen statistischen Integration des Forschungsstandes. Es war der amerikanische Statistiker **Gene Glass**, der in Abgrenzung von Primäranalyse und Sekundäranalyse den **Begriff der Metaanalyse** prägte und folgendermaßen definierte (Glass, 1976, S. 3):

» Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature.

Zusammen mit der amerikanischen Erziehungswissenschaftlerin Mary Lee Smith veröffentlichte Glass eine viel zitierte **Metaanalyse** (Smith & Glass, 1977), die auf der Basis einer statistischen Aggregierung von rund 400 Primärstudien aus der Psychotherapieforschung zu dem Ergebnis kam, dass sich bei psychotherapeutisch behandelten Patienten eine **viel stärkere Besserung** zeigte als bei unbehandelten Patienten.
Dieser Befund widersprach den Studien von Eysenck, der seinerseits die Metaanalyse als „**Mega-Silliness**“ bezeichnete (Eysenck, 1978). Doch inhaltlich setzte sich die Anerkennung der empirisch nachweisbaren Wirksamkeit von Psychotherapie ebenso durch wie die Methodik der Metaanalyse, zu deren Weiterentwicklung Glass zahlreiche wichtige Beiträge lieferte (Glass, 1999).
Ein zweiter wichtiger Impuls zur Entwicklung der Metaanalyse ging von der **Forschung zu interpersonalen Erwartungseffekten** aus. Der amerikanische Psychologe **Robert Rosenthal** entdeckte in den 1960er-Jahren, dass die Erwartungen von Versuchsleitern das Verhalten von Versuchspersonen und die Erwartungen von Lehrern die Leistungen von Schülern systematisch beeinflussen können. Der von ihm postulierte Erwartungseffekt, der in der Praxis auf die Verfälschung von Experimenten sowie die systematische Benachteiligung bestimmter Schülergruppen hinausläuft, war anfangs sehr umstritten. Rosenthal publizierte ab den 1970er-Jahren eine ganze Serie von Metaanalysen, in denen er die Gültigkeit des Effektes belegte (▶ Abschn. 16.4). Gleichzeitig lieferte Rosenthal (1978, 1984, 1993) zahlreiche Beiträge zur methodischen Weiterentwicklung der Metaanalyse und auch eine Zusammenfassung der Geschichte der Metaanalyse und ihrer Kritik (Rosenthal & DiMatteo, 2001).

16

als „heiliger Gral“ der Schul- und Unterrichtsforschung gehandelt, fachintern jedoch neben der Würdigung der wichtigen Synthese-Leistung inhaltlich und methodisch auch kritisch diskutiert (Snook et al., 2009).

Neben Metaanalysen mit inhaltlichen Fragestellungen existieren auch **methodenorientierte Metaanalysen**, die es sich z. B. zum Ziel setzen, Studien zu Testgütekriterien eines Instruments zu aggregieren und auf dieser Basis eine Gesamtschätzung der Reliabilität oder Validität eines Testverfahrens vorzulegen (für eine Metaanalyse zu den Testgütekriterien des in ▶ Abschn. 10.4.2 angesprochenen Wartegg-Zeichen-Tests: Soilevuo Grønnerød & Grønnerød, 2012). Auch andere Methodenfragen wie z. B. die Rücklaufquoten verschiedener Umfragemodi lassen sich metaanalytisch untersuchen (für einen metaanalytischen Vergleich der Rücklaufquoten bei Web-Umfragen im Vergleich zu postalischen, telefonischen oder sonstigen Umfragetypen: Manfreda et al., 2008).

16.1.2 Abgrenzung zwischen Metaanalyse und narrativem Review

Ein **Übersichtsartikel** bzw. **narratives Review** („narrative review“, „ review of the literature“, „review of the state of research“) fasst den aktuellen Forschungsstand in einem Gebiet zusammen, indem er die einschlägige Literatur recherchiert, strukturiert vorstellt und bewertet. Dabei können theoretische, empirische und methodische Stärken und Schwächen des fraglichen Forschungsgebietes diskutiert werden, v. a. aber sollen die wichtigsten Befunde herausgearbeitet und Perspektiven für die zukünftige Forschung aufgezeigt werden. Vertreter der Metaanalyse kritisieren am narrativen Review v. a. seine **Subjektivität**: Auswahl, Gewichtung und Bewertung der zu integrierenden Untersuchungen obliegen weitgehend den Einschätzungen der Reviewerin bzw. des Reviewers und erfolgen oft nicht strikt systematisch bzw. werden nicht im De-

Tabelle 16.1 Zusammenfassung berichteter Ergebnisse von Psychotherapie. (Mod. nach Eysenck (1952). The Effects of Psychotherapy: An Evaluation. *Journal of Consulting Psychology, 16*(5), 319–324. The content is in the public domain)

	Zusammenfassung der Berichte der Psychotherapieergebnisse					
	Stichprobenumfang je Studie	Geheilt; stark verbessert	Verbessert	Leicht verbessert	Nicht verbessert; verstorben; Behandlung abgebrochen	Geheilt; stark verbessert; verbessert
	[*N*]	[absolute Häufigkeit]	[absolute Häufigkeit]	[absolute Häufigkeit]	[absolute Häufigkeit]	[%]
Psychoanalytisch						
1. Fenichel (1930)	484	104	84	99	197	39
2. Kessel & Hyman (1933)	34	16	5	4	9	62
3. Jones (1936)	59	20	8	28	3	47
4. Alexander (1937)	141	28	42	23	48	50
5. Knight (1941)	42	8	20	7	7	67
Alle Fälle	760	335		425		44

tail dokumentiert (ausführlicher hierzu Rustenbach, 2003, oder – bezogen auf medizinische Forschung – Sauerbrei & Blettner, 2003). Insbesondere bei einer großen Zahl zu verarbeitender Studien sei es kaum möglich, durch pure Experteneinschätzung nach Sichtung der Literatur das resultierende Gesamtergebnis zuverlässig abzuleiten.

Demgegenüber seien **Metaanalysen** bzw. **metaanalytische Reviews** objektiver und transparenter, weil die Integration von Forschungsergebnissen hier nicht auf der sprachlichen Ebene, sondern auf der Ebene statistischer Indikatoren ansetzt (zum Vergleich von narrativem Review und Metaanalyse Beaman, 1991; Borenstein et al., 2009). Diese stärkere „Objektivität" wird jedoch dadurch erkauft, dass der Fokus der Metaanalyse meist viel enger ist als der des narrativen Reviews: Die Metaanalyse läuft auf eine **statistische Effektgrößenschätzung und Moderatorvariablen-Analyse** hinaus und resultiert also in der mehr oder minder gut gesicherten Aussage, ob ein fraglicher Effekt (z. B. Unterschied in der Lebenszufriedenheit zwischen Singles und Verheirateten; Zusammenhang zwischen Passivrauchen und Lungenkrebs; Reduktion des Gewichts durch eine bestimmte Diät) in der Population existiert und wie groß er ist sowie unter welchen Bedingungen er stärker oder schwächer auftritt. Dabei kann eine Metaanalyse eine sehr große Zahl von Primärstudien schlüssig integrieren.

Demgegenüber kann in einem narrativen Review meist nur eine kleinere Zahl von Beiträgen berücksichtigt werden. Der Schwerpunkt liegt nicht so stark auf der Aggregierung der Befunde. Stattdessen werden **theoretische, methodische und empirische Aspekte** ausführlich behandelt (z. B. Lebensform „ Single" im historischen Wandel, im interkulturellen Vergleich sowie im Selbst- und Fremdbild; Passivrauchen als soziales und medizinisches Problem etc.). Ein narratives Review liefert also z. B. Anhaltspunkte dazu, welche Unterschiede in der Konzeptualisierung des fraglichen Phänomens gemäß verschiedener Theorierichtungen bestehen, welche Bezugspunkte zu anderen Fachdisziplinen es gibt, welcher Theorie- und Methodenwandel sich im Forschungsgebiet abzeichnet etc. Derartige konzeptionelle Fragen spielen zwar auch in einer Metaanalyse eine Rolle, allerdings in anderer Gewichtung.

Es erscheint wenig sinnvoll, eine Konkurrenz zwischen den Leistungen einer Metaanalyse (v. a. quantitative Ergebnisintegration in einem Forschungsfeld) und narrativem Review (breite Darstellung und Reflexion theoretischer, methodischer und empirischer Aspekte eines Forschungsfeldes) aufzubauen, stattdessen ist eher von einem **Ergänzungsverhältnis** auszugehen. Denn beide Verfahren liefern, wenn sie sorgfältig umgesetzt werden, wichtige Beiträge zur Weiterentwicklung des wissenschaftlichen Kenntnisstandes. Wer eine Metaanalyse durchführen will, ist gut beraten, zunächst ein möglichst aktuelles narratives Review zu studieren, um sich im Forschungsfeld zu orientieren und die durch die Metaanalyse zu beantwortende Forschungsfrage nach einem bestimmten Effekt zu präzisieren. Umgekehrt sollte man in einem narrativen Review selbstverständlich die Ergebnisse von Metaanalysen gegenüber Einzelstudien bevorzugt referieren. Statistische Integrationen und narrative Zusammenfassungen des Forschungsstandes sind im Zuge des

Systematische Reviews in Medizin und Sozialwissenschaften

Als Oberbegriff für wissenschaftliche Arbeiten mit dem Ziel der Forschungssynthese hat sich insbesondere in der Medizin seit den 1990er-Jahren die Bezeichnung **systematisches Review** („systematic review") eingebürgert, um einen Forschungsüberblick zu kennzeichnen, der im Unterschied zum einfachen narrativen Review sehr viel systematischer und transparenter erfolgt (z. B. präzise Dokumentation aller Einzelschritte der Literaturrecherche, Vorabdefinition von Einschlusskriterien für die zu berücksichtigenden Studien) und oft eine quantitative Metaanalyse enthält.

Die Erstellung, regelmäßige Aktualisierung und Verbreitung von systematischen Reviews zu medizinischen Fragestellungen (z. B. Wirkt Vitamin C gegen Erkältungen? Wie wirken Antidepressiva bei Parkinson-Patienten?) wird seit 1993 weltweit über die **Cochrane Collaboration** (kurz: CC) organisiert. Das Netzwerk ist nach dem britischen Epidemiologen Archibald Leman Cochrane (1909–1988) benannt. Systematische Reviews zu schreiben und in der internetbasierten **Cochrane-Bibliothek** zu veröffentlichen (http://www.thecochranelibrary.com), fördert die wissenschaftliche Reputation der Forschenden und unterliegt strengen Qualitätsanforderungen (Higgins & Green, 2009).

Die Zusammenfassung und Online-Bereitstellung aktueller medizinischer Forschungsergebnisse in Form systematischer Reviews im Rahmen des Cochrane-Netzwerks soll nicht nur den Erkenntnisfortschritt in der medizinischen Forschung vorantreiben, sondern v. a. auch den praktisch tätigen Ärztinnen und Ärzten helfen, diagnostische und therapeutische Entscheidungen – unter Berücksichtigung der Patientenwünsche – nicht ausschließlich aus ihrer naturgemäß stets begrenzten klinischen Erfahrung heraus zu treffen (z. B. selektierter Patientenkreis der eigenen Praxis), sondern ergänzend auch externe Evidenz, d. h. den aktuellen internationalen Forschungsstand mit einzubeziehen (sog. **evidenzbasierte Medizin**). Dabei sollten v. a. die aussagekräftigsten Forschungsergebnisse herangezogen werden, und das sind bei vielen Interventionen Metaanalysen, die die Befunde von randomisierten Kontrollgruppenstudien zusammenfassen. Sie haben **„Evidenzklasse Ia"** nach dem GRADE-System („Grading of Recommendations Assessment, Development and Evaluation"; http://www.gradeworkinggroup.org) und rechtfertigen auf der Basis valider Wirksamkeitsnachweise klare Empfehlungen für oder gegen eine Behandlungsform.

Nach dem Vorbild der medizinischen Cochrane Collaboration wurde im Jahr 2000 die sozialwissenschaftliche **Campbell Collaboration** – kurz: C2 (http://www.campbellcollaboration.org) – gegründet, benannt nach dem US-amerikanischen Psychologen Donald T. Campbell (1916–1996). Inhaltlich geht es der C2 um die Erstellung und Publikation hochwertiger systematischer Reviews aus den Sozialwissenschaften, die sich mit **Fragen der Erziehung sowie des Rechts- und Sozialwesens** befassen. Ebenso wie bei der Cochrane Collaboration zielt auch die Campbell Collaboration durch die Synthese des aktuellen Forschungsstandes zum einen auf eine Förderung des wissenschaftlichen Erkenntnisfortschritts ab, zum anderen auf eine Unterstützung praktischer Entscheidungen, etwa im Sinne **evidenzbasierter Bildungs- oder Sozialreformen**.

Sich bei der Konzeption von Bildungs- und Sozialreformen auf systematische Reviews zu stützen, ist eine Möglichkeit, um wissenschaftliche Erkenntnisse für die Praxis fruchtbar zu machen. Ergänzend soll Evaluationsforschung (► Kap. 18) dazu beitragen, Entscheidungen für oder gegen praktische Maßnahmen auf der Basis empirischer Daten zu treffen.

16

wissenschaftlichen Erkenntnisgewinns stets wechselseitig aufeinander angewiesen (als Beispiel für eine kombinierte Anwendung beider Techniken anlässlich der Evaluation von Summer Schools s. Cooper, Charlton, Valentine, & Muhlenbruck, 2000). Im heutigen Verständnis wird das sog. narrative Review ohnehin zunehmend durch das anspruchsvollere **systematische Review** abgelöst, das meist auch eine Metaanalyse enthält (► Systematische Reviews in Medizin und Sozialwissenschaften).

16.1.3 Probleme bei der Durchführung von Metaanalysen und Gegenmaßnahmen

Bei der Durchführung von Metaanalysen stellen sich in der Praxis v. a. folgende sechs Probleme: 1. die notwendige Mindestanzahl an Primärstudien, 2. der Publication-Bias, 3. das Äpfel-und-Birnen-Problem, 4. das Garbage-in-Garbage-out-Problem, 5. das Problem abhängiger Untersuchungsergebnisse und 6. das Missing-Data-Problem.

Dass wir hier zunächst auf Probleme bei der Durchführung von Metaanalysen eingehen, bevor das typische metaanalytische Vorgehen schrittweise geschildert wird (► Abschn. 16.2), hat zwei Gründe: Die Sensibilisierung für Probleme soll von vorne herein eine kritische Lektüre vorliegender Metaanalysen fördern und zudem verdeutlichen, auf welche besonderen Herausforderungen man sich bei der Durchführung einer eigenen Metaanalyse einzustellen hat.

Die notwendige Mindestanzahl an Primärstudien

Ebenso wie wir bei Primärstudien von einer Mindestzahl an Untersuchungseinheiten ausgehen, um eine sinnvolle statistische Auswertung vorzunehmen (zum optimalen Stichprobenumfang ► Abschn. 14.3), ist auch eine Metaanalyse umso aussagekräftiger, je mehr Primärstudien eingehen. In der Logik des Modells fester Effekte („fixed-effect model", ► Abschn. 16.1.1), in dem jede Primärstudie einen Schätzwert derselben Populationseffektgröße liefert, ist die metaanalytische Zusammenfassung von

2 Studien bereits sinnvoll (Borenstein et al., 2009, S. 363). Im Modell zufallsbedingter Effekte („random-effects model“) besteht das Problem, dass bei geringer Studienzahl die Unterschiedlichkeit der Studien nicht gut geschätzt werden kann und somit die Gefahr großer Ergebnisverzerrungen besteht.

Mit einer **A-priori-Teststärkeanalyse** kann im Zuge der Planung einer Metaanalyse prinzipiell bestimmt werden, wie groß der Studienpool mindestens sein muss, damit der Signifikanztest des Gesamteffektes und v. a. die zur Voraussetzungsprüfung notwendigen Homogenitätstests jeweils die notwendige Teststärke aufweisen. Dabei sind jedoch zwei Probleme zu beachten: Zum einen können teilweise unrealisierbar hohe Studienzahlen resultieren. Zum anderen stehen Teststärkeanalysen in der Planungsphase einer Metaanalyse oft auf äußerst „tönernen Füßen“. Es müssen nämlich Annahmen über die Größe des zu erwartenden Gesamteffekts und über die Größe der studienspezifischen Effekte gemacht werden, die oft nicht gut begründbar sind. Falls dies doch möglich sein sollte, ist damit eigentlich das Ergebnis der Metaanalyse vorweggenommen – d. h., die Metaanalyse wäre letztlich überflüssig. „Simply taking a wild guess and inserting values into the formulas … is little better than simply guessing the power“ (Hedges & Pigott, 2001, S. 216). Zur Teststärke von Metaanalysen s. Hedges und Pigott (2001), Cornwell und Ladd (1993), Sackett, Harris und Orr (1986), Schulze (2004) oder Spector und Levine (1987).

Forschungspragmatisch mag man argumentieren, dass der Aufwand für eine Metaanalyse erst ab einer bestimmten Mindeststudienzahl von z. B. einem Dutzend oder mehr Studien lohnenswert erscheint – auch mit Blick auf die Publikationsmöglichkeiten. Andererseits ist es v. a. bei lebenswichtigen Fragestellungen (z. B. Überlebenschancen oder Komplikationsrisiken bei einer bestimmten Therapiemethode) durchaus wünschenswert, auch eine noch geringe Studienzahl bereits metaanalytisch zu aggregieren. Werden die Studienergebnisse nämlich einfach so nebeneinander gestellt, nehmen die Leserinnen und Leser in der Regel selbst eine **Ad-hoc-Zusammenfassung** vor, indem sie z. B. die Signifikanzen auszählen oder einen ungewichteten Mittelwert aus den Effektgrößen der Primärstudien bilden. Im Vergleich zu dieser Vorgehensweise sind die formalen metaanalytischen Ergebnisse, inklusive einer kritischen Diskussion der Grenzen ihrer Aussagekraft hilfreicher (Borenstein et al., 2009, S. 364).

Um ein sich gerade entwickelndes Forschungsfeld metaanalytisch zu begleiten, kann eine **prospektive kumulative Metaanalyse** durchgeführt werden, d. h. mit der jeweils vorhandenen (kleinen) Studienzahl wird eine Metaanalyse realisiert und diese wird wiederholt, sobald eine neue Studie hinzukommt. Die chronologisch entstehenden einzelnen (dann allerdings voneinander abhängigen) Metaanalysen werden in ihren Ergebnissen im Überblick grafisch dargestellt (Borenstein et al., 2009, S. 375). In der Cochrane-Bibliothek (▶ Systematische Reviews in Medizin und Sozialwissenschaften) sind systematische Reviews zu finden, die auch Metaanalysen über **sehr kleine Studienzahlen** enthalten – hier ist dann der Befund, wie klein der auffindbare Studienpool (noch) ist, eine relevante Information. Beispiel: Laut einem Cochrane-Review aus dem Jahr 2009 lagen zur Wirksamkeit von Homöopathie bei der Aufmerksamkeitsdefizit-Hyperaktivitätsstörung ADHS bis dato nur 4 Kontrollgruppenstudien mit einem Gesamtstichprobenumfang von $n = 168$ vor, die dennoch metaanalytisch zusammengefasst wurden (Heirs & Dean, 2007). Angesichts der Tatsache, dass die Wirkprinzipien der Homöopathie bei aktuellem Wissensstand nicht erklärbar sind, ist ein methodisch sauberer Wirksamkeitsnachweis wichtig, um jenseits ideologischer Positionierungen (Homöopathie-Gläubige vs. Homöopathie-Skeptiker; ▶ Abschn. 3.1.1), eine wissenschaftlich fundierte Position einnehmen zu können.

Der Publication-Bias

Das Anfertigen einer wissenschaftlichen Publikation ist in jedem Fall arbeitsaufwändig und aufgrund des Peer-Review-Prozesses auch nicht immer von Erfolg gekrönt. Viele empirische Studien bleiben deswegen unveröffentlicht in der sprichwörtlichen Schreibtischschublade liegen (Schubladenproblem; „file drawer problem“) bzw. landen im Papierkorb (Papierkorbproblem; „waste basket problem“). So mögen einerseits die Forschenden selbst von der Veröffentlichung einer Studie Abstand nehmen, wenn diese theoriekonträre oder nicht-signifikante Ergebnisse zeigte. Wenn sie trotz der vermeintlich „ schlechten“ Befunde ein Manuskript anfertigen und z. B. in einer nationalen Fachzeitschrift einreichen, so wird dieses mit höherer Wahrscheinlichkeit von den Gutachtenden abgelehnt als ein Manuskript mit statistisch signifikanten bzw. theoriekonformen Befunden. Die Tendenz von Forschenden sowie von Gutachtenden, theoriekonforme Befunde als publikationswürdiger anzusehen, wird auch als **Confirmation-Bias** bezeichnet (MacKay, 1993, S. 231). Im Ergebnis entsteht ein **Publication-Bias** dergemäß, dass empirische Studien mit wunschgemäßen Befunden häufiger publiziert werden, während Studien mit theoriekonträren oder nicht-signifikanten Ergebnissen häufiger unveröffentlicht bleiben. Eine Studie von Cooper, De Neve und Charlton (1997) zeigte, dass in von ihnen untersuchten psychologischen Instituten **ca. zwei Drittel aller durchgeführten Untersuchungen nicht publiziert** wurden, obwohl die Untersuchungen keine gravierenden

methodischen Mängel aufwiesen. Neben einer ergebnisbezogenen Vorselektion im Sinne des Confirmation-Bias werden Publikationsentscheidungen dabei natürlich noch durch andere Faktoren beeinflusst (z. B. Zeitmangel, Personalmangel, Karriereentwicklung der Forschenden etc.).

Festzuhalten bleibt: Eine Metaanalyse, die sich nur auf publizierte Studien stützt, ist systematisch in Richtung signifikanter Ergebnisse verzerrt und überschätzt tendenziell den interessierenden Effekt, weil alle unpublizierten Studien ignoriert werden, bei denen es sich häufiger um Studien mit nicht-signifikanten oder theoriekonträren Befunden handelt. Die drei wichtigsten **Gegenmaßnahmen gegen den Publication-Bias** sind:

- Maßnahmen zur Beschaffung unpublizierter Studien,
- Verfahren zur Abschätzung der Verzerrung durch den Funnel-Plot oder die Fail-Safe-N-Methode und deren Weiterentwicklungen sowie
- Verfahren für eine um die Verzerrung korrigierte Schätzung des Gesamteffekts, etwa mittels Trim-and-Fill-Prozedur (zu all diesen und weiteren Maßnahmen ▶ Abschn. 16.2.2 „Maßnahmen gegen den Publication-Bias“).

Man beachte, dass jegliche Literaturrecherchen (▶ Abschn. 6.1) und Erarbeitungen des Forschungsstandes (z. B. Theorieteil einer empirischen Studie, narratives Review) vom Publication-Bias betroffen sind, ohne dass dieses Problem diskutiert wird. Demgegenüber wird es im Rahmen der Metaanalyse ausdrücklich berücksichtigt.

Während sich der Publication-Bias im engeren Sinne auf die Problematik nichtpublizierter Studien bezieht, werden im weiteren Sinne auch **andere Verzerrungen im Zuge der Literaturrecherche** unter diesem Oberbegriff behandelt. Das methodische Problem besteht dabei nicht darin, dass einzelne relevante Studien übersehen werden und in der Metaanalyse fehlen, sondern dass systematisch v. a. diejenigen Studien unterrepräsentiert sind, die theoriekonträre oder nur schwache theoriekonforme Ergebnisse liefern. Somit führen entsprechende Publikationsverzerrungen zu einer **Begünstigung des untersuchten Effektes**, dem deswegen aktiv entgegengearbeitet werden muss. Unter anderem können folgende Verzerrungen wirken (Borenstein et al., 2009, S. 279f.):

- **Language-Bias**: Forschende verfügen meist nur über Kenntnisse ihrer Muttersprache sowie der Wissenschaftssprache Englisch und können somit einen Großteil der wissenschaftlichen Publikationen gar nicht rezipieren **(„tower-of-babel-problem“)**. Bei der Auswahl von Primärstudien für eine Metaanalyse werden somit sprachliche Auswahlkriterien angelegt: Typischerweise konzentriert man sich auf **internationale englischsprachige Journale**, die wissenschaftlich die höchste Reputation genießen (▶ Abschn. 13.1). Die Fokussierung auf englischsprachige Publikationen hat jedoch inhaltliche Konsequenzen, denn in den hoch angesehenen internationalen Zeitschriften werden meist nur besonders „gute“ Forschungsergebnisse publiziert. Nicht-englischsprachige Forschende haben nämlich die Tendenz, ihre „ besten“ Ergebnisse in internationalen, englischsprachigen Journalen zu veröffentlichen (Ellis, 2010). Die weniger prägnanten oder nicht-signifikanten Ergebnisse werden hingegen bevorzugt bei Zeitschriften in der Landessprache der Autoren eingereicht (Heres, Wagenpfeil, Hamann, Kissling, & Leucht, 2004). Grégoire, Derderian und LeLorie (1995) untersuchten dieses Phänomen anhand von 16 Metaanalysen und fanden heraus, dass sprachbezogene Ausschlusskriterien für Studien zu einer Verzerrung von Aussagen der Metaanalyse führen können. In der Konsequenz ergibt sich daraus die Empfehlung, sich bei Metaanalysen wenn möglich nicht auf englischsprachige Veröffentlichungen zu beschränken, sondern unterschiedliche Nationalsprachen zu berücksichtigen, was jedoch stark erschwert ist, wenn keine entsprechenden Datenbanken vorliegen (z. B. existiert keine spanischsprachige wissenschaftliche Literaturdatenbank). Nicht zuletzt ist auch zu beachten, dass Publikationssprache und Kulturkreis miteinander verknüpft sind. Eine Beschränkung auf englischsprachige Literatur impliziert meist, dass Studien aus bestimmten Kulturkreisen systematisch unterrepräsentiert sind, was im Hinblick auf die Generalisierbarkeit der Befunde dann diskutiert werden muss.
- **Citation-Bias**: Studien mit prägnanten bzw. signifikanten Ergebnissen werden häufiger zitiert. Eine Literatursuche nach dem Schneeballprinzip (▶ Abschn. 6.1.3) vernachlässigt also Studien mit nicht-signifikanten Befunden und sollte deswegen stets durch andere Suchstrategien flankiert werden (z. B. Datenbankrecherche).
- **Availability-Bias** und **Cost-Bias**: Studien werden eher gefunden und in Metaanalysen verarbeitet, wenn sie leicht bzw. kostengünstig zugänglich sind. Das betrifft v. a. Open-Access-Online-Publikationen, seien es entsprechende Fachzeitschriften, die mit Open-Access arbeiten oder auch die Autorinnen und Autoren selbst, die ihre besten Artikel auf ihrer persönlichen Website anbieten (▶ Abschn. 13.4.1). Um die eigene Metaanalyse auf einer weniger verzerrten Stichprobe von Studien basieren zu lassen, sollte man also auch schwer zugängliche und ggf. kostenpflichtige Beiträge aktiv suchen und nutzen.

Das Äpfel-und-Birnen-Problem

Eine grundlegende Kritik an der Metaanalyse lautet, hier würden letztlich „Äpfel und Birnen“ verglichen („apples and oranges problem“; formaler: **Uniformitäts-Problem**, **Inkommensurabilitäts-Problem**): Wenn sich Studien zum gleichen Sachverhalt hinsichtlich ihrer abhängigen und/oder unabhängigen Variablen, ihres Designs (z. B. experimentell vs. nicht-experimentell) oder ihrer Stichprobenzusammensetzungen (Kontrollvariablen) unterscheiden und dennoch in einer Metaanalyse miteinander verrechnet werden, entstünden sinnlose Ergebnisse. Tatsächlich muss bei jeder Metaanalyse definiert werden, welche Kriterien Studien zu erfüllen haben, damit sie **einander inhaltlich und methodisch ähnlich genug sind**, um zur Beantwortung derselben Forschungsfrage zusammengefasst zu werden. Nur exakt identische Studien zu einer Metaanalyse zulassen zu wollen, ist indessen praktisch nicht möglich (reale Studien unterscheiden sich immer) und auch inhaltlich nicht sinnvoll, denn der zusätzliche Informationsnutzen der Metaanalyse entsteht ja gerade dadurch, dass ermittelt wird, welcher Gesamteffekt sich aus leicht variierenden Studien ergibt. Um dem Äpfel-und-Birnen-Problem zu begegnen, müssen also ähnliche bzw. vergleichbare Studien ausgewählt werden (► Abschn. 16.2.3), zudem ist die Ähnlichkeit ihrer Effekte durch statistische **Homogenitätstests** zu prüfen (► Abschn. 16.2.7). Wenn der Äpfel-und-Birnen-Vorwurf genutzt wird, um die Metaanalyse grundsätzlich abzulehnen, erwidern deren Vertreter zuweilen, dass es durchaus sinnvoll sei, Äpfel und Birnen zu vergleichen (anstatt nur Äpfel mit Äpfeln), um auf dieser Basis Aussagen über Obst zu treffen.

Der kompetente Umgang mit dem Inkommensurabilitätsproblem stellt eine der größten Herausforderungen bei der Metanalyse dar und verlangt von den Forschenden sehr viel methodische und inhaltliche Erfahrung in dem jeweiligen Forschungsfeld. Denn es geht neben der sachgerechten Auswahl geeigneter Studien v. a. auch um die **Konstruktion einer einheitlichen Strukturierungs- und Auswertungsschablone** für den gesamten Pool der Primärstudien. Welche Dimensionen der Unterschiedlichkeit und/oder Ähnlichkeit der Studien hierbei anhand welcher Indikatoren zu berücksichtigen sind, damit eine aussagekräftige Metaanalyse resultiert, kann nur vor dem Hintergrund sehr detaillierter Sachkenntnisse über den Forschungsgegenstand und die Modalitäten seiner empirischen Untersuchung entschieden werden.

Das Garbage-in-Garbage-out-Problem

Eine weitere Kritik an der Auswahl von Primärstudien für Metaanalysen ist das Garbage-in-Garbage-out-(GIGO)-Problem: Die Hoffnung, durch eine Metaanalyse zu besonders aussagekräftigen Resultaten über einen Effekt zu gelangen, sei irrig, da sich methodische Schwächen der einzelnen Primärstudien durch deren rechnerische Zusammenfassung nicht ausgleichen, sondern vielmehr vermehren. Wenn fehlerhafte Primärstudien in die Metaanalyse eingehen („garbage in“), dann kommt notgedrungen auch ein fehlerhaftes metaanalytisches Ergebnis heraus („garbage out“). Das Problem ist umso stärker ausgeprägt, je mehr methodische Schwächen die verwendeten Primärstudien aufweisen (Borenstein et al., 2009). Diesem Problem ist dadurch zu begegnen, dass **die methodische Qualität aller Primärstudien** anhand unterschiedlicher Indikatoren genau spezifiziert wird (für Details ► Abschn. 16.2.3). Studien, die Mindeststandards wissenschaftlicher Qualität nicht erfüllen, können dann aus der Metaanalyse ausgeschlossen werden. Zudem lassen sich Studien unterschiedlicher methodischer Qualität bei der Bestimmung des Gesamteffekts unterschiedlich gewichten. Nicht zuletzt sind Indikatoren der methodischen Qualität im Kontext der Moderatorvariablen-Analyse dahingehend zu untersuchen, ob sie mit einem stärkeren oder schwächeren Effekt einhergehen. Zu beachten ist, dass die methodische Qualität einer Studie ein mehrdimensionales Konstrukt darstellt (zu Qualitätskriterien in der Wissenschaft ► Kap. 3) und dass die Beurteilung methodischer Qualität (z. B. Auswahl, Messung und Kodierung von Qualitätsindikatoren) auch auf den jeweiligen Untersuchungsgegenstand abgestimmt werden muss.

Das Problem abhängiger Untersuchungsergebnisse

Geht man davon aus, dass für die Gesamteffektschätzung bei der Metaanalyse die Befunde von unabhängigen Primärstudien zusammengefasst werden und sich somit die einzelnen Stichproben der Primärstudien zur Gesamtstichprobe der Metaanalyse addieren sollen, so kann pro Studie eigentlich nur eine Effektgröße berücksichtigt werden. Jeweils **mehrere Effektgrößen pro Studie** einzubeziehen würde bedeuten, dass abhängige (korrelierte) Ergebnisse eingehen und dass die Stichprobe der jeweiligen Primärstudie mehrfach in der Gesamtstichprobe der Metaanalyse enthalten ist. Diese Problematik abhängiger Untersuchungsergebnisse ist in der Praxis sehr verbreitet. Denn es ist eher die Regel als die Ausnahme, dass pro Primärstudie nicht nur eine Effektgröße, sondern mehrere relevante Effektgrößen in die Metaanalyse aufgenommen werden sollen. Für das Problem abhängiger Effektgrößen gibt es verschiedene Lösungsstrategien – von der Auswahl der wichtigsten Effektgröße pro Studie bis zur Berechnung multivariater Metaanalysen (zum Umgang mit abhängigen Ergebnissen ► Abschn. 16.2.3).

Das Missing-Data-Problem

Im Kern läuft eine Metaanalyse auf eine Zusammenfassung der Effektgrößen einzelner, vergleichbarer Studien zu einer Gesamteffektgrößenschätzung hinaus, ergänzt um eine Analyse relevanter Moderatorvariablen, die die Höhe des Gesamteffekts beeinflussen. Probleme entstehen, wenn in der Ergebnisdarstellung der Primärstudien nur **unvollständige Angaben zu den Effektgrößen** zu finden sind (Missing-Data-Problem; Pigott, 1994). Manche Untersuchungsberichte begnügen sich mit der Nennung des Stichprobenumfanges und des Ausganges der statistischen Signifikanzprüfung im Sinne der Mitteilung, ob ein statistisch signifikantes oder nicht-signifikantes Ergebnis vorliegt. Zunehmend häufiger werden jedoch zumindest exakte Irrtumswahrscheinlichkeiten angegeben, die Statistik-Softwarepakete routinemäßig berechnen.

Eine erste Lösungsmöglichkeit des Missing-Data-Problems (wobei es sich im Kontext der Metaanalyse um fehlende Angaben zur Effektgröße einer Studie handelt, nicht um fehlende Einzelwerte wie bei Primärstudien; zur Behandlung von Missing Data in Primärstudien ▶ Abschn. 11.4.4) besteht darin, Details der betreffenden empirischen Studie, die in der Publikation fehlen, direkt **bei den Autorinnen und Autoren in Erfahrung** zu bringen. Im günstigen Fall lässt sich das Problem dadurch beheben und eine vollständige Datenbasis herstellen.

Lässt sich die unvollständige Datenbasis indessen nicht durch Zusatzinformationen komplettieren, so existieren als Notbehelf mehrere **Ansätze zur Aggregierung der Befunde der Primärstudien ohne gemeinsame Effektgrößenschätzung** (▶ Abschn. 16.3), wobei dann jeweils nur berücksichtigt wird, wie viele Studien hypothesenkonforme oder signifikante Ergebnisse liefern bzw. wie hoch die Irrtumswahrscheinlichkeiten jeweils sind. Diese Aggregationsansätze sind gegenüber der eigentlichen Metaanalyse mit erheblichen Informationsverlusten verbunden und liefern insbesondere keine Resultate zur Größe des interessierenden Gesamteffekts. Sie laufen stattdessen nur auf eine mehr oder minder unsichere Aussage darüber hinaus, ob der Studienpool eher für oder gegen die Existenz des interessierenden Effekts spricht. Sie sind somit mit Vorsicht einzusetzen und zu interpretieren (▶ Abschn. 16.3).

16.2 Arbeitsschritte einer Metaanalyse

Die Durchführung einer Metaanalyse lässt sich in einzelne Schritte unterteilen – von der Formulierung der Forschungshypothese der Metaanalyse bis zur tabellarischen und/oder grafischen Darstellung der Ergebnisse sowie der abschließenden Interpretation. **Neun Arbeitsschritte** werden im Folgenden beschrieben. Dabei zeigen wir zum einen exemplarisch konkrete Berechnungsweisen auf und illustrieren zum anderen das Vorgehen anhand ausgewählter publizierter Metaanalysen.

Wir möchten betonen, dass die Metaanalyse zwar auf eine statistische Effektgrößen-Aggregierung hinausläuft, diese aber den **geringsten Teil der praktischen Forschungsarbeit** ausmacht: Die statistische Auswertung ist computergestützt binnen weniger Stunden durchführbar. Dass eine Metaanalyse oft viele Monate Zeit beansprucht, liegt daran, dass die Hauptarbeit in der sorgfältigen **Recherche und Beschaffung aller einschlägigen Studien** besteht sowie in der **Extraktion der für die Forschungshypothese relevanten Effektgrößen aus den Publikationstexten** (oft sind dazu wegen unklarer oder unvollständiger Angaben mühsame Hintergrundrecherchen, Umrechnungen und Schätzungen sowie Rückfragen bei den Autorinnen und Autoren notwendig). Besonders problematisch ist es, wenn einige der zu integrierenden Studien die berichteten Effekte um unterschiedliche Anzahlen und Arten von Kontrollvariablen bereinigt haben, andere Studien dagegen unbereinigte Effekte berichten. Eine einfache Zusammenfassung dieser Effektgrößen ist nicht zulässig, denn sie würde „statistischen Fruchtsalat“ (Brüderl, 2004) erzeugen, da unterschiedliche Populationseffekte geschätzt werden. Einfache Lösungen für derartige Probleme liegen nicht vor, stattdessen müssen diese Schwierigkeiten in jeder einzelnen Metaanalyse erneut in nachvollziehbarer und transparenter Weise angegangen werden, was im Zweifelsfall auch bedeuten kann, dass Studien aus methodischen Gründen ausgeschlossen werden müssen. Um entsprechenden Problemen bei der Durchführung von Metaanalysen vorzubeugen, wird zunehmend vehementer im Sinne des wissenschaftlichen Gütekriteriums der Präsentationsqualität (▶ Abschn. 3.1.2) eine vereinheitlichte Darstellung von Forschungsergebnissen (inklusive aller relevanten Effektgrößen) in der Fachliteratur gefordert. Auch wird an die Kooperationsbereitschaft der Forschenden appelliert, zu publizierten Primärstudien auf Anfrage die fehlenden Effektgrößenangaben oder auch den Primärdatensatz bereitzustellen (Bosnjak & Viechtbauer, 2009, S. 332).

Abgesehen davon, dass die statistische Verrechnung der Effektgrößen aus den Primärstudien in der Metaanalyse nur ein Arbeitsschritt neben anderen ist, muss auch erwähnt werden, dass unterschiedliche statistische Ansätze oder „Schulen“ innerhalb der Metaanalyse existieren:

- Mit Abstand am weitesten verbreitet in den Human- und Sozialwissenschaften ist die auf Hedges und Olkin (1985) sowie Rosenthal und Rubin (1986) zurückgehende **homogenitätsbasierte Metaanalyse** (HO-Analyse; „homogeneity based meta-analysis“) und deren Weiterentwicklungen. Deswegen stellen wir in diesem Kapitel den HO-Ansatz vor. Er analysiert die

Effektgrößen unter dem Aspekt der Varianz innerhalb der Studien (Fixed-Effect-Modell) und/oder innerhalb und zwischen den Studien (Random-Effects-Modell).

- Darüber hinaus verfolgt die sog. **psychometrische Metaanalyse** („psychometric meta-analysis“) nach Hunter und Schmidt (2004) das Ziel, nicht nur Stichprobenfehler, sondern auch andere Methodenartefakte zu berücksichtigen, also die zu aggregierenden Effektgrößen um Störfaktoren zu bereinigen (z. B. geringe Reliabilität und Validität von Messinstrumenten, künstliche Dichotomisierung oder Beschränkung des Wertebereichs von Variablen etc.). Die psychometrische Metaanalyse kommt in der Psychologie sowie den Wirtschaftswissenschaften regelmäßig zum Einsatz, bislang jedoch selten in der Medizin (Bosnjak & Viechtbauer, 2009, S. 323). Der Name spielt auf die im Rahmen der Psychometrie verwendete klassische Testtheorie (KTT) an, die postuliert, dass sich ein empirischer Testwert aus wahrem Wert plus Fehlerkomponente zusammensetzt (▶ Abschn. 10.4.4). Analog sieht die psychometrische Metaanalyse ihre Aufgabe darin, anhand der empirischen Effektgrößen der Primärstudien die „wahren“ Effekte durch Bereinigung um Fehlerkomponenten zu bestimmen. Auf die Unterschiede und Gemeinsamkeit der beiden homogenitätsbasierten Schulen nach Hedges und Olkin und Rosenthal und Rubin sowie der psychometrischen Schule nach Hunter und Schmidt wird hier nicht näher eingegangen (für einen systematischen Vergleich s. Johnson, Mullen, & Salas, 1995)
- Auch der Ansatz der Bayes-Statistik, die empirische Befunde im Lichte von Vorwissen analysiert (▶ Abschn. 12.2.1), wird auf die Metaanalyse angewendet. Die **Bayes’sche Metaanalyse** („baysian meta-analysis“; Sutton & Abrams, 2001) ist in der Psychologie nicht sehr verbreitet, dafür in der Medizin nahezu dominierend.

Ausführlicher erörtert wird die Methodik der Metaanalyse z. B. bei Borenstein et al. (2009); Ellis (2010); Fricke und Treinies (1985); Glass, McGraw und Smith (1981); Green und Hall (1984); Hedges und Olkin (1985); Hedges, Cooper und Bushman (1992); Hunter und Schmidt (2004); Hunter, Schmidt und Jackson (1982); Lipsey und Wilson (2001); Lösel und Breuer-Kreuzer (1990); Mullen (1989); Mullen und Rosenthal (1985); Rosenthal (1995); Rosenthal und Rubin (1986); Rustenbach (2003); Schulze (2004); Schulze, Holling, Großmann, Jütting, & Brocke (2003); Wachter und Straf (1990) sowie Wolf (1987). Auf ausgewählte einschlägige Monografien zur Vertiefung wird am Ende dieses Kapitels verwiesen. Wer eine deutsch- oder englischsprachige Erläuterung der Metaanalyse in Artikel- bzw. Kapitellänge sucht, sei z. B. auf Bosnjak und Viechtbauer (2009) sowie Beelmann & Lipsey (in press) verwiesen.

16.2.1 Formulierung der Forschungsfrage(n) der Metaanalyse

Am Beginn einer Metaanalyse steht – genau wie bei jeder anderen empirischen Arbeit – die Wahl des Forschungsthemas, die Eingrenzung des Forschungsproblems und schließlich die Formulierung konkreter Forschungsfragen bzw. Forschungshypothesen (▶ Kap. 5). Bei der Formulierung der Forschungsfrage(n) bzw. Hypothese(n) der Metaanalyse ist zu beachten, dass es sich um **empirisch bereits geprüfte Fragestellungen** handeln muss, da man ja definitionsgemäß vorhandene Evidenzen aggregieren möchte. Die Leistung besteht bei der Metaanalyse in dieser Arbeitsphase – im Unterschied zum Vorgehen bei der Primärstudie – also nicht darin, aus einer Theorie neue Hypothesen abzuleiten, sondern darin, vielfach untersuchte, relevante Forschungsfragen oder Forschungshypothesen eines konkreten Untersuchungsfeldes zu identifizieren und zu **explizieren**: Für die Metaanalyse muss im Sinne einer **Konzeptspezifikation** (▶ Abschn. 8.2) genau geklärt werden, über welche theoretischen Konstrukte Aussagen getroffen werden sollen und welche Indikatoren für diese Konstrukte sprechen und welche nicht. Denn diese theoretischen Klärungen sind notwendig, um die Auswahlkriterien zu definieren, nach denen Primärstudien letztlich in die Metaanalyse aufgenommen werden sollen, um die betreffende Fragestellung zu klären.

Beispielhaft seien die Hypothesen von drei aktuellen Metaanalysen aus der Zeitschrift *Psychological Bulletin* vorgestellt:

- Eine Metaanalyse beschäftigt sich mit dem sog. **Zuschauer- bzw. Bystander-Effekt**, d. h. dem sozialpsychologischen Phänomen, dass Menschen, die Zeugen einer kritischen Situation Dritter werden (z. B. Autopanne, Asthmaanfall, Taschendiebstahl), seltener helfend eingreifen, wenn sich weitere Zuschauer vor Ort befinden, als wenn sie die einzigen Augenzeugen sind (Fischer et al., 2011). Dieses relativ breite Forschungsthema wurde zunächst zugespitzt auf das Forschungsproblem, ob sich der Bystander-Effekt in gefährlichen Situationen (z. B. Präsenz eines Angreifers) anders ausprägt als in ungefährlichen (z. B. technische Panne). Auf der Basis verschiedener theoretischer Modelle wurde als Forschungshypothese (H_1) eine **gerichtete Unterschiedshypothese** formuliert, dergemäß die Anwesenheit von Zuschauern in ungefährlichen Situationen die Wahrscheinlichkeit von Hilfeverhalten stärker im Sinne des Bystander-Effekts reduziert als in gefährlichen Situationen (z. B. weil

Augenzeugen in gefährlichen Situationen leichter erkennen, dass ihre Hilfe benötigt wird und weil die Anwesenheit weiterer Augenzeugen die Angst mindert, selbst aktiv zu werden).

- Eine andere Metaanalyse befasst sich mit den **Effekten von gewalthaltigen Videospielen** auf Aggression, Empathie und prosoziales Verhalten (Anderson et al., 2010). Geprüft wird als Forschungshypothese jeweils eine **gerichtete Zusammenhangshypothese**, nach der das Spielen gewalthaltiger Videospiele positiv mit Aggression und negativ mit Empathie sowie negativ mit prosozialem Verhalten korreliert. Zur Explizierung dieser Forschungshypothese gehört eine Konzeptspezifikation, die u. a. exakt angibt, was unter „gewalthaltigen Videospielen" zu verstehen ist und woran man erkennt, dass eine Primärstudie sich mit diesem Gegenstand beschäftigt und nicht etwa mit einem Videospiel ohne Gewaltinhalte. Ebenso wichtig ist die Konzeptspezifikation für alle relevanten Wirkungskonstrukte, in diesem Fall etwa die Aggression: Was genau ist unter diesem Konstrukt zu verstehen, welche Operationalisierungen in unterschiedlichen Primärstudien werden als valide Messungen anerkannt (und gehen somit in die Metaanalyse ein) und welche nicht?
- Eine weitere aktuelle Metaanalyse untersucht **Geschlechtsunterschiede in der Sexualität** (Petersen & Hyde, 2010). Es wird als Forschungshypothese einerseits eine **gerichtete Unterschiedshypothese** postuliert, nach der Männer in einigen Variablen ein aktiveres Sexualverhalten und liberalere sexuelle Einstellungen zeigen als Frauen. Für andere Variablen wird andererseits die **Nullhypothese** als Forschungshypothese unterstellt, d. h. es werden keine oder minimale Geschlechtsunterschiede erwartet. Schließlich wird postuliert, dass Geschlechtsunterschiede in der Sexualität durch eine Reihe von **Moderatorvariablen** beeinflusst werden (z. B. historischer Wandel sowie kulturelle Unterschiede in den Geschlechtsrollen).

Der **Bedarf nach einer Metaanalyse** wird – neben der Relevanz des untersuchten Phänomens – jeweils auch damit begründet, dass seit der letzten einschlägigen Metaanalyse zahlreiche neue Studien erschienen sind, die einer aktuellen Forschungssynthese bedürfen. Zudem zeigt sich in einem Forschungsgebiet mit der Zeit auch oft eine Tendenz zu methodisch und inhaltlich verbesserten Studien (z. B. Berücksichtigung einer größeren Zahl an Moderatorvariablen; höherwertige Stichproben etc.), so dass neuere Metaanalysen die Befunde älterer Metaanalysen revidieren oder differenzieren können. Die zu prüfenden Forschungshypothesen einer Metaanalyse sind auf der Basis des früheren Forschungsstandes (insbesondere früherer Metaanalysen) sowie theoretischer Modelle zu begründen, wobei die Metaanalyse letztlich auf die Theoriemodelle der Primärstudien angewiesen ist, da auf der Basis dieser Theorien die Auswahl und Operationalisierung der Variablen in den Primärstudien fußt.

Zusammen mit der Auswahl und Explizierung der Forschungshypothese(n) ist auch über die **avisierte Aussagekraft** der Metaanalyse zu entscheiden, da sich aus ihr die inhaltlichen und methodischen **Einschluss- bzw. Ausschlusskriterien für die berücksichtigten Primärstudien** ableiten (▶ Abschn. 16.2.3). Wer etwa den Anspruch hat, eine Metaanalyse durchzuführen, deren Befund in einer intern validen Kausalaussage besteht, wird dazu nur solche Primärstudien aggregieren, die ihrerseits hohe interne Validität aufweisen (z. B. experimentelle Kontrollgruppenstudien).

16.2.2 Suche nach inhaltlich einschlägigen Primärstudien

Das Ergebnis einer Metaanalyse ist maßgeblich von der **Menge und Qualität der einbezogenen Primäruntersuchungen** abhängig, die im Kontext einer Metaanalyse die zu aggregierenden Untersuchungseinheiten darstellen. Akribische Bemühungen um eine **möglichst vollständige Erfassung aller thematisch einschlägigen Arbeiten** (Monografien, Buchkapitel, Zeitschriftenartikel, Dissertationen etc.) sind hierbei unerlässlich.

Literaturrecherche

Das übliche Vorgehen der Studienbeschaffung umfasst folgende Strategien: Unverzichtbar ist die Suche über mindestens eine einschlägige **elektronische Literaturdatenbank** (z. B. PsychINFO, PubMed, Web of Science etc.), wobei im Bericht der Arbeit dann sowohl die Datenbanken als auch die verwendeten Suchbegriffe und Suchbegriffskombinationen zu nennen sind (zur Literaturrecherche ▶ Abschn. 6.1). Empfehlenswert ist ergänzend eine Suche in den **Literaturverzeichnissen** der bereits gefundenen Primärstudien (**Schneeballprinzip**).

Unter Beachtung des Problems des **Language-Bias** (▶ Abschn. 16.1.3) sollte man bei kulturübergreifenden Fragestellungen eine Beschränkung auf deutsch- und englischsprachige Quellen vermeiden und für eine globale Betrachtung z. B. auch Literaturdatenbanken aus dem asiatischen Raum einbeziehen oder – sofern die notwendigen Ressourcen für entsprechende Gegenmaßnahmen (z. B. Übersetzungskosten, Mitarbeit von Personen mit entsprechenden Sprachkompetenzen) fehlen – diese Problematik zumindest ausdrücklich diskutieren. Es gibt jedoch auch metaanalytische Fragestellungen, die auf

einen Sprachraum oder ein spezifisches Land konzentriert sind. So fassten Beelmann und Schneider (2003) z. B. ausdrücklich nur den Stand der deutschsprachigen Psychotherapieforschung bei Kindern und Jugendlichen metaanalytisch zusammen und identifizierten insgesamt 47 Kontrollgruppenstudien.

Maßnahmen gegen den Publication-Bias

Der Publication-Bias beschreibt die Tendenz, dass v. a. theoriekonforme Ergebnisse publiziert werden und Studien, die theoriekonträre, kleine bzw. nicht-signifikante Effekte nachweisen, eher unpubliziert in der Schublade liegen bleiben (**Schubladenproblem**). Eine Metaanalyse auf der Basis publizierter Primärstudien läuft deswegen Gefahr, den untersuchten Effekt zu überschätzen. Zu beachten ist, dass die Neigung, v. a. besonders „gute" (d. h. hypothesenkonforme) Ergebnisse zu publizieren, nicht nur dazu führt, dass ganze Studien unveröffentlicht bleiben, sondern auch dazu, dass in einer publizierten Studie die Ergebnisse zuweilen nur ausschnitthaft oder „geschönt" dargestellt werden. Diese Form der verzerrten Darstellung durch **partielle Nichtveröffentlichung von Teilergebnissen** ist ebenfalls ein Problem für Metaanalysen. Es ist in der Regel noch schwerer zu identifizieren und zu kompensieren ist als die Verzerrung durch vollständige Nichtpublikation ganzer Studien.

Um eine Verzerrung durch vollständige Nichtpublikation zu vermeiden, gibt es die Empfehlung, bei der Suche nach Studien auch ganz bewusst **(noch) nicht publizierte Untersuchungen einzubeziehen**. Diese können z. B. über **fachliche Communitys und akademische Netzwerke** recherchiert und beschafft werden. Relativ üblich sind zu diesem Zweck z. B. Anfragen bei ausgewiesenen Fachleuten sowie in Mailinglisten der wissenschaftlichen Fachgesellschaften. Detaillierte Informationen über diesen für Metaanalysen wichtigen Arbeitsschritt findet man u. a. bei White (1994) oder Rustenbach (2003). Das spezielle Problem, an sog. **graue Literatur** („grey literature") heranzukommen (z. B. Masterarbeiten, Institutsberichte), wird bei M. C. Rosenthal (1994) behandelt. Die Aufnahme nichtpublizierter Studien in den Studienpool einer Metaanalyse ist einerseits sinnvoll zur Kompensation des Publication-Bias, andererseits aber auch mit Problemen behaftet. Nichtpublizierte Studien unterliegen keiner formalen Qualitätskontrolle und der Zugriff auf sie erfolgt sehr unsystematisch, so dass hier weitere bzw. noch größere Ergebnisverzerrungen entstehen können. Manche Autoren raten deswegen davon ab, das Einbeziehen einzelner, mehr oder minder willkürlich herausgegriffener unpublizierter Studien als effektive Maßnahme gegen den Publication-Bias zu betrachten und empfehlen die **Beschränkung auf publizierte Arbeiten**, wozu in jedem Fall auch **Dissertationen** zählen (Cook et al., 1993; Smith & Egger, 1998). Disserationen lassen sich u. a. über das internationale Print-Dissertations-Verzeichnis „Dissertation Abstracts International" (DAI), einem Auszug aus der „ProQuest Dissertations & Theses" PQDT Online-Datenbank (http://www.proquest.umi.com/login), recherchieren. DAI ist z. B. in die Psychologie-Literatur-Datenbank der APA PsycInfo (▶ Abschn. 6.1.2) bereits integriert.

Unabhängig davon, ob man für eine Metaanalyse nun einen Pool an ausschließlich publizierten oder an publizierten plus nichtpublizierten Studien zusammengestellt hat – es besteht in jedem Fall die Möglichkeit, den **Grad der Verzerrung des Studienpools durch den Publication-Bias** statistisch abzuschätzen und im Ergebnisbericht der Metaanalyse zu dokumentieren (Ferguson & Brannick, 2012). Drei einschlägige Methoden zur Abschätzung des Publication-Bias – der Funnel-Plot, die Fail-Safe-N-Methode und die Trim-and-Fill-Prozedur – werden im Folgenden beschrieben, wobei zu beachten ist, dass diese Auswertungen es voraussetzen, dass die Effektgrößen aller Primärstudien bereits in ein **einheitliches Effektgrößenmaß** transformiert wurden (▶ Abschn. 16.2.5).

Der Funnel-Plot

Eine populäre Technik zur Abschätzung des Ausmaßes des Publication-Bias bzw. des Schubladenproblems ist der sog. Funnel-Plot bzw. die **Trichtergrafik**. Bei diesem auf Light und Pillemer (1984) zurückgehenden grafischen Verfahren trägt man auf der Abszisse eines Koordinatensystems die Effektgrößenschätzungen der zu aggregierenden Studien ab und auf der Ordinate die studienspezifischen Stichprobenumfänge (bzw. häufiger: die entsprechenden Standardfehler). Es ist nun bekannt, dass die Streuung der Effektgrößenschätzungen um die wahre Effektgröße mit wachsendem Stichprobenumfang (bzw. sinkendem Standardfehler) abnimmt, so dass die grafische Darstellung der Studien die Form eines umgedrehten symmetrischen Trichters („funnel") annimmt (◘ Abb. 16.1).

Bei fehlenden Studien hingegen, die – weil nicht hypothesenkonform bzw. nicht-signifikant – unveröffentlichte Schubladenuntersuchungen geblieben sind, resultiert ein **asymmetrischer Funnel-Plot** (◘ Abb. 16.1b). Im (fiktiven) Beispiel deutet sich ein wahrer Effekt-(Korrelations-)Parameter von $\rho = .30$ an. Untersuchungen mit (hypothesenkonträren) positiven Korrelationen wurden offenbar nicht publiziert. Einen **Signifikanztest der Funnel-Plot-Asymmetrie** haben z. B. Egger, Smith, Schneider und Minder (1997) entwickelt (Eggers Regression-Test), wobei bei kleiner Studienzahl die limitierte Teststärke des Verfahrens zu beachten ist, d. h. ein nicht-signifikantes Ergebnis darf dann nicht ohne weiteres als

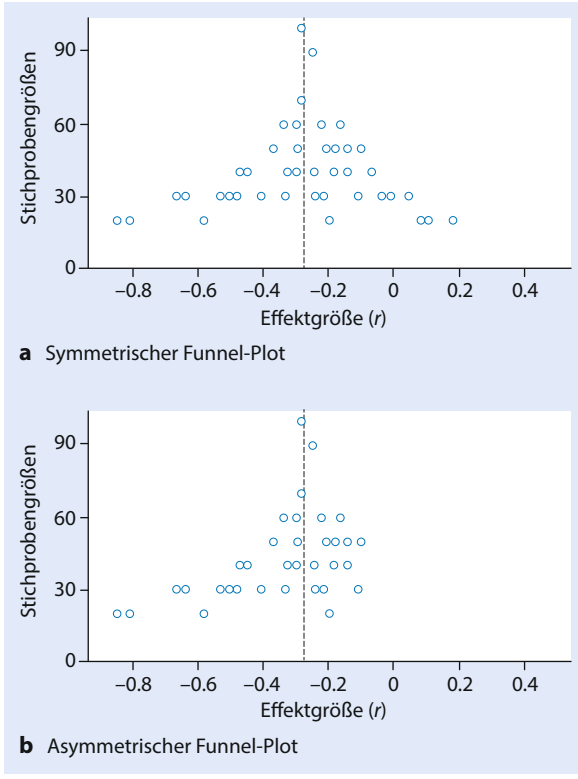


Abb. 16.1 Funnel-Plot zur Prüfung von Auswahlverzerrungen („file drawer bias"). **a** symmetrischer Funnel-Plot, **b** asymmetrischer Funnel-Plot

Beleg für Symmetrie interpretiert werden. Mittels der verfügbaren Metaanalyse-Software (▶ Abschn. 16.5) ist es bequem möglich, sich Funnel-Plots erstellen zu lassen. Ein weiterer grafischer Ansatz zur Identifizierung fehlender Untersuchungen, der **Normal-Quantile-Plot**, wurde von Wang und Bushman (1998) vorgestellt. Zusammenfassende Informationen über grafische Methoden findet man bei Sterne, Egger, & Davey Smith (2001).

Erweist sich der *Funnel-Plot* als weitgehend symmetrisch, ist nicht von einem relevanten Publication-Bias auszugehen. Wird eine nennenswerte Asymmetrie deutlich, so ist zur Abschätzung von deren Einfluss auf das Ergebnis der Metaanalyse ergänzend die Fail-Safe-*N*-Methode oder auch die Trim-and-Fill-Prozedur sinnvoll.

Die Fail-Safe-*N*-Methode („fail-safe number")

Dieses auf Rosenthal (1979) zurückgehende, oft genutzte, aber – wie noch erklärt wird – nicht unproblematische Verfahren schätzt ab, wie viele nicht-signifikante Studien erforderlich wären, um einen signifikanten Gesamteffekt statistisch unbedeutend werden zu lassen. Diese Frage nach der **widerlegungssicheren Untersuchungszahl** lässt sich unter Verwendung der **Stouffer-Methode**, die exakte Irrtumswahrscheinlichkeiten aggregiert (▶ Abschn. 16.3.3), relativ einfach beantworten: Man erhöht k (bzw. N in der Fail-Safe-Terminologie) um so viele Untersuchungen ohne Effekt (mit $z = 0$), bis ein Z_S-Wert resultiert, dem eine Irrtumswahrscheinlichkeit von $P > \alpha$ entspricht.

Angenommen, es wurden die exakten Irrtumswahrscheinlichkeiten von $k = 5$ Studien ($P_1 = .04$; $P_2 = .17$; $P_3 = .18$; $P_4 = .02$ und $P = .25$ bzw. die zugehörigen z-Werte) metaanalytisch gemäß der **Stouffer-Prüfgröße** Z_S (▶ Abschn. 16.3.3) zusammengefasst:

$$Z_S = \frac{-6.25}{\sqrt{5}} = -2.80$$

Diesem Z_S-Wert entspricht eine Wahrscheinlichkeit (Fläche unter der Standardnormalverteilung) von $P(_{Z_S}) = .0023 = 0.2\,\%$. Da 0.2 % unter der Signifikanzgrenze von $\alpha = 5\,\%$ liegt, ist die Nullhypothese zu verwerfen und von einem überzufälligen Gesamteffekt auszugehen.

Wir fragen uns jetzt, wie viele weitere Studien ohne Effekt (d. h. mit $z = 0$) in die Metaanalyse einzubeziehen wären, damit ein nicht-signifikanter Gesamteffekt entsteht. Der Signifikanzgrenze von $\alpha = 5\,\%$ entspricht ein z-Wert von -1.645. Z_S-Werte, die dem Betrag nach unter 1.645 liegen, sind somit nicht signifikant.

$$Z_S = \frac{-6.25}{\sqrt{5 + X}} = -1.645$$

Die Gleichung muss also nach X bzw. der Anzahl der notwendigen Studien mit nicht-signifikantem Ergebnis aufgelöst werden: **Fail-Safe-*N*** bzw. N_{FS}. Allgemein errechnet man das N_{FS} wie folgt:

$$N_{FS} = \frac{\left(\sum_{i=1}^{k} z_i\right)^2 - k \cdot z_\alpha^2}{z_\alpha^2}. \qquad (16.1)$$

In dieser Gleichung steht z_α für denjenigen Wert, der bei einseitigem Test $\alpha\,\%$ der Standardnormalverteilung abschneidet. Für $\alpha = .05$ resultiert $z_\alpha = 1.645$, d. h., wir erhalten nach Gl. 16.1:

$$N_{FS} = \frac{(-6.34)^2 - 5 \cdot (-1.645)^2}{-1.645^2} = 9.85.$$

In unserem Beispiel wären also rund **10 weitere Untersuchungen ohne Effekt** erforderlich, um das Gesamtergebnis nicht-signifikant werden zu lassen. Denn wenn wir $k = 10$ in die Stouffer-Formel einsetzen resultiert ein

Studienbeispiel

Metaanalyse zur mentalen Beeinflussung von Würfelergebnissen

Radin und Ferrari (1991) fassten 148 **parapsychologische Studien** zusammen, die sich mit der mentalen Beeinflussung von Würfelergebnissen beschäftigten. Es zeigte sich ein stabiler Effekt: Nur wenn $N_{FS} = 17\,974$ Studien mit nicht-signifikanten Ergebnissen vorliegen würden, wäre die bislang dokumentierte Evidenz, die für die mentale Beeinflussbarkeit des Würfelns spricht, widerlegt. Da $N_{FS} = 17\,974$ weit über der Daumenregel $N_{FS} \geq 5 \cdot k + 10 = 5 \cdot 148 + 10 = 750$ liegt, kann der metaanalytische Befund in diesem Fall als sehr gesichert gelten. Mit Blick auf das Schubladenproblem hat das *Journal of Parapsychology* bereits im Jahr 1975 in seinen Herausgeberrichtlinien festgelegt, dass auch nicht-signifikante Ergebnisse publiziert werden sollen, um damit einer Überschätzung von Effekten entgegenzuwirken. In welchen Größenordnungen sich das „Fail-Safe-*N*" für unterschiedliche parapsychologische Effekte bewegt, berichtet Utts (1991) in ihrem kritischen Review diverser einschlägiger Metaanalysen.

nicht-signifikanter Befund:

$$Z_S = \frac{-6.34}{\sqrt{5+10}} = -1.637.$$

Generell gilt, dass mit größer werdendem N_{FS} die Stabilität des metaanalytischen Ergebnisses steigt. Für $\boldsymbol{N_{FS} \geq 5 \cdot k + 10}$ ist nach Rosenthal (1993) ein signifikanter metaanalytischer Effekt als stabil anzusehen bzw. eine nennenswerte Verzerrung durch den Publication-Bias auszuschließen. Für unser Beispiel wäre dies $N_{FS} \geq 5 \cdot 5 + 10 = 35$. Daraus ist abzuleiten: Anhand der vorliegenden Evidenz aus 5 Studien mit einem Fail-Safe $N = 10$ ist das Ergebnis **noch nicht abgesichert** genug, um aus metaanalytischer Sicht von einem tatsächlichen Populationsgesamteffekt auszugehen. Die Gefahr ist zu groß, dass den fünf publizierten Studien, die zusammengenommen für den Effekt sprechen, bereits genügend widerlegende unveröffentlichte Studien gegenüberstehen, die jedoch im Zuge der Recherche nach einschlägigen Primärstudien mangels Publikation nicht zugänglich waren (für ein weiteres ▶ Studienbeispiel „Metaanalyse zur mentalen Beeinflussung von Würfelergebnissen").

Die Fail-Safe-*N*-Methode birgt zwei Probleme:

1. Die widerlegungssichere Untersuchungszahl wird für einen **Nulleffekt** bestimmt und somit systematisch zu hoch angesetzt, wenn man davon ausgeht, dass der Nulleffekt ohnehin unwahrscheinlich ist und man sinnvoller gegen **Minimumeffekte** (d. h. theoretisch, praktisch und/oder klinisch bedeutungslos kleine Effekte) testen sollte (zu Minimum-Effektgrößen-Tests ▶ Kap. 15). Ein entsprechendes Verfahren stammt z. B. von Orwin (1983).
2. Die Verwendung der **Stouffer-Prüfgröße** entspricht nicht dem Ziel der Metaanalyse, die Effektgrößen der Primärstudien zu kombinieren, da nur die Irrtumswahrscheinlichkeiten eingehen. Letztere wiederum sind stark durch die Teststärke bzw. den Stichprobenumfang jeder einzelnen Studie beeinflusst. Die Stouffer-Methode gilt als reiner Notbehelf zur Aggregierung von Studien, deren Effektgrößen nicht bekannt sind. Manche Autoren monieren deswegen, dass die Fail-Safe-*N*-Methode nach Rosenthal (d. h. auf der Basis der Stouffer-Prüfgröße) angesichts ihrer Grenzen „übergebührende Beachtung findet" (Bosnjak & Viechtbauer, 2009, S. 328). **Weiterentwicklungen der Fail-Safe-*N*-Methode**, bei denen zur Bestimmung der widerlegungssicheren Untersuchungszahl nicht mehr die Irrtumswahrscheinlichkeiten, sondern die **Effektgrößen** genutzt werden, stammen z. B. von Carson, Schriesheim und Kinicki (1990) sowie von Rosenberg (2005).

Wird mit der *Fail-Safe-N-Methode* eine widerlegungssichere Studienzahl ermittelt, die deutlich höher liegt als die Zahl der identifizierten Studien (k) oder die sogar höher liegt als $N_{FS} \geq 5 \cdot k + 10$, gilt der metaanalytische Befund als inhaltlich gesichert und nicht als Artefakt des Publication-Bias.

▪ Die Trim-and-Fill-Prozedur

Dieses von Duval und Tweedie (2000a, b) entwickelte, auf dem Funnel-Plot basierende **Imputationsverfahren** liefert eine vom Publication-Bias **bereinigte Effektgrößenschätzung**. Dabei werden auf der überrepräsentierten Seite des asymmetrischen Funnel-Plots iterativ so viele kleine Studien mit extremen Effekten aus der Berechnung des Gesamteffekts ausgeschlossen („trimming"), bis ein symmetrischer Funnel-Plot resultiert. Auf der Basis dieses reduzierten Studienpools wird dann eine neue Schätzung der Gesamteffektgröße vorgenommen. Da die Eliminierung von Studien die Varianz reduziert und somit zu verzerrten Konfidenzintervallen des Gesamteffekts führt, werden im zweiten Berechnungsschritt alle eliminierten Studien wieder eingesetzt und dabei jeweils durch einen spiegelsymmetrischen Wert ergänzt („filling"); dadurch verändert sich die Punktschätzung des neuen Gesamteffekts nicht, nur die Varianz wird erhöht. Bezogen auf den Funnel-Plot in ◘ Abb. 16.1b würde dies z. B. bedeuten, dass für Extremstudien wie die *links unten* repräsentierte Studie mit der Effektgröße $r = -.83$ (bei $n \approx 20$) auf der anderen Seite des asymmetrischen Funnel-Plots die „fehlende" Studie mit $r = +.83$ eingesetzt wird. Die Trim-and-Fill-Prozedur führt zu einer korrigierten Gesamteffektschätzung, die mehr oder minder stark von der

ursprünglichen Schätzung abweicht. Eine grafische Veranschaulichung in einem Funnel-Plot, in den die gemäß Trim-and-Fill-Prozedur imputierten „fehlenden" Studien und die neue Gesamteffektschätzung eingetragen werden, ist mit gängiger Metaanalyse-Software (▶ Abschn. 16.5) möglich.

Weichen die einfache Gesamteffektschätzung und die durch die *Trim-and-Fill-Prozedur* korrigierte Schätzung stark voneinander ab, weist dies auf einen relevanten Einfluss des Publication-Bias auf das metaanalytische Ergebnis hin, wobei auch dessen Größenordnung sichtbar wird.

Eine Zusammenfassung, Kritik und Weiterentwicklungen der Methoden zur Kontrolle des Publikation-Bias haben Vevea und Woods (2005) zusammengestellt. Hingewiesen sei ferner auf Rustenbach (2003) und Borenstein et al. (2009) zum Thema Publication-Bias. Interessante Alternativen zur Behandlung des Schubladenproblems haben Darlington und Hayes (2000) entwickelt. Statistische Methoden zur Korrektur des Publikation-Bias finden sich bei Hedges und Vevea (1996) bzw. Vevea und Hedges (1995).

16.2.3 Auswahl und Kodierung der methodisch geeigneten Primärstudien

Der Metaanalyse wird gelegentlich vorgehalten, ihre Ergebnisse seien wenig aussagekräftig, weil unkritisch jede thematisch passende Studie unabhängig von ihrer methodischen Qualität berücksichtigt wird (Garbage-in-Garbage-out-Problem) bzw. weil Studien verwendet werden, deren inhaltliche Kohärenz nicht überzeugt (Äpfel-und-Birnen-Problem) oder weil die berücksichtigten Studienergebnisse nicht unabhängig voneinander sind (Problem abhängiger Untersuchungsergebnisse). Den Vertretern dieser Argumentation (z. B. Eysenck, 1952; Mansfield & Busse, 1977) ist natürlich Recht zu geben, wenn Studien mit offensichtlichen methodischen Mängeln bzw. wenigen inhaltlichen Gemeinsamkeiten metaanalytisch verarbeitet werden. Im Übrigen hat sich jedoch die Auffassung durchgesetzt, dass bei der Auswahl von Studien eher **liberale Kriterien** von Nutzen sind (z. B. Bangert-Drowns, 1986).

Der gesammelte Studienpool wird jedoch in jedem Fall vor der metaanalytischen Auswertung noch einmal hinsichtlich inhaltlicher, v. a. aber methodischer Kriterien gründlich gesichtet. Dabei werden **ungeeignete Studien ausgeschlossen**. So wurden aus dem Studienpool der Metaanalyse zu Geschlechtsunterschieden in der Sexualität (Petersen & Hyde, 2010) u. a. alle Studien ausgeschlossen, die nur auf der Untersuchung einer Geschlechtsgruppe basierten, nur Teilnehmende aus besonderen Populationen beinhalteten (z. B. nur HIV-Positive) oder die nicht die interessierenden abhängigen Variablen erfassten.

Der mittels Literaturrecherche (▶ Abschn. 16.2.2) gebildete Studienpool wird somit im Zuge der Auswahl geeigneter Untersuchungen anhand von **theoretisch begründeten Inklusions- und Exklusionskriterien** noch mehr oder minder stark reduziert. Einschluss- und Ausschlusskriterien dürfen nicht beliebig angewendet werden, sondern müssen bereits mit der Formulierung der Forschungshypothese (▶ Abschn. 16.2.1) konkretisiert und inhaltlich begründet worden sein. So kann z. B. vorab begründet werden, dass die betreffende Metaanalyse auf einen Wirksamkeitsnachweis hoher interner Validität abzielt und somit nur randomisierte Kontrollgruppenstudien berücksichtigt werden (dies ist z. B. bei Cochrane-Reviews üblich).

Generell differieren die Studienanzahlen, auf denen Metaanalysen basieren, je nach Forschungsthema erheblich: Die Metaanalyse zu aggressionsfördernden Effekten von gewalthaltigen Videospielen (Anderson et al., 2010) basierte z. B. auf einem Pool aus 381 Studien, von denen schließlich 221 Studien gemäß Kriterien der methodischen Qualität in die Analyse einbezogen wurden – dies entsprach einem Gesamtstichprobenumfang von rund 62 000 Probanden. Die Metaanalyse zu Geschlechtsunterschieden in der Sexualität (Petersen & Hyde, 2010) war sogar noch umfassender und basierte auf 730 Studien, 834 unabhängigen Stichproben und einem Gesamtstichprobenumfang von rund 1.4 Millionen Untersuchungspersonen. Dagegen war aus der experimentellen Bystander-Effekt-Forschung nur ein Studienpool von 53 Artikeln zu bilden, die entsprechende Metaanalyse basierte auf Daten von knapp 8 000 Versuchspersonen (Fischer et al., 2011). Dass manche Metaanalysen in Cochrane-Reviews auf weniger als 5 Studien und einem Gesamtstichprobenumfang im unteren dreistelligen Bereich basieren, wurde bereits erwähnt.

Alle Studien im endgültigen Studienpool werden hinsichtlich ihrer für die Berechnung der Metaanalyse relevanten Merkmale kodiert (z. B. Stichprobenumfang, Effektgrößenmaß). Dies schließt eine **Kodierung von Moderatorvariablen** ein, bei denen es sich um inhaltliche (z. B. Kulturkreis der Studienteilnehmenden; Durchschnittsalter der Stichprobe; historischer Zeitpunkt der Datenerhebung), aber auch um methodische Merkmale der Studien handeln kann (z. B. Untersuchungsdesign: experimentell, quasi-experimentell, nicht-experimentell; Stichprobe: probabilistisch, nicht-probabilistisch etc.). In der Metaanalyse zur Wirkung von gewalthaltigen Videospielen (Anderson et al., 2010) wurde u. a. das Untersuchungsdesign als Moderatorvariable kodiert mit den

Ausprägungen: experimentelle Studie, nicht-experimentelle Querschnittstudie und Längsschnittstudie.

Maßnahmen gegen das Garbage-in-Garbage-out-Problem

Liberale Selektionskriterien werden gegenüber Vertretern des Garbage-in-Garbage-out-Arguments damit begründet, dass sich der Einfluss der methodischen Qualität (typischerweise sind hiermit Beeinträchtigungen der internen Validität gemeint; Gottfredson, 1978; Wortmann, 1994) auf die Ergebnisse einer Metaanalyse empirisch kontrollieren lässt. Hierzu werden die einzelnen Studien bzgl. relevant erscheinender **Qualitätskriterien anhand von Ratingskalen** bewertet (z. B. Einsatz einer Kontrollgruppe; Größe, Art und Auswahl der Stichprobe; Reliabilität der verwendeten Messinstrumente; Fehlerkontrollen; Genauigkeit der Treatmentimplementierung; Angemessenheit der eingesetzten statistischen Verfahren etc.; Stock, 1994; Stock et al., 1982, für die Entwicklung von Formblättern zur Beschreibung der Studien). Die Qualitätskriterien werden anschließend mit den studienspezifischen Effektgrößen in Beziehung gesetzt. Systematische Zusammenhänge zwischen Studienmerkmalen und Effektgrößen erleichtern eine – eventuell aufgrund eines signifikanten Homogenitätstests erforderliche – Einteilung der Untersuchungen in homogene Subgruppen (► Abschn. 16.2.7 und ► Abschn. 16.2.8). Allerdings weist Hedges (1982) darauf hin, dass sich methodisch gute bzw. schlechte Untersuchungen kaum in ihren Effektgrößen unterscheiden, dass jedoch die Variabilität der Effektgrößen von schlecht kontrollierten Studien (z. B. keine Randomisierung) größer sei als bei methodisch anspruchsvolleren Studien.

Vor die Alternative gestellt, bestimmte Untersuchungen wegen mangelnder methodischer Qualität von einer Metaanalyse auszuschließen oder alle Untersuchungen einzubeziehen, die zumindest einem **minimalen methodischen Standard** genügen, ist dem zweiten Vorgehen der Vorrang zu geben. Wichtig ist hierbei jedoch, dass die Auswahl der Qualitätskriterien – bzw. der Inklusions- und Exklusionskriterien – nachvollziehbar begründet wird und dass die Bewertung der Untersuchungen anhand dieser Kriterien möglichst objektiv erfolgt (mehrere Urteiler mit Nennung der Urteilerübereinstimmung, so dass die Metaanalyse prinzipiell replizierbar ist; Orwin, 1994).

Wichtig ist nach Kraemer, Gardner, Brooks III, & Yesavage, (1998) eine **Kontrolle der Teststärke** der zu aggregierenden Studien. Wenn man Studien mit geringer Teststärke („underpowered“) von der Metaanalyse ausschließt, so sei dies gleichzeitig eine Strategie, Ergebnisverzerrungen durch das sog. „Schubladenproblem“ (File-Drawer-Problem) zu reduzieren. Rossi (1997) erläutert an einem Beispiel, dass die Ergebnisse von thematisch ähnlichen Studien wegen zu geringer Teststärke inkonsistent sein können, so dass Metaanalysen kontraindiziert sind.

In eine Metaanalyse sollten nur Primärstudien eingehen, die *methodischen Mindeststandards* genügen (also insbesondere eine hinreichend hohe interne Validität und ausreichende Teststärke aufweisen). Diese Mindeststandards sollten bereits zusammen mit der Forschungshypothese der Metaanalyse als Inklusions- bzw. Exklusionskriterien definiert werden (Beispiel für eine Inklusionsregel: berücksichtigt werden sollen nur randomisierte Kontrollgruppenstudien mit jeweils einer Mindestgruppengröße von 20).

Maßnahmen gegen das Äpfel-und-Birnen-Problem

Die Diskussion des sog. Äpfel-und-Birnen-Arguments hat drei Aspekte zu berücksichtigen: Die Homogenität der Untersuchungen in Bezug auf 1.) die unabhängigen Variablen, 2.) die Kontrollvariablen und 3.) die abhängigen Variablen.

Bezogen auf die **unabhängige(n) Variable(n)** ist zunächst festzustellen, dass das mit einer Metaanalyse verbundene Forschungsinteresse meist umfassender ist als in der Primärforschung. Metaanalytische Fragestellungen thematisieren z. B. die Wirkung psychotherapeutischer Maßnahmen oder den Einfluss der Lehrkraft auf das Schülerverhalten, wobei Abstraktionen von konkreten Ausprägungen der unabhängigen Variablen (z. B. konkrete Therapieformen oder bestimmte Unterrichtsstile) häufig bewusst angestrebt werden. Hier eine objektivierbare Abgrenzungsstrategie zu entwickeln, ist schwierig und sollte den spezifischen Besonderheiten der inhaltlichen Fragestellung bzw. den Erkenntnisinteressen der einzelnen Forschenden überlassen bleiben. Sehr heterogene Varianten der Operationalisierung einer unabhängigen Variablen können im Nachhinein in getrennt zu analysierende, homogene Untersuchungsgruppen aufgeteilt werden (► Abschn. 16.2.8 sowie das Rechenbeispiel in ► Abschn. 16.4).

Es sollte soweit wie möglich sichergestellt werden, dass die zu integrierenden Primärstudien vergleichbare unabhängige Variablen überprüfen und auch bzgl. der eingesetzten **Kontrollvariablen** einigermaßen vergleichbar sind. Andernfalls – d. h. bei unterschiedlichen Kontrollvariablen – ergibt sich eine studienbedingte Variation der Effektgrößen, womit Metaanalysen erschwert werden. Es können dann entweder nur Effektgrößen aggregiert werden, die um identische Kontrollvariablen bereinigt wurden oder man bemüht sich darum, den Studien jeweils

die unbereinigten Effektgrößen zu entnehmen bzw. diese bei den Autoren zu erfragen.

Für die Auswahl der Operationalisierungsvarianten der **abhängigen Variablen** sollten strenge Kriterien gelten, denn eine Zusammenfassung von Untersuchungsergebnissen ist nur sinnvoll, wenn die verschiedenen abhängigen Variablen Indikatoren eines gemeinsamen inhaltlichen Konstruktes und damit hoch korreliert sind. Würde man z. B. die Beeinflussung kognitiver und sozialer Schülerfähigkeiten durch das Lehrerverhalten metaanalytisch integrieren, könnte ein nichtssagender Gesamteffekt resultieren, obwohl jede dieser Schülerfähigkeiten für sich genommen sehr wohl, und zwar unabhängig voneinander oder gar kompensatorisch, vom Lehrerverhalten abhängt. Auf eine genaue Definition der abhängigen Variablen einschließlich der Ausgrenzung nicht zulässiger Operationalisierungsvarianten durch Exklusionskriterien ist deshalb großen Wert zu legen.

In eine Metaanalyse sollten nur Primärstudien eingehen, die untereinander gut vergleichbare unabhängige, abhängige und Kontrollvariablen untersuchen. Dabei ist eine *Vergleichbarkeit der Operationalisierungs-Varianten* bei den abhängigen Variablen besonders wichtig (z. B. könnte als Einschlusskriterium definiert werden, dass die interessierende abhängige Variable mit einem standardisierten Testverfahren erfasst wurde).

Maßnahmen bei abhängigen Untersuchungsergebnissen

Es ist eher die Regel als die Ausnahme, dass in einer Publikation mehrere Teilergebnisse zu einer Fragestellung dargestellt werden. Solange diese **Teilergebnisse an ein- und derselben Stichprobe** gewonnen wurden, sind sie voneinander abhängig und sollten deshalb nicht als Einzelbefunde metaanalytisch verarbeitet werden. Abhängige Untersuchungsergebnisse liegen auch vor, wenn in mehreren Publikationen unter unterschiedlichen Blickwinkeln auf dieselbe Ursprungsstudie Bezug genommen wird und die Daten jeweils auf derselben Stichprobe basieren. Statistisch abhängige Ergebnisse entstehen z. B. auch in Messwiederholungsdesigns (▶ Abschn. 7.8) oder in (quasi-)experimentellen Studien mit mehr als zwei Gruppen, etwa wenn drei verschiedene Treatmentvarianten jeweils untereinander und mit der Kontrollgruppe zu vergleichen sind.

Generell gilt, dass metaanalytische Aussagen auf einer **Gesamtstichprobe** basieren, die sich additiv aus den Stichprobenumfängen der Einzelstudien zusammensetzt, bei der also keine Teilstichprobe doppelt oder gar mehrfach gezählt wird (zur Begründung dieser Forderung siehe Kraemer, 1983; weitere Gefährdungen der Unabhängigkeitsforderung für Metaanalysen erörtern Landman & Dawes, 1982).

Liegen mehrere abhängige bzw. korrelierte Teilergebnisse einer Primärstudie vor, so gibt es verschiedene Möglichkeiten, diese metaanalytisch einzubeziehen (Beelmann & Bliesener, 1994, S. 219ff.; weitere Informationen zur metaanalytischen Behandlung abhängiger Ergebnisse findet man bei Gleser & Olkin, 1994; Rosenthal & Rubin, 1986 und Tracz, Elmore, & Pohlmann, 1992):

a. Alle verfügbaren und relevanten Teilergebnisse der Primärstudie werden zunächst aggregiert und nur dieses Gesamtergebnis der Primärstudie **(mittlere Studieneffektgröße)** geht mit dem Stichprobenumfang in die Metaanalyse ein. Bestehen die Einzelergebnisse einer Studie z. B. aus bivariaten Korrelationen, so verwendet man das arithmetische Mittel der Korrelationskoeffizienten als Schätzwert für das Gesamtergebnis (Hunter et al., 1982): Soll etwa der Zusammenhang zwischen dem Erziehungsverhalten der Eltern und den schulischen Leistungen ihrer Kinder in einer Primärstudie durch 3 Korrelationen beschrieben werden (der Betreuungsaufwand bei Hausaufgaben korreliert mit der Deutschnote zu $r = .40$, mit der Mathematiknote zu $r = .60$ und der Musiknote zu $r = .30$), ergäbe sich als geschätzter Gesamteffekt der Wert $\bar{r} = (.40 + .60 + .30)/3 = .43$. Dieser Wert wäre also unter Zugrundelegung des einfachen Stichprobenumfanges der Untersuchung metaanalytisch weiterzuverarbeiten (genauer bestimmt man den Durchschnittswert von Korrelationen über Fishers Z-Werte; Bortz & Schuster, 2010, S. 161f.). Basieren die abhängigen Ergebnisse auf verschiedenen statistischen Tests (z. B. t-Werte, F-Wert, χ^2-Werte etc.), so sollten diese in einheitliche standardisierte Effektgrößenmaße – z. B. Korrelationsäquivalente – transformiert werden (▶ Abschn. 14.2), die anschließend zusammengefasst werden können. Neben der einfachen arithmetischen Mittelung der Teilstudienergebnisse kann bei der Zusammenfassung auch die Höhe der Interkorrelationen berücksichtigt werden. Der Nachteil des Arbeitens mit einem mittleren Studieneffekt ist der damit verbundene Informationsverlust hinsichtlich der einzelnen Effektgrößen und ihrer Varianz innerhalb der Studie.

b. Es wird von allen verfügbaren Teilergebnissen nur **das wichtigste Teilergebnis** der Primärstudie genutzt und der Stichprobenumfang der Primärstudie geht einmal in die Metaanalyse ein. Dieses Verfahren erfordert keinerlei Rechenaufwand, führt aber zu noch größerem Informationsverlust als die Zusammenfassung aller Effektgrößen einer Studie. Zudem ist fraglich, ob sich z. B. aus Evaluationsstudien, die verschiedene Outcome-Variablen berichten, überhaupt stringent

ein im Hinblick auf die Forschungsfrage „wichtigstes" Teilergebnis herauslösen lässt.

c. Anstatt alle Teilergebnisse einer Studie zu einem einzigen Studieneffekt zu aggregieren oder nur ein einzelnes Teilergebnis zu berücksichtigen, lassen sich die Einzeleffekte einer Studie auch auf eine kleine Zahl theoretisch wichtiger Konstrukte hin zusammenfassen. Diese **wenigen aggregierten Effektgrößen** pro Studie werden dann wie unabhängige Effektgrößen behandelt. Dieses Vorgehen verlangt plausible theoretische Vorüberlegungen zur Bildung sinnvoller Aggregatwerte, ist aber oft praktikabel. Es verringert das Problem der Abhängigkeit multipler Effektgrößen ohne mit zu großem Informationsverlust einherzugehen.

d. Dem Problem korrelierter Effektgrößen kann auch begegnet werden, indem für alle interessierenden Effektgrößen jeweils **separate univariate Metaanalysen** durchgeführt werden. Da jedoch bestimmte Effektgrößen immer nur in einem Teil des Studienpools vorkommen, sind für die separaten Analysen oft nicht genügend Studien vorhanden.

e. Die korrelierten Effektgrößen lassen sich schließlich auch mittels einer **multivariaten Metaanalyse** auswerten, welche die Interkorrelationen der Effektgrößen einbezieht. Dies erfordert komplexe und voraussetzungsvolle statistische Analysetechniken, die jedoch in jüngster Zeit an Akzeptanz gewinnen (Jackson, Riley, & White, 2011).

In eine Metaanalyse sollten nur *unabhängige Einzelergebnisse* eingehen, das sind Einzelergebnisse aus Primärstudien, die jeweils aus unabhängigen Stichproben stammen. Werden in einer Primärstudie mehrere Teilergebnisse derselben Stichprobe aufgeführt, so geht in die Metaanalyse meist entweder nur das wichtigste Teilergebnis der Primärstudien-Stichprobe ein oder man fasst die relevanten Teilergebnisse der Primärstudien-Stichprobe zu einem Gesamtwert zusammen. Darüber hinaus existieren weitere Verfahren zur Behandlung abhängiger Untersuchungsergebnisse (Zusammenfassung abhängiger Ergebnisse zu übergeordneten Konstrukten; Berechnung mehrerer univariater Metaanalysen; Durchführung einer multivariaten Metaanalyse).

16.2.4 Entscheidung für ein metaanalytisches Modell

Je nachdem, welches metaanalytische Modell aufgrund der Fragestellung gewählt wird, ergibt sich eine andere Berechnung der Gewichtungsfaktoren, mit denen die Effektgrößen der einzelnen Studien in die Schätzung der Gesamteffektgröße eingehen (Borenstein et al., 2009, S. 183; Lipsey & Wilson, 2001, S. 119).

- Das einfachste Modell ist das **Fixed-Effect-Modell**, dessen Wahl gerechtfertigt ist, wenn man davon ausgehen kann, dass alle Studien inhaltlich **denselben Populationseffekt** erfassen und dass die Unterschiedlichkeit der Effektgrößen zwischen den Studien allein auf Stichprobenfehler („sampling error", „within-study variance") zurückgehen. Dementsprechend werden Studien mit geringem Stichprobenumfang (d. h. hohem Stichprobenfehler) im Fixed-Effect-Modell geringer gewichtet. Das Fixed-Effect-Modell prüft die Nullhypothese, dass der **geschätzte Populationseffekt für jede einzelne Studie Null ist**. Inhaltlich sinnvoll anwendbar ist dieses Modell v. a. dann, wenn sich die aggregierten Primärstudien hinsichtlich Operationalisierung und Dosierung der unabhängigen Variablen sehr stark ähneln. Der statistische Beleg für die Anwendbarkeit des Fixed-Effect-Modells ist ein statistischer **Homogenitätstest**, der prüft, ob die Effektgrößen der einzelnen Studien signifikant voneinander abweichen (Heterogenität) oder nicht (Homogenität; ▶ Abschn. 16.2.7). Im Falle von Heterogenität der Studien kann das Fixed-Effect-Modell weiter genutzt werden, sofern es gelingt, durch eine **Moderatorvariablenanalyse** die Unterschiedlichkeit zwischen den Studien aufzuklären. Dann werden die Studien nämlich in **homogene Subgruppen** aufgeteilt und pro Subgruppe wird eine Gesamteffektschätzung nach dem Fixed-Effect-Modell durchgeführt.
- Im Unterschied zum Fixed-Effect-Modell geht das **Random-Effects-Modell** davon aus, dass die Studien **unterschiedliche Populationseffekte** messen. Die Unterschiedlichkeit der Effektgrößen zwischen den Studien wird im Random-Effects-Modell also nicht nur auf Stichprobenfehler, sondern auch auf systematische Unterschiede zwischen den Studien zurückgeführt, ohne dass diese inhaltlich genau erklärt werden müssen. Dementsprechend werden Studien mit geringerem Stichprobenumfang weiterhin geringer gewichtet, allerdings wird eine Studie gleichzeitig höher gewichtet, wenn ihre Effektgröße stark von den anderen Effektgrößen abweicht und sie somit eine neue Information in die Metaanalyse einbringt. Die Studiengewichte im Random-Effects-Modell sind ausgeglichener als im Fixed-Effect-Modell. Das Random-Effects-Modell prüft die Nullhypothese, dass der **Mittelwert der Effektgrößenschätzungen der einzelnen Studien Null** ist (Borenstein et al., 2009, S. 83). Es ist anwendbar, wenn von inhaltlich **heterogenen Studien** auszugehen ist, jedoch die Ursachen der Un-

terschiedlichlichkeit nicht im Einzelnen durch eine Moderatorvariablenanalyse aufgeklärt werden sollen oder können. Es ist zu beachten, dass das Random-Effects-Modell zu breiteren Konfidenzintervallen führt als das Fixed-Effect-Modell und damit seltener zu statistisch signifikanten Ergebnissen.

- Das **Mixed-Effects-Modell** ist indiziert, wenn die Studien inhaltlich-konzeptuell (z. B. unterschiedliche Operationalisierungen der unabhängigen Variablen) sowie statistisch (d. h. signifikantes Ergebnis des Homogenitätstests) als heterogen einzustufen sind und wenn zumindest ein Teil der Unterschiedlichkeit der Ergebnisse der einzelnen Studien durch Moderatorvariablen aufgeklärt werden kann, während ein anderer Teil der Unterschiedlichkeit wie im Random-Effects-Modell als Zufallsfehler berücksichtigt wird (Borenstein et al., 2009, S. 183f.; Lipsey & Wilson, 2001, S. 124ff.).

In der bisherigen Forschungspraxis dominieren in der Psychologie Metaanalysen gemäß Fixed-Effect-Modell, wobei Random-Effects-Modelle an Popularität gewinnen und sowohl konzeptuell als auch statistisch oft als angemessener eingestuft werden (Cafri, Kromrey, & Brannick, 2010). In welchen Situationen eine Moderatorvariablenanalyse im Fixed-Effect-Modell oder im Mixed-Effects-Modell zuverlässigere Gesamteffektschätzungen liefert, wird teilweise noch kontrovers diskutiert (Lipsey & Wilson, 2001, S. 125). Die bereits zitierte Metaanalyse zu aggressionssteigernden Effekten gewalthaltiger Videospiele nutzte ein Fixed-Effect-Modell mit Moderatorvariablenanalyse (Anderson et al., 2010), die Metaanalyse zu Geschlechtsunterschieden in der Sexualität griff dagegen auf ein Mixed-Effects-Modell zurück (Petersen & Hyde, 2010). In der schon angesprochenen Metaanalyse zum Bystander-Effekt wurden wiederum sowohl die Befunde des Fixed-Effect- als auch des Random-Effects-Modells berechnet und vergleichend berichtet (Fischer et al., 2011).

16.2.5 Berechnung einheitlicher Effektgrößenmaße für alle Primärstudien

Um die in empirischen Studien ermittelten Effekte zu quantifizieren und nicht nur hinsichtlich ihrer statistischen Signifikanz, sondern auch ihrer praktischen Bedeutsamkeit vergleichend einordnen zu können, liegen unterschiedliche standardisierte **Effektgrößenmaße** vor (▶ Abschn. 14.2). Eine Zusammenstellung und Kommentierung der gebräuchlichsten Effektgrößen findet man bei Kirk (1996) oder auch bei Olejnik und Algina (2000). Hinweise zu Berechnung und Berechnungs-Tools sind außerdem auf der Website der Campbell Collaboration zu finden (z. B. http://www.campbellcollaboration.org/resources/effect_size_input.php).

Wir unterscheiden bei den Effektgrößen grob zwischen **Gruppendifferenzmaßen** (z. B. Cohens d, auch: d-Familie), **Zusammenhangsmaßen** (z. B. Korrelationskoeffizient r, auch: r-Familie), **Varianzaufklärungsmaßen** (z. B. Determinationskoeffizient r^2 oder das aus der Varianzanalyse stammende η^2-Maß) sowie **Risikomaßen** (z. B. der Odds Ratio [OR] oder das relative Risiko [RR]). Die Wahl des Effektgrößenmaßes einer Primärstudie hängt in der Regel v. a. vom Untersuchungsdesign sowie vom Skalenniveau der Variablen ab.

Üblicherweise berichten Primärstudien die jeweiligen Effektgrößen mit **unterschiedlichen Effektgrößenmaßen**. Zuweilen wird auch gar kein standardisiertes Effektgrößenmaß berichtet, sondern dieses muss zunächst aus den vorhandenen Angaben rekonstruiert oder bei den Autorinnen und Autoren erfragt werden. Zudem besteht nicht selten das Problem, dass im Studienpool einer Metaanalyse teils unbereinigte, teils um unterschiedliche Kontrollvariablen bereinigte Effektgrößen berichtet werden. Hier muss dann zunächst vor dem Hintergrund der Forschungshypothese entschieden werden, welchen inhaltlichen Effekt (unbereinigt oder in spezifischer Weise bereinigt) die Metaanalyse aggregieren soll. Dann sind für möglichst alle Studien im Studienpool die entsprechenden Effektgrößenmaße herauszufinden (anhand der Angaben in der Publikation bzw. durch Rückfragen bei den Autoren).

Liegen die inhaltlich interessierenden, vergleichbaren Effektgrößenmaße aller zusammenzufassenden Studien vor, so müssen diese statistisch „auf einen Nenner" gebracht werden, indem man sie alle in ein **einheitliches Effektgrößenmaß** transformiert. Dies ist mit **Umrechnungsregeln** relativ problemlos möglich (Tab. 14.3). In der Praxis erfolgt die Vereinheitlichung der Effektgrößenmaße ohnehin softwaregestützt mit entsprechenden Metaanalyse-Programmen (▶ Abschn. 16.5).

Wenn die Forschungshypothese der Metaanalyse z. B. auf eine Gruppendifferenz zwischen 2 Gruppen abzielt, werden die einzelnen Effektgrößen in der Regel auch auf ein einheitliches **standardisiertes Gruppendifferenzmaß** transformiert. Die bereits angesprochene Metaanalyse zu Geschlechterdifferenzen in der Sexualität nutzte Cohens d (Petersen & Hyde, 2010), die Metaanalyse zum Bystander-Effekt in gefährlichen und ungefährlichen Situationen (Fischer et al., 2011) dagegen Hedges g (▶ Kap. 14). Gelegentlich werden auch **unstandardisierte Gruppendifferenzen** (d. h. einfache Mittelwertdifferenzen) als Effektgrößen metaanalytisch aggregiert. Das unstandardisierte Effektgrößenmaß ist zu präferieren,

wenn die abhängigen Variablen in einem Forschungsgebiet keine Intervallskalen mit arbiträrem Ursprung sind, sondern Verhältnisskalen mit natürlichem Nullpunkt wie z. B. Körpergewicht, Währungseinheiten, Blutdruck oder Zeit, so dass die unstandardisierten Differenzen direkt aussagekräftig sind (ausführlicher hierzu Bond, Wiitala, & Dan Richard, 2003).

Bei metaanalytischen Hypothesen über mehr als 2 Gruppen und bei varianzanalytischer Auswertung wird man für die metaanalytische Aggregierung ein **Varianzaufklärungsmaß** als einheitliches Effektgrößenmaß wählen. Zielt die Forschungshypothese der Metaanalyse auf die Verbreitung oder Veränderung eines Risikos, so sind **Risiko-Effektgrößenmaße** auch für die Metaanalyse einschlägig (▶ Abschn. 14.2).

Häufig sollen mit Metaanalysen Zusammenhangshypothesen getestet werden, dann sind wiederum **Zusammenhangs-Effektgrößenmaße** indiziert wie Korrelationskoeffizienten bzw. Korrelationsäquivalente.

Δ-Maß – Das Δ-Maß ist ein universelles Effektgrößenmaß, das als **Korrelationsäquivalent** der bivariaten Produkt-Moment-Korrelation r entspricht. Es dient dazu, unterschiedliche standardisierte Effektgrößenmaße vergleichbar und aggregierbar zu machen. Praktisch jede testspezifische Effektgröße lässt sich in einen Δ-Wert transformieren.

Die folgenden Transformationsregeln gehen auf Kraemer (1985) bzw. Kraemer und Thiemann (1987) zurück. Man beachte, dass die nach diesen Regeln errechneten Δ-**Werte** die „wahren", über r ermittelten Merkmalszusammenhänge nur approximativ schätzen, wobei jedoch die Schätzgenauigkeit für metaanalytische Zwecke ausreichend ist. Leider stimmen die von Kraemer (1985) genannten Effektgrößen nur teilweise mit den in ▶ Abschn. 14.2.2 genannten, auf Cohen (1988) zurückgehenden Effektgrößen überein. Eine Umrechnung in die Kraemer'sche Terminologie ist jedoch in den meisten Fällen problemlos. Für weitere Formeln zur Berechnung und wechselseitigen Umrechnung von Effektgrößenmaßen für Metaanalysen sei auf die einschlägigen Handbücher verwiesen (Borenstein et al., 2009; Lipsey & Wilson, 2001).

Wir bevorzugen im Folgenden für die didaktische Darstellung das Δ-Maß, weil nicht nur Teststatistiken der „r-Familie" (Produkt-Moment-Korrelation, punktbiseriale Korrelation, Phi-Koeffizient, Spearmans rho) als Δ-Äquivalente dargestellt werden können, sondern auch andere Teststatistiken wie t, F oder χ^2, und weil die **Interpretation von Korrelationen** vielen Forschenden und Adressaten von Forschungsberichten geläufiger ist als Interpretationen von δ-Maßen (R. Rosenthal, 1994, S. 234ff.). Speziell in der Interventionsforschung (Vergleich von Behandlungs- und Kontrollgruppe) sind Effektgrößen der d-Familie allerdings verbreiteter als Zusammenhangsmaße.

Die folgenden Transformationsregeln werden später anhand einerls didaktisch vereinfachten Beispiel-Metaanalyse numerisch erläutert (▶ Abschn. 16.4).

Produkt-Moment-Korrelation

Die Produkt-Moment-Korrelation r als Schätzer eines Populationszusammenhanges ρ (griech: rho) entspricht direkt dem Δ-Maß:

$$\Delta = r. \tag{16.2}$$

t-Test für unabhängige Stichproben

Sind die zu vergleichenden Stichproben gleich groß ($n_1 = n_2$), ermitteln wir

$$r_{\text{pbis}} = \frac{\hat{\delta}}{\sqrt{\hat{\delta}^2 + 4}}. \tag{16.3}$$

Die Bestimmungsgleichung für $\hat{\delta}$ findet man in Tab. 14.5 unter Ziffer 1. Die in Gl. 16.3 genannte Transformation ist bei Gilpin (1993) tabelliert.

Bei ungleich großen Stichproben errechnet man nach Kraemer und Thiemann (1987):

$$r_{\text{pbis}} = \frac{\hat{\delta}}{\sqrt{\frac{\hat{\delta}^2 + 1}{(p \cdot q)}}}, \tag{16.4}$$

mit $p = n_1/n$ und $q = n_2/n \cdot (n_1 + n_2 = n)$.

Um eine bessere Vergleichbarkeit mit der Produkt-Moment-Korrelation herzustellen, empfehlen Glass et al. (1981, S. 149), die punktbiseriale Korrelation für metaanalytische Zwecke in eine biseriale Korrelation zu überführen:

$$\Delta = r_{\text{bis}} = r_{\text{pbis}} \cdot \frac{\sqrt{n_1 \cdot n_2}}{\upsilon \cdot n}. \tag{16.5}$$

υ (griech. ypsilon) ist hierbei die Ordinate (Dichte) desjenigen z-Wertes der Standardnormalverteilung, der die Grenze zwischen den Teilflächen $p = n_1/n$ und $q = n_2/n$ markiert (Tab. 1 im ▶ Webangebot zu diesem Buch). Für $0.2 < p < 0.8$ kann Gl. 16.5 nach Magnusson (1966) wie folgt approximiert werden:

$$\Delta = r_{\text{bis}} = 1.25 \cdot r_{\text{pbis}}. \tag{16.6}$$

Falls die Ergebnisdarstellung in einer Untersuchung nicht genügend Angaben enthält, um die für Gl. 16.6 benötigte Effektgröße berechnen zu können, stattdessen aber ein t-Wert genannt wird, errechnet man r_{pbis} nach Cohen (1988, S. 545) wie folgt:

$$r_{\text{pbis}} = \sqrt{\frac{t^2}{t^2 + \text{df}}}, \tag{16.7}$$

mit $\text{df} = n_1 + n_2 - 2$.

Die Transformation der als klein, mittel oder groß klassifizierten δ-Werte (0.20, 0.50, 0.80; Tab. 14.5) in Δ-Werte nach Gl. 16.3 bis Gl 16.6 führt zu Δ-Werten, die mit den als klein, mittel oder groß klassifizierten Korrelationseffekten (.10, .30, .50) nur ungefähr übereinstimmen. Wir erhalten (für gleich große Stichproben)

$$\Delta_{\text{klein}} = 1.25 \cdot \frac{0.2}{\sqrt{0.2^2 + 4}} = 0.12,$$

$$\Delta_{\text{mittel}} = 1.25 \cdot \frac{0.5}{\sqrt{0.5^2 + 4}} = 0.30,$$

$$\Delta_{\text{groß}} = 1.25 \cdot \frac{0.8}{\sqrt{0.8^2 + 4}} = 0.46.$$

Die Kompatibilität der **Effektgrößenklassifikation** ist für mittlere Effekte nahezu perfekt (Abweichungen treten erst in der dritten Nachkommastelle auf) und für kleine bzw. große Effekte für praktische Zwecke akzeptabel (ausführlicher hierzu Cohen, 1988, Abschn. 3.2.2 oder auch die Anmerkung Nr. 781 bei Westermann, 2000, S. 366)

U-Test von Mann-Whitney

Das verteilungsfreie Pendant zum *t*-Test für unabhängige Stichproben ist der *U*-Test von Mann-Whitney (Bortz & Lienert, 2008). Ein *U*-Wert lässt sich nach Rustenbach (2003, S. 100) wie folgt in ein (approximatives) Korrelationsäquivalent transformieren:

$$\Delta = r \approx \frac{1 - 2 \cdot U}{n_1 \cdot n_2}. \tag{16.8}$$

t-Test für abhängige Stichproben

Das Ergebnis eines *t*-Tests für abhängige Stichproben wird nach Kraemer (1985, S. 183) folgendermaßen in ein Korrelationsäquivalent transformiert:

$$\Delta = \frac{\hat{\delta}'}{\sqrt{\hat{\delta}'^2 + 1}}$$

$$\text{mit } \hat{\delta}' = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma} \cdot \sqrt{2 \cdot (1 - r)}}. \tag{16.9}$$

Die Parameter μ_1 und μ_2 werden hier durch $\bar{x}_1$ und $\bar{x}_2$ geschätzt. In Gl. 16.9 ist vorausgesetzt, dass die Korrelation r zwischen den beiden Messwertreihen sowie die geschätzte Populationsstreuung $\hat{\sigma}$ des untersuchten Merkmals bekannt sind. Bei bekannten Streuungen der beiden Messwertreihen ($\hat{\sigma}_{x_1}$ und $\hat{\sigma}_{x_2}$) lässt sich $\hat{\sigma}$ nach Gl. 14.3 schätzen.

Abweichung eines Anteilwertes *P* von π_0 (Binomialtest)

Es wird geprüft, wie stark ein Anteilswert P (bzw. ein Prozentwert) als Schätzer für π von einem bekannten Populationswert π_0 abweicht. Eine korrelationsäquivalente Effektgröße Δ kann hier wie folgt berechnet werden:

$$\Delta = \frac{e^{2d} - 1}{e^{2d} + 1}, \tag{16.10}$$

$$\text{mit } d = 2 \cdot (\arcsin\ \pi^{1/2} - \arcsin\ \pi_0^{1/2}). \tag{16.11}$$

Eine Tabelle für die Arkussinustransformation ($\phi = 2 \arcsin \sqrt{\pi}$) findet man in Tab. 10 im ▶ Webangebot zu diesem Buch.

Mit diesem Ansatz lässt sich auch eine Effektgröße für den **McNemar-Test** bestimmen (das Verfahren wird bei Bortz & Schuster, 2010, S. 146f. erklärt) Die Nullhypothese behauptet, dass der Anteil der Personen mit Änderung von „+“ nach „−“ gleich dem Anteil der Veränderung von „−“ nach „+“ ist ($\pi_0 = 0.50$). Der in Gl. 16.11 genannte π-Wert entspricht dann dem empirischen Anteil der Veränderer von „+“ nach „−“ (bzw. von „−“ nach „+“) an der Gesamtzahl aller Veränderer.

Eine weitere Anwendungsvariante von Gl. 16.10 stellt der **Vorzeichentest** dar (Bortz & Lienert 2008). Bei der Bildung von Differenzen der Messungen zweier abhängiger Messwertreihen werden gemäß H_0 zu gleichen Teilen positive und negative Vorzeichen erwartet ($\pi_0 = 0.50$). Der Parameter π wird hier durch den Anteil positiver (bzw. negativer) Vorzeichen geschätzt.

Vergleich von Anteilswerten aus zwei unabhängigen Stichproben (Vierfeldertafel)

Man ermittelt für 2 Stichproben 1 und 2 jeweils den Anteil einer Merkmalsalternative als Schätzwerte für π_1 und π_2 sowie $p = n_1/N$ und $q = n_2/N$ mit $N = n_1 + n_2$. Hiervon ausgehend wird das Korrelationsäquivalent wie folgt berechnet:

$$\Delta = \frac{e^{2d} - 1}{e^{2d} + 1}, \tag{16.12}$$

$$\text{mit } d = 2 \cdot (p \cdot q)^{1/2} \cdot (\arcsin \pi_1^{1/2} - \arcsin \pi_2^{1/2}). \tag{16.13}$$

Δ entspricht dem **ϕ-Koeffizienten** für eine Vierfeldertafel mit den Häufigkeiten für 2 Merkmalsalternativen in den verglichenen Stichproben (zum ϕ-Koeffizienten Bortz & Schuster, 2010, S. 174ff.). Handelt es sich bei den Merkmalsalternativen um künstliche Dichotomien zweier normalverteilter Merkmale, ist die **tetrachorische Korrelation** als Effektgröße vorzuziehen (Bortz & Schuster, 2010, S. 176). Eine vergleichende Zusammenstellung der gebräuchlichsten Effektgrößen für Vierfeldertafeln (Risk Ratio, Odds Ratio, biserialer ϕ-Koeffizient etc.) findet man bei Sánchez-Meca, Marin-Martinez und Chacón-Moscoso (2003).

Tabelle 16.2 Reduzierte *rxc*-Tafel

		Merkmal b	
		b_j	Nicht b_j
Merkmal a	a_i		
	Nicht a_i		

rxc-Kontingenztafel

In der Praxis kommt es nur selten vor, dass in verschiedenen Untersuchungen Kontingenztafeln mit identischen Merkmalskategorien analysiert werden, die sich metaanalytisch zusammenfassen lassen. Gelegentlich sind Kontingenztafeln jedoch teilidentisch, weil eine der r Kategorien des Merkmals a und eine der c Kategorien des Merkmals b in den zu integrierenden Untersuchungen vorkommen. Sind dies die Kategorien a_i und b_j, lässt sich eine reduzierte Kontingenztafel nach Art von Tab. 16.2 erstellen.

Das Korrelationsäquivalent für diese Vierfeldertafel errechnet man nach Gl. 16.12.

Ergibt es im Kontext einer Metaanalyse Sinn, den Effekt auf die gesamte Kontingenztafel zu beziehen, lässt sich der χ^2-Wert einer rx_2-Tafel wie folgt in eine **(multiple) Korrelation *r*** überführen (z. B. Bortz & Schuster, 2010, Abschn. 22.2.11):

$$R = \sqrt{\frac{\chi^2}{n}}. \tag{16.14}$$

Im allgemeinen Fall einer *rxc*-Tafel kann ein Korrelationsäquivalent zum χ^2-Wert über die „set correlation" bestimmt werden (Bortz & Schuster, 2010, S. 510f.; zur Ermittlung der hierfür erforderlichen kanonischen Korrelation s. Bortz & Schuster, 2010, Kap. 28). Effektgrößen, Poweranalyse und optimale Stichprobenumfänge im Zusammenhang mit der „set correlation" behandelt Cohen (1988, Kap. 10).

Varianzanalyse

Hat ein varianzanalytischer Effekt nur einen Zählerfreiheitsgrad, kann der entsprechende F-Wert über

$$t = \sqrt{F_{(1,df)}} \tag{16.15}$$

in einen t-Wert überführt werden, der wiederum über Gl. 16.7 in eine punktbiseriale und weiter über Gl. 16.5 bzw. Gl. 16.6 in eine biseriale Korrelation zu transformieren wäre. Bei varianzanalytischen Effekten mit mehr als einem Freiheitsgrad ist es für metaanalytische Zwecke oft ausreichend, wenn nur die zur metaanalytischen Fragestellung passenden **Einzelvergleiche** bzw. Kontraste (mit jeweils einem Freiheitsgrad) ausgewertet werden Hall, Tickle-Degnerz, Rosenthal, & Mosteller, 1994, Kap. 3; zur Überprüfung von Einzelvergleichen im Kontext der einfaktoriellen Varianzanalyse Bortz & Schuster, 2010, Abschn. 13.1).

Die auf mehr als 2 Gruppen bezogenen Ergebnisse **einfaktorieller Varianzanalysen** können über das Korrelationsäquivalent η zusammengefasst werden:

$$\hat{\eta} = \sqrt{\frac{QS_{treat}}{QS_{tot}}} \tag{16.16}$$

(zur Beziehung von η und der Effektgröße E Gl. 14.34).

$\hat{\eta}^2$ entspricht dem Varianzanteil der abhängigen Variablen, der durch die unabhängige Variable erklärt wird (zur Terminologie und Durchführung von Varianzanalysen s. Bortz & Schuster, 2010).

Falls die für Gl. 16.16 erforderlichen Quadratsummen nicht genannt werden, kann $\hat{\eta}$ auch über den F-Wert der einfaktoriellen Varianzanalyse mit den dazugehörenden Freiheitsgraden bestimmt werden:

$$\hat{\eta} = \sqrt{\frac{df_{treat} \cdot F}{df_{treat} \cdot F + df_{Fehler}}}. \tag{16.17}$$

Interessiert in einer **mehrfaktoriellen Varianzanalyse** der mit einem Faktor j verbundene Effekt, bestimmt man nach Glass et al. (1981, S. 150):

$$\hat{\eta}_p = \sqrt{\frac{QS_j}{QS_j + QS_{Fehler}}}, \tag{16.18}$$

bzw.

$$\hat{\eta}_p = \sqrt{\frac{df_j \cdot F_j}{df_j \cdot F_j + df_{Fehler}}}. \tag{16.19}$$

In zweifaktoriellen Plänen mit p Stufen für Faktor A und q Stufen für Faktor B gelten $df_A = p - 1$, $df_B = q - 1$ und $df_{Fehler} = p \cdot q \cdot (n - 1)$.

Oft benötigt man auch bei mehrfaktoriellen Varianzanalysen lediglich einen auf Faktor j bezogenen **Einzelvergleich** $D(j)$. Für die in einem Einzelvergleich (z. B. $\bar{A}_1$ vs. $\bar{A}_2$) gebundene Quadratsumme errechnet man im Kontext einer zweifaktoriellen Varianzanalyse (Bortz & Schuster, 2010, Abschn. 15.5.1 bzw. in etwas anderer Darstellung Bortz, 2005, Abschn. 8.2):

$$QS_{D(A)} = \frac{n \cdot q \cdot (\bar{A}_1 - \bar{A}_2)^2}{2}, \tag{16.20}$$

mit n = Stichprobenumfang pro Faktorstufenkombination und q = Anzahl der Stufen des Faktors B.

Hieraus bestimmt man in Analogie zu Gl. 16.18 bzw. Gl. 16.19 folgenden, durch einen Einzelvergleich erklärten Varianzanteil $\hat{\eta}_p^2$:

$$\hat{\eta}_p^2 = \frac{QS_{D(A)}}{QS_{D(A)} + QS_{Fehler}} \tag{16.21}$$

oder

$$\hat{\eta}_p^2 = \frac{F_D}{F_D + df_{Fehler}}. \tag{16.22}$$

Der Wert für $\hat{\eta}_p$ entspricht dem Korrelationsäquivalent Δ.

Die in den Gl. 16.18, Gl. 16.19, Gl. 16.21 und Gl. 16.22 definierten $\hat{\eta}$-Koeffizienten wurden in ▶ Abschn. 14.2.3 als partielles $\hat{\eta}$ problematisiert Gl. 14.49. Im Zentrum der Überlegungen stand die Frage nach der Vergleichbarkeit von Effekten aus einfaktoriellen und mehrfaktoriellen Varianzanalysen. Werden in einer mehrfaktoriellen Varianzanalyse neben einem Treatment auch organismische Variablen wie Alter, Geschlecht, Schulbildung etc. kontrolliert, schätzt die Fehlervarianz (Varianz innerhalb der Zellen) nicht mehr die Merkmalsvarianz, sondern eine um den Beitrag der organismischen Variablen reduzierte Merkmalsvarianz.

Dies wäre bei der einfaktoriellen Varianzanalyse mit einem Treatmentfaktor (z. B. Behandlung vs. Kontrolle) anders, denn hier kann man davon ausgehen, dass – Varianzhomogenität vorausgesetzt – die Varianz innerhalb der Gruppen der Merkmalsvarianz entspricht. Für metaanalytische Zwecke ist es sinnvoll oder erforderlich, die Varianzaufklärung unabhängig vom Studiendesign auf die Merkmalsvarianz zu beziehen, so dass die $\hat{\eta}$-Werte untereinander vergleichbar sind. Ausführlicher mit dieser Thematik befasst sich eine Arbeit von Gillett (2003). Olejnik und Algina (2000) schlagen eine Modifikation von η^2 vor, die diesen Sachverhalt berücksichtigt.

Ein besonderes Problem ergibt sich, wenn man Resultate aus Studien mit unabhängigen und abhängigen Stichproben (z. B. Messwiederholung) „auf einen Nenner" bringen will. Hierzu haben – allerdings auf der Basis des δ-Maßes – Morris und De Shon (2002) eine Reihe wichtiger Vorschläge formuliert. Weitere Hinweise, wie aus unvollständigen varianzanalytischen Ergebnisdarstellungen die für metaanalytische Zwecke benötigten Effektgrößen zu bestimmen sind, findet man bei Seifert (1991).

Spearmans rho (r_s)

Für die Rangkorrelation r_s (Bortz & Lienert 2008, Abschn. 5.2.1) erhält man nach folgender Beziehung eine zur Produkt-Moment-Korrelation äquivalente Effektgröße (Hager, 2004, S. 415):

$$\Delta = \frac{3}{\pi} \cdot 2 \arcsin\left(\frac{r_s^2}{4}\right). \tag{16.23}$$

Kendalls tau (τ)

Für diese Rangkorrelation (Bortz & Lienert 2008, Abschn. 5.2.5) gilt die in Gl. 16.24 genannte Transformation (Hager, 2004, S. 417):

$$\Delta = \frac{1}{\pi} \cdot 2 \arcsin(\tau^2). \tag{16.24}$$

Eine Transformationstabelle für r_s, τ und r findet man bei Gilpin (1993).

16.2.6 Berechnung und Signifikanzprüfung des Gesamteffektes

Liegen für alle Primärstudien die interessierenden Effektgrößenmaße in vereinheitlichter Form vor (z. B. als Gruppendifferenzmaß d_i oder als Zusammenhangsmaß Δ_i), so gilt es, sie im nächsten Schritt zu einem gewichteten Mittelwert zusammenzufassen. Der so geschätzte Gesamteffekt als Ergebnis der Metaanalyse ist anschließend auf statistische Signifikanz zu prüfen.

Berechnung des Gesamteffekts $\bar{Z}$

Zur metaanalytischen Zusammenfassung der Primärstudien ist deren **Gewichtung** gemäß dem gewählten metaanalytischen Modell notwendig. Im einfachsten Fall handelt es sich um ein Fixed-Effect-Modell (▶ Abschn. 16.2.4), bei dem jede Studie gemäß ihrem Stichprobenumfang in der Weise gewichtet wird, dass Studien mit kleinen Stichproben geringer und Studien mit großen Stichproben stärker gewichtet werden. Der Sinn dieser Gewichtung ist offensichtlich: Eine Studie mit geringem Stichprobenumfang ist weniger präzise bzw. bietet eine ungenauere Effektgrößenschätzung und soll somit den Gesamteffekt weniger beeinflussen als eine Studie, die – bei Gleichheit aller anderen Studienmerkmale – dank größerem Stichprobenumfang **höhere Präzision** aufweist. Konkret werden die **Gewichtungsfaktoren w_i** im Fixed-Effect-Modell jeweils als **Kehrwert des quadrierten Standardfehlers v_i** jeder Studie berechnet.

Zur Schätzung des Gesamteffektes werden die Δ_i-Werte über Tab. 9 (▶ Webangebot zu diesem Buch) zunächst in **Fishers Z_i-Werte** transformiert. Jeder Z_i-Wert wird mit seinem Gewichtungsfaktor w_i multipliziert und alle Studien werden aufsummiert, abschließend wird durch die Summe der Gewichtungsfaktoren dividiert. Es

ergibt sich als **Gesamteffektschätzung** das gewichtete arithmetische Mittel aller Z_i als $\bar{Z}$:

$$\bar{Z} = \frac{\sum_{i=1}^{k} w_i \cdot Z_i}{\sum_{i=1}^{k} w_i} \tag{16.25}$$

$$\text{mit } w_i = \frac{1}{v_i} \quad \text{und } v_i = \frac{1}{(n_i - 3)} \tag{16.26}$$

Der $\sqrt{v_i}$-Wert entspricht der Streuung (genauer: dem Standardfehler) des entsprechenden Z_i-Wertes. Je größer der Stichprobenumfang n_i, auf dem ein Z_i-Wert basiert, desto weniger streuen die Z_i-Werte um den „wahren" Effektparameter ζ (Zeta), der durch $\bar{Z}$ geschätzt wird. Z_i-Werte, die auf großen Stichproben basieren, werden also bei der Durchschnittsbildung stärker gewichtet als Z_i-Werte aus kleineren Stichproben. Das Studiengewicht für die Z-transformierte Effektgröße jeder Studie lautet im **Fixed-Effect-Modell**: $w_i = 1/v_i = n_i - 3$ (für ein Berechnungsbeispiel ▶ Abschn. 16.4.3). Man beachte, dass bei der Wahl eines anderen metaanalytischen Modells (Random-Effects-Modell, Mixed-Effects-Modell, ▶ Abschn. 16.2.4) andere Gewichtungsfaktoren gelten.

Signifikanzprüfung des Gesamteffekts $\bar{Z}$

Um zu überprüfen, ob $\bar{Z}$ als Schätzwert der wahren Effektgröße ζ (Zeta) signifikant von Null abweicht (ob also ein überzufälliger Gesamteffekt in der Population existiert), wird folgender Test durchgeführt (Shadish & Haddock, 1994, S. 266):

$$z = \frac{\bar{Z}}{1/\sqrt{\sum_{i=1}^{k} w_i}} = \bar{Z} \cdot \sqrt{\sum_{i=1}^{k} w_i}. \tag{16.27}$$

Der Gesamteffektschätzer $\bar{Z}$ wird mit der Wurzel der Summe aller Studiengewichte multipliziert und es ergibt sich eine **standardnormalverteilte Prüfgröße** bzw. ein ***z*-Wert**. Der Gesamteffekt ist auf dem 5 %-Niveau signifikant, wenn bei einseitigem Test auf positiven Zusammenhang $z \geq 1.65$ ist. Der $\bar{Z}$-Wert sollte über Tab. 9 (▶ Webangebot zu diesem Buch) in einen $\bar{\Delta}$-Wert transformiert werden, der als Korrelationsäquivalent gemäß Tab. 14.5 zu klassifizieren wäre.

Das Konfidenzintervall des Gesamteffekts $\bar{Z}$ bestimmt man wie folgt:

$$\text{KI}_\zeta = \bar{Z} \pm z_{(1-\alpha/2)} \cdot \sqrt{1/\sum_{i=1}^{k} w_i}, \tag{16.28}$$

mit $z_{(1-\alpha/2)} = 1.96$ (für $\alpha = .05$) bzw. .258 (für $\alpha = .01$). Zur besseren Veranschaulichung des Ergebnisses kann das Konfidenzintervall dann wiederum über Korrelationsäquivalente ausgedrückt werden, indem man die obere und untere Grenze über Tab. 9 im ▶ Webangebot zu diesem Buch jeweils in einen $\bar{\Delta}$- bzw. ***r*-Wert** transformiert (die Abweichungen sind allerdings meist gering).

Fällt das Ergebnis des Tests des Gesamteffekts signifikant aus, so kann in der Regel ausreichende Teststärke als gegeben betrachtet werden. Bei einem nicht-signifikanten Ergebnis stellt sich jedoch die Frage, ob die Teststärke ausreichend war, was mit einer **A-posteriori-Teststärkeanalyse** geklärt werden kann (für entsprechende Formeln und deren Anwendung in einem Rechenbeispiel ▶ Abschn. 16.4.3).

Der Signifikanztest für den metaanalytischen Gesamteffekt (falls dieser wegen homogener Primärstudien sinnvoll ist) hat schon bei kleinen bis mittleren Studienzahlen mit Stichprobengrößen $n > 50$ eine *hohe bis sehr hohe Teststärke* (Cohn & Becker, 2003).

16.2.7 Prüfung der Homogenität der einzelnen Effektgrößen der Primärstudien

Die metaanalytische Zusammenfassung von k Effektgrößen aus k unabhängigen Untersuchungen bzw. Stichproben ist im Rahmen des **Fixed-Effect-Modells** (▶ Abschn. 16.2.4) nur sinnvoll, wenn die einzelnen Effektgrößen Schätzungen einer gemeinsamen Populationseffektgröße ζ (Zeta) darstellen, was **Homogenität** der untersuchungsspezifischen Effektgrößen impliziert.

Der Homogenitätstest über die Prüfgröße Q

Für die Überprüfung dieser Voraussetzung verwenden wir den folgenden, von Shadish und Haddock (1994) vorgeschlagenen **Homogenitätstest** auf der Basis der Prüfgröße Q (weitere Homogenitätstests findet man z. B. bei Schulze et al., 2003, Tab. 2.1):

$$Q = \sum_{i=1}^{k} \left[(Z_i - \bar{Z})^2 / v_i \right] = \sum_{i=1}^{k} w_i \cdot Z_i^2 - \frac{\left(\sum_{i=1}^{k} w_i \cdot Z_i \right)^2}{\sum_{i=1}^{k} w_i}. \tag{16.29}$$

In die **Prüfgröße Q** gehen die in Fishers Z-Werte transformierten ζ_i-Effektgrößenmaße der Einzelstudien sowie

die Gewichtungsfaktoren (w_i) ein. Die Prüfgröße Q ist approximativ χ^2-verteilt mit $k - 1$ Freiheitsgraden. Ein **signifikanter Q-Wert** weist darauf hin, dass die Streuung der Z_i-Werte größer ist als die stichprobenbedingte Zufallsstreuung, die man erwarten würde, wenn alle Z_i-Werte Schätzwerte desselben Populationsparameters ζ wären. In diesem Falle müsste man also von dem meist unerwünschten Ergebnis heterogener Z_i-Werte bzw. **heterogener Effektparameter** ausgehen. Dies ist bei der Mehrzahl der Metaanalysen in der Psychologie der Fall (Cafri et al., 2010) und erfordert eine anschließende Moderatorvariablen-Analyse unter Beibehaltung des Fixed-Effect-Modells oder die Verwendung eines Random- oder Mixed-Effects-Modells.

Es ist jedoch zu beachten, dass ein signifikanter Q-Wert auch bei augenscheinlich ähnlichen Effektgrößen als Folge **sehr großer Stichproben** resultieren kann. Wir haben bereits exemplarisch auf Metaanalysen hingewiesen, die Hunderte von Studien und Hundertausende von Untersuchungseinheiten beinhalten. In diesem Fall wird empfohlen, trotz signifikantem Homogenitätstest $\bar{Z}$ zu berechnen und als Schätzwert des Gesamteffekts zu interpretieren.

Die Teststärke des Homogenitätstests

Wie bei jedem Homogenitätstest ist auch beim Homogenitätstest für Effektgrößen die **Nullhypothese die Wunschhypothese** (zur Annahme der Nullhypothese ▶ Abschn. 15.3), denn man möchte ja die Homogenität (d. h. Nicht-Unterschiedlichkeit = Nulleffekt) der Effektgrößen belegen, die sich in einem **nicht-signifikanten Q-Wert** widerspiegelt (Hedges & Pigott, 2001, S. 445). Um bei einem nicht-signifikanten Q-Wert die Nullhypothese akzeptieren und Homogenität annehmen zu können, müssen wir sicherstellen, dass der Test eine **genügend große Teststärke** hatte (z. B. $1 - \beta = .95$), um im Falle der Heterogenität auch signifikant werden zu können. Bei zu geringer Teststärke bliebe nämlich unklar, ob der Test tatsächlich vorhandene Heterogenität der Effektgrößen bloß nicht nachweisen konnte oder ob tatsächlich keine Heterogenität (sondern stattdessen Homogenität) vorliegt.

Da bei einer Entscheidung für die Nullhypothese die β-Fehler-Wahrscheinlichkeit gering sein soll (z. B. β-Fehler-Niveau: 5 %) und sich α- und β-Fehler gegenläufig verhalten, muss das nicht-signifikante Ergebnis ein **α-Fehler-Niveau von 10 %** überschreiten ($p > .10$).

Des Weiteren soll die Alternativhypothese eine **schwache Heterogenität** postulieren, da bereits gering ausgeprägte Heterogenität die Zusammenfassung der Einzeleffekte zu einem Gesamteffekt problematisch erscheinen lässt. Die Heterogenität orientiert sich – ähnlich wie die 75 %-Regel von Hunter und Schmidt (1989) – am Verhältnis der Effektvarianz zwischen den Studien und der stichprobenbedingten Zufallsvarianz der Effekte. Mit k = Anzahl der Studien lauten diese Kategorien (Hedges & Pigott, 2001, S. 209):

- schwache Heterogenität: $0.33 \cdot (k - 1)$,
- mittlere Heterogenität: $0.67 \cdot (k - 1)$,
- starke Heterogenität: $1 \cdot (k - 1)$.

Dies sind gleichzeitig die **Nichtzentralitätsparameter der nichtzentralen χ^2-Verteilungen**, über die man die Teststärke des Homogenitätstests ermittelt. (Genauer wird der Nichtzentralitätsparameter über Gl. 16.29 geschätzt.)

Im Statistikprogramm SPSS ermittelt man die Teststärke über

$$(1 - \beta) = 1 - \text{NCDF.CHISQ (quant, df, nc)}, \quad (16.30)$$

mit q = kritischer χ^2-Wert für ein vorgegebenes α-Fehler-Niveau der zentralen χ^2-Verteilung mit df Freiheitsgraden, df = $k - 1$ = Anzahl der Freiheitsgrade und nc = Nichtzentralitätsparameter der nichtzentralen χ^2-Verteilung gem. Gl. 16.29 oder gemäß oben genannter Klassifikation.

Der *Homogenitätstest* der Effektgrößen der einzelnen Primärstudien hat bei kleiner bis mittlerer Studienanzahl und/oder kleinen bis mittleren Fallzahlen nur eine *geringe Teststärke*.

Die Homogenitätshypothese sollte bei nicht-signifikantem Ergebnis nur dann als „bestätigt" gelten, wenn

- **von der H_1 „schwache Heterogenität" ausgegangen wird,**
- **der Homogenitätstest für $\alpha = .10$ nicht signifikant wird und**
- **der Homogenitätstest mindestens eine Teststärke von 90 % aufweist.**

Die Implikationen dieser Argumentation für eine **A-posteriori-Teststärkeanalyse** seien an einem kleinen Beispiel (nach Hedges & Pigott, 2001, S. 209f.) verdeutlicht (zu den rechnerischen Details ▶ Abschn. 16.4.3). Eine Metaanalyse möge auf $k = 10$ einschlägigen, unabhängigen Studien mit jeweils $n = 25$ Untersuchungsteilnehmenden basieren. Geht man von der H_1: „schwache Heterogenität" aus, hätte der Homogenitätstest eine Teststärke von 17 %. Die Ablehnung der H_1 zugunsten von H_0 wäre also mit einem β-Fehler-Risiko von $\beta = .83$ verbunden (für $\alpha = .05$). Dies ist zweifellos ein Wert, der es verbietet, auf Homogenität zu plädieren. Die Situation verbessert sich ein wenig, wenn wir $\alpha = .10$ setzen: Die H_1 „schwache Heterogenität" kann dann bei

einem nicht-signifikanten Ergebnis mit einem β-Fehler-Risiko von $\beta = .73$ abgelehnt werden. Da wir jedoch ein β-Fehler-Risiko von 5 % fordern müssen, ist der Homogenitätstest in diesem Beispiel als nicht aussagekräftig zu betrachten.

A-priori-Teststärkeanalysen von Homogenitätstests im Rahmen einer Metaanalyse gestalten sich als schwierig. Nicht nur, dass vorab bekannt sein (oder eine plausible Schätzung vorliegen) muss, wie viele Studien Gegenstand der Metaanalyse sein werden und wie viele Untersuchungspersonen durchschnittlich an diesen Studien teilnahmen; es muss auch möglich sein, das Ausmaß der Heterogenität der Studien vorab zu schätzen, um dann die für einen aussagekräftigen Homogenitätstest notwendige Mindeststudienzahl zu berechnen.

16.2.8 Moderatorvariablen-/ Subgruppen-Analyse bei Heterogenität der Effektgrößen

Bei heterogenen Δ_i-Maßen sollte der Einfluss von Moderatorvariablen geprüft werden, die die Unterschiede in den Δ_i-Maßen erklären könnten (Czienskowski, 2003). Bei diesen Moderatorvariablen handelt es sich um theoretisch begründet ausgewählte Merkmale der zusammengefassten Studien (zur Kodierung der Studien ▶ Abschn. 16.2.3). Wie wichtig z. B. die Art des Studiendesigns und der Operationalisierung von abhängigen Variablen für die Größe des Studieneffektes ist, haben Wilson und Lipsey (2001) in einer Synthese von 319 Metaanalysen gezeigt. Die Studienmerkmale erklärten nahezu identische Anteile der Varianz der Studieneffekte wie die Art der geprüften Interventionen.

Hat man keine Hypothese zur Bedeutung einer speziellen Moderatorvariable, kann man zwischen den potenziellen Moderatorvariablen (nominal- oder ordinalskalierte Merkmale werden zuvor in Indikatorvariablen überführt; Kodierung eines polytom-nominalen Merkmals durch dichotome Indikatorvariablen in ▶ Abschn. 12.5.2 und den studienspezifischen Δ_i-Maßen eine multiple Korrelation bzw. Regression berechnen (zur Meta-Regression s. z. B. Borenstein et al., 2009, Kap. 21), deren Höhe über die Bedeutung der Studienmerkmale für die Heterogenität der Δ_i-Maße informiert. Signifikante β-Gewichte in der Meta-Regression weisen auf Studienmerkmale hin, die als Moderatorvariablen zur Bildung homogener Subgruppen in Betracht kommen (ausführlicher hierzu Hedges & Pigott, 2004, S. 439ff.)

Ist eine potenzielle Moderatorvariable identifiziert, die die Unterschiedlichkeit der Effektgrößen zwischen den Studien aufklären soll, so ist im Fixed-Effect-Modell bei der **Subgruppen-Analyse** in 2 Schritten vorzugehen:

1. Überprüfung der Bedeutung der potenziellen Moderatorvariable (hierzu muss der Omnibustest für Gruppendifferenzen über Q_{zw} ein signifikantes Ergebnis liefern),
2. Überprüfung der Homogenität der Studien innerhalb der durch die Moderatorvariable gebildeten Subgruppen (hierzu muss der Test bei ausreichender Teststärke auf Innerhalb-Gruppen-Homogenität über Q_{in} jeweils ein nicht-signifikantes Ergebnis liefern).

Der Omnibustest für Gruppendifferenzen über Q_{zw}

Die Überprüfung der Bedeutung einer Moderatorvariable kann nach Hedges (1994) varianzanalytisch erfolgen. Hierzu werden die Studien zunächst nach den p Stufen der zu prüfenden Moderatorvariable gruppiert. Die Unterschiedlichkeit zwischen den p Gruppen wird mit folgendem **Q_{zw}-Wert** getestet:

$$Q_{zw} = \sum_{j=1}^{p} w_{j.} \cdot (\bar{Z}_{j.} - \bar{Z})^2, \qquad (16.31)$$

mit

$\bar{Z}$ = durchschnittlicher Fisher-Z-Wert für alle k Studien gem. Gl. 16.25,

$\bar{Z}_j$ = durchschnittlicher Fisher-Z-Wert der k_j Studien in Gruppe j analog zu Gl. 16.25,

$w_i = \sum_{i=1}^{k_j} w_{ij} = \sum_{i=1}^{k_j} (1/v_{ij})$,

$v_{ij} = 1/(n_{ij} - 3)$ (quadrierter Standardfehler des Z_{ij}-Wertes der i-ten Studie in der j-ten Gruppe),

n_{ij} = Stichprobenumfang der i-ten Studie in der j-ten Gruppe.

Der Q-Wert entspricht dem Omnibus-F-Wert der einfaktoriellen Varianzanalyse. Er ist bei Gültigkeit von H_0: „keine Gruppenunterschiede“ mit $p - 1$ Freiheitsgraden approximativ χ^2 verteilt. Ein **signifikantes Ergebnis dieses Omnibustests** bedeutet, dass die gewählte Moderatorvariable einen Teil der Unterschiedlichkeit der Effektgrößenschätzungen der Primärstudien erklärt. Ein nicht-signifikantes Ergebnis darf nur dann als Hinweis auf die Bedeutungslosigkeit der betreffenden Moderatorvariable gewertet werden, wenn a posteriori ausreichende Teststärke nachweisbar ist (zur Teststärke des Omnibustests über alle Stufen der Moderatorvariable sowie von Kontrasten, d. h. von Vergleichen ausgewählter Stufen s. Hedges & Pigott, 2004)

Der Test auf Innerhalb-Gruppen-Homogenität über Q_{in}

Die Unterschiedlichkeit der studienspezifischen Gruppen-Z_{ij}-Werte innerhalb der p Gruppen überprüft der folgende

Q_{in}-Wert:

$$Q_{in} = \sum_{j=1}^{p} Q_{in\,j}, \qquad (16.32)$$

$$\text{mit } Q_{in\,j} = \sum_{j=1}^{k_j} w_{ij} \cdot (Z_{ij} - \bar{Z}_{j.})^2.$$

Der Q_{in}-Wert ist bei Gültigkeit von H: „keine Unterschiede innerhalb der Gruppen" approximativ χ^2 verteilt mit $k - p$ Freiheitsgraden. Die Homogenität der Studien in Gruppe j testet man über $Q_{in\,j}$ an der χ^2-Verteilung mit $k_j - 1$ Freiheitsgraden (hierzu auch ▶ Abschn. 16.4.4).

Der Gesamt-Q-Wert nach Gl. 16.29 ergibt sich additiv aus Q_{zw} und Q_{in}, d. h., es gilt

$$Q = Q_{zw} + Q_{in}. \qquad (16.33)$$

Eine Moderatorvariable unterteilt die k Studien in p **homogene Subgruppen**, wenn a) Q_{zw} signifikant ist (s. vorigen Abschnitt) und b) $\boldsymbol{Q_{in}}$ **nicht-signifikant** ist, d. h. die Nullhypothese: „identische Effektparameter innerhalb der p Gruppen" gilt. Dabei kann ein nicht-signifikanter Q_{in}-Wert nur dann als Beleg für die Homogenität der Studien innerhalb der durch die Moderatorvariable gebildeten Subgruppen gelten, wenn er über ausreichende Teststärke verfügt, was meist erst bei einer größeren Zahl an Studien mit großen Stichproben der Fall ist (Hedges & Pigott, 2004).

16.2.9 Präsentation und Interpretation der Ergebnisse der Metaanalyse

Die Darstellung der statistischen Befunde einer Metaanalyse kann **tabellarisch** erfolgen, ist aber auch als grafische Präsentation in Form eines **Forest-Plots** empfehlenswert, sofern die Studienzahl überschaubar ist. Der Forest-Plot präsentiert untereinander die Konfidenzintervalle der Effektgrößen der einzelnen Primärstudien und zudem das Konfidenzintervall der Schätzung der Gesamteffektgröße (▶ Präsentation der Ergebnisse einer Metaanalyse als Forest-Plot), die Bezeichnung „Forst-Grafik" ist aus der optischen Anmutung der Darstellung abgeleitet. Gängige Metaanalyse-Programme fertigen Forest-Plots an (▶ Abschn. 16.5).

Neben einer übersichtlichen Präsentation ist bei der Metaanalyse – genau wie bei jeder Primärstudie – eine sorgfältige Interpretation der Befunde von besonderer Wichtigkeit. Dabei wird das Ergebnis oftmals zunächst mit Blick auf seine **theoretische Bedeutung** diskutiert: Konnte die Forschungshypothese der Metaanalyse – und damit indirekt und partiell auch das zugrunde liegende Theoriemodell – bestätigt werden oder nicht? Inwiefern bekräftigt, modifiziert oder widerlegt die Metaanalyse den bisherigen empirischen Forschungsstand? In der Metaanalyse zum Bystander-Effekt in gefährlichen versus ungefährlichen Situationen wird z. B. als Befund zweierlei konstatiert (Fischer et al., 2011):

Präsentation der Ergebnisse einer Metaanalyse als Forest-Plot

Angenommen, es seien 6 Primärstudien metaanalytisch zu aggregieren, die als Effektgrößenmaße jeweils einen Korrelationskoeffizienten angeben. Die Studien lassen sich im Überblick tabellarisch darstellen mit Autorinnen- und Autorennamen und Jahreszahl in der ersten Spalte, gefolgt von Effektgröße und Stichprobenumfang (Tab. 16.3).

Sollen diese 6 Studien auf der Basis des Fixed-Effect-Modells zusammengefasst werden, so sind die Korrelationskoeffizienten in Fishers Z-Werte (hier: „effect size" Y) zu transformieren, zudem sind anhand der Stichprobenumfänge bzw. der Standardfehler der einzelnen Studien jeweils die Studiengewichte W zu bestimmen (Tab. 16.4). Als Gesamteffektschätzung ergibt sich $\bar{Z} = 0.358$ (▶ Abschn. 16.2.6 „Berechnung des Gesamteffektes").

Mittels des Metaanalyse-Programms CMA (Borenstein et al., 2009, Abschn. 16.5.3) lässt sich nicht nur der Gesamteffekt berechnen, sondern auch ein Forest-Plot erstellen, der sowohl die eingegangenen Effekte, die relativen Gewichte der Studien also auch deren Konfidenzintervalle grafisch wiedergibt (Abb. 16.2). Die relativen Gewichte ergeben sich, indem man das Studiengewicht W der jeweiligen Studie an der Summe aller Gewichtungsfaktoren (hier: 647) relativiert. Es wird deutlich, dass die Studie von Grant mit $N = 25$ nur ein relatives Gewicht von 3.4 % erhält, während die Studie von Granger mit $N = 400$ mit einem relativen Gewicht von 61.36 % die Gesamteffektschätzung im Fixed-Effect-Modell stark beeinflusst.

Tabelle 16.3 Basisdaten für eine Metaanalyse mit sechs Korrelationsstudien. (Mod. nach Borenstein et al., 2009, Tab. 14.7. Mit freundlicher Genehmigung von John Wiley and Sons)

	Effect Size Measure	
Study	Correlation	*N*
Fonda (1998)	.50	40
Newman (2000)	.60	90
Grant (2003)	.40	25
Granger (2005)	.20	400
Milland (2009)	.70	60
Finch (2012)	.45	50

Tabelle 16.4 Studiengewichte für eine Metaanalyse der sechs Korrelationsstudien nach dem Fixed-Effect-Modell. (Mod. nach Borenstein et al., 2009, Tab. 14.8. Mit freundlicher Genehmigung von John Wiley and Sons)

				Calculated quantities		
Study	Effect Size Y	Variance within V_Y	Weight W	WY	WY^2	W^2
Fonda	0.5493	0.0270	37.0000	20.3243	11.1643	1 369.0000
Newman	0.6931	0.0115	87.0000	60.3038	41.7994	7 569.0000
Grant	0.4236	0.0455	22.0000	9.3203	3.9485	484.0000
Granger	0.2027	0.0025	397.0000	80.4848	16.3169	157 609.0000
Milland	0.8673	0.0175	57.0000	49.4361	42.8760	3 249.0000
Finch	0.4847	0.0213	47.0000	22.7809	11.0419	2 209.0000
Sum			647.0000	242.6503	127.1470	172 489.0000

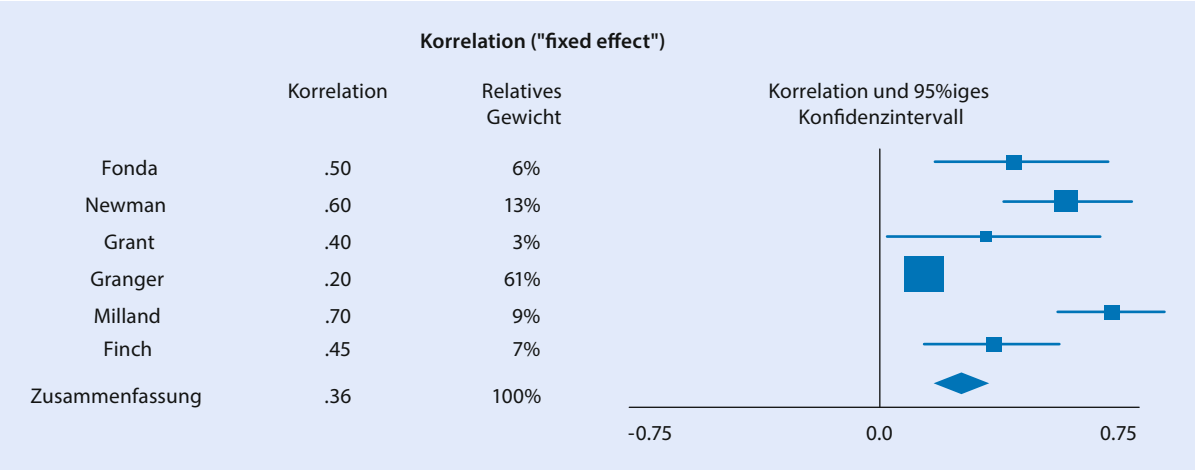


Abb. 16.2 Forest-Plot einer Metaanalyse über 6 Korrelationsstudien nach dem Fixed-Effect-Modell. (Mod. nach Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. [2009]. Introduction to Meta-Analysis. Chichester: Wiley., Abbildung 14.5. Mit freundlicher Genehmigung von John Wiley and Sons)

1. dass in Übereinstimmung mit dem bisherigen Forschungsstand die Existenz des Bystander-Effekts **zu bestätigen ist** (die Gesamteffektschätzung gemäß Fixed-Effect-Modell beträgt $g = -0.35$ und ist statistisch signifikant, wobei allerdings Heterogenität der Einzeleffekte vorliegt) und
2. dass die Studie außerdem den bisherigen Forschungsstand **ergänzt**, weil die Moderatorvariablen- bzw. Subgruppen-Analyse belegt, dass der Bystander-Effekt in gefährlichen Situationen hypothesenkonform tatsächlich signifikant geringer ausgeprägt ist als in ungefährlichen bzw. dass Bystander unter bestimmten Bedingungen gerade in gefährlichen Situationen ein helfendes Eingreifen sogar fördern.

Von Bedeutung ist weiterhin auch eine Interpretation des metaanalytischen Befundes im Hinblick auf dessen **praktische Bedeutsamkeit** (▶ Abschn. 14.2). Dabei geht es v. a. auch um eine Einordnung der Gesamteffektgröße bzw. der Effektgrößen der Subgruppen. Als erster Anhaltspunkt wird dabei oft auf die **Effektgrößenklassifikation von Cohen** (▶ Abschn. 14.2.3) zurückgegriffen. Es zeigt sich, dass die Befunde von Metaanalysen in der Psychologie häufig als kleine Effekte zu klassifizieren sind (Cafri et al., 2010). Noch aussagekräftiger ist es, die Effektgrößenschätzungen einer Metaanalyse nicht mit Cohens groben Klassifikationskriterien, sondern mit den für das untersuchte Phänomen laut Forschungsstand **typischen Effektgrößen** oder mit **Effektgrößen aus**

verwandten Forschungsgebieten interpretativ in Relation zu setzen. So argumentieren Anderson et al. (2010, S. 170) bei der Interpretation ihrer Metaanalyse zum Zusammenhang zwischen der Nutzung gewalthaltiger Videospiele und Aggressivität, dass der in Längsschnittstudien nachweisbare Gesamtzusammenhang von $r = +.15$ (was einer Varianzaufklärung von 2.3 % entspricht) laut Cohen zwar als **kleiner Effekt** zu klassifizieren, aber **praktisch dennoch sehr bedeutsam** sei. Denn dieser Effekt a) könne sich über die Lebenszeit kumulieren, b) betreffe sehr viele Menschen und c) habe sehr negative soziale Konsequenzen. Zudem sei der in der Metaanalyse nachgewiesene Einfluss gewalthaltiger Videospiele auf die Aggression mit einer Varianzaufklärung von gut 2 % sogar stärker ausgeprägt als der Einfluss anderer aggressionsfördernder Risikofaktoren (z. B. Armut, Drogenmissbrauch oder Misshandlung durch die Eltern).

Die Befunde von Metaanalyen, die Dutzende oder gar Hunderte von Studien systematisch aggregieren, haben für den wissenschaftlichen Erkenntnisfortschritt eine deutlich höhere Bedeutung als einzelne Primärstudien. Dennoch sind auch die Befunde von Metaanalysen ihrerseits Gegenstand kontroverser öffentlicher und wissenschaftlicher Debatten. Dabei wird in der Regel sowohl eine fehlerhafte Bestimmung des metaanalytischen Gesamteffektes (z. B. aufgrund eines lückenhaften oder verzerrten Studienpools) als auch dessen **Überinterpretation** moniert. Das gilt nicht zuletzt im umstrittenen Bereich der Wirkungen gewalthaltiger Medieninhalte. So wird der Metaanalyse von Anderson et al. (2010) u. a. methodisch vorgeworfen, auf einem durch Publication-Bias verzerrten Studienpool zu basieren, Studien mit mangelhafter Operationalisierung von Aggression einzubeziehen und durch ungenügende Berücksichtigung von Moderatorvariablen den Gesamteffekt zu überschätzen. Zudem wird eine Überinterpretation des Gesamteffektes von $r = +.15$ kritisiert: Die Metaanalyse erfasse keine gravierende, alltagsrelevante Aggression, zudem seien andere aggressionsfördernde Faktoren (z. B. Drogen, Gewalt im Elternhaus) tatsächlich laut vorliegender Effektgrößenschätzungen sehr viel einflussreicher als gewalthaltige Videospiele, deren seit Jahren wachsende Verbreitung mit sinkender Gewaltkriminalität laut nationaler Kriminalstatistiken einhergehe (Ferguson & Kilburn, 2010).

Dass auch Ergebnisse von Metaanalysen Gegenstand kontroverser Debatten sind und nicht einfach für sich beanspruchen können, die Existenz und Größe eines Populationseffekts final zu belegen oder zu widerlegen, ist aus wissenschaftstheoretischer Sicht nicht verwunderlich. Denn die fortwährende kritische Reflexion der Befunde sowie des Forschungsprozesses selbst ist ja gerade ein zentrales Kennzeichen von Wissenschaftlichkeit (▶ Abschn. 1.1.1).

16.3 Aggregierung von Primärstudien ohne gemeinsame Effektgrößenschätzung

Untersuchungen, deren Effektgrößen nicht vollständig in Publikationen berichtet werden (was als mangelnde Präsentationsqualität gelten muss; ▶ Abschn. 3.1.2), erzeugen das bereits erwähnte **Missing-Data-Problem** (▶ Abschn. 16.1.3) und erschweren metaanalytisches Arbeiten erheblich. Dennoch hat die metaanalytische Entwicklung Wege aufgezeigt, wie auch unvollständige Untersuchungsberichte dieser Art zu aggregieren sind. Bei den hierfür einschlägigen Verfahren handelt es sich um sog. **kombinierte Signifikanztests**, die aus den einfachen Signifikanzaussagen eine Gesamtaussage über die Existenz eines Effektes zu formulieren suchen. Die Kritik an diesen Verfahren richtet sich im Wesentlichen auf zwei Aspekte (Becker, 1987, 1994; Borenstein et al., 2009; Fricke & Treinies, 1985, Kap. 5):

1. Die Aussage, dass ein Zusammenhang oder Unterschied statistisch signifikant sei, enthält **keinerlei Informationen über die Größe des zugrunde liegenden Effekts**. Einerseits wird jeder auch noch so kleine Effekt statistisch signifikant, wenn die untersuchte Stichprobe genügend groß ist, und andererseits können beachtliche Effekte wegen zu kleiner Stichproben (bzw. zu geringer Teststärke) nicht-signifikant bleiben. Die metaanalytische Zusammenfassung einzelner Signifikanzprüfungen kann deshalb bestenfalls in die Aussage münden, dass mit hoher Wahrscheinlichkeit von einem irgendwie gearteten Effekt auszugehen ist.
2. Wenn die Effekte der Primärstudien nicht bekannt sind, lässt sich deren **Homogenität nicht überprüfen**, das heißt, die Metaanalyse fasst möglicherweise heterogene Einzelergebnisse zusammen, was nicht zulässig ist (zu diesem „Äpfel-und-Birnen-Problem“ ▶ Abschn. 16.1.3, ▶ Abschn. 16.2.7).

Nach diesen Vorbemerkungen sind kombinierte Signifikanztests nur als ein **selten einzusetzender Notbehelf** anzusehen. Sie bereiten allerdings weniger Probleme, wenn die metaanalytisch zu integrierenden Untersuchungen auf ähnlichen Stichprobenumfängen beruhen und wenn nach sorgfältiger Inspektion der Operationalisierungen der interessierenden Variablen von gleichartigen Untersuchungen ausgegangen werden kann. Dennoch muss man sich darüber im Klaren sein, dass kombinierte Signifikanztests an der eigentlichen Zielsetzung der Metaanalyse, nämlich der summarischen Abschätzung der Stärke des Populationsgesamteffektes durch gewichtete Mittelwertsbildung der einzelnen Effektgrößenschätzungen der Studien, letztlich vorbeigehen.

Werden in den Primärstudien die relevanten Effektgrößen nicht oder unvollständig angegeben (Missing-Data-Problem), kann man die vorhandenen Informationen für eine Metaanalyse behelfsweise mittels kombinierter Signifikanztests zusammenfassen:

- **Aggregierung statistisch signifikanter Ergebnisse:**
  - **Auszählen signifikanter vs. nicht-signifikanter Ergebnisse (Vote-Counting).**
  - **Überprüfung der Anzahl signifikanter Ergebnisse auf Zufälligkeit (Binomialtest).**
- **Aggregierung positiver (deskriptiv-hypothesenkonformer) Ergebnisse:**
  - **Statistischer Vergleich positiver/hypothesenkonform gerichteter und negativer/hypothesenkonträr gerichteter Ergebnisse unabhängig von ihrer Signifikanz (Vorzeichentest).**
- **Aggregierung exakter Irrtumswahrscheinlichkeiten:**
  - **Statistische Zusammenfassung exakter Irrtumswahrscheinlichkeiten (Stouffer-Methode).**
  - **Zusammenfassung exakter Irrtumswahrscheinlichkeiten und Stichprobenumfänge bei Zweigruppenvergleichen durch ein Korrelationsäquivalent ($r_{equivalent}$) und Konfidenzintervallbestimmung.**

Die im Folgenden behandelte Verfahrensgruppe stellt eine als Notbehelf zu verstehende und **zunehmend seltener eingesetzte** Alternative zur herkömmlichen (► Abschn. 16.2 in ihren Arbeitsschritten detailliert beschriebenen) Metanalyse dar. Die Verfahren beziehen sich auf Untersuchungen ohne vollständige Angaben zu den relevanten Effektgrößen, aus denen lediglich hervorgeht, ob das Untersuchungsergebnis im Sinne der Forschungshypothese signifikant ist oder nicht bzw. auf Untersuchungen, in denen exakte Irrtumswahrscheinlichkeiten genannt werden. Beiträge mit derart unvollständigen Ergebnisangaben werden heute kaum noch publiziert. Dementsprechend sind die im Folgenden behandelten Verfahren v. a. dann von Interesse, wenn eine Forschungssynthese verstärkt aus älteren, unvollständig berichteten Primärstudien erstellt werden soll oder wenn eine ältere Forschungssynthese rezipiert wird, die diese Methoden nutzt.

16.3.1 Aggregierung statistisch signifikanter Ergebnisse

Zur Aggregierung statistisch signifikanter Ergebnisse wird zum einen das einfache Auszählen (Vote Counting) eingesetzt, zum anderen der Binomialtest.

Vote Counting (Auszählungen, „box score review“)

Für metaanalytische Zwecke am wenigsten geeignet sind Untersuchungen, die lediglich nach dem Kriterium „Ergebnis signifikant“ bzw. „Ergebnis nicht signifikant“ klassifiziert werden können und keine detaillierteren Befunde berichten. Werden solche Studien in eine Metaanalyse einbezogen, müssen bei der Aggregierung der Ergebnisse auch die Untersuchungen, die ausführlichere Informationen enthalten, auf ihre Signifikanzaussage reduziert werden. Die Aggregierung läuft hierbei auf eine einfache Auszählung aller Studien mit signifikanten Ergebnissen hinaus. Zeigt der Studienpool **mehrheitlich signifikante Ergebnisse**, gilt das als Hinweis auf die Existenz des Effektes bzw. die Bestätigung der fraglichen Unterschieds-, Zusammenhangs- oder Veränderungshypothese. Überwiegen jedoch die nicht-signifikanten Ergebnisse, so wird dies als Hinweis auf die Nichtexistenz des Effektes bzw. eine Entscheidung für die Nullhypothese gewertet (Bushman, 1994; Light & Smith, 1971). Die Problematik dieses Verfahrens besteht nun darin, dass ein statistisch nichtsignifikantes Ergebnis nur dann als Bestätigung der Nullhypothese bzw. der Nichtexistenz des Effektes aufgefasst werden kann, wenn die Teststärke genügend groß ist (► Abschn. 14.1). Ein Vote Counting, das ungeachtet der Teststärke der Metaanalyse nicht-signifikante Ergebnisse zählt und als Belege für die Nullhypothese wertet, führt nicht zu sinnvollen Ergebnissen, insbesondere wenn der Studienpool klein und somit die Teststärke gering ist (Borenstein et al., 2009, S. 251ff.).

Man beachte, dass ein simples Vote Counting oft der **intuitive Ansatz** ist, auf den bei der Rezeption und Bewertung eines Studienpools zurückgegriffen wird, um sich schnell einen Gesamtüberblick zu verschaffen. Möglicherweise basieren viele narrative Reviews in ihrem Gesamtfazit letztlich auf Vote Counting (Borenstein et al., 2009, S. 301ff.).

Binomialtest („binomial test“)

Während beim Vote Counting der Effekt bereits als bestätigt angesehen wird, wenn nur mehr signifikante als nichtsignifikante Ergebnisse vorliegen, wird mittels Binomialtest darüber hinaus geprüft, ob es sich um überzufällig mehr signifikante als nicht-signifikante Resultate der Primärstudien handelt. Mit dem Binomialtest kann man überprüfen, wie groß die Wahrscheinlichkeit P ist, dass von k Untersuchungen mindestens x Untersuchungen zufällig signifikant werden (Wilkinson, 1951). Die H_0 „Die Anzahl der signifikanten Ergebnisse entspricht dem Zufall“ ist gemäß Logik des klassischen Signifikanztests zu verwerfen, wenn $P \leq \alpha$ ist. Der Binomialtest ist z. B. bei Bortz, Lienert und Boehnke (2008) vertafelt, was die

rechnerische Durchführung per Hand erheblich erleichtert.

16.3.2 Aggregierung positiver (deskriptiv-hypothesenkonformer) Ergebnisse

Eine Globalentscheidung für die Alternativhypothese kann nicht nur an der Mehrzahl oder Überzufälligkeit signifikanter Primärstudien festgemacht werden, sondern auch an der signifikanten Überzahl positiver bzw. hypothesenkonform gerichteter Ergebnisse (unabhängig von ihrer statistischen Signifikanz). Dies ist insofern sinnvoll, als eine pure Signifikanzbetrachtung bei den Primärstudien – ohne Berücksichtigung von Stichprobenumfang bzw. Teststärke der Primärstudien – sehr irreführend sein kann. Liegt der Fokus auf der Zahl der deskriptiv-hypothesenkonformen Ergebnisse der Primärstudien, so lässt sich zu deren statistischer Aggregierung der Vorzeichentest einsetzen.

Vorzeichentest („sign test“)

Will man entscheiden, ob positive/hypothesenkonforme Ergebnisse ungeachtet ihrer Signifikanz statistisch überwiegen (z. B. gerichtete Alternativhypothese im Sinne der Überlegenheit einer Experimentalgruppe gegenüber einer Kontrollgruppe), kann hierfür der Vorzeichentest eingesetzt werden (Bortz et al., 2008, Abschn. 6.2.1). Gemäß H_0 nimmt man an, dass positive und negative Ergebnisse mit gleicher Wahrscheinlichkeit (50 %) auftreten und ermittelt die Irrtumswahrscheinlichkeit p, dass unter k Untersuchungen bei Gültigkeit von H_0 mindestens x Untersuchungen positiv ausfallen. Ausgehend von einer gerichteten Alternativhypothese der Metaanalyse, ist die H_0 bei $P \leq \alpha$ zu verwerfen (einseitiger Test). Man beachte, dass dieser Ansatz nicht zwischen signifikanten und nicht-signifikanten Ergebnissen unterscheidet, d. h., bei vielen nicht-signifikanten negativen Ergebnissen und wenigen signifikanten positiven Ergebnissen könnte zugunsten der negativen Ergebnisse entschieden werden.

16.3.3 Aggregierung exakter Irrtumswahrscheinlichkeiten

Zur Aggregierung exakter Irrtumswahrscheinlichkeiten ist die Stouffer-Methode einschlägig. Beim Zweigruppenvergleich kann auch ein Korrelationsäquivalent bestimmt werden.

Stouffer-Methode („stouffer method“)

Die am häufigsten eingesetzte Methode, exakte Irrtumswahrscheinlichkeiten aus mehreren Untersuchungen – und ggf. für unterschiedliche statistische Kennwerte – zu aggregieren, geht auf Stouffer et al. (1949) zurück.

Man transformiert bei der **Stouffer-Methode** zunächst die untersuchungsspezifischen Irrtumswahrscheinlichkeiten (P-Werte) in z-**Werte der Standardnormalverteilung**, aus denen folgende Stouffer-Prüfgröße ermittelt wird (mit k: Anzahl der Studien):

$$Z_S = \frac{\sum_{i=1}^{k} z_i}{\sqrt{k}}. \tag{16.34}$$

Z_S ist bei Gültigkeit von H_0 mit $\mu = 0$ und $\sigma = \sqrt{k}$ standardnormalverteilt, so dass anhand Tab. 1 im ▶ Webangebot zu diesem Buch die zu Z_S gehörende Wahrscheinlichkeit $P(Z_S)$ abgelesen werden kann. Das metaanalytische Gesamtergebnis ist für $P(Z_S) \leq \alpha$ signifikant. Über die Teststärke dieses Verfahrens berichten Hedges et al. (1992).

Gelegentlich befinden sich unter den zu aggregierenden Untersuchungen Studien mit extremen Irrtumswahrscheinlichkeiten, die aus dem Niveau der anderen Irrtumswahrscheinlichkeiten deutlich herausfallen und deshalb auf außergewöhnliche Untersuchungsumstände schließen lassen. Für diesen Fall hat Saner (1994) eine **Trimmed Statistic** entwickelt, die dieser Ausreißer-Problematik gerecht wird.

Effektgrößenschätzung bei Zweigruppenvergleichen durch ein Korrelationsäquivalent ($r_{equivalent}$)

Für den Fall, dass pro Studie nicht nur eine exakte Irrtumswahrscheinlichkeit (ein P-Wert) berichtet wird, sondern zusätzlich der Stichprobenumfang (n) der Untersuchung, kann man nach Rosenthal und Rubin (2003) eine Effektgröße berechnen, die dem Δ-Maß (▶ Abschn. 16.2.5) bzw. der bivariaten Korrelation entspricht. Das als **Korrelationsäquivalent** ($r_{equivalent}$) bezeichnete Effektgrößenmaß setzt allerdings voraus, dass das Studiendesign einen Zweigruppenvergleich impliziert (z. B. Kontroll- vs. Experimentalgruppe oder ein Einzelvergleich im Rahmen einer Varianzanalyse).

Über einschlägige Tabellen von t-Verteilungsfunktionen (t-Tabellen) oder geeignete Statistik-Software wird der (einseitige) P-Wert in einen t-Wert mit $n-2$ Freiheitsgraden (df) transformiert. Für diesen t-Wert errechnet man über folgende Gleichung den $r_{equivalent}$-Wert (zur Umrechnung von unterschiedlichen Effektgrößenmaßen auch ◘ Tab. 14.3):

$$r_{equivalent} = \sqrt{\frac{t^2}{t^2 + (n-2)}}. \tag{16.35}$$

$r_{equivalent}$ entspricht einer punktbiserialen Korrelation (Gl. 16.7).

Wenn statt eines exakten P-Wertes nur das Signifikanzniveau ($P < .05$, $P < .01$ etc.) markiert wird, transformiert man das entsprechende Signifikanzniveau und errechnet über Gl. 16.35 eine untere Grenze für $r_{\text{equivalent}}$.

Beispiel: In einem Zweigruppenvergleich zwischen einer Experimentalgruppe (E) und einer Kontrollgruppe (K) sowie insgesamt 3 Untersuchungspersonen ($n_E = n_K = 15$) wird ein signifikantes Ergebnis berichtet mit $P < 01$. Der t-Tabelle (z. B. Tab. 3 im ▶ Webangebot zu diesem Buch) entnimmt man (für $1-P = .990$ und $df = n-2 = 28$) $t = 2.467$ und errechnet über Gl. 16.35

$$r_{\text{equivalent}} = \sqrt{\frac{2.467^2}{2.467^2 + 28}} = .42.$$

Dieser Wert markiert die untere Grenze des Korrelationsäquivalentes.

Falls der P-Wert für einen Einzelvergleich bestimmt wurde, bei dem mehrere Bedingungen (insgesamt k Bedingungen) in 2 zu kontrastierende Gruppen zusammengefasst sind, ersetzt man in Gl. 16.35 $(n-2)$ durch $(n-k)$. Das 95 %ige Konfidenzintervall für $r_{\text{equivalent}}$ kann wie folgt bestimmt werden:

$$KI = \frac{Z \pm 1.96}{\sqrt{n-3}}. \tag{16.36}$$

Z ist hierbei Fishers Z-Wert für $r_{\text{equivalent}}$ (Tab. 9 im ▶ Webangebot zu diesem Buch). Für das Beispiel erhält man mit $Z(r = .42) = 0.448$

$$KI = \frac{0.448 \pm 1.96}{\sqrt{30-3}} = 0.448 \pm 0.377.$$

Das Korrelationsäquivalent befindet sich nach Rücktransformation der Z-Wert-Grenzen in r-Grenzen also mit 95 %iger Konfidenz im Bereich 0.07–0.68.

Man beachte, dass $r_{\text{equivalent}}$ auch für verteilungsfreie Tests (z. B. Fisher-Yates-Test, U-Test, Vorzeichentest; Bortz & Lienert, 2003) berechnet werden kann, für die sich bislang noch keine allgemein akzeptierten Effektgrößen etabliert haben (zur Kritik und Weiterentwicklung von $r_{\text{equivalent}}$ s. Hsu, 2005; Kraemer, 2005).

16.4 Beispiel: Eine Fixed-Effect-Metaanalyse zum Lehrererwartungs-Effekt

Im Folgenden möchten wir eine didaktisch vereinfachte Metaanalyse zum Lehrererwartungs-Effekt schrittweise präsentieren. Damit begeben wir uns auf ein Forschungsfeld, in dem die Anwendung von Metaanalysen traditionell eine sehr wichtige Rolle spielt (▶ Abschn. 16.1.1 „Geschichte der Metaanalyse“). Bevor wir mit den Berechnungen beginnen, soll zunächst das **Forschungsfeld der Erwartungseffekte** skizziert werden, das in der Psychologie (Versuchsleitereffekte in der Experimentalforschung) sowie in der Erziehungswissenschaft (Lehrererwartungseffekte auf das Schülerverhalten) von großer praktischer Bedeutung ist.

Gemäß der Hypothese der **selbsterfüllenden Prophezeiung** („self-fulfilling prophecy“) bzw. der **interpersonalen Erwartungen** („interpersonal expectancy effects“) können Erwartungen, die wir gegenüber anderen Menschen hegen, insbesondere wenn wir für sie „Autoritätspersonen“ darstellen, deren Verhalten in erwartungskonformer Richtung beeinflussen, auch wenn wir diese Erwartungen gar nicht ausdrücklich aussprechen. Der Erwartungseffekt ist sogar auf Tiere übertragbar: So konnte der amerikanische Psychologe **Robert Rosenthal** nachweisen, dass Laborratten ein Labyrinth schneller lernten, wenn die Versuchsleiter glaubten, eine besonders lernbegabte Rattenart vor sich zu haben, obwohl es sich nur um ganz normale Laborratten handelte (Rosenthal & Fode, 1963).

Ein **Versuchsleiter(erwartungs)effekt** („experimenter expectancy effect“) tritt bei empirischen Untersuchungen mit Menschen auf, weil offenbar Versuchsleiter bei Kenntnis der Forschungshypothese die Versuchspersonen in der Experimentalbedingung unterschwellig in hypothesenkonformer Richtung beeinflussen (Rosenthal, 1968: Metaanalyse von 35 Studien zum Versuchsleitereffekt mittels Stouffer-Methode: $Z = 9.25$, $p < .001$). Um diesen Versuchsleitereffekt zu vermeiden, wird in der Forschung, insbesondere der Experimentalforschung, bevorzugt mit „blinden“ Versuchsleitern gearbeitet, die die Forschungshypothese nicht kennen bzw. nicht wissen, ob sie gerade Personen der Experimental- oder der Kontrollgruppe vor sich haben (zur Verblindung ▶ Abschn. 7.6). Die Existenz von Versuchsleiterartefakten wird heute allgemein akzeptiert. In den 1960er-Jahren reagierte man jedoch auf die ersten Studien von Rosenthal äußerst skeptisch – schließlich wurde durch den damals neu postulierten Versuchsleitererwartungs-Effekt die interne Validität aller früheren (nicht verblindeten) Experimente infrage gestellt.

Erwartungseffekte spielen nicht nur im Forschungslabor, sondern auch im Klassenzimmer eine Rolle: Wenn Lehrkräfte z. B. aufgrund ethnischer Zugehörigkeit bestimmte Schüler für besonders intelligent (bzw. wenig intelligent) halten, schlägt sich dies nicht selten in einer besonders positiven (bzw. negativen) Leistungsentwicklung der betreffenden Kinder nieder – auch wenn die Schüler sich ursprünglich in ihrer tatsächlichen Intelligenz gar nicht von den anderen Schülern unterschieden (selbsterfüllende Prophezeiung im Klassenzimmer bzw.

Pygmalion-Effekt nach Rosenthal & Jacobson, 1968; Rosenthal, 1995). Konkret kann dies z. B. bedeuten, dass Lehrkräfte asiatischen Schülerinnen und Schülern besonderen Fleiß und besondere Begabung zuschreiben oder den Kindern von Hartz-IV-Empfängern eben diese Eigenschaften absprechen und von ihnen schlechtere Leistungen erwarten. Ebenso wie der Versuchsleitereffekt wurde auch der **Lehrererwartungs-Effekt** („teacher expectancy effect") von der Forschungsgemeinschaft zunächst sehr kritisch aufgenommen, gleichzeitig wurden zahlreiche Studien zum Phänomen durchgeführt.

Um die Kontroverse zu entscheiden, nutzte Robert Rosenthal die **Methodik der Metaanalyse**, indem er regelmäßig den aktuellen Forschungsstand zu Versuchsleiter- und Lehrererwartungs-Effekten aggregierte und publizierte (Rosenthal, 1969, 1973a, b, 1976; Rosenthal & Rubin, 1971, 1978). Insbesondere mit der letzten Metaanalyse dieser Serie (Rosenthal & Rubin, 1978), die 345 Studien zusammenfasste, ließ sich die Existenz des Effektes nicht mehr abstreiten: Die Metaanalyse zeigte einen starken Gesamteffekt von $d = .70$, der statistisch signifikant war ($Z = 6.62$, $p < .001$). Die inhaltliche Kontroverse um Erwartungseffekte trug wesentlich zur **Popularisierung und Weiterentwicklung der Metaanalyse** bei (für eine anschauliche Zusammenfassung der Entwicklung s. Harris, 1991): Zahlreiche statistische Methoden im Bereich der Metaanalyse wurden von Rosenthal selbst vorgeschlagen, der auch eines der ersten Lehrbücher zum Thema verfasste (Rosenthal, 1984).

Während sich Versuchsleiter-Effekte im Labor durch Verblindung völlig ausschließen lassen, ist der **praktische Umgang mit Lehrererwartungs-Effekten** in der Praxis sehr viel schwieriger. Dementsprechend verwundert es nicht, dass die Erforschung von Lehrererwartungseffekten nach wie vor relativ umfassend betrieben wird. Hierbei geht es dann v. a. auch um die Untersuchung der Prozesse, die im Einzelnen zur Bildung von (v. a. negativen) Erwartungen auf Seiten der Lehrerinnen und Lehrer führen sowie um die Prozesse, die im Einzelnen die Verhaltensänderungen der betroffenen Schülerinnen und Schüler auslösen (für aktuelle narrative Reviews dieses Forschungsfeldes s. Jussim & Harber, 2005; Jussim, Robustelli, & Cain, 2009).

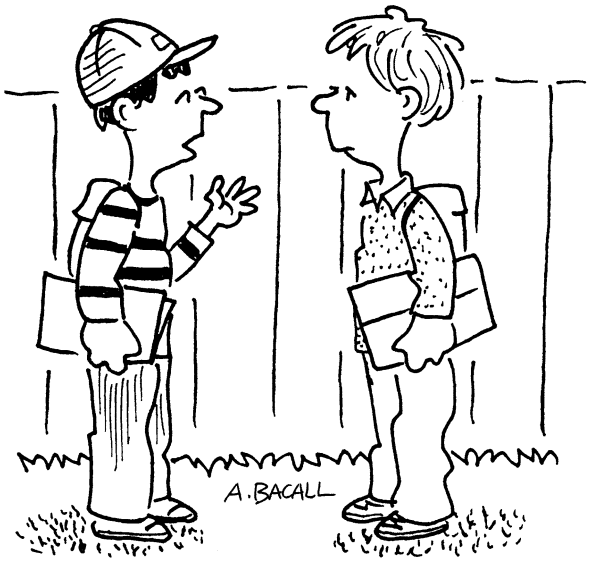
Nach diesen Bemerkungen zum Forschungsfeld soll nun die rechnerische Durchführung einer Metaanalyse an einem didaktisch vereinfachten Beispiel demonstriert werden: Es gilt, fünf fiktive Untersuchungen zur Frage der Beeinflussung von Schülerleistungen durch Lehrererwartungen sachgerecht metaanalytisch zu aggregieren.

16.4.1 Die Primärstudien: Fünf Untersuchungen zum Lehrererwartungs-Effekt

Für eine Metaanalyse zum Lehrererwartungs-Effekt müssten zunächst alle einschlägigen Studien recherchiert werden. Dabei greift man idealerweise mit entsprechenden Suchbegriffskombinationen auf Publikationsdatenbanken zurück (▶ Abschn. 6.1.2), denen man englischsprachige sowie möglichst auch sonstige nationalsprachige Beiträge entnimmt (Maßnahme gegen den Language-Bias). Zudem nutzt man aber auch weitere Recherchewege wie z. B. Kollegennetzwerke, um unveröffentlichte Untersuchungen und graue Literatur zu finden (Maßnahme gegen den Publication-Bias). Da es sich hier um ein sehr großes Forschungsfeld handelt, in dem bereits zahlreiche Metaanalysen vorliegen, wäre eine Einschränkung des Fokus sinnvoll und notwendig, etwa eine Betrachtung nur der neuesten Studien der letzten drei Jahre oder eine Beschränkung auf Studien, die sich auf eine spezielle Zielgruppe (z. B. Gymnasiallehrer; Schülerinnen und Schüler bestimmter ethnischer Minoritäten) beziehen. Möglich wäre auch ein Anknüpfen an aktuelle öffentliche Debatten um schulische Geschlechtseffekte in der Weise, dass der oft behaupteten These, weibliche Lehrkräfte würden Jungen gegenüber Mädchen bewusst oder unbewusst – z. B. im Sinne von Erwartungseffekten – benachteiligen, metaanalytisch nachgegangen wird. (Der aktuelle Forschungsstand widerspricht laut Primärstudien und narrativen Reviews bislang recht einhellig der These einer geschlechtsspezifischen Benachteiligung von Jungen durch weibliche Lehrkräfte: Hadjar, 2011; Hadjar & Lupatsch, 2011; Hannover & Kessels, 2011; Neugebauer, Helbig, & Landmann, 2011, ◘ Abb. 16.3.)

Auf die gezielte Studiensuche folgt die Studienselektion: Die gefundenen Studien werden hinsichtlich ihrer methodischen Qualität beurteilt. Nur Studien, die methodische Mindeststandards der internen Validität und Teststärke erfüllen, sollten in die Metaanalyse eingehen.

In unserem didaktisch vereinfachten Beispiel wurden schließlich folgende fünf Studien zum Lehrererwartungs-Effekt als inhaltlich einschlägig und methodisch fundiert für die Metaanalyse ausgewählt (in Wirklichkeit umfassen Metaanalysen meist deutlich mehr Studien). Man beachte hinsichtlich des Äpfel-und-Birnen-Problems der Metaanalyse, dass die Studien sich alle gleichermaßen auf Lehrererwartungs-Effekte beziehen, aber die unabhängigen und abhängigen Variablen dabei leicht variieren. So wird eine positive Lehrererwartung (Ausprägung der unabhängigen Variablen) auf ganz unterschiedlichem We-



"My teacher says I'm an underachiever, but I think she's an overexpecter."

Abb. 16.3 Dass Lehrererwartungs-Effekte existieren, ist unstrittig. Ihre genaue Wirkungsweise jedoch teilweise noch unbekannt. (© Aaron Bacall/Search ID: aba0085, Rights Available from CartoonStock.com)

ge hergestellt bzw. operationalisiert (z. B. Information über vermeintlichen IQ, über Zeitungsleseverhalten, über Gymnasialempfehlung oder über gute Leistungen älterer Geschwister). Ebenso variieren die Operationalisierungen der abhängigen Variablen (z. B. Beurteilung von Aufsätzen, Schätzung von Noten, Bewertung eines Wissenstests, Benotung im Fach Mathematik). Diese Variationen werden hier nicht als Äpfel-und-Birnen-Problem aufgefasst, sondern stellen gerade die Stärke der Metaanalyse dar: Nämlich durch Aggregierung von Einzelstudien mit jeweils ganz spezifischen Operationalisierungen desselben Effektes zu einer verallgemeinerbaren Aussage zu kommen. Letztlich ist die Grenze für eine angemessene Vergleichbarkeit der Studien inhaltlich zu begründen und kann durchaus kontrovers bewertet werden. Die Studien dieses überspitzten didaktischen Beispiels zeichnen sich durch vergleichsweise große Unterschiedlichkeit aus.

Untersuchung 1

100 Schülern wird zufällig ein IQ-Wert (X) im Bereich $80 \leq \text{IQ} \leq 120$ zugeordnet. Die fiktiven IQ-Werte von jeweils 20 Schülern werden 5 Lehrern als „wahre" Intelligenzwerte mitgeteilt, die mit diesen Hintergrundinformationen Aufsätze der Schüler auf 10-Punkte-Skalen beurteilen (Y). Man ermittelt zwischen den fiktiven IQ-Werten und den Aufsatzbewertungen eine durchschnittliche Korrelation von $r_{xy} = .40$.

Untersuchung 2

4 Klassenlehrer von 86 Abiturienten werden vor Bekanntwerden der Abschlussnoten gebeten, die Leistungen ihrer Schüler (Y: Durchschnittsnoten) zu schätzen. Sie erhalten zusätzlich die Ergebnisse eines Intelligenztests, die allerdings bei 50 % der Schüler um 10 IQ-Punkte zu hoch (X_+) und bei 50 % um 10 IQ-Punkte zu niedrig angegeben werden (X_-). Als mittlere Durchschnittsnoten errechnet man $\bar{y} = 2.50$ für die X_+-Gruppe und $\bar{y} = 3.10$ für die X_--Gruppe bei einer Streuung von $\hat{\sigma} = 1.10$ für die geschätzten Durchschnittsnoten. Die Hypothese, nach der die Leistungen der vermeintlich intelligenteren Schüler besser eingeschätzt werden als die Leistungen der vermeintlich weniger intelligenten Schüler, konnte über einen t-Test für unabhängige Stichproben bestätigt werden ($t = -2.53$, $\text{df} = 84$).

Untersuchung 3

60 Schüler werden in politischer Weltkunde getestet, und 2 Lehrer A und B beurteilen die Testprotokolle anhand einer 10-Punkte-Skala (X). Lehrer A erhält zusätzlich die Information, dass es sich bei den 60 Schülern um eine einfache Zufallsauswahl von Schülern handele, und Lehrer B wird mitgeteilt, dass sich die Stichprobe nur aus Schülern zusammensetzt, die regelmäßig eine Tageszeitung lesen.

Zwischen den Urteilen der beiden Lehrer ermittelt man eine Korrelation von $r_{AB} = .63$. Als Mittelwerte und Varianzen werden genannt: $\bar{X}_A = 7.20$; $\hat{\sigma}^2_{X(A)} = 4.80$; $\bar{X}_B = 7.90$; $\hat{\sigma}^2_{X(B)} = 4.00$. Die Hypothese, nach der die (vermeintlich) zeitungslesenden Schüler besser beurteilt werden als „ normale" Schüler, konnte mit einem t-Test für abhängige Stichproben bestätigt werden ($\alpha = .01$).

Untersuchung 4

In dieser Studie geht es um die Benotung der Hausarbeitshefte von 180 Schülern der 6. Schulklasse. 90 zufällig ausgewählte Hausarbeitshefte werden 5 Lehrern und die restlichen Hefte 5 Lehrerinnen zur Beurteilung vorgelegt. Per Zufall wird jeweils ein Drittel der von einem Lehrer oder einer Lehrerin beurteilten Hefte mit dem Vermerk „mit Gymnasialempfehlung", „ohne Gymnasialempfehlung" oder „Gymnasialempfehlung fraglich" versehen. Abhängige Variable (X) ist pro Schüler der Durchschnittswert der 5 Lehrer- bzw. Lehrerinnenurteile (als Schulnoten).

Die varianzanalytische Auswertung der Daten (Faktor A: Geschlecht des Lehrers, Faktor B: Art der Empfehlung) führte zu einem signifikanten Haupteffekt B ($F_B = 7.20$; $\text{df}_{\text{Zähler}} = 2$, $\text{df}_{\text{Nenner}} = 174$). Die Mittelwerte für die 3 Stufen des Faktors B lauten: Mit Gymnasialempfehlung: $\bar{X}_1 = 2.20$; ohne Gymnasialempfehlung:

Tabelle 16.5 Ergebnis der 5. Untersuchung

	Positive Information	Negative Information
Leistung verbessert	10	10
Leistung verschlechtert	30	50
keine Leistungsänderung	60	40
Summe	100	100

$\bar{X}_2 = 3.40$; Gymnasialempfehlung fraglich: $\bar{X}_3 = 2.90$. Die 180 Noten haben eine Streuung von $\hat{\sigma} = 1.74$.

Untersuchung 5

10 Mathematiklehrer haben nach Ablauf eines Schulhalbjahres die Leistungen ihrer 200 Schüler benotet. Zu Beginn des zweiten Schulhalbjahres berichten vom Untersuchungsleiter instruierte Fachkollegen den „Experimentalkollegen" über ihre Erfahrungen mit den schulischen Leistungen von älteren Geschwistern der 200 Schüler. Nach deren Angaben haben 100 zufällig ausgewählte Schüler einen Bruder bzw. eine Schwester mit hervorragenden schulischen Leistungen (positive Informationen) und die restlichen Schüler Geschwister mit schlechten Leistungen (negative Informationen). Es interessiert die Frage, ob das fiktive Urteil der Kollegen über die Leistungsfähigkeit der Geschwister die Benotung der 200 Schüler durch die 10 Lehrer beeinflusst.

Hierfür wurden die Mathematiknoten der 200 Schüler nach dem ersten Schulhalbjahr mit den Noten nach dem zweiten Halbjahr verglichen, um festzustellen, bei welchen Schülern die Mathematiklehrer eine Leistungsverbesserung, Leistungsverschlechterung bzw. keine Leistungsveränderung konstatiert hatten. Die Ergebnisse zeigt Tab. 16.5.

Für diese Kontingenztafel wird $\chi^2 = 9.40$ (df $= 2$) errechnet, was die Hypothese bestätigt, dass positive bzw. negative Hintergrundinformationen über die Schüler das Lehrerurteil beeinflussen.

16.4.2 Berechnung der Δ-Maße für alle fünf Studien

Um die von der Metaanalyse angezielte Bestimmung des Gesamteffektes vorzubereiten, müssen zunächst die statistischen Befunde aller fünf Primärstudien in ein einheitliches Effektgrößenmaß – hier: das **Δ-Maß** – überführt werden (▶ Abschn. 16.2.5).

Effektgröße von Untersuchung 1

In Untersuchung 1 wurde für $n_1 = 100$ Schüler eine Produkt-Moment-Korrelation errechnet ($r_{xy} = .40$), die nach Gl. 16.2 dem Δ-Wert direkt entspricht. Wir erhalten also $\mathbf{\Delta_1 = .40}$.

Effektgröße von Untersuchung 2

Diese Untersuchung vergleicht die Lehrerurteile über zwei gleich große Stichproben ($n_+ = n_- = 43$ bzw. $n = 86$) anhand eines t-Tests für unabhängige Stichproben. Wir errechnen zunächst $\hat{\delta}$ nach Ziffer 1 in Tab. 14.5

$$\hat{\delta} = \frac{2.5 - 3.1}{1.1} = -0.55$$

und bestimmen für diesen Wert die punktbiseriale Korrelation nach Gl. 16.3

$$r_{pbis} = \frac{-0.55}{\sqrt{0.55^2 + 4}} = -.26.$$

Wegen $p = q = 0.5$ ermittelt man dann nach Gl. 16.6 die folgende, dem Δ-Wert entsprechende biseriale Korrelation:

$$\Delta_2 = r_{bis} = 1.25 \cdot -.26 = -.33.$$

Da das Ergebnis der Untersuchung hypothesenkonform ist (positive Informationen führen zu besseren Noten), ignorieren wir das negative Vorzeichen und setzen $\mathbf{\Delta_2 = .33}$.

Die punktbiseriale Korrelation ermittelt man – bis auf Rundungsungenauigkeiten – auch über Gl. 16.7

$$r_{pbis} = \sqrt{\frac{-2.53^2}{-2.53^2 + 84}} = .266.$$

Effektgröße von Untersuchung 3

Das Ergebnis des t-Tests für abhängige Stichproben in der 3. Untersuchung ($n_3 = 60$) wird wie folgt transformiert:

Wir errechnen zunächst nach Gl. 14.3 die durchschnittliche Streuung

$$\hat{\sigma} = \sqrt{\frac{4.8 + 4.0}{2}} = 2.1.$$

Mit diesem Wert und $r = .63$ ergibt sich analog nach Gl. 14.1

$$\hat{\delta}' = \frac{7.2 - 7.9}{2.1 \cdot \sqrt{2 \cdot (1 - 0.63)}} = -.39$$

bzw. nach Gl. 16.9

$$\Delta_3 = \frac{-0.39}{\sqrt{0.39^2 + 1}} = -.36.$$

Auch dieses Ergebnis unterstützt die metaanalytische Hypothese, d. h., wir setzen $\mathbf{\Delta_3 = .36}$.

Effektgröße von Untersuchung 4

Aus der vierten Untersuchung ($N = 180$) benötigen wir für die metaanalytische Integration die den Faktor B (Art der Gymnasialempfehlung) betreffenden Informationen. In der Metaanalyse interessiert jedoch vorrangig das Ausmaß der Voreingenommenheit bzw. der Urteilsbeeinflussung durch positive bzw. negative Schülerinformationen. Deshalb ist der mit dem Einzelvergleich B_1 (mit Gymnasialempfehlung) vs. B_2 (ohne Gymnasialempfehlung) verbundene Effekt zu ermitteln. Die auf diesen Einzelvergleich bezogene Quadratsumme ergibt sich nach Gl. 16.20 zu

$$QS_{D(B)} = \frac{30.20 \cdot (2.20 - 3.40)^2}{2} = 43.20.$$

Da sich der Einzelvergleich auf Faktor B bezieht, wurde q in Gl. 16.20 durch $p = 2$ ersetzt. Wir könnten nun über Gl. 16.21 ein partielles $\hat{\eta}_p^2$ für den Einzelvergleich errechnen, der jedoch den „realistischen Varianzanteil" des Einzelvergleiches überschätzen würde (▶ Abschn. 14.2.3). Wir relativieren deshalb – in Analogie zu Gl. 16.16 – den Einzelvergleich an der Merkmalsstreuung bzw. der totalen Quadratsumme, die sich wie folgt aus der Merkmalsstreuung $\hat{\sigma} = .174$ ergibt:

$$QS_{tot} = df_{tot} \cdot \hat{\sigma}^2 = 179 \cdot 1.74^2 = 541.94.$$

Damit ergibt sich

$$\Delta_4 = \hat{\eta}_D = \sqrt{\frac{43.2}{541.94}} = .28.$$

Da die Mittelwerte $\bar{B}_1$ und $\bar{B}_2$ jeweils auf 60 Schülern basieren, liegen dem Einzelvergleich $\mathbf{\Delta_4 = .28}$ insgesamt $n_4 = 120$ Schüler zugrunde.

Effektgröße von Untersuchung 5

In der fünften Untersuchung resultierte für die 3×2-Kontingenztafel $\chi^2 = 9.49$ ($n_5 = 200$). Will man das auf die gesamte Tafel bezogene Ergebnis metaanalytisch integrieren, erhält man nach Gl. 16.14:

$$R = \sqrt{\frac{9.49}{200}} = .22.$$

Man könnte jedoch auch argumentieren, dass die 100 Schüler, die sich im Lehrerurteil nicht (oder nicht merklich) verändert haben, eigentlich gegen die metaanalytische Hypothese sprechen. Um diese Fälle zu „neutralisieren", werden sie – getrennt für die Kategorien „positive Informationen" und „negative Informationen" – zu gleichen Teilen in den Kategorien verbessert und verschlechtert mitgezählt. Damit resultiert die in Tab. 16.6 dargestellte Vierfeldertafel.

Tabelle 16.6 Korrigiertes Ergebnis der 5. Untersuchung

	Positive Informationen	Negative Informationen
Verbessert	40	30
Verschlechtert	60	70
Summe	100	100

Auch in dieser Tafel deutet sich die Tendenz an, dass v. a. negative Informationen das Lehrerurteil hypothesengemäß beeinflussen. Den mit dieser Vierfeldertafel verbundenen Effekt ermitteln wir nach Gl. 16.12, wobei wir – um die in Tab. 10 im ▶ Webangebot zu diesem Buch tabellierten Werte ($\phi = 2 \arcsin \sqrt{x}$) benutzen zu können – Gl. 16.13 wie folgt umstellen

$$d = (p \cdot q)^{1/2} \cdot (2 \arcsin \pi_1^{1/2} - 2 \arcsin \pi_2^{1/2}).$$

Wir erhalten $p = q = 100/200 = 0.5$ und schätzen den Anteil für „verbessert" in der 1. Stichprobe mit $\pi_1 = 40/100 = 0.4$ und in der 2. Stichprobe mit $\pi_2 = 30/100 = 0.3$. Nach Tab. 10 (▶ Webangebot zu diesem Buch) ergeben sich $2 \arcsin \sqrt{0.40} = 1.3694$ und $2 \arcsin \sqrt{0.30} = 1.1593$, was zu

$$d = \sqrt{0.5 \cdot 0.5} \cdot (1.3694 - 1.1593) = .1051$$

führt. Für das Korrelationsäquivalent Δ erhält man also

$$\Delta_5 = \frac{e^{2 \cdot 0.1051} - 1}{e^{2 \cdot 0.1051} + 1} = .105$$

($e = 2.7183$).

Dieser Δ-Wert $\mathbf{\Delta_5 = .10}$ entspricht dem ϕ-Koeffizienten der Vierfeldertafel, für die man $\chi^2 = 2.198$ ermittelt: $\sqrt{\frac{2.198}{200}} = 0.105$.

16.4.3 Berechnung und Signifikanzprüfung des Lehrererwartungs-Gesamteffektes

Als besten Schätzwert für den Populationsgesamteffekt Zeta ζ berechnen wir $\bar{Z}$ nach Gl. 16.25 als gewichteten Mittelwert der einzelnen Effektgrößenschätzungen Z_i mit entsprechenden Gewichtungsfaktoren nach dem Fixed-Effect-Modell. Die einzelnen Bestandteile der Formel lassen sich übersichtlich tabellarisch darstellen (Tab. 16.7): Für jede Studie benötigen wird die vereinheitlichten Effektgrößenmaße (Δ_i), die Stichprobenumfänge

Tabelle 16.7 Rechenschritte zur Gesamteffektschätzung und für den Homogenitätstest

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Unters.-Nr. (i)	Δ_i	n_i	Z_i	v_i	w_i	$w_i \cdot Z_i$	$W_i \cdot Z_i^2$
1	0.40	100	0.424	0.0103	97	41.128	17.4383
2	0.33	86	0.343	0.0120	83	28.469	9.7649
3	0.36	60	0.377	0.0175	57	21.489	8.1014
4	0.28	120	0.288	0.0085	117	33.696	9.7044
5	0.10	200	0.100	0.0051	197	19.700	1.9700
Summe		$N = 566$			**551**	**144.482**	**46.9789**

(n_i), die Z-transformierten Effektgrößenmaße (Z_i), die Varianz innerhalb der Studien bzw. den quadrierten Standardfehler (v_i) und die daraus im Fixed-Effect-Modell abgeleiteten Gewichtungsfaktoren für die Studien (w_i). Es ist erkennbar, dass Studie 3 die vergleichsweise größte interne Varianz bzw. den größten Stichprobenfehler aufweist ($v_3 = 0.0175$) und somit mit dem vergleichsweise geringsten Gewichtungsfaktor ($w_3 = 57$) in die Gesamteffektschätzung eingeht, während Studie 5 mit der geringsten Varianz den größten Gewichtungsfaktor ($w_5 = 197$) erhält.

Anhand der in Tab. 16.7 dargestellten Werte berechnen wir $\bar{Z}$ nach Gl. 16.25:

$$\bar{Z} = \frac{144.482}{551} = 0.2622.$$

Ob dieser gewichtete Durchschnittswert signifikant von Null abweicht, überprüfen wir mit Gl. 16.27

$$Z = 0.2622 \cdot \sqrt{551} = 6.15.$$

Dieser Wert ist gem. Tab. 1 im ▶ Webangebot zu diesem Buch signifikant. Der **signifikante Effekt** besagt rechnerisch, dass wir davon ausgehen können, dass die Leistungsbeurteilung von Schülern überzufällig von den Erwartungen der Lehrer abhängt. Tab. 9 im ▶ Webangebot zu diesem Buch entnehmen wir, dass $\bar{Z} = 0.2622$ einer Korrelation von $r = .26$ entspricht. Dieser Zusammenhang wäre nach Tab. 14.5 als nahezu **mittlere Effektgröße** ($r = .30$) zu klassifizieren.

Für das **95 %ige Konfidenzintervall der Gesamteffektgröße** $\bar{Z}$ – bzw. der Korrelation r (beide Werte unterscheiden sich laut Tab. 9, ▶ Webangebot zu diesem Buch, nicht nennenswert) – errechnen wir nach Gl. 16.28

$$\begin{aligned} \mathrm{KI}_\zeta(95\,\%) &= 0.2622 \pm 1.96 \cdot \sqrt{\frac{1}{551}} \\ &= 0.2622 \pm 0.0835 = 0.18 \leq \zeta \leq 0.35. \end{aligned}$$

Da das Konfidenzintervall den Wert Null **nicht** umschließt, wird das Ergebnis des Signifikanztestes bestätigt. Insgesamt wäre auf der Basis der fünf Primärstudien – Homogenität vorausgesetzt (▶ Abschn. 16.4.4) – somit ein **kleiner bis mittlerer metaanalytischer Gesamtzusammenhang zwischen Lehrererwartung und Schülerleistung** in der Population zu konstatieren.

Die Ergebnisse dieser Metaanalyse wurden außerdem mit dem Metaanalyseprogramm **„Comprehensive Meta Analysis" CMA** (Borenstein et al., 2005) berechnet. Dazu wurden die Z-Werte der einzelnen Studien als Rechenbasis herangezogen. Die Rechenergebnisse und der Forest-Plot sind in Abb. 16.4 im Überblick dargestellt.

Da der Signifikanztest für den Gesamteffekt signifikant ausgefallen ist, können wir schließen, dass er vermutlich über ausreichende Teststärke verfügte. Eine A-posteriori-Teststärkeanalyse ist jedoch immer dann indiziert, wenn ein Signifikanztest zu einem nicht-signifikanten Ergebnis führt. In diesem Fall ist sicherzustellen, dass das nicht-signifikante Ergebnis tatsächlich einen Hinweis auf die Nichtexistenz des Effektes liefert und nicht nur darauf basiert, dass der Test aufgrund geringer Stichprobengröße (d. h. in der Metaanalyse: zu geringer Primärstudienanzahl) zu schwach war, um einen Populationseffekt nachweisen zu können. Aus Gründen der Anschaulichkeit wird das Vorgehen bei der A-posteriori-Teststärkeanalyse hier wiederum an unserem Zahlenbeispiel demonstriert, obwohl wir im Vorfeld **bereits wissen**, dass die Teststärke relativ hoch gewesen sein muss.

Name der Studie	Statistik für jede Studie						
	Fisher's *Z*	Standardfehler	Varianz	Untergrenze	Obergrenze	*Z*-Wert	*p*-Wert
Untersuchung 1	10.424	0.102	0.010	0.225	0.623	4.176	< .0001
Untersuchung 2	20.343	0.110	0.012	0.128	0.558	3.125	.002
Untersuchung 3	30.377	0.132	0.018	0.117	0.637	2.846	.004
Untersuchung 4	40.288	0.092	0.009	0.107	0.469	3.115	.002
Untersuchung 5	50.100	0.071	0.005	–0.040	0.240	1.404	.160
Gesamt	0.262	0.043	0.002	0.179	0.346	6.155	< .0001

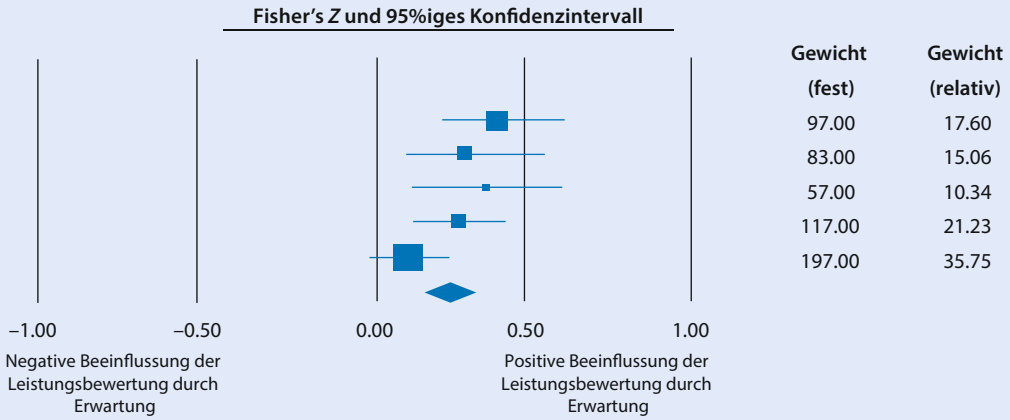


Abb. 16.4 Tabellarische Übersicht und Forest-Plot der Metaanalyse der fünf Studien zum Lehrererwartungs-Effekt (Fixed-Effect-Modell)

Für die A-posteriori-Teststärkeanalyse des Signifikanztestes für den Gesamteffekt benötigen wir folgenden Nichtzentralitätsparameter λ:

$$\lambda = \frac{\zeta - \zeta_0}{\sqrt{V.}} \quad \text{mit } V. = 1/\sum_{i=1}^{k} W_i. \tag{16.37}$$

ζ entspricht dem gem. H_1 angenommenen Fisher-Z-Wert und ζ_0 dem H_0-Parameter, der üblicherweise Null gesetzt wird. Im Beispiel (mit $k = 5$) ergibt sich (Tab. 16.7):

$$V. = \frac{1}{551} = 0.0018$$

Wenn wir $\zeta = 0.3$ setzen, erhält man

$$\lambda = \frac{0.3 - 0}{\sqrt{0.0018}} = 7.07.$$

Bei Gültigkeit von H_1 ist z (gem. Gl. 16.27) normalverteilt mit dem Erwartungswert $\lambda = 7.07$ und einer Streuung von 1 (Hedges & Pigott, 2001, S. 206).

Die Teststärke ergibt sich zu

$$1 - \beta = 1 - \phi(C_\alpha - \lambda) \tag{16.38}$$

mit c_α = kritischer z-Wert der Standardnormalverteilung für $\alpha = .05$ (.01) und einseitigem Test. $\phi(X)$ ist die Verteilungsfunktion der Standardnormalverteilung an der Stelle x. Im Beispiel setzen wir $c_\alpha = 1.64$ (einseitiger Test, $\alpha = .05$), d. h., wir erhalten

$$1-\beta = 1-\phi(1.64-7.07) = 1-\phi(-5.43) \approx 1-0.00 \approx 1.$$

Der Wert -5.43 schneidet von der Standardnormalverteilung eine Fläche von nahezu (aber nicht exakt!) Null ab, d. h., wir erhalten $1 - \beta \approx 1.00$. Der Signifikanztest für den Gesamteffekt hat bei den gegebenen Rahmenbedingungen also eine beobachtete Teststärke von nahezu 100 %. Die hohe Teststärke trotz sehr geringer Studienzahl ($k = 5$) erklärt sich dadurch, dass von einer mittleren Effektgröße in der Population ausgegangen wurde ($\zeta = 0.3$), was im sozialwissenschaftlichen Bereich inhaltlich als recht beachtlich zu betrachten ist.

16.4.4 Prüfung der Homogenität der fünf Δ-Maße

Als metaanalytisches Modell wird im vorliegenden didaktischen Beispiel aus Gründen der Einfachheit ein Fixed-Effect-Modell (▶ Abschn. 16.2.4) unterstellt. Dieses geht von Homogenität der Studien aus, d. h. davon, dass alle Effektgrößen der fünf Studien denselben Populationseffekt schätzen und Unterschiede zwischen den Studien nur auf Stichprobenfehler zurückgehen bzw. nicht statistisch signifikant sind.

Die angenommene Homogenität der Studien ist zu prüfen. Alle Informationen, die wir für die Durchführung des **Homogenitätstestes** nach Gl. 16.29, 2. Teil benötigen, sind in Tab. ◘ Tab. 16.7 enthalten.

Mit diesen Werten ergibt sich als Q-Wert für den Homogenitätstest der Effektgrößenmaße:

$$Q = 46.9789 - \frac{144.482^2}{551} = 9.0931.$$

Um auf der Basis dieses Q-Wertes die Nullhypothese, dass sich die einzelnen Effektgrößen **nicht** unterscheiden bzw. dass Homogenität vorliegt, annehmen zu können, muss das Ergebnis auf dem α-Fehler-Niveau von 10 % nicht-signifikant sein. Zudem muss eine ausreichende Teststärke gegeben sein.

Für ein nicht-signifikantes Ergebnis müsste der empirische Q-Wert, der einer χ^2-Verteilung folgt, kleiner sein als der kritische χ^2-Wert. Mit $Q = 9.09$ und χ^2_{krit} ($\alpha = .10, df = 4$) $= 7.78$ zeigt der Homogenitätstest leider ein **signifikantes Ergebnis**: Die fünf einzelnen Effektgrößen sind als **heterogen** zu betrachten.

Mittels des bereits genannten Metaanalyse-Programms CMA wird $Q = 9.093$, $df = 4$ und $p = .059$ ermittelt. Da CMA die Teststatistiken für eine zweiseitige Testung berechnet, ist auch bei der softwaregestützten Durchführung der Homogenitätstest signifikant.

Selbst wenn das Ergebnis nicht-signifikant ausgefallen wäre, hätte man dennoch nicht auf Homogenität schließen können. Denn die **A-posteriori-Teststärkeanalyse des Homogenitätstests** (zur A-posteriori-Teststärkeanalyse ▶ Abschn. 14.2.3) ist bei $N = 566$ offensichtlich viel zu gering: Wenn wir von der H_1: $|\zeta - \zeta_i| \leq 0.1$ ausgehen (d. h. gemäß Alternativhypothese geringe Heterogenität vermuten), ergibt sich nach Gl. 16.39

$$\lambda = 551 \cdot 0.1^2 = 5.51.$$

Hierfür errechnet man (mit $q = .778$, $df = 4$ und $nc = \lambda = 5.51$) über Gl. 16.30 eine Teststärke von $1 - \beta = .48$. Diese Teststärke wäre **viel zu klein**, um bei einem nichtsignifikanten Ergebnis die Nullhypothese (Homogenität der Studieneffekte) mit einer akzeptablen β-Fehler-Wahrscheinlichkeit annehmen zu können.

Zusammenfassend ist also davon auszugehen, dass die Studieneffekte heterogen sind und deshalb nicht einfach zu einem Gesamteffekt aggregiert werden sollten. Stattdessen sind folgende Alternativen für das vorliegende Beispiel möglich:

- Im Rahmen einer Moderatorvariablen-Analyse könnte man prüfen, ob sich die fünf Studien in **homogene Subgruppen** einteilen lassen, für die dann jeweils ein gruppenspezifischer Gesamteffekt auf der Basis des Fixed-Effect-Modells zu berechnen und auf Signifikanz zu prüfen wäre. Das entsprechende Vorgehen wird unten erläutert.
- Es könnte eine Metaanalyse nach dem **Random-Effects-Modell** oder dem **Mixed-Effects-Modell** durchgeführt werden, wobei die Modellvoraussetzungen konzeptuell zu begründen sind (▶ Abschn. 16.2.4). Diese Vorgehensweisen werden hier nicht vertieft (zur Bestimmung und Interpretation der Heterogenität von Studieneffekten siehe auch Borenstein et al., 2009, Kap. 16).

Dass der Effektgrößen-Homogenitätstest zu geringe Teststärke aufweist, ist ein Problem, das in der Praxis leider häufiger auftritt. Deswegen ist eigentlich für Metaanalysen – ebenso wie für alle empirischen Studien – eine **A-priori-Teststärkeanalyse des Homogenitätstestes** empfehlenswert (zur A-priori-Teststärkeanalyse ▶ Abschn. 14.1.2): Statt sich im Nachhinein von zu geringer Teststärke „überraschen" zu lassen, sollte man idealerweise bereits die Planung einer Metaanalyse so anlegen, dass am Ende ausreichende Teststärke vorhanden ist. Das dafür notwendige Vorgehen sei kurz skizziert:

Angenommen, aufgrund von Recherchen im Vorfeld der Metaanalyse zum Lehrererwartungs-Effekt seien 20 einschlägige Untersuchungen realistisch. Ferner soll davon ausgegangen werden, dass durchschnittlich pro Studie 50 Schüler untersucht wurden. Eine weitere Annahme besagt, dass die Abweichungen der studienspezifischen Effektparameter (ζ_i) vom Gesamtparameter (ζ) nicht größer als 0.1 sein dürfen (H_1: $|\zeta - \zeta_i| \leq 0.1$). In Analogie zu Gl. 16.29 errechnen wir

$$\lambda = \sum_{i=1}^{k} \left[\frac{(\zeta_i - \zeta^2)}{V_i} \right] = 20 \cdot 47 \cdot 0.1^2 = 9.40. \quad (16.39)$$

Dieser Wert entspricht nach der Klassifikation von ▶ Abschn. 16.2.7 einer schwachen ($0.33 \cdot 19 = .627$) bis mittleren ($0.67 \cdot 19 = 12.73$) Heterogenität.

Die Teststärke des Homogenitätstestes errechnen wir über Gl. 16.30 mit $q = 27.20$ (Tab. 8 in Anhang F im ▶ Webangebot zu diesem Buch, für $\alpha = .10$ und $df =$

19), df = 19 und nc = λ = 9.40. Es resultiert $1 - \beta = 1 - 0.48 = .52$.

Der Homogenitätstest hat also mit den genannten Vorgaben eine Teststärke von 52 %. Sollte der Homogenitätstest mit α = .10 nicht signifikant werden, würde man bei Ablehnung der H_1 mit einer Wahrscheinlichkeit von 48 % einen β-Fehler riskieren. Auf der Grundlage dieses β-Fehler-Risikos sollte die H_0 (identische Studieneffekte) nicht als „bestätigt" gelten. Anders formuliert: Die Planungsphase hätte deutlich gemacht, dass der Homogenitätstest mit den genannten Vorgaben (λ = 9.40; α = .10; df = 19) nicht geeignet ist, eine Entscheidung zugunsten von H_0 zu rechtfertigen, d. h., auf eine Metaanalyse sollte verzichtet werden.

Bei mittlerer Heterogenität als H_1 (λ = 12.73) ergibt sich eine Teststärke von 66 % (β = .34) und bei starker Heterogenität (λ = 19) von 85 % (β = .15). Auch diese Werte sind also nicht geeignet, eine Metaanalyse zu rechtfertigen. Erst wenn **mehr als 100 Studien** zu aggregieren sind, hat der Homogenitätstest eine akzeptable Teststärke und damit – bei einem nicht-signifikanten Ergebnis und schwacher Heterogenität als H_1 – eine relativ geringe β-Fehler-Wahrscheinlichkeit, die für eine fälschliche „Bestätigung" von H_0 tolerierbar wäre (für df = 100 und λ = 47.47 ergibt sich eine Teststärke von $1 - \beta = .94$ bzw. β = .06). Ähnlich wie auch bei Primärstudien stellt sich heraus, dass der notwendige Mindeststichprobenumfang für ausreichende Teststärke oft über den forschungsökonomisch realisierbaren Stichprobenumfängen liegt (▶ Abschn. 14.3).

16.4.5 Moderatorvariablen-Analyse wegen Heterogenität der einzelnen Effektgrößen

Ein mittelstarker und signifikanter Gesamteffekt, wie er im letzten Abschnitt berechnet wurde, kann für das vorliegende Beispiel wegen mangelnder Homogenität der Einzeleffekte bzw. der Δ_i-Werte (▶ Abschn. 16.4.4) **nicht** in Anspruch genommen werden. Stattdessen muss eine **Moderatorvariablen-Analyse** durchgeführt werden, um herauszufinden, ob sich die fünf heterogenen Studien möglicherweise in **homogene Subgruppen** einteilen lassen, für die man dann wiederum gruppenspezifische Gesamteffekte nach dem Fixed-Effect-Modell berechnen könnte. (Auf die Darstellung möglicher alternativer Berechnungen der Metaanalyse nach dem Random-Effects- oder Mixed-Effects-Modell wird hier verzichtet.)

Als Moderatorvariable scheint die **„Art der experimentellen Manipulation der Lehrererwartung"** vielversprechend zu sein. Zur Begründung dieser Auswahl führen wir uns noch einmal vor Augen, wie in den einzelnen Untersuchungen die Erwartungshaltungen der Lehrer beeinflusst wurden: Dies geschah in den Untersuchungen 1, 2 und 4 durch Manipulation der fiktiven Intelligenz bzw. Leistungsfähigkeit der Schüler (Intelligenz in den Untersuchungen 1 und 2, Gymnasialempfehlung in der Untersuchung 4). Es soll deshalb geprüft werden, ob diese 3 Untersuchungen eine homogene Teilgruppe bilden. Die beiden anderen Untersuchungen fallen hier heraus (Untersuchung 3: Tageszeitung lesen; Untersuchung 5: Leistung der Geschwister); sie sollen deshalb zwei weitere „Gruppen" mit je einer Untersuchung bilden, so dass die **Moderatorvariable 3-stufig** ist.

Die inhaltlich begründete Subgruppeneinteilung ist nun statistisch zu prüfen: Mindestens 2 Subgruppen müssen sich untereinander signifikant unterscheiden (signifikanter Q-zwischen-Wert) und die Effekte innerhalb jeder Subgruppe müssen homogen sein (nicht-signifikanter Q-innerhalb-Wert).

Wir berechnen zunächst Q_{zw} nach Gl. 16.31. Dies bereitet unter Zuhilfenahme von ◘ Tab. 16.7 keine größeren Probleme. Wir erhalten

$$W_1 = 97 + 83 + 117 = 297$$

und nach Gl. 16.25

$$\bar{Z}_1 = \frac{97 \cdot 0.424 + 83 \cdot 0.343 + 117 \cdot 0.288}{97 + 83 + 117} = \frac{103.293}{297} = 0.348.$$

Mittels des Metaanalyseprogramms CMA wird $\bar{Z}_1 = 0.364$ und $W = 297$ berechnet.

Die w-Gewichte und $\bar{Z}$-Werte der beiden anderen „Gruppen" sind wegen $k_2 = k_3 = 1$ mit den w- und Z-Werten der Untersuchungen 3 und 5 identisch;

$$W_{2.} = 57,$$
$$\bar{Z}_{2.} = 0.377,$$
$$W_{3.} = 107,$$
$$\bar{Z}_{3.} = 0.100.$$

Eingesetzt in Gl. 16.31 resultiert also:

$$Q_{ZW} = 297 \cdot (0.348 - 0.2622)^2 + 57 \cdot (0.377 - 0.2622)^2 + 197 \cdot (0.100 - 0.2622)^2 = 8.1097.$$

Dieser Wert ist für df = 2 und α = .01 gem. Tab. 8 im ▶ Webangebot zu diesem Buch **signifikant**. Die Berechnung mittels des Metaanalyse-Programms CMA ergibt

Post-hoc-Teststärkenanalyse für den Test auf Homogenität *zwischen* den Gruppen im Rahmen der Subgruppen-Analyse einer Metaanalyse

Im Beispiel ist der Omnibustest für Gruppendifferenzen über den Q-zwischen-Wert signifikant geworden, dementsprechend war die **Teststärke vermutlich ausreichend** (möglich wäre aber auch mangelnde Teststärke und ein signifikantes Ergebnis in Folge eines Alpha-Fehlers).

Im Falle eines **nicht-signifikanten Ergebnisses** ist in jedem Fall eine Post-hoc-Teststärkeanalyse notwendig, um zu prüfen, ob die Teststärke ausreichend war. Sie wäre anhand des Zahlenbeispiels folgendermaßen zu berechnen:

Im Beispiel wurden im Zuge der Moderatorvariablen- bzw. Subgruppen-Analyse 3 Gruppen mit $k_1 = 3$, $k_2 = 1$ und $k_3 = 1$ gebildet mit $w_{1.} = 297$, $w_{2.} = 57$ und $w_{3.} = 197$. Man errechnet also ex post für λ_{zw} über Gl. 16.31

$$\begin{aligned} Q_{ZW} = \lambda_{ZW} &= 297 \cdot (0.2 - 0.275)^2 + 57 \cdot (0.3 - 0.275)^2 \\ &\quad + 197 \cdot (0.4 - 0.275)^2 \\ &= 1.67 + 0.04 + 3.08 = 4.79 \end{aligned}$$

und über Gl. 16.30

$$1 - \beta = 1 - 0.51 = 0.49.$$

Die Teststärke wäre mit 49 % also zu niedrig für einen aussagekräftigen Hypothesentest über die Unterschiede im Lehrererwartungs-Effekt zwischen den drei betrachteten Gruppen, sofern das Ergebnis nicht-signifikant ausgefallen wäre.

Post-hoc-Teststärkenanalyse für den Test auf Homogenität *innerhalb* der Gruppen im Rahmen der Subgruppen-Analyse einer Metaanalyse

Können wir anhand des nicht-signifikanten **Innerhalb-Gruppen-Homogenitätstests** behaupten, die drei Untersuchungen in der einen Subgruppe seien homogen?

Im Beispiel wurden die Studien 1, 2 und 4 zusammengefasst ($k_1 = 3$; $n_{11} = 100$; $n_{21} = 86$; $n_{31} = 120$) und die beiden übrigen Studien bildeten jeweils eine „Gruppe“ ($k_2 = 1$; $n_{12} = 60$; $k_3 = 1$; $n_{13} = 200$). Die „Gruppen“ 2 und 3 brauchen also beim **Innerhalb-Homogenitätstest** nicht berücksichtigt zu werden, da eine „Gruppe“ mit nur einer Studie natürlich homogen ist.

Mit diesen Werten und den oben genannten, gem. H_1 erwarteten Abweichungen der studienspezifischen ζ_{ij}-Werte vom gruppenspezifischen Durchschnittswert ($\bar{\zeta}_j$) errechnet man für die Ex-post-Analyse über Gl. 16.32:

$$\lambda_{in} = Q_{in} = Q_{in1} = 97 \cdot 0.05^2 + 83 \cdot 0.1^2 + 117 \cdot 0.1^2 = 2.24.$$

Wir entnehmen Tab. 8 im ▶ Webangebot zu diesem Buch für df = 2 und $\alpha = .10$ den Wert $\chi^2 = q = 4.61$ und errechnen über Gl. 16.30 $1 - \beta = 1 - 0.64 = .36$. Der Test auf Homogenität der Effekte innerhalb der Gruppe hat mit $1 - \beta = 36\,\%$ also eine **sehr geringe Teststärke**.

Die im Beispiel genannten fünf Studien erwiesen sich für eine Metaanalyse gemäß Fixed-Effect-Modell insgesamt als zu heterogen (▶ Abschn. 16.4.4). Dies gilt nun möglicherweise auch für die Gruppierung nach der Moderatorvariablen „Art der Lehrermanipulation“: Ob die drei in eine Gruppe sortierten Studien homogen sind, kann aufgrund mangelnder Teststärke nicht festgestellt werden.

$Q_{zw} = 9.228$, df = 2, $p = .01$. Die Differenzen in den Berechnungen sind auf Rundungsabweichungen zwischen händischer und computergestützter Berechnung zurückzuführen. Die durch die Moderatorvariable gebildeten Untergruppen unterscheiden sich gemäß dem **Omnibustest für Gruppendifferenzen** überzufällig. Damit ist die erste Voraussetzung für die Subgruppenanalyse erfüllt (zur Teststärke des Verfahrens ▶ Post-hoc-Teststärkenanalyse für den Test auf Homogenität *zwischen* den Gruppen im Rahmen der Subgruppen-Analyse einer Metaanalyse).

Als zweite Voraussetzung für die Subgruppen-Analyse ist über den Q-innerhalb-Wert zu prüfen, ob die drei Untersuchungen in Gruppe 1 homogen sind. Hierfür berechnen wir nach Gl. 16.32 $Q_{in} = Q_{in1} + Q_{in2} + Q_{in3}$. Da die „Gruppen“ 2 und 3 jeweils aus nur einer Untersuchung bestehen, sind $Q_{in2} = Q_{in3} = 0$, so dass in diesem Beispiel $Q_{in} = Q_{in1}$ ist. Für Q_{in1} ergibt sich:

$$\begin{aligned} Q_{in} = Q_{in1} &= 97 \cdot (0.424 - 0.348)^2 \\ &\quad + 83 \cdot (0.343 - 0.348)^2 \\ &\quad + 117 \cdot (0.288 - 0.348)^2 \\ &= 0.9835. \end{aligned}$$

Dieser Wert ist für df = 2 und $\alpha = .05$ nach Tab. 8 im ▶ Webangebot zu diesem Buch **nicht signifikant**.

Zur Kontrolle prüfen wir nach Gl. 16.33:

$$\begin{aligned} Q &= 9.0931 \\ Q_{zw} + Q_{in} &= 8.1097 + 0.9835 = 9.0932. \end{aligned}$$

Die Gleichung ist bis auf Rundungsungenauigkeiten erfüllt.

Das gleiche Bild zeigt sich bei der Berechnung mittels der Metaanalysesoftware CMA: $Q_{in} = 1.299$, df = 2, $p = .522$.

Ein nicht-signikantes Ergebnis besagt hier wunschgemäß, dass die **gebildeten Subgruppen homogene Effektgrößen beinhalten** – allerdings nur, sofern **ausreichende Teststärke** gegeben ist. Andernfalls ist das nicht-signifikante Ergebnis kein Anlass, die H_0 anzunehmen und Homogenität zu postulieren (zur Teststärke des Verfahrens ▶ Post-hoc-Teststärkenanalyse für den Test auf Homogenität *innerhalb* der Gruppen im Rahmen der Subgruppen-Analyse einer Metaanalyse).

Wären die Voraussetzungen nachweislich erfüllt, dass sich mindestens zwei der Subgruppen überzufällig unterscheiden und gleichzeitig die Effekte innerhalb der Subgruppen homogen sind, könnte abschließend im Rahmen der Subgruppenanalyse noch über Gl. 16.27 getestet werden, ob die gruppenspezifischen $\bar{Z}$-Werte signifikant sind bzw. ob in den 3 Subgruppen jeweils ein überzufälliger Lehrererwartungseffekt vorliegt:

$$Z_1 = 0.348 \cdot \sqrt{297} = 6.00^{**}$$
$$Z_2 = 0.377 \cdot \sqrt{57} = 2.85^{**}$$
$$Z_3 = 0.100 \cdot \sqrt{197} = 1.40 \text{ (n.s.)}.$$

Die Metaanalyse-Software CMA errechnet – abgesehen von Rundungsungenauigkeiten – identische Werte:

$$Z_1 = 6.28; \quad p < .0001$$
$$Z_2 = 2.994; \quad p = .003$$
$$Z_3 = 1.408; \quad p = .159.$$

Der durchschnittliche Zusammenhang in der ersten Gruppe ist signifikant ($\alpha = .01$). Für $\bar{Z}_1 = 0.348$ entnehmen wir Tab. 9 im ▶ Webangebot zu diesem Buch eine Korrelation von .335, die nach Tab. 14.5 als mittlerer bis großer Effekt zu klassifizieren wäre. Die zweite „Gruppe“ mit Untersuchung 3 zeigt – wie in ▶ Abschn. 16.4.3 bereits erwähnt – ebenfalls einen signifikanten Zusammenhang, während der Zusammenhang in der dritten „Gruppe“ (Untersuchung 5) nicht signifikant ist.

Inhaltlich würde dieses Ergebnis besagen, dass das Lehrerurteil über die Schülerleistung tatsächlich durch fiktive Intelligenz- bzw. Leistungsunterschiede und durch eine freiwillige, schulrelevante Beschäftigung der Schüler (Zeitunglesen) beeinflusst werden kann. Bezogen auf fiktive Leistungsunterschiede der Geschwister konnte ein entsprechender Effekt nicht nachgewiesen werden. Um hier die Nullhypothese für Gruppe 3 annehmen zu können (bzw. Unterschiede zwischen den 3 Gruppen postulieren zu können), muss wiederum ausreichende Teststärke auch für diese Analyse gesichert sein (dies wird hier nicht weiter ausgeführt). Nicht zuletzt wären neben der Signifikanzaussage hier Konfidenzintervalle für die Effektgrößenschätzungen pro Gruppe relevant.

16.4.6 Interpretation der Ergebnisse der fiktiven Metaanalyse zum Lehrererwartungs-Effekt

Da anhand des fiktiven Zahlenbeispiels der fünf Primärstudien zum Lehrererwartungs-Effekt exemplarisch diverse Berechnungsgänge schrittweise „per Hand“ präsentiert wurden, soll abschließend die Studie noch einmal mit Blick auf die Forschungspraxis reflektiert werden.

Es bleibt festzuhalten, dass die fiktive Metaanalyse auf der Basis von nur fünf Primärstudien vergleichsweise klein angelegt ist und dass bei der Studienplanung eine A-priori-Teststärkeanalyse nicht durchgeführt wurde. Für die Metaanalyse wurde ein einfaches Fixed-Effect-Modell zugrunde gelegt, dessen inhaltliche Plausibilität in diesem Anwendungsfall diskussionswürdig ist (generell werden Random-Effects-Modelle heute konzeptuell und statistisch oft als angemessener eingeschätzt). Aufgrund mangelnder Homogenität der Studien, erwies sich das Fixed-Effect-Modell als auf die fünf Primärstudien nicht anwendbar. Eine Moderatorvariablen- bzw. Subgruppen-Analyse wurde durchgeführt und eine Aufteilung der Studien in 3 Gruppen gemäß der Moderatorvariable „Art der experimentellen Manipulation der Lehrererwartung“ vorgeschlagen. Dabei enthielten 2 „Subgruppen“ je eine Studie und eine Subgruppe 3 Studien. Inhaltlich ist diese Form der Gruppenbildung mit Ein-Element-Gruppen problematisch. Statistisch ließ sich mangels Teststärke die Voraussetzung homogener Subgruppen nicht eindeutig nachweisen.

In der Praxis hätte man die Metaanalyse ausschließlich softwaregestützt durchgeführt (▶ Abschn. 16.5) und die Primärstudienergebnisse sowie die Gesamteffektschätzung auch grafisch in einem Forest-Plot (Abb. 16.4) dargestellt. Zudem wäre es notwendig gewesen, die Grenzen der Aussagekraft der Metaanalyse ausdrücklich zu diskutieren, das umfasst eine kritische Bewertung der Anzahl und Qualität der genutzten Primärstudien sowie des metaanalytischen Vorgehens (z. B. Begründung der Wahl eines Fixed-Effect-Modells, Prüfung der Homogenität bzw. Heterogenität der Studien, Teststärkeproblematik bei allen verwendeten Signifikanztests, vor allem denjenigen, die nicht-signifikante Ergebnisse liefern etc.).

16.5 Software für Metaanalysen

Im Zusammenhang mit quantitativer Datenanalyse wurden bereits drei Ansätze der computergestützten Datenauswertung unterschieden (▶ Abschn. 12.2.3), die auch auf den Bereich der Metaanalyse anzuwenden sind: Das Arbeiten mit Bürosoftware, mit Statistikprogrammpaketen sowie mit spezialisierten Statistikprogrammen (eine Link-Sammlung mit Software und anderen Informationen zur Metaanalyse wie z. B. Weiterbildungsangeboten findet man z. B. auf der Website http://www.meta-analysis.eu von Michael Bosnjak).

16.5.1 Metaanalyse-Add-Ons für Excel

Die Durchführung einer Metaanalyse mit einem Tabellenkalkulationsprogramm wie **MSSoftware Excel** ist heute dank professioneller und kostenloser **Metaanalyse-Add-Ons**, also Programmerweiterungen, möglich. Vorteilhaft ist, dass MS Excel in vielen Institutionen und Privathaushalten bereits vorliegt und somit keine Zusatzkosten für die Beschaffung von Statistiksoftware anfallen.

Ein gut dokumentiertes Metanalyse-Add-On mit großem Funktionsumfang inklusive automatischer Vereinheitlichung von unterschiedlichen Effektgrößenmaßen sowie Erstellung von Forest-Plots ist das an der Universität Manchester, UK, entwickelte **MetaEasy** (Kontopantelis & Reeves, 2009; http://www.statanalysis.co.uk/meta-analysis.html).

Ein weiteres professionelles und mächtiges Metaanalyse-Add-On für Excel ist **MIX** (Meta-analysis In eXcel), das in einer kostenlosen Lite-Version (mittels derer jedoch nur vorgegebene Datensätze analysiert werden können) und einer kostenpflichtigen Professional-Version vertrieben wird (Bax, Yu, Ikeda, Tsuruta, & Moons, 2006; http://www.meta-analysis-made-easy.com).

16.5.2 Metaanalyse-Makros für Statistik-Programmpakete

Statistik-Programmpakete wie **SPSS, SAS oder Stata** enthalten (bislang) keine eigenen Metaanalyse-Funktionen. Allerdings wurden von der Nutzer-Community eine Reihe von **Metaanalyse-Makros/-Kommandos** entwickelt und im Internet bereitgestellt (z. B. von David Wilson: http://mason.gmu.edu/~dwilsonb/ma.html). Die Qualität dieser Makros ist heterogen. Eigenentwicklungen, die nicht von ausgewiesenen Metaanalyse-Fachleuten stammen, sollten mit besonderer Skepsis betrachtet werden.

Für das **Open-Source-Statistikpaket *R*** (http://www.r-project.org) liegen Metaanalyse-Module vor (z. B. „meta“ von Guido Schwarzer: http://cran.r-project.org/web/packages/meta/ und „rmeta“ von Thomas Lumley: http://cran.r-project.org/web/packages/rmeta/) sowie „metafor“ von Wolfgang Viechtbauer: http://www.metafor-project.org).

16.5.3 Spezielle Metaanalyse-Programme

Unter den speziellen, eigenständigen Metaanalyse-Programmen sind zwei kommerzielle besonders bekannt:

- Comprehensive Meta-Analysis **CMA** von Borenstein, Hedges, Higgins und Rothstein (2005): http://www.meta-analysis.com (mit der Testversion dieses Programms können die Rechenbeispiele im vorliegenden Kapitel nachvollzogen werden) und
- **MetaWin** von Rosenberg, Adams und Gurevitch (2007): http://www.metawinsoft.com.

Darüber hinaus existiert als kostenloses Tool:

- der Review Manager **RevMan** der Cochrane Collaboration (2011): http://ims.cochrane.org/revman/ (zur Cochrane Collaboration ▶ Abschn. 16.1.2 „Systematische Reviews in Medizin und Sozialwissenschaften“).

RevMan bietet neben statistischen Funktionen der Metaanalyse auch allgemeine Unterstützung zur Erstellung einer Übersichtsarbeit (z. B. Literaturorganisation) und liegt allen Cochrane-Reviews zugrunde.

Lernquiz

1. Grenzen Sie vier verschiedene Formen der wissenschaftlichen Forschungssynthese charakterisierend voneinander ab!
2. Was versteht man im Kontext der Metaanalyse unter abhängigen Untersuchungsergebnissen? Wie ist mit ihnen zu verfahren?
3. Was ist das Δ-Maß und wozu wird es benötigt?
4. Was sind kombinierte Signifikanztests?
5. Was versteht man unter dem Publication-Bias? Wie geht man mit diesem Problem im Kontext der Metaanalyse um?
6. Wozu dient eine A-priori-Teststärkeanalyse im Kontext der Metaanalyse, welche Probleme sind damit verbunden und wie ist mit ihnen umzugehen?
7. Was unterscheidet eine Metaanalyse von einer Primärstudie und was sind die Vorteile der Metaanalyse?
8. Welche Aussagen sind richtig?
   a. Das Fixed-Effect-Modell geht davon aus, dass sich die zu aggregierenden Primärstudien systematisch unterscheiden.
   b. Das Random-Effects-Modell postuliert, dass in jeder Primärstudie eine andere Populationseffektgröße geschätzt wird.
   c. Ob ein Fixed-Effect- oder ein Random-Effects-Modell zum Einsatz kommen soll, ist vor der Durchführung der Metaanalyse inhaltlich-konzeptuell zu entscheiden.
   d. Im Fixed-Effect-Modell und im Random-Effects-Modell werden Studien mit geringem Stichprobenumfang geringer gewichtet.

9. Welche Aussagen sind falsch?
   a. Die Teststärke einer Metaanalyse übertrifft durch den größeren Gesamtstichprobenumfang die Teststärke der Primärstudien.
   b. Werden die unabhängigen Variablen in den Primärstudien unterschiedlich operationalisiert, sollte die Art der Operationalisierung als Moderatorvariable in der Metaanalyse berücksichtigt werden.
   c. Eine Zusammenfassung von Effektgrößen im Fixed-Effect-Modell ist nur sinnvoll, wenn die untersuchungsspezifischen Effektgrößen zumindest schwach heterogen sind.
   d. Eine vergleichbare Operationalisierung der abhängigen Variablen ist bei der Metaanalyse im Random-Effects-Modell nicht notwendig.
10. In welchen inhaltlichen Forschungsfeldern wurde die quantitative Metaanalyse ursprünglich entwickelt und wo kommt sie heute vermehrt zum Einsatz?

Übungsaufgaben

1. Laden Sie sich die Metaanalyse-Software Comprehensive Meta Analysis (CMA) als Trialversion herunter (http://www.meta-analysis.com). Führen Sie das Tutorial für die Dateneingabe und die Datenanalyse durch und legen Sie einen Datensatz mit den Basisdaten für die 6 Korrelationsstudien in ► Abschn. 16.2.9 „Präsentation der Ergebnisse einer Metaanalyse als Forest-Plot“ an. Erstellen Sie einen Funnel-Plot und interpretieren Sie diesen!
2. Ihnen liegen drei Untersuchungen zum Zusammenhang zwischen Neurotizismus und Unfallhäufigkeit vor, die folgende Ergebnisse berichten: $r_1 = .44$ ($n_1 = 240$), $r_2 = .35$ ($n_2 = 26$) und $r_3 = .26$ ($n_3 = 118$). Schätzen Sie den Gesamteffekt, testen Sie ihn auf Signifikanz und bestimmen Sie das 95 %ige Konfidenzintervall. Beurteilen Sie die Größe des Effektes. Sind die drei Untersuchungsergebnisse homogen?
3. Sie entwickeln Entspannungsübungen, bei denen auch die Biofeedback-Methode zum Einsatz kommt. Ihr fertiges Übungsprogramm wollen Sie in einer Evaluationsstudie auf seine Wirksamkeit prüfen. Dazu untersuchen Sie 6 Gruppen (à 30 Personen) hinsichtlich ihres Spannungszustandes vor Beginn und nach Abschluss des Trainings. Es ergeben sich die in ◘ Tab. 16.8 aufgeführten Gruppenmittelwerte und Irrtumswahrscheinlichkeiten für die Testung der Vorher-Nachher-Mittelwertdifferenzen (t-Test für abhängige Stichproben). Fassen Sie diese Ergebnisse durch kombinierte Signifikanztests zusammen (Vote Counting, Vorzeichentest, Binomialtest und Stouffer-Methode)!

◘ Tabelle 16.8 Ergebnisse einer fiktiven Evaluationsstudie eines Entspannungs-Trainings (Indexwert von 1 = völlig angespannt bis 5 = völlig entspannt)

Übungsgruppen (jeweils $n = 30$)	Vorher-Messung (Mittelwert)	Nachher-Messung (Mittelwert)	p (einseitig)
1	3.10	3.70	.022
2	2.50	2.30	.387
3	1.90	2.40	.219
4	2.20	3.60	.001
5	3.10	2.90	.046
6	2.60	3.50	.042

Weiterführende Literatur

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester: Wiley
- Cooper, H. (2009). *Research synthesis and meta-analysis. A step-by-step-approach*. Thousand Oaks: Sage.
- Cooper, H. M., Hedges, L. V., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage.
- Kunz, R., Khan, K., Kleijnen, J. & Antes, G. (2009). *Systematische Übersichtsarbeiten und Meta-Analysen. Einführung in Instrumente der evidenzbasierten Medizin für Ärzte, klinische Forscher und Experten im Gesundheitswesen* (2. Aufl.). Bern: Huber.
- Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.
- Littell, J. H., Corcoran, J. & Pillai, V. (Eds.) (2008). *Systematic reviews and metaanalysis*. New York: Oxford University Press.
- Sandelowski, M. & Barroso, J. (2006). *Handbook for synthesizing qualitative research*. Berlin Heidelberg: Springer.

Literatur

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., et al. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin, 136*(2), 151–173.

Bangert-Drowns, R. L. (1986). Review of development in meta-analytic method. *Psychological Bulletin, 99*(3), 388–399.

Bax, L., Yu, L. M., Ikeda, N., Tsuruta, H., & Moons, K. G. (2006). Development and validation of MIX: Comprehensive free software for meta-analysis of casual research data. *BMC Medial Research Methodology, 6*. Retrieved [November 6, 2013], from http://www.biomedcentral.com/14712288/6/50/

Beaman, A. L. (1991). An empirical comparison of meta-analytic and traditional reviews. *Personality and Social Psychology Bulletin, 17*(3), 252–257.

Becker, B. J. (1987). Applying tests of combined significance in meta-analysis. *Psychological Bulletin, 102*(1), 164–172.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 215–230). Thousand Oaks: Sage.

Beelmann, A. & Bliesener, T. (1994). Aktuelle Probleme und Strategien der Metaanalyse. *Psychologische Rundschau, 45*, 211–233.

Beelmann, A. & Lipsey, M. W. (in press). Meta-analysis of effect eestimates from multiple studies. In M. W. Lipsey & D. S. Cordray (Eds.), *Field experimentation: Methods for evaluating what works, for whom, under what circumstances, how, and why*. Thousand Oaks: Sage.

Beelmann, A. & Schneider, N. (2003). Wirksamkeit der Psychotherapie bei Kindern und Jugendlichen. Eine Übersicht und Meta-Analyse zum Stand und zu Ergebnissen der deutschsprachigen Effektivitätsforschung. *Zeitschrift für Klinische Psychologie und Psychotherapie, 32*(2), 129–143.

Bond, C. F. J., Wiitala, W. L., & Dan Richard, F. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8*(4), 406–418.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). Comprehensive meta-analysis [Computer Program] (Version 2). Englewood Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Bortz, J. (2005). *Statistik* (6. Aufl.). Berlin: Springer.

Bortz, J. & Lienert , G. A. (2003). Kurzgefaßte Statistik für die klinische Forschung. Ein praktischer Leitfaden für die Analyse kleiner Stichproben (2. Aufl.). Heidelberg: Springer.

Bortz, J. & Lienert , G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben.* (3. Aufl.). Berlin: Springer.

Bortz, J., Lienert , G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Berlin: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.

Bosnjak, M. & Viechtbauer, W. (2009). Die Methode der Meta-Analyse zur Evidenzbasierung von Gesundheitsrisiken: Beiträge der Sozial-, Verhaltens- und Wirtschaftswissenschaften. *Zentralblatt für Abeitsmedizin, Abeitsschutz & Ergonomie, 11*, 322–333.

Brüderl, J. (2004). Meta-Analyse in der Soziologie: Bilanz der deutschen Scheidungsursachenforschung oder „statistischer Fruchtsalat"? *Zeitschrift für Soziologie, 33*(1), 84–86.

Bushman, B. J. (1994). Vote Counting Procedures in Meta-Analysis. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 193–213). Thousand Oaks: Sage.

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A Meta-Meta-Analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*(2), 239–270.

Carson, K. P., Schriesheim, C. A., & Kinicki, A. J. (1990). The Usefullness of the „Fail-Safe'" statistic in meta-analysis. *Educational and Psychological Measurement 50*(2), 233–243.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum.

Cohn, L. D. & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*(3), 243–253.

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., et al. (1993). Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA The Journal of the American Medical Association, 269*(21), 2749–2753.

Cooper, H., Charlton, K., Valentine, J. C., & Muhlenbruck, L. (2000). *Making the most of summer school: A meta-analytic and narrative review* (Vol. 65). United Kingdom: Blackwell Publishing.

Cooper, H., De Neve, K., & Charlton, K. (1997). Finding the missing science. The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2*(4), 447–452.

Cooper, H. M. (2009). *Research synthesis and meta-analysis: A step-by-step approach*. Thousand Oaks: Sage.

Cooper, H. M. & Hedges, L. V. (Eds.). (1993). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cooper, H. M., Hedges, L. V., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage.

Cornwell, J. M. & Ladd, R. T. (1993). Power and accuracy of the Schmidt and Hunter meta-analytic procedures. *Educational and Psychological Measurement, 53*(4), 877–895.

Czienskowski, U. (2003). Meta-analysis – not just research synthesis. In R. Schulze, H. Holling & D. Böhning (Eds.), *Metaanalysis. New developments and aplications in medical and social science* (pp. 141–152). Göttingen: Hogrefe & Huber.

Darlington, R. B. & Hayes, A. F. (2000). Combining Independent p values: Extensions of the Stouffer and Binomial Methods. *Psychological Methods, 5*, 496-515.

Duval, S. & Tweedie, R. (2000a). A nonparametric „trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–98.

Duval, S. & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *The British Medical Journal, 315*(7109), 629–634.

Ellis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: University Press.

Eysenck, H. J. (1952). The effects of psychotherapy: An Evaluation. *Journal of Consulting Psychology, 16*(5), 319–324.

Eysenck, H. J. (1978). An Exercise in Mega-Silliness. *American Psychologist, 33*, 517.

Ferguson, C. J. & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120–128.

Ferguson, C. J. & Kilburn, J. (2010). Much ado about nothing: The misestimation and overinterpretation of violent video game effects in eastern and western nations: Comment on Anderson et al. *Psychological Bulletin, 136*(2), 174–178.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., et al. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517–537.

Fricke, R. & Treinies, G. (1985). *Einführung in die Metaanalyse*. Bern: Huber.

Gillett, R. (2003). The metric comparability of meta-analytic effectsize estimators from factorial designs. *Psychological Methods, 8*(4), 419–433.

Gilpin, A. R. (1993). Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude effect for meta-analysis. *Educational and Psychological Measurement, 53*(1), 87–92.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Glass, G. V. (1999). *Meta-analysis at 25*. Paper presented at the Office of Special Education Programs Research Project Directors' Conference, U.S. Department of Education. Retrieved November 6, 2013, from http://www.gvglass.info/papers/meta25.html

Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta analysis in social research*. Thousand Oaks: Sage.

Gleser, L. J. & Olkin, J. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). Thousand Oaks: Sage.

Gottfredson, S. D. (1978). Evaluating psychological research reports. Dimensions, reliability, and correlates of quality judgements. *American Psychologists, 33*(10), 920–934.

Green, B. F. & Hall, J. A. (1984). Quantitative methods for literature review's. *Annual Revue of Psychology, 35*, 37–53.

Grégoire, G., Derderian, F., & LeLorie, J. (1995). Selecting the language of the publications included in meta,analysis: is there a tower-of-babel-bias? *Journal of Clinical Epidemiology, 48*(1), 159–163.

Hadjar, A. (2011). *Geschlechtsspezifische Bildungsungleichheiten*. Wiesbaden: VS.

Hadjar, A. & Lupatsch, J. (2011). Geschlechterunterschiede im Schulerfolg: Spielt die Lehrperson eine Rolle? *ZSE Zeitschrift für Soziologie der Erziehung und Sozialisation, 31*(1), 79–94.

Hager, W. (2004). *Testplanung zur statistischen Prüfung psychologischer Hypothesen*. Göttingen: Hogrefe.

Hall, J. A., Tickle-Degnerz, L., Rosenthal, R., & Mosteller, F. (1994). Hypothesis and problems in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 18–27). Thousand Oaks: Sage.

Hannover, B. & Kessels, U. (2011). Sind Jungen die neuen Bildungsverlierer? Empirische Evidenz für Geschlechterdisparitäten zuungunsten von Jungen und Erklärungsansätze. *Zeitschrift für Pädagogische Psychologie, 25*(2), 89–103.

Harris, M. J. (1991). Controversy and cumulation: Meta-analysis and research on interpersonal expectancy effects. *Personality and Social Psychology Bulletin, 17*(3), 316–322.

Hattie, J. A. (2008). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hedges, L. V. (1982). *Statistical methodology in meta-analysis*. Princeton: Educational Testing Service.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 286–298). Thousand Oaks: Sage.

Hedges, L. V., Cooper, H., & Bushman, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin, 111*(1), 188–194.

Hedges, L. V. & Olkin, J. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Hedges, L. V. & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*(3), 203–217.

Hedges, L. V. & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426–445.

Hedges, L. V. & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299–332.

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

Heirs, M. & Dean, M. E. (2007). Homeopathy for attention deficit/hyperactivity disorder or hyperkinetic disorder. *Cochrane Database of Systematic Reviews, 4*.

Heres, S., Wagenpfeil, S., Hamann, J., Kissling, W., & Leucht, S. (2004). Language bias in neuroscience – is the tower of babel located in Germany? *European Psychiatry, 19*(4), 230–232.

Heyvaert, M., Maes, B., & Onghena, P. (2011). Mixed methods research synthesis: definition, framework, and potential. *Quality & Quantity*(online first), 1–18.

Higgins, J. P. T. & Green, S. (2009). Cochrane handbook for systematic reviews of interventions version 5.0.2. Retrieved November 6, 2013, from http://www.cochranehandbook.org

Hsu, L. M. (2005). Some properties of *r* equivalent: a simple effect size indicator. *Psychological Methods, 10*(4), 420–427.

Hunter, J. E. & Schmidt, F. L. (1989). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks: Sage.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (2nd ed.). Thousand Oaks: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis cumulating research finding across studies*. Thousand Oaks: Sage.

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30*(20), 2481–2498.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80*(1), 94–106.

Jussim, L. & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131–155.

Jussim, L., Robustelli, S. L., & Cain, T. R. (2009). Teacher expectations and self-fulfilling prophecies. In K. R. Wenzel & A. Wigfield (Eds.), *Handbook of motivation at school. (Educational Psychology Handbook Series)* (pp. 349–380). New York: Routledge/Taylor & Francis Group.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational Psychological Measurement, 56*(5), 746–759.

Kontopantelis, E. & Reeves, D. (2009). MetaEasy: A meta-analysis add-in for Microsoft Excel. *Journal of Statistical Software, 30*(7), 1–25.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics, 8*(2), 93–101.

Kraemer, H. C. (1985). A strategy to teach the concept and application of power of statistical tests. *Journal of Educational Statistics, 10*(3), 173–195.

Kraemer, H. C. (2005). A simple effect size indicator for two-group comparisons? A comment on *r* equivalent. *Psychological Methods, 10*(4), 413–419.

Kraemer, H. C., Gardner, C., Brooks III, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist vs. exclusionist view points. *Psychological Methods, 3*(1), 23–31.

Kraemer, H. C. & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Thousand Oaks: Sage.

Kunz, R., Khan, K. S., Kleijnen, J., & Antes, G. (2009). *Systematische Übersichtsarbeiten und Meta-Analysen* (2. Aufl.). Bern: Huber.

Landman, J. R. & Dawes, R. M. (1982). Psychotherapy outcome: Smith and Glass' Conclusions stand up under scrutinity. *American Psychologist, 37*(5), 504–516.

Light, R. J. & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge: Harvard University Press.

Light, R. J. & Smith, P. V. (1971). Accumulating evidence: Procedure for resolving contradictions among different research studies. *Harvard Educational Review, 41*(4), 429–471.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Littell, J., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York: Oxford University Press.

Lösel, F. & Breuer-Kreuzer, D. (1990). Metaanalyse in der Evaluationsforschung: Allgemeine Probleme und eine Studie über den Zusammenhang zwischen Familienmerkmalen und psychischen Auffälligkeiten bei Kindern und Jugendlichen. *Zeitschrift für Pädagogische Psychologie, 4*, 253–268.

MacKay, D. G. (1993). The theoretical epistemology: A new perspective on some long-standing methodological issues in psychology. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues* (pp. 229–255). Hillsdale: Erlbaum.

Magnusson, D. (1966). *Test theory*. Reading: Addison-Wesley.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., Vehovar, V., & Berzelak, N. (2008). Web surveys vs. other survey modes. A meta-analysis comparing response rates. *Source International Journal of Market Research, 50*(1), 79–104.

Mansfield, R. S. & Busse, T. V. (1977). Meta-analysis of research: A rejoinder to Glass. *Educational Researcher, 6*(9), 3.

Morris, S. B. & De Shon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measure and independent groups designs. *Psychological Methods, 7*(1), 105–125.

Mullen, B. (1989). *Advanced basic meta-analysis*. Hillsdale: Erlbaum.

Mullen, B. & Rosenthal, R. (1985). *BASIC meta-analysis: Procedures and program*. Hillsdale: Erlbaum.

Neugebauer, M., Helbig, M., & Landmann, A. (2011). Unmasking the myth of the same-sex teacher advantage. *European Sociological Review, 27*(5), 669–689.

Olejnik, S. & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations and limitations. *Contemporary Educational Psychology, 25*(3), 241–286.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157–159.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139–162). Thousand Oaks: Sage.

Paterson, B., Dubouloz, C.-J., Chevrier, J., Ashe, B., King, J., & Moldoveanu, M. (2009). Conducting qualitative metasynthesis research: Insights from a metasynthesis project. *International Journal of Qualitative Methods, 8*(3), 22–33.

Petersen, J. L. & Hyde, J. S. (2010). A meta-analytic review of research on gender differences in sexuality. *Psychological Bulletin, 136*(1), 21–38.

Petticrew, M. & Roberts, H. (Eds.). (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell.

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 164–174). Thousand Oaks: Sage.

Popay, J. (2006). *Moving beyond effectiveness in evidence synthesis: Methodological issues in the synthesis of diverse sources of evidence*. London: NICE.

Pope, C., Mays, N., & Popay, J. (2007). *Synthesising qualitative and quantitative health research: A guide to methods* maidenhead: Open University Press.

Radin, D. I. & Ferrari, D. C. (1991). Effects of consciousness on the fall of dice: a meta-analysis. *Journal of Scientific Exploration, 5*(3), 61–83.

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution, 59*(2), 464–468.

Rosenberg, M. S., Adams, D. C., & Gurevitch, J. (2007). MetaWin [Computer Program] (Version 2.0): self-distributed.

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 85–94). Thousand Oaks: Sage.

Rosenthal, R. (1968). Experimenter expectancy and the reassuring natur of the null hypothesis decision procedure. *Psychological Bulletin, 70*(6), 30–47.

Rosenthal, R. (1969). Interpersonal expectations. In R. Rosenthal & R. L. Rosnow (Hrsg.), *Artifact in behavioral research* (S. 181–277). Orlando: Academic Press.

Rosenthal, R. (1973a). the mediation of pygmalion effects: A four factor „ theory". *Papau New Guinea Journal of Education, 9*(1), 1–12.

Rosenthal, R. (1973b). On the social psychology of the self-fulfilling prophecy: Further evidence for pygmalion effects and their mediating mechanisms. *MSS Modular Publications, Module 53*, 1–28.

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Halsted Press.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*(1), 185–193.

Rosenthal, R. (1979). The „file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Thousand Oaks: Sage.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioural sciences. Methodological issues* (pp. 519–559). Hillsdale: Erlbaum.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (S. 232–243). Thousand Oaks: Sage.

Rosenthal, R. (1995). Critiquing Pygmalion: A 25-year perspective. *Current Directions in Psychological Science, 4*(6), 171–172.

Rosenthal, R. & DiMatteo, M. R. (2001). Meta analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Rosenthal, R. & Fode, K. (1963). The effect of experimenter bias on performance of the albino rat. *Behavioral Science, 8*(3), 183–189.

Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart & Winston.

Rosenthal, R. & Rubin, D. B. (1971). Pygmalion Reaffirmed. In J. D. Elashoff & R. E. Snow (Eds.), *Pygmalion reconsidered* (pp. 139–155). Worthington: Jones.

Rosenthal, R. & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 1*(3), 377–386.

Rosenthal, R. & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*(3), 400–406.

Rosenthal, R. & Rubin, D. B. (2003).*r*equivalent: A simple effect size indicator. *Psychological Methods, 8*(4), 492–496.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175–197). Mahwah: Erlbaum.

Rustenbach, S. J. (2003). *Metaanalyse. Eine anwendungsorientierte Einführung*. Bern: Huber.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A monte carlo investigation of statistical power and resistence to type I error. *Journal Applied Psychology, 71*(2), 302–310.

Sánchez-Meca, J., Marin-Martinez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*, 448–467.

Sandelowski, M. & Barroso, J. (2006). *Handbook for synthesising qualitative research*. New York: Springer.

Saner, H. (1994). A conservative inverse normal test procedure for combining *p*-values in integrative research. *Psychometrika, 59*(2), 253–267.

Sauerbrei, W. & Blettner, M. (2003). Issues of traditional reviews and meta-analysis of observational studies in medical research. In R. Schulze, H. Holling & D. Böhning (Eds.), *Metaanalysis. new developments and applications in medical and social sciences* (pp. 79–98). Göttingen: Hogrefe & Huber.

Schulze, R. (2004). *Meta-analysis. A comparison of approaches*. Göttingen: Hogrefe & Huber.

Schulze, R., Holling, H., Großmann, H., Jütting, A., & Brocke, M. (2003). Differences in the results of two meta-analytical approaches. In R. Schulze, H. Holling & D. Böhning (Eds.), *Meta-analysis. New developments and applications in medical and social sciences* (pp. 19–39). Göttingen: Hogrefe & Huber.

Seifert, T. L. (1991). Determining effect sizes in various experimental designs. *Educational and Psychological Measurement, 51*(2), 341–347.

Shadish, W. R. & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 262–280). Thousand Oaks: Sage.

Smith, G. & Egger, M. (1998). Meta-analysis: Unresolved issues and future developments. *British Medical Journal, 316*(7126), 221–225.

Smith, M. L. & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752–760.

Snook, I., O'Neill, J., Clark, J., O'Neill, A.-M., & Openshaw, R. (2009). Invisible learnings? A Commentary on John Hattie's book – Visible learnings? A synthesis of over 800 meta-analyses relating to achievement. *New Zealand Journal of Educational Studies, 44*(1), 93–106.

Soilevuo Grønnerød, J. & Grønnerød, C. (2012). The Wartegg Zeichen Test: A literature overview and a meta-analysis of reliability and validity. *Psychological Assessment, 24*(2), 476–489.

Spector, P. E. & Levine, E. L. (1987). Meta-analysis for integrating study outcomes. A Monte Carlo study of its susceptibility to type I and type II errors. *Journal of Applied Psychology, 72*(1), 3–9.

Steiner, D. D., Lane, J. M., Dobbins, G. H., Schnur, A., & McConnell, S. (1991). A review of meta-analysis in organizational behavior and human resources management: An empirical assessment. *Educational and Psychological Measurement, 51*(3), 609–626.

Sterne, J. A. C., Egger, M., & Davey Smith, G. (2001). Investigating and dealing with publication and other biases. In M. Egger, G. Davey Smith & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd. ed., pp. 189–208). London: BMJ Books.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 125–138). Thousand Oaks: Sage.

Stock, W. A., Okun, M. A., Haring, M. J., Miller, W., Kinney, C., & Ceur vorst, R. W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11*(6), 10–14.

Stouffer, S. A., Suchman, E. A., de Vinney, L. C., Star, S. A., & Williams, R. M. j. (1949). *The American soldier: Adjustment during army life* (Vol. 1). Princeton: Princeton University Press.

Sutton, A. J. & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research, 10*(4), 277–303.

The Cochrane Collaboration. (2011). Review manager (RevMan) [Computer program] (Version 5.1). Copenhagen: The Nordic Cochrane Centre.

Timulak, L. (2009). Meta-analysis of qualitative studies: a tool for reviewing qualitative research findings in psychotherapy. *Psychotherapy Research, 19*(4–5), 591–600.

Tracz, S. M., Elmore, P. B., & Pohlmann, J. T. (1992). Correlational meta-analysis. Independent and nonindependent cases. *Educational and Psychological Measurement, 52*(4), 879–888.

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science, 6*(4), 363–403.

Vevea, J. L. & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419–435.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis. Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*(4), 428–443.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26*(1), 37–52.

Wachter, K. M. & Straf, M. L. (Eds.). (1990). *The future of meta-analysis*. New York: Sage.

Wang, M. C. & Bushman, B. J. (1998). Using the normal quantile plot to explore meta–analytic data sets. *Psychological Methods, 3*(1), 46–54.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur Psychologischen Methodenlehre*. Göttingen: Hogrefe.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of research synthesis* (pp. 41–55). Thousand Oaks: Sage.

Wilkinson, B. (1951). Statistical consideration in psychological research. *Psychological Bulletin, 48*(2), 156–158.

Wilson, D. B. & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*(4), 413–429.

Wolf, F. M. (1987). *Meta-analysis: Quantitative methods for research synthesis*. Thousand Oaks: Sage.

Wortmann, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (S. 97–109). Thousand Oaks: Sage.

Zhao, S. (1991). Metatheory, metamethod, qualitative meta-analysis: What, why and how? *Sociological Perspectives, 34*(3), 377–390.

Strukturgleichungsmodelle

Christina S. Werner, Karin Schermelleh-Engel, Carla Gerhard, Jana C. Gäde

17.1 Motivation: Wozu Strukturgleichungsmodelle? – 946

17.2 Aufbau von Strukturgleichungsmodellen – 949

17.3 Analyse von Strukturgleichungsmodellen – 955

17.4 Anwendungsbeispiel: Ein Strukturgleichungsmodell zu Burnout – 958

17.5 Praktischer Einsatz, Herausforderungen und Probleme – 963

17.6 Software für Strukturgleichungsmodelle – 969

Literatur – 972

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_17, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Vorteile von Strukturgleichungsmodellen für Datenanalysen kennen.
- Den Aufbau der Modelle und die dahinterstehenden Gleichungen verstehen.
- Verstehen, wie empirische Daten mit Strukturgleichungsmodellen analysiert werden.
- Einfache Analysen selbst durchführen und interpretieren können.
- Voraussetzungen und Probleme des Einsatzes von Strukturgleichungsmodellen kennen.
- Ergebnisse aus Analysen kritisch beurteilen können.

Dieses Kapitel stellt Strukturgleichungsmodelle als Methoden der statistischen Datenanalyse vor. Strukturgleichungsmodelle kann man als Kombinationen aus Faktorenanalysen und Regressions- bzw. Pfadanalysen auffassen. Mit Strukturgleichungsmodellen kann man Zusammenhänge zwischen Konstrukten untersuchen, also zwischen nicht direkt beobachtbaren Variablen, wie sie in den Sozialwissenschaften häufig vorkommen (z. B. Intelligenz, Lebenszufriedenheit, Burnout). Werden solche Konstrukte durch messbare Variablen (z. B. psychologische Tests, standardisierte Fragebögen) operationalisiert, so können in Strukturgleichungsmodellen statistische Zusammenhänge und Effekte auf der Ebene der nicht direkt beobachtbaren Konstrukte analysiert werden.

Vorteile von Strukturgleichungsmodellen für die Sozialwissenschaften werden erklärt (► Abschn. 17.1), der Aufbau der Modelle dargestellt (► Abschn. 17.2) und Grundsätze empirischer Analysen an einem Beispiel veranschaulicht (► Abschn. 17.3; ► Abschn. 17.4). Wichtige Punkte des praktischen Einsatzes werden erläutert und Analyseprogramme vorgestellt (► Abschn. 17.5; ► Abschn. 17.6).

17.1 Motivation: Wozu Strukturgleichungsmodelle?

Strukturgleichungsmodelle („structural equation models“, SEM) sind Methoden für relativ komplexe Datenanalysen, die besonders in den Sozialwissenschaften nützlich sind (z. B. Psychologie, Erziehungs-, Gesellschafts- und Wirtschaftswissenschaften). Es gibt mehrere Gründe, warum sich sozialwissenschaftliche Theorien und Daten mit Strukturgleichungsmodellen oftmals gut handhaben lassen.

1. **Validität**: Viele sozialwissenschaftliche Theorien beziehen sich auf Konstrukte, die nicht direkt beobachtbar sind: Intelligenz, Einstellungen, Lernerfolg oder Berufseignung können nur indirekt aus beobachteten Variablen erschlossen werden, z. B. aus Lösungen von Testaufgaben, aus Antworten in Fragebögen, aus Verhalten in bestimmten Situationen. Dabei kommen oft verschiedene beobachtete Variablen infrage, von denen keine für sich allein optimal ist, z. B. verschiedene Tests oder Fragebögen zum gleichen Konstrukt, Vorgesetzten-, Kollegen- oder Kundenurteile zu Berufserfolg, etc. In Strukturgleichungsmodellen können für jedes Konstrukt mehrere Variablen (mehrere Operationalisierungen) gleichzeitig berücksichtigt werden. Dadurch kann die Validität (inhaltliche Gültigkeit) von Aussagen umfassender sein als bei der Analyse nur einzelner beobachteter Variablen.
2. **Reliabilität**: Sozialwissenschaftliche Daten enthalten oft nennenswerte Anteile Messfehler, z. B. durch Raten bei Testaufgaben oder je nach Tagesform schwankende Antworten auf Fragebogenitems. Zufällige Messfehleranteile vermindern die Reliabilität (Messgenauigkeit) von Variablen und schwächen damit die Beziehungen zwischen diesen Variablen ab, so dass z. B. Korrelationen unterschätzt werden. In Strukturgleichungsmodellen können Messfehlereinflüsse in Variablen systematisch berücksichtigt werden. Schätzungen der Beziehungen zwischen inhaltlich interessierenden Konstrukten sind dann unbeeinträchtigt von zufälligen Messfehleranteilen.
3. **Komplexität**: Sozialwissenschaftliche Theorien beziehen meist viele Variablen gleichzeitig ein: Verhalten ist z. B. abhängig von verschiedenen Persönlichkeitsvariablen und verschiedenen Eigenschaften einer Situation. Variablen können dabei komplexe Beziehungen haben, z. B. gleichzeitig abhängige und unabhängige Variable sein (Mediatoreffekt). Beispiel: Wenn Intelligenz Berufserfolg begünstigt und Berufserfolg Lebenszufriedenheit begünstigt, so besteht damit ein indirekter Effekt von Intelligenz über die Mediatorvariable Berufserfolg auf Lebenszufriedenheit. In Strukturgleichungsmodellen können komplexe Beziehungen vieler Variablen simultan berücksichtigt und indirekte Effekte einbezogen werden, für die sonst mehrere einzelne Regressionsanalysen durchgeführt werden müssten.
4. **Konfirmatorische Analyse**: Bei der Prüfung sozialwissenschaftlicher Theorien interessiert häufig nicht nur, ob bestimmte Variablen zusammenhängen oder einander beeinflussen, sondern man möchte wissen, ob aufgestellte Theorien die Beziehungen zwischen vielen Variablen insgesamt richtig beschreiben, also z. B. auch keine weiteren Beziehungen übersehen wurden. Mit Strukturgleichungsmodellen kann man konfirmatorisch (d. h. das gesamte Modell bestätigend oder ablehnend) prüfen, ob präzise theoretische Annahmen mit allen insgesamt gefundenen Beziehungen zwischen beobachteten Variablen zu vereinbaren sind oder nicht.

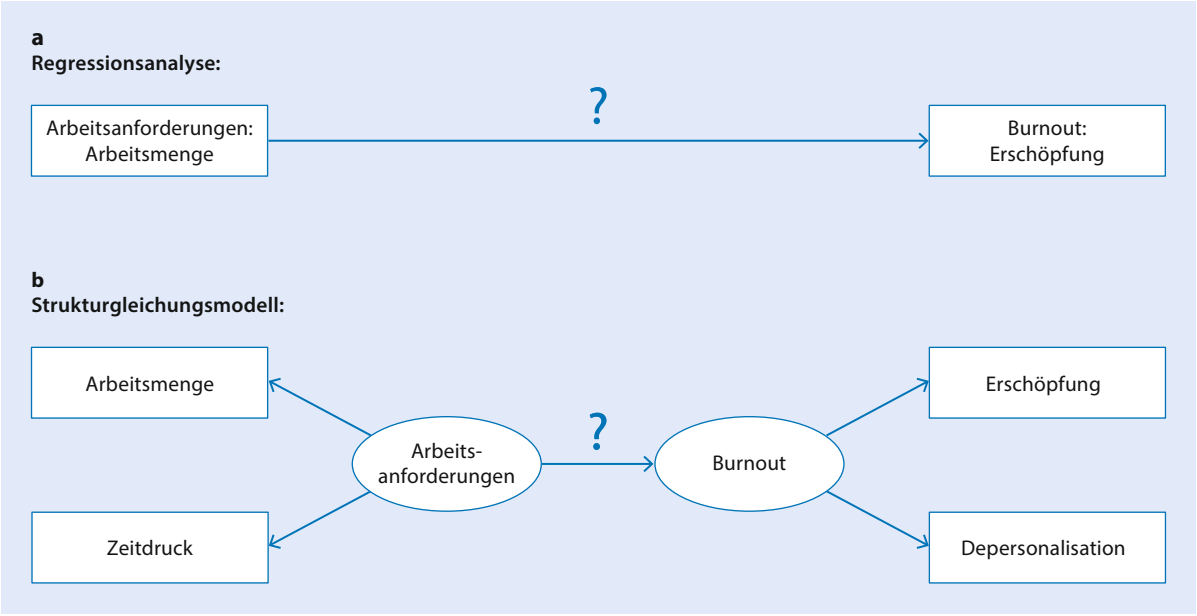


Abb. 17.1 Schema einer Regressionsanalyse **(a)** und eines Strukturgleichungsmodells **(b)** am Beispiel eines vermuteten Effektes quantitativer Arbeitsanforderungen auf Burnout: Regression analysiert Zusammenhänge zwischen beobachteten Variablen; Strukturgleichungsmodelle können Zusammenhänge zwischen Konstrukten analysieren, die über mehrere beobachtete Variablen erfasst werden

Einfache statistische Methoden, wie z. B. Hypothesenprüfungen durch Regressionsanalysen, sind bei vielen sozialwissenschaftlichen Fragen dagegen weniger aussagekräftig. Angenommen, man möchte prüfen, ob hohe quantitative Arbeitsanforderungen (gemessen z. B. über Fragen zur Arbeitsmenge) zu Burnout führen (gemessen z. B. über eine Skala zur emotionalen Erschöpfung). Stellt man nun mittels Regressionsanalyse fest, dass die Arbeitsmenge die Erschöpfung nicht vorhersagen kann, bleibt unklar, ob a) quantitative Arbeitsanforderungen grundsätzlich nichts mit Burnout zu tun haben, ob b) die Skala zu emotionaler Erschöpfung nicht gut zur Messung von Burnout geeignet ist, oder ob c) die verwendeten Fragen zur Arbeitsmenge die quantitativen Arbeitsanforderungen nicht gut messen können – während mit anderen Maßen der Arbeitsanforderungen oder anderen Maßen für Burnout vielleicht eine Vorhersage möglich wäre.

Strukturgleichungsmodelle können mehrere Messungen (z. B. mehrere Tests, mehrere Fragebogenskalen, Angaben unterschiedlicher Beurteiler etc.) kombinieren, so dass man Aussagen über Beziehungen zwischen Konstrukten trennen kann von Aussagen über die Qualität einzelner Messungen. Dies ist in Abb. 17.1 schematisch dargestellt. In diesem einfachen Strukturgleichungsmodell entspricht die Beziehung zwischen den beiden Konstrukten einer Regressionsanalyse mit Arbeitsanforderungen als Prädiktor und Burnout als Kriterium. Zur Messung der Konstrukte werden aber jeweils mehrere beobachtete Variablen verwendet. Das Konstrukt quantitative Arbeitsanforderungen wird über Skalen zu Arbeitsmenge und Zeitdruck gemessen und Burnout über Skalen zu emotionaler Erschöpfung und Depersonalisation. Die Frage, ob Arbeitsanforderungen zu Burnout beitragen, kann dann unter Berücksichtigung all dieser Messungen auf der Ebene der beiden Konstrukte geprüft werden.

Außerdem lassen sich in Strukturgleichungsmodellen komplexe Beziehungen simultan prüfen: Hätte man z. B. die Annahmen, dass

- einerseits quantitative Arbeitsanforderungen zu Burnout führen können,
- andererseits aber auch ein übermäßiges Arbeitsengagement (Overcommitment) Burnout fördert,
- sowie dass quantitative Arbeitsanforderungen ihrerseits zu Overcommitment beitragen,
- aber auch Perfektionismus als Persönlichkeitsvariable zu Overcommitment beiträgt,

könnte man all diese zusammenhängenden Annahmen in einem einzigen Strukturgleichungsmodell analysieren (Abb. 17.2) statt mit separaten Regressionsanalysen. Mit dem Strukturgleichungsmodell werden dann sowohl die angenommenen Beziehungen geprüft, als auch, dass nicht angenommene Beziehungen tatsächlich nicht bestehen (z. B. kein direkter Effekt von Perfektionismus auf Burnout, sondern nur ein indirekter über die Mediatorvariable Overcommitment).

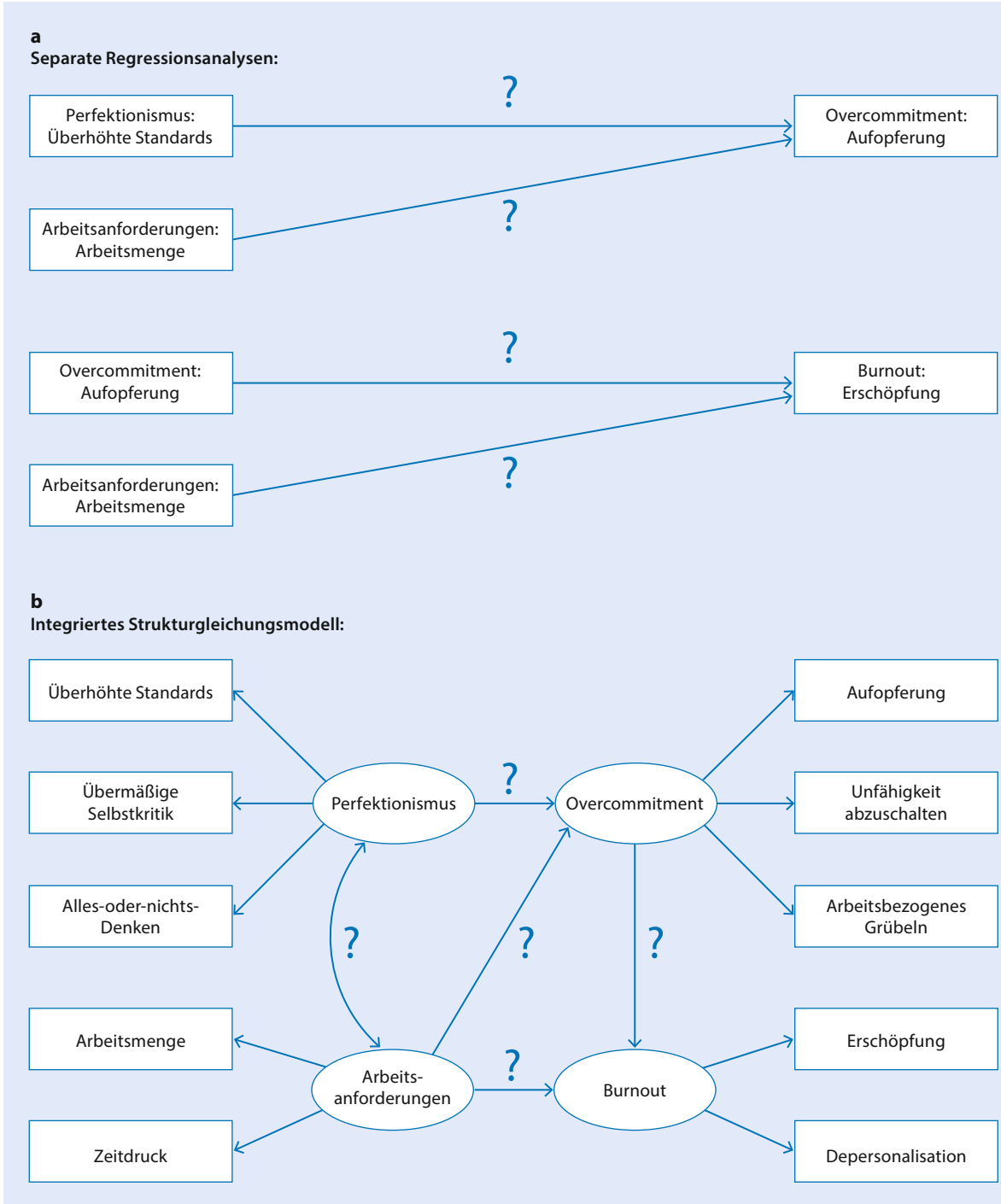


Abb. 17.2 Schema der Analyse komplexer Beziehungen zwischen mehreren Variablen mittels separater Regressionsanalysen (**a**) oder in einem einzigen Strukturgleichungsmodell (**b**) am Beispiel vermuteter Effekte von Perfektionismus und Arbeitsanforderungen über Overcommitment auf Burnout

17.2 Aufbau von Strukturgleichungsmodellen

Strukturgleichungsmodelle kann man als Kombinationen aus Faktorenanalysen und Regressions- bzw. Pfadanalysen auffassen (vgl. Bollen, 1989; Kline, 2011). **Faktorenanalysen** entsprechen den Beziehungen zwischen den Konstrukten und den beobachteten Variablen, die diese Konstrukte operationalisieren und messen sollen. **Regressions- bzw. Pfadanalysen** entsprechen den Beziehungen zwischen den Konstrukten untereinander. Wir stellen zuerst faktorenanalytische Modelle zur Messung von Konstrukten vor, danach folgen vollständige Strukturgleichungsmodelle zur Analyse von Effekten zwischen Konstrukten (► Pfaddiagramm).

Pfaddiagramm

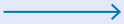
Strukturgleichungsmodelle können grafisch in Form von Pfaddiagrammen dargestellt werden. Hierzu werden 4 Symbole verwendet:



Ein Kreis (oder eine Ellipse) symbolisiert ein Konstrukt (latente Variable, ► Abschn. 17.2.1).



In einem Quadrat (oder einem Rechteck) steht eine beobachtete Variable (manifeste Variable, ► Abschn. 17.2.1).



Ein Pfeil signalisiert einen gerichteten Effekt (z. B. Vorhersage, Kausaleffekt).



Ein gebogener Doppelpfeil weist auf einen ungerichteten Zusammenhang (Korrelation, Kovarianz) hin.

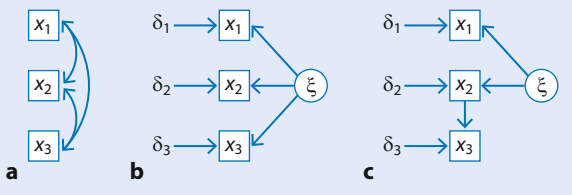
17.2.1 Messung von Konstrukten: Faktorenanalyse

Wie bei einer Faktorenanalyse werden für jedes interessierende Konstrukt möglichst mehrere beobachtete Variablen als Operationalisierungen genutzt. Jede der beobachteten Variablen soll das entsprechende Konstrukt messen, kann dabei aber auch Anteile zufälliger Messfehler enthalten. Neben Messfehleranteilen können in beobachteten Variablen außerdem spezifische Anteile enthalten sein. So werden z. B. Selbst- und Fremdeinschätzungen der Persönlichkeit normalerweise ähnlich sein, also positiv zusammenhängen, aber nicht perfekt übereinstimmen: Selbst- und Fremdeinschätzungen beruhen auf unterschiedlichen Erfahrungen und können deshalb neben dem, was sie gemeinsam haben, jeweils auch spezifische Perspektiven widerspiegeln.

In Strukturgleichungsmodellen wird ein Konstrukt als das aufgefasst, was die verschiedenen beobachteten Variablen eines Konstruktes gemeinsam haben. Es wird angenommen, dass die Messfehleranteile mehrerer beobachteter Variablen zufällig und deshalb voneinander unabhängig sind, also nicht miteinander korrelieren. Daher kann alle gemeinsame Varianz der beobachteten Variablen als messfehlerfreie, wahre Varianz des interessierenden Konstruktes aufgefasst werden.

Latente und manifeste Variablen

Präzisiert wird diese Vorstellung von Konstrukten und beobachteten Variablen in Strukturgleichungsmodellen durch die Konzepte der **latenten** und **manifesten Variablen**. Grundsätzlich wird angenommen, dass Konstrukte ξ (xi) nicht direkt beobachtbar sind, aber einen Einfluss auf die ihnen zugeordneten beobachteten Variablen haben (diese sind im Folgenden mit x bezeichnet; ◘ Abb. 17.3).



◘ **Abb. 17.3** Strukturgleichungsmodelle führen Zusammenhänge zwischen beobachteten Variablen (**a**) auf Effekte dahinterstehender Konstrukte zurück. Erklärt ein unbeobachtbares Konstrukt die Zusammenhänge vollständig, so bleiben nur unsystematische Einflüsse (Messfehlervariablen δ) übrig. Die beobachteten Variablen sind dann Indikatorvariablen (Messungen) der latenten Variable ξ (**b**). Dagegen könnte man nicht auf Wirkungen eines Konstruktes schließen, wenn direkte Effekte zwischen beobachteten Variablen bestünden (**c**)

Ein Zusammenhang zwischen mehreren beobachteten Variablen, die das gleiche Konstrukt erfassen sollen, also deren gemeinsame Varianz (Kovarianz), kann dann auf die Wirkung des hinter ihnen stehenden, unbeobachtbaren Konstruktes ξ zurückgeführt werden. Beispiel: Weil eine Person intelligent ist (hohe Ausprägung in ξ), kann sie mehrere Intelligenztestaufgaben lösen (hohe Ausprägungen in x_1, x_2, x_3).

Wenn es gelingt, Zusammenhänge beobachteter Variablen vollständig auf die Wirkung eines unbeobachtbaren Konstruktes zurückzuführen, so bedeutet dies, dass der Zusammenhang der beobachteten Variablen verschwinden würde, wenn man die Ausprägung des Konstruktes konstant hält (also z. B. Auswahl von Personen mit gleicher Intelligenz, die alle mit gleicher Wahrscheinlichkeit die Aufgaben lösen). Die beobachteten Variablen würden nur deshalb kovariieren, weil sie das gleiche Konstrukt messen, so dass darüber hinaus lediglich unsys-

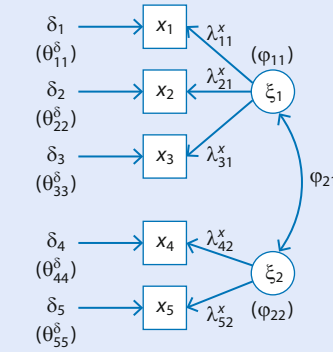
tematische Messfehleranteile δ (delta) verbleiben. Diese Bedingung bezeichnet man als **lokale stochastische Unabhängigkeit** der beobachteten Variablen ◘ Abb. 17.3b). Wenn sie erfüllt ist, so stellen die beobachteten Variablen im Strukturgleichungsmodell **manifeste Indikatorvariablen** einer gemeinsam hinter ihnen stehenden, nicht direkt beobachtbaren **latenten Variable** dar. Ein inhaltlich definiertes psychologisches Konstrukt kann also dann als latente Variable eines Strukturgleichungsmodells aufgefasst werden, wenn es mit lokal stochastisch unabhängigen manifesten Indikatorvariablen gemessen werden kann.

Die Bedingung lokaler stochastischer Unabhängigkeit ist nicht trivial: Es ist durchaus möglich, dass die gemeinsame Varianz mehrerer beobachteter Variablen nicht vollständig auf ein dahinterstehendes Konstrukt zurückgeführt werden kann. Wenn z. B. bei mehreren Rechenaufgaben das Lösen einer Aufgabe logisch auch zum Lösen einer weiteren Aufgabe führt, so kann unabhängig davon, dass alle Aufgaben Rechenfähigkeit messen, ein direkter Effekt zwischen den beiden logisch voneinander abhängigen Aufgaben bestehen. Die Bedingung der lokalen stochastischen Unabhängigkeit wäre dann nicht erfüllt (◘ Abb. 17.3c).

Messmodelle

Das Modell der angenommenen Beziehungen zwischen einer latenten Variable und ihren manifesten Indikatorvariablen bildet ein **Messmodell** für diese latente Variable. Messmodelle können im Rahmen von Strukturgleichungsmodellen nicht nur für ein einzelnes Konstrukt aufgestellt werden, sondern für mehrere zugleich – wie bei einer Faktorenanalyse von Messungen zu mehreren Konstrukten, z. B. der Faktorenanalyse eines mehrdimensionalen Fragebogens (◘ Abb. 17.4). Dabei kann gezielt festgelegt werden, ob zwischen zwei latenten Variablen jeweils ein Zusammenhang angenommen werden soll oder nicht.

Rechnerisch wird im Messmodell wie in einer explorativen Faktorenanalyse („Konfirmatorische und explorative Faktorenanalyse“ im nächsten Abschnitt) angenommen, dass die beobachteten Variablen x linear von den Konstrukten ξ abhängen. Üblicherweise wird jede beobachtete Variable auf genau ein Konstrukt zurückgeführt, grundsätzlich können beobachtete Variablen aber auch von mehreren Konstrukten gleichzeitig abhängen. Der von den Konstrukten nicht erklärte Anteil jeder beobachteten Variablen wird durch eigene **Messfehlervariablen** δ dargestellt. Dies ähnelt der Annahme im Rahmen der klassischen Testtheorie, dass jeder Messwert aus einem wahren Wert (entsprechend dem auf das Konstrukt zurückführbaren Anteil) und einem Fehleranteil besteht (▶ Abschn. 10.4.4 sowie Moosbrugger, 2011). Überträgt man die Beziehungen des in ◘ Abb. 17.4 dargestellten



◘ Abb. 17.4 Beispiel-Pfaddiagramm einer konfirmatorischen Faktorenanalyse mit zwei latenten Variablen (Konstrukten) ξ_1 und ξ_2, operationalisiert durch die manifesten Indikatorvariablen x_1, x_2, x_3 bzw. x_4, x_5. Im Modell sind verschiedene Parameter enthalten (zu den verwendeten griechischen Buchstaben siehe Box ▶ Variablen und Parameter): Faktorladungen λ^x, Messfehlervarianzen θ^δ, die Varianzen $\varphi_{11}, \varphi_{22}$ der latenten Variablen und deren Kovarianz φ_{21}. Varianzen als Modellparameter sind hier *in Klammern* notiert

Variablen und Parameter

In Strukturgleichungsmodellen werden Variablen (◘ Tab. 17.1) und Parameter (◘ Tab. 17.2; Pfadkoeffizienten, Varianzen/Kovarianzen) oft mit griechischen Buchstaben benannt. Die Bezeichnungen stammen aus dem Strukturgleichungsmodell-Programm LISREL (Jöreskog & Sörbom, 1996; ▶ Abschn. 17.6), werden aber ähnlich auch in anderen Programmen und in methodischer Literatur verwendet (z. B. Bollen, 1989).

◘ Tabelle 17.1 Variablen in Strukturgleichungsmodellen

Variablen		
ξ	xi	Latente exogene (Prädiktor-)Variablen
η	eta	Latente endogene (Kriteriums-/Mediator-) Variablen
x	x	Manifeste Indikatorvariablen der latenten exogenen (Prädiktor-)Variablen ξ
y	y	Manifeste Indikatorvariablen der latenten endogenen (Kriteriums-/Mediator-)Variablen η
δ	delta	Messfehlervariablen der manifesten Indikatorvariablen x
ε	epsilon	Messfehlervariablen der manifesten Indikatorvariablen y
ζ	zeta	Residualvariablen (unerklärte Anteile) latenter endogener Variablen η

Tabelle 17.2 Parameter in Strukturgleichungsmodellen

Parameter		Matrizen	
γ	gamma	Γ	Effekte (Pfadkoeffizienten) latenter exogener Variablen ξ auf latente endogene Variablen η
β	beta	B	Effekte (Pfadkoeffizienten) latenter endogener Variablen η auf andere latente endogene Variablen η
λ^x	lambda-x	Λ^x	Effekte (Pfadkoeffizienten) latenter exogener Variablen ξ auf manifeste Indikatorvariablen x (Faktorladungen der x-Variablen)
λ^y	lambda-y	Λ^y	Effekte (Pfadkoeffizienten) latenter endogener Variablen η auf manifeste Indikatorvariablen y (Faktorladungen der y-Variablen)
θ^δ	theta-delta	Θ^δ	Varianzen der Messfehlervariablen δ (Fehlervarianzen/nicht durch latente Variablen erklärbare Anteile der x-Indikatorvariablen)
θ^ε	theta-epsilon	Θ^ε	Varianzen der Messfehlervariablen ε (Fehlervarianzen/nicht durch latente Variablen erklärbare Anteile der y-Indikatorvariablen)
φ	phi	Φ	Varianzen und Kovarianzen der latenten exogenen (Prädiktor-)Variablen ξ
ψ	psi	Ψ	Varianzen der Residualvariablen ζ (nicht durch andere latente Variablen erklärbare Anteile von η-Variablen)

Modells in lineare Gleichungen, so erhält man als Messmodell:

$$\begin{aligned} x_1 &= \lambda^x_{11}\xi_1 + \delta_1 & x_4 &= \lambda^x_{42}\xi_2 + \delta_4 \\ x_2 &= \lambda^x_{21}\xi_1 + \delta_2 & x_5 &= \lambda^x_{52}\xi_2 + \delta_5. \\ x_3 &= \lambda^x_{31}\xi_1 + \delta_3 \end{aligned} \qquad (17.1)$$

Die Faktorladungen λ^x (lambda-x) geben dabei an, wie eine manifeste Variable x durch eine latente Variable ξ erklärt wird: Steigt bei einer Person der Wert der latenten Variable ξ um 1, so würde der Wert der manifesten Variablen um λ^x steigen. Unter sonst gleichen Bedingungen (gleiche Maßstäbe aller Variablen) würde also eine betraglich hohe Faktorladung auf eine bessere Messung/einen engeren Zusammenhang der Messung zum Konstrukt hindeuten als eine betraglich niedrigere Faktorladung. Der erste Index der Ladung bezeichnet dabei immer die Variable, auf die der Effekt wirkt (auf die der Pfeil in einem Pfaddiagramm hinführt), der zweite Index die Variable, von der der Effekt ausgeht (von der der Pfeil wegführt): λ^x_{42} ist also der Effekt der latenten Variable ξ_2 auf die manifeste Variable x_4.

Neben den Faktorladungen gibt es weitere Parameter im Modell: Die Messfehlervarianzen θ^δ (theta-delta; Varianzen der δ-Variablen) geben an, wie groß Messfehlereinflüsse in den beobachteten Variablen sind. Die Varianzen und Kovarianzen der latenten Variablen ξ werden mit φ (phi) bezeichnet und geben an, wie groß interindividuelle Unterschiede in den latenten Variablen (Varianzen) bzw. Zusammenhänge zwischen den latenten Variablen sind (Kovarianzen).

Konfirmatorische und explorative Faktorenanalyse

Die Analyse eines Messmodells alleine (wie in Abb. 17.4) ist ein einfacher Sonderfall der Analyse von Strukturgleichungsmodellen, nämlich eine **konfirmatorische Faktorenanalyse**. Wie bei explorativen Faktorenanalysen, z. B. Hauptachsenanalysen (▶ Abschn. 10.4), wird hier versucht, Zusammenhänge zwischen beobachteten Variablen auf wenige, dahinterstehende unbeobachtbare Konstrukte (Faktoren) zurückzuführen. In explorativen Faktorenanalysen kann aber nicht statistisch festgelegt werden, welche beobachteten Variablen auf welchen Faktoren laden sollen, und man kann nur darauf hoffen, dass sich bei der Analyse eine möglichst eindeutige Zuordnung von beobachteten Variablen zu Faktoren ergibt, die sog. Einfachstruktur. Ebenso kann bei explorativen Faktorenanalysen nur global festgelegt werden, ob alle Faktoren miteinander zusammenhängen dürfen oder alle unabhängig voneinander sein sollen. Bei konfirmatorischen Faktorenanalysen wird dagegen a priori festgelegt, welche beobachteten Variablen auf welchen Faktoren laden sollen und auf welchen Faktoren nicht, und zwischen welchen Faktoren man Zusammenhänge erwartet und zwischen welchen nicht. Diese Annahmen lassen sich dann exakt prüfen, wie unten in ▶ Abschn. 17.3 beschrieben wird.

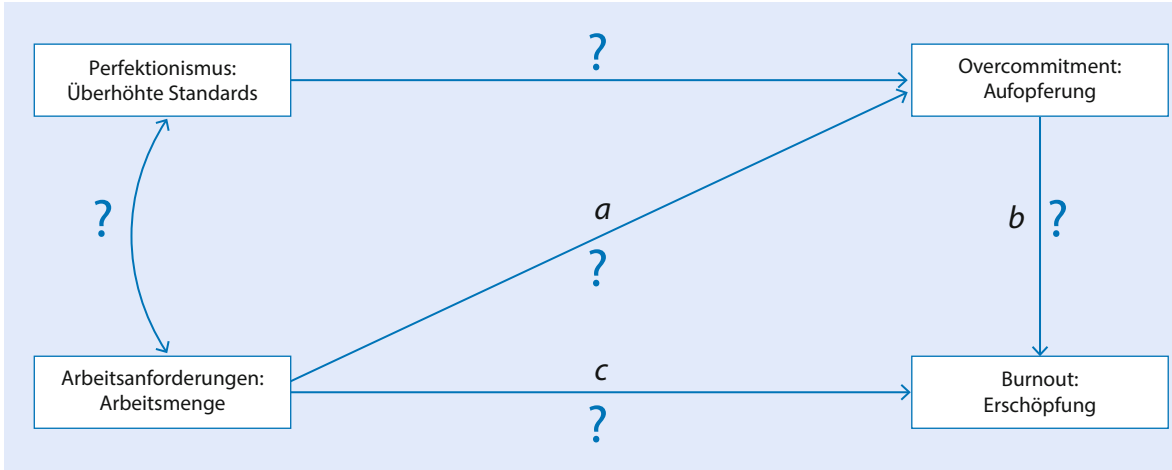


Abb. 17.5 Schema einer Pfadanalyse komplexer Beziehungen zwischen mehreren beobachteten Variablen am Beispiel vermuteter Effekte von *Perfektionismus (überhöhte Standards)* und *Arbeitsanforderungen (Arbeitsmenge)* über *Overcommitment (Aufopferung)* auf *Burnout (Erschöpfung)*. Wie in einem vollständigen Strukturgleichungsmodell können neben direkten Effekten (z. B. *Pfad c*) auch indirekte, über eine Mediatorvariable vermittelte Effekte analysiert werden (*Pfad* $a \times b$), jedoch nur auf der Ebene beobachteter Variablen

17.2.2 Effekte zwischen Konstrukten: Regressions-/Pfadanalyse

Wenn über Messmodelle im Sinne der Faktorenanalyse festgelegt ist, wie die interessierenden Konstrukte als latente Variablen durch manifeste Indikatorvariablen operationalisiert werden, so lassen sich vermutete Effekte zwischen verschiedenen Konstrukten in Strukturgleichungsmodellen direkt auf der Ebene der latenten Variablen analysieren. Im einfachsten Fall entspricht dies einer Regressionsanalyse mit einer oder mehreren latenten Variablen als Prädiktoren, und einer latenten Variable als Kriterium (Abb. 17.1). Zwischen latenten Variablen in Strukturgleichungsmodellen können aber nicht nur Prädiktor-Kriteriums-Beziehungen im Sinne einer einfachen oder multiplen Regression modelliert werden. Wie eingangs erwähnt, können Variablen auch gleichzeitig Kriteriums- und Prädiktorvariablen sein, sog. **Mediatorvariablen** (Baron & Kenny, 1986), womit indirekte Effekte analysiert werden können.

Direkter und indirekter Effekt – Ein direkter Effekt liegt vor, wenn eine Prädiktorvariable eine Kriteriumsvariable unmittelbar beeinflusst. Als indirekter Effekt wird der Einfluss einer Prädiktorvariable auf eine Kriteriumsvariable bezeichnet, der durch eine dritte Variable (Mediatorvariable) vermittelt (mediiert) wird: Die Prädiktorvariable hat einen Effekt auf die Mediatorvariable, und diese hat einen Effekt auf die Kriteriumsvariable.

Auf der Ebene beobachteter Variablen kann man solche indirekten Beziehungen in einer Pfadanalyse untersuchen (Wright, 1918, 1934). Pfadanalysen kann man als Kombination mehrerer simultaner Regressionsanalysen auffassen (Duncan, 1966, S. 6; Wright, 1934, S. 170ff.), womit sich in einer einzigen Analyse direkte Effekte und über eine dritte Variable vermittelte (mediierte), indirekte Effekte untersuchen lassen (Abb. 17.5). Gehen wir z. B. davon aus, dass Arbeitsanforderungen zu Burnout führen, so würden wir auf der Ebene beobachteter Variablen einen direkten Effekt der Arbeitsmenge auf die Erschöpfung annehmen (Abb. 17.5, Pfad c). Wenn die Arbeitsmenge darüber hinaus auch zu Aufopferung führt (Pfad a), und Aufopferung ihrerseits zu Erschöpfung (Pfad b), so bestünde auf diesem Weg auch ein indirekter, durch Aufopferung vermittelter Effekt der Arbeitsmenge auf die Erschöpfung. Dieser indirekte Effekt von Arbeitsmenge auf Erschöpfung ergibt sich dann als Produkt der beiden Pfadkoeffizienten ($a \times b$), und der totale Effekt (Gesamteffekt) der Arbeitsmenge auf die Erschöpfung als Summe von direktem Effekt und indirekten Effekten (hier: $c + a \times b$, wobei in komplexeren Modellen auch mehrere indirekte Effekte über verschiedene Mediatorvariablen bestehen könnten).

Wie in Abb. 17.2 dargestellt, können in Strukturgleichungsmodellen solch komplexe und indirekte Beziehungen zwischen Konstrukten wie in einer Pfadanalyse, aber auf der Ebene der latenten Variablen untersucht werden. Umgekehrt kann man Pfadanalysen als einfache Sonderfälle von Strukturgleichungsmodellen auffassen, bei denen jedes beteiligte Konstrukt nur durch eine einzige manifeste Variable gemessen wird. Ausführlichere Erläuterungen zu Pfadanalysen manifester Variablen finden sich z. B. bei Reinecke (2014). Kritik am ursprünglichen Mediatoransatz von Baron und Kenny (1986) findet sich

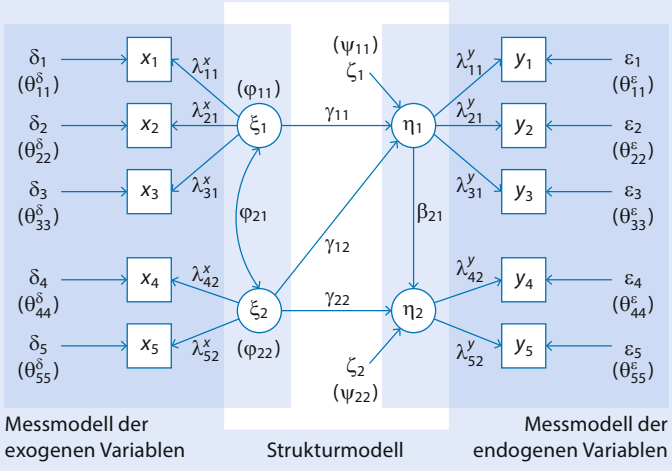


Abb. 17.6 Pfaddiagramm eines Strukturgleichungsmodells mit zwei latenten Prädiktorvariablen ξ_1, ξ_2 (operationalisiert durch die manifesten Variablen x_1, x_2, x_3 bzw. x_4, x_5), einer Mediatorvariablen η_1 (operationalisiert durch y_1, y_2, y_3) und einer Kriteriumsvariablen η_2 (operationalisiert durch y_4, y_5). In diesem Beispiel würden direkte Effekte von ξ_1 und ξ_2 auf η_1 modelliert (Parameter γ_{11}, γ_{12}), von ξ_2 auf η_2 (γ_{22}), ein direkter Effekt von η_1 auf η_2 (β_{21}) und damit auch indirekte, durch η_1 mediierte (vermittelte) Effekte von ξ_1 und ξ_2 über η_1 auf η_2. Zu schätzende Varianzen als Modellparameter sind bei den jeweiligen Variablen *in Klammern* dargestellt

z. B. bei Gelfand, Mensinger und Tenhave (2009) sowie Zhao, Lynch und Chen (2010).

Strukturmodell

Das **Strukturmodell** ist die Gesamtheit aller theoretisch angenommenen Beziehungen zwischen den latenten Variablen. Im Strukturmodell werden latente Variablen durch andere latente Variablen vorhergesagt (Abb. 17.6 *Mitte*). Die latenten unabhängigen Variablen (von denen nur Effekte ausgehen) werden auch als **exogene Variablen** bezeichnet (d. h. nicht innerhalb des Modells erklärte Variablen) und üblicherweise mit ξ symbolisiert, ihre Indikatorvariablen mit x und die zugehörigen Messfehlervariablen mit δ. Die latenten Mediator- und abhängigen Variablen, die im Strukturmodell vorhergesagt werden sollen, werden als **endogene Variablen** bezeichnet (d. h. innerhalb des Modells erklärte Variablen) und mit η (eta) symbolisiert, ihre Indikatorvariablen mit y und die zugehörigen Messfehlervariablen mit ε (epsilon). Parameter für die Effekte von exogenen auf endogene Variablen, also von ξ- auf η-Variablen, werden üblicherweise mit γ (gamma) symbolisiert, Effekte zwischen latenten endogenen Variablen (wie bei Mediatoreffekten) mit β (beta).

Ebenso wie in den Messmodellen angenommen wird, dass manifeste Variablen nicht perfekt durch latente Variablen erklärbar sind, sondern auch unerklärbare Anteile Messfehler (und ggf. spezifische Varianz) enthalten, geht man analog im Strukturmodell davon aus, dass latente Variablen nicht perfekt durch andere latente Variablen erklärbar sind, sondern unerklärbare Anteile verbleiben können. Da man latente Variablen nicht direkt beobachten kann, sind unerklärbare Anteile nicht auf der Ebene individueller Werte bestimmbar. In Strukturgleichungsmodellen kann allerdings geschätzt werden, wie groß der insgesamt im Modell nicht erklärbare Varianzanteil einer endogenen Variablen ist. Hierzu wird für jede endogene Variable η der jeweils nicht auf andere latente Variablen zurückführbare Anteil durch eine **Residualvariable** (Modellgleichungsfehlervariable) ζ (zeta) symbolisiert. Die Varianzen dieser Residualvariablen entsprechen damit den nicht durch andere latente Variablen erklärbaren Anteilen der jeweiligen endogenen Variablen. Als Modellparameter werden diese Varianzen geschätzt und üblicherweise mit ψ (psi) bezeichnet (Abb. 17.6).

Vollständiges Strukturgleichungsmodell

Ein vollständiges Strukturgleichungsmodell besteht aus dem Strukturmodell mit den angenommenen Beziehungen zwischen latenten Variablen und Messmodellen zur Operationalisierung der latenten Variablen über manifeste Indikatorvariablen. Dabei ist es grundsätzlich möglich, ein einziges Messmodell für alle latenten Variablen zu formulieren, oder aber man verwendet separate Messmodelle für die latenten Prädiktorvariablen (exogene Variablen) und für die latenten Mediator- und Kriteriumsvariablen (endogene Variablen). Ein vollständiges Strukturgleichungsmodell mit Strukturmodell und Messmodellen für die exogenen und endogenen latenten Variablen ist in

Abb. 17.6 dargestellt. Dieses Beispielmodell entspricht den schon in Abb. Abb. 17.2 schematisch dargestellten Annahmen.

Die rechnerische Analyse von Strukturgleichungsmodellen erfolgt, wie schon oben für ein Messmodell dargestellt, über Systeme linearer Gleichungen, die die angenommenen Beziehungen zwischen den beteiligten Variablen beschreiben – vergleichbar zu Regressionsgleichungen. Hierbei kann man für jede abhängige Variable (x-, y-, η-Variablen) eine einzelne Gleichung formulieren oder aber, wenn man die Matrixschreibweise bevorzugt, diese Gleichungen integriert in Matrixschreibweise darstellen. Einzelgleichungen und Matrixschreibweise sind inhaltlich gleichwertig: Wenn man mit Matrixmultiplikation nicht vertraut ist, kann man genauso anhand der Einzelgleichungen nachvollziehen, welche rechnerischen Beziehungen in einem Strukturgleichungsmodell angenommen werden. In Matrixform beschreibt z. B. für das Strukturmodell eine einzige Gleichung (Gl. 17.2), wie alle latenten endogenen Variablen auf latente exogene sowie andere endogene Variablen zurückgeführt werden. Auch die Messmodelle lassen sich jeweils in einer Matrixgleichung zusammenfassen (Gl. 17.3; Gl. 17.4; potenziell sogar sämtliche Beziehungen des Modells in einer einzigen, komplexen Matrixgleichung). Auf diese Weise werden nicht nur die tatsächlich angenommenen Beziehungen zwischen Variablen in Form zu schätzender Parameter angegeben, sondern es wird auch ausdrücklich – nämlich durch die auf den Wert 0 festgesetzten, nicht zu schätzenden Parameter – angegeben, zwischen welchen Variablen **keine** Beziehung angenommen wird. Für das in Abb. 17.6 dargestellte Modell ergäben sich z. B. die folgenden Modellgleichungen:

Strukturmodell in Einzelgleichungen

$$\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1$$
$$\eta_2 = \gamma_{22}\xi_2 + \beta_{21}\eta_1 + \zeta_2.$$

Strukturmodell in Matrixschreibweise (die griechischen Großbuchstaben Γ (Gamma) und B (Beta) bezeichnen die Matrizen der Pfadkoeffizienten γ und β; der erste Index jedes Koeffizienten entspricht dabei der Zeile, der zweite Index der Spalte der Matrix):

$$\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}\gamma_{11} & \gamma_{12}\\ 0 & \gamma_{22}\end{pmatrix}\cdot\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} + \begin{pmatrix}0 & 0\\ \beta_{21} & 0\end{pmatrix}\cdot\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} + \begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix} \qquad (17.2)$$
$$\eta = \Gamma\cdot\xi + \mathrm{B}\cdot\eta + \zeta$$

Messmodelle in Einzelgleichungen

$$\begin{aligned} x_1 &= \lambda^x_{11}\xi_1 + \delta_1 & y_1 &= \lambda^y_{11}\eta_1 + \varepsilon_1\\ x_2 &= \lambda^x_{21}\xi_1 + \delta_2 & y_2 &= \lambda^y_{21}\eta_1 + \varepsilon_2\\ x_3 &= \lambda^x_{31}\xi_1 + \delta_3 & y_3 &= \lambda^y_{31}\eta_1 + \varepsilon_3\\ x_4 &= \lambda^x_{42}\xi_2 + \delta_4 & y_4 &= \lambda^y_{42}\eta_2 + \varepsilon_4\\ x_5 &= \lambda^x_{52}\xi_2 + \delta_5 & y_5 &= \lambda^y_{52}\eta_2 + \varepsilon_5 \end{aligned}$$

Messmodelle in Matrixschreibweise (die griechischen Großbuchstaben Λ^x (Lambda-x) und Λ^y (Lambda-y) bezeichnen die Matrizen der Faktorladungen λ^x und λ^y):

$$\begin{pmatrix}x_1\\ x_2\\ x_3\\ x_4\\ x_5\end{pmatrix} = \begin{pmatrix}\lambda^x_{11}=1 & 0\\ \lambda^x_{21} & 0\\ \lambda^x_{31} & 0\\ 0 & \lambda^x_{42}=1\\ 0 & \lambda^x_{52}\end{pmatrix}\cdot\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} + \begin{pmatrix}\delta_1\\ \delta_2\\ \delta_3\\ \delta_4\\ \delta_5\end{pmatrix} \qquad (17.3)$$
$$x = \Lambda^x\cdot\xi + \delta$$

$$\begin{pmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\end{pmatrix} = \begin{pmatrix}\lambda^y_{11}=1 & 0\\ \lambda^y_{21} & 0\\ \lambda^y_{31} & 0\\ 0 & \lambda^y_{42}=1\\ 0 & \lambda^y_{52}\end{pmatrix}\cdot\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} + \begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\end{pmatrix} \qquad (17.4)$$
$$y = \Lambda^y\cdot\eta + \varepsilon$$

Skalierung

Latente Variablen wie Perfektionismus oder Burnout sind nicht direkt beobachtbar, daher gibt es für sie keine eindeutigen Maßeinheiten und damit auch keine eindeutigen Varianzen. Zum Schätzen der Parameter in den Gleichungssystemen ist es aber erforderlich, für die latenten Variablen Maßeinheiten festzulegen: Nur wenn man ein Maß dafür hat, wie stark sich Personen auf einer Variable unterscheiden, lässt sich auch eine Aussage machen, wie groß der Effekt dieser Variable auf eine andere Variable ist.

Um Maßeinheiten für die latenten Variablen auf eine sinnvolle Weise festzulegen, ist daher in den obigen Gleichungssystemen **pro latenter Variable jeweils eine Faktorladung auf den Wert 1 festgesetzt** worden ($\lambda^x_{11}, \lambda^x_{42}, \lambda^y_{11}, \lambda^y_{42}$). Die entsprechenden manifesten Variablen (x_1, x_4, y_1, y_4) dienen dadurch als sog. **Skalierer** zur Festlegung der Varianzen der latenten Variablen ($\xi_1, \xi_2, \eta_1, \eta_2$). Durch die Festsetzung von jeweils einer Faktorladung pro latenter Variable auf den Wert 1 wird die Varianz der latenten Variable gleichgesetzt mit der erklärten Varianz dieser Indikatorvariablen. Auf diese Weise wird die

Maßeinheit der latenten Variablen interpretierbar: Eine Erhöhung der latenten Variable um einen bestimmten Betrag führt dann zu einer Erhöhung dieser Skalierervariable um genau den gleichen Betrag (Bollen, 1989, Kap. 6; Jöreskog & Sörbom, 1996, S. 339f.). In der Praxis verwendet man dazu pro Konstrukt möglichst die reliabelste und valideste Indikatorvariable oder, wenn nichts Näheres bekannt ist, meist einfach die jeweils erste.

Eine grundsätzliche Alternative bestände darin, die Varianzen latenter Variablen direkt auf einen festen Wert zu setzen, z. B. auf den Wert 1. Damit wären die Variablen von vornherein standardisiert und Zusammenhänge zwischen ihnen betraglich wie Korrelationskoeffizienten interpretierbar. Dies funktioniert allerdings nur für faktorenanalytische Modelle (ohne endogene Variablen): Die Varianzen der latenten unabhängigen (exogenen) Variablen sind Modellparameter, die man auf einen festen Wert setzen könnte. Dagegen setzen sich die Varianzen der latenten abhängigen (endogenen) Variablen aus erklärten und unerklärten Anteilen zusammen, was bedeutet, dass sie sich aus anderen Modellparametern ergeben, sich aber nicht direkt festlegen lassen.

17.3 Analyse von Strukturgleichungsmodellen

Die Analyse von Strukturgleichungsmodellen ist viel rechenaufwendiger als Analysemethoden auf der Ebene manifester Variablen, wie z. B. Regressionsanalysen. Die Methoden dafür wurden ab Ende der 1960er-Jahre aus dem Kontext der Pfadanalyse (Duncan, 1966, S. 16) und der konfirmatorischen Faktorenanalyse entwickelt (Jöreskog, 1970, 1973; einen Überblick zur methodischen Entstehung geben Goldberger & Duncan, 1973).

Strukturgleichungsmodelle sind aber erst mit der allgemeinen Verbreitung leistungsfähiger PCs und entsprechender Analyseprogramme in den letzten 20 Jahren von Großrechenanlagen unabhängig geworden und werden seitdem immer häufiger eingesetzt (Hershberger, 2003). Heute steht eine Reihe von Programmpaketen zur Analyse von Strukturgleichungsmodellen auf PCs zur Verfügung (► Abschn. 17.6).

17.3.1 Kovarianzstrukturanalyse

Die Analyse eines Strukturgleichungsmodells anhand empirischer Daten erfolgt in der Regel kovarianzanalytisch: Aus empirischen Daten der an einer Stichprobe erhobenen manifesten Indikatorvariablen werden alle Kovarianzen (Zusammenhänge) dieser Indikatorvariablen bestimmt und in der **empirischen Kovarianzmatrix** S der Indikatorvariablen zusammengefasst. Analog ergeben sich auch aus den im Modell angenommenen Beziehungen zwischen den Variablen bestimmte (von den Parameterwerten abhängige) Zusammenhangsmuster, zusammengefasst in der **modelltheoretischen Kovarianzmatrix** $\hat{\Sigma}$ (Sigma-Dach). Um Schätzungen für die gesuchten Modellparameter zu erhalten und die aufgestellten Annahmen zu prüfen, werden diejenigen Werte für die Parameter gesucht, bei denen das Modell die Zusammenhänge in den empirischen Daten so gut wie möglich wiedergeben kann. Das heißt, die Abweichungen zwischen empirischer und modelltheoretischer Kovarianzmatrix sollten möglichst klein werden.

Ein einfaches Beispiel: Drei beobachtete Variablen x_1, x_2, x_3 (z. B. Testaufgaben), die das gleiche Konstrukt messen sollen, haben in einer Stichprobe die folgende empirische Kovarianzmatrix S (var: Varianz; cov: Kovarianz):

$$S = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1.25 & .70 & .60 \\ .70 & 1.00 & .42 \\ .60 & .42 & 1.00 \end{pmatrix} \end{array} = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) & \mathrm{cov}(x_1, x_3) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) & \mathrm{cov}(x_2, x_3) \\ \mathrm{cov}(x_3, x_1) & \mathrm{cov}(x_3, x_2) & \mathrm{var}(x_3) \end{pmatrix} \tag{17.5}$$

Im Sinne eines einfachen faktorenanalytischen Modells würde man die drei beobachteten Variablen als Indikatoren einer einzigen latenten Variablen ansehen und sie auf diese latente Variable sowie Messfehleranteile zurückführen (die Faktorladung λ_{11}^x ist zur Skalierung auf den Wert 1 gesetzt):

$$\begin{aligned} x_1 &= \quad 1\xi_1 + \delta_1 \\ x_2 &= \lambda_{21}^x \xi_1 + \delta_2 \\ x_3 &= \lambda_{31}^x \xi_1 + \delta_3 \end{aligned} \tag{17.6}$$

Um zu sehen, wie die modelltheoretischen Zusammenhänge dieser Variablen von den Parameterwerten abhängen, werden für die Indikatorvariablen x_1, x_2, x_3 deren Messmodellgleichungen in die Formeln für die Varianzen und Kovarianzen eingesetzt und diese ausmultipliziert. Als Beispiel ist dies hier gezeigt für die modelltheoretische Kovarianz der Variablen x_2 und x_3. Dabei wird die Annahme verwendet, dass Messfehler untereinander und

mit den latenten Variablen unkorreliert sind:

$$\begin{aligned}\text{cov}(x_2, x_3) &= \text{cov}(\lambda^x_{21}\xi_1 + \delta_2, \lambda^x_{31}\xi_1 + \delta_3)\\ &= \text{cov}(\lambda^x_{21}\xi_1, \lambda^x_{31}\xi_1) + \text{cov}(\lambda^x_{21}\xi_1, \delta_3)\\ &\quad + \text{cov}(\delta_2, \lambda^x_{31}\xi_1) + \text{cov}(\delta_2, \delta_3)\\ &= \lambda^x_{21}\lambda^x_{31}\text{cov}(\xi_1, \xi_1) + \lambda^x_{21}\text{cov}(\xi_1, \delta_3)\\ &\quad + \lambda^x_{31}\text{cov}(\delta_2, \xi_1) + \text{cov}(\delta_2, \delta_3)\\ &= \lambda^x_{21}\lambda^x_{31}\text{var}(\xi_1) + 0 + 0 + 0\\ &= \lambda^x_{21}\lambda^x_{31}\varphi_{11} \qquad (17.7)\end{aligned}$$

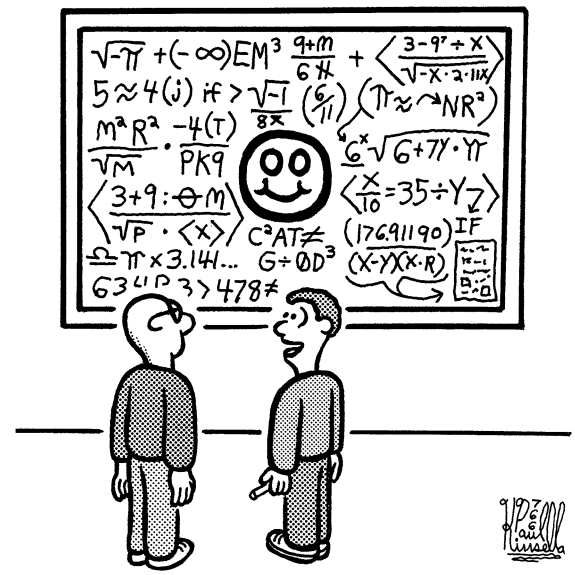
Die vom Modell erklärte Kovarianz zwischen x_2 und x_3 wird also umso größer, je größer die Faktorladungen λ^x_{21} und λ^x_{31} sind und je größer die Varianz φ_{11} der latenten Variable ist. Für die anderen modelltheoretischen Varianzen und Kovarianzen funktioniert dies analog. Insgesamt erhält man so die folgende modelltheoretische Kovarianzmatrix $\hat{\Sigma}$:

$$\begin{aligned}\hat{\Sigma} &= \begin{pmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_3)\\ \text{cov}(x_2, x_1) & \text{var}(x_2) & \text{cov}(x_2, x_3)\\ \text{cov}(x_3, x_1) & \text{cov}(x_3, x_2) & \text{var}(x_3)\end{pmatrix}\\ &= \begin{pmatrix} \varphi_{11} + \theta^\delta_{11} & \lambda^x_{21}\varphi_{11} & \lambda^x_{31}\varphi_{11}\\ \lambda^x_{21}\varphi_{11} & (\lambda^x_{21})^2\varphi_{11} + \theta^\delta_{22} & \lambda^x_{21}\lambda^x_{31}\varphi_{11}\\ \lambda^x_{31}\varphi_{11} & \lambda^x_{31}\lambda^x_{21}\varphi_{11} & (\lambda^x_{31})^2\varphi_{11} + \theta^\delta_{33}\end{pmatrix}\end{aligned} \qquad (17.8)$$

In diesem einfachen Fall gibt es Werte für die Modellparameter, mit denen die empirischen Zusammenhänge sogar perfekt reproduziert werden können (wie Nachrechnen zeigt):

$$\begin{aligned}\Lambda^x &= \begin{pmatrix}\lambda^x_{11}\\ \lambda^x_{21}\\ \lambda^x_{31}\end{pmatrix} = \begin{pmatrix}1.00\\ .70\\ .60\end{pmatrix}\\ \Phi &= (\varphi_{11}) = (1.0) \qquad (17.9)\\ \Theta^\delta &= \begin{pmatrix}\theta^\delta_{11} & 0 & 0\\ 0 & \theta^\delta_{22} & 0\\ 0 & 0 & \theta^\delta_{33}\end{pmatrix} = \begin{pmatrix}.25 & 0 & 0\\ 0 & .51 & 0\\ 0 & 0 & .64\end{pmatrix}\end{aligned}$$

Dies funktioniert hier aber nur deshalb perfekt, weil es sich um ein sehr einfaches Modell handelt (▶ Abschn. 17.5.2; ◘ Abb. 17.7). Für komplexere, theoretisch sinnvolle Modelle gibt es in aller Regel keine Parameterwerte, mit denen die empirischen Zusammenhänge perfekt reproduziert werden können ($S = \hat{\Sigma}$), sondern es können nur Parameterwerte gefunden werden, bei denen die Unterschiede zwischen S und $\hat{\Sigma}$ möglichst gering werden,



"Oh that?... It has nothing to do with the formula; it just makes the whole thing seem less grumpy."

◘ **Abb. 17.7** Wegen der Vielzahl der Symbole wirken Strukturgleichungsmodelle oft auf den ersten Blick etwas abschreckend. (© Paul Kinsella/Search ID: pknn17, Rights Available from CartoonStock.com)

mit denen also die empirischen Zusammenhänge vom Modell so gut wie möglich wiedergegeben werden.

17.3.2 Schätzmethoden

Analysemethoden für Strukturgleichungsmodelle versuchen generell, Parameter so zu schätzen, dass die Unterschiede zwischen den empirischen Zusammenhängen (empirische Kovarianzmatrix S) und den im Modell reproduzierten Zusammenhängen (modelltheoretische Kovarianzmatrix $\hat{\Sigma}$) möglichst gering werden. Während beim klassischen Signifikanztest gegen das H_0-Modell getestet wird und eine schlechte Passung der Daten zum H_0-Modell (Signifikanz) das Wunschergebnis darstellt, ist bei Strukturgleichungsmodellen die Nullhypothese, dass Daten und Modell gut zusammenpassen, das Wunschergebnis (▶ Abschn. 15.3). Verschiedene Analysemethoden unterscheiden sich darin, auf welche Weise das Maß für die Abweichungen zwischen empirischer und modelltheoretischer Kovarianzmatrix, die sog. **Fit-Funktion** der Anpassung des Modells an die Daten, bestimmt wird (Bollen, 1989; Jöreskog & Sörbom, 1996). Entsprechend können sich auch die mit den verschiedenen Methoden erhaltenen Parameterschätzungen für das gleiche Modell mehr oder weniger unterscheiden.

Die in Strukturgleichungsmodellen am häufigsten eingesetzte Analysemethode zur Parameterschätzung ist die **Maximum-Likelihood**-(ML-)Schätzung.

Maximum-Likelihood-Schätzung – Eine Maximum-Likelihood-(ML-)Schätzung in Strukturgleichungsmodellen erfolgt so, dass bei Gültigkeit der – von den Parametern abhängigen – modelltheoretischen Kovarianzmatrix in der Population die Likelihood (Plausibilität, „Wahrscheinlichkeit") maximal wird, in einer Stichprobe die gegebene empirische Kovarianzmatrix vorzufinden.

ML-Schätzungen haben theoretische Vorteile: So sind die Schätzungen asymptotisch (für große Stichproben) erwartungstreu, d. h. unverzerrt, und sie besitzen im Vergleich zu anderen Methoden maximale Effizienz (Bollen, 1989), d. h. sie sind bei gleichem Stichprobenumfang theoretisch genauer als die anderer Methoden (zu Gütekriterien für Schätzer ► Abschn. 12.4). Unter bestimmten Bedingungen können ML-Schätzungen außerdem auch bei unvollständigen Daten (Fälle mit fehlenden Werten) angewendet werden (Schafer & Graham, 2002)

Die Durchführung von ML-Schätzungen in Strukturgleichungsmodellen ist jedoch an Voraussetzungen gebunden: Für die Schätzungen werden, vereinfacht gesprochen, Wahrscheinlichkeiten benötigt, bei einer gegebenen (modelltheoretischen) Populationskovarianzmatrix bestimmte (empirische) Stichprobenkovarianzmatrizen zu beobachten. Dies basiert zum einen auf der Annahme, dass das Modell tatsächlich richtig ist, und funktioniert außerdem nur auf der Basis von Annahmen über die Verteilung der Variablen. Für die Anwendung von ML-Schätzungen besteht daher formal die Voraussetzung **multivariat normalverteilter Indikatorvariablen**.

Alternativ gibt es verschiedene Kleinstquadrate-Schätzverfahren. Das einfachste Beispiel hierfür ist die **Unweighted-Least-Squares**-(ULS-)Methode, bei der die Parameter so geschätzt werden, dass die quadrierten Differenzen zwischen den sich entsprechenden Einträgen aus empirischer und modelltheoretischer Kovarianzmatrix minimiert werden, wofür keine Annahmen über Verteilungsformen der beteiligten Variablen nötig sind (zu Anwendungsaspekten der verschiedenen Schätzmethoden ► Abschn. 17.5.5).

17.3.3 Modellgüte und Test der Annahmen

Selbst bei bestmöglicher Anpassung zwischen Modell und empirischen Daten verbleiben in praktischen Analysen Abweichungen zwischen empirischer und modelltheoretischer Kovarianzmatrix, die **Residuen** $S - \hat{\Sigma}$.

Residuen – Residuen sind Differenzen zwischen den empirischen Kovarianzen und den durch das Modell reproduzierten (modelltheoretischen) Kovarianzen. Je kleiner diese Residuen sind, desto besser kann das Modell insgesamt die empirischen Zusammenhänge erklären, d. h. desto besser stimmen Modell und Daten überein. Je größer die Residuen sind, desto schlechter ist die Passung zwischen Modell und Daten. Die Residuen $S - \hat{\Sigma}$ sollten nicht verwechselt werden mit den Residualvariablen im Modell, die unerklärte Anteile endogener latenter Variablen darstellen.

Diese Residuen entstehen durch 2 Einflüsse: einerseits durch zufällige Schwankungen der Zusammenhänge von Stichprobe zu Stichprobe, selbst wenn das Modell in der Population richtig wäre (Stichprobenfehler), andererseits aber auch durch systematische Fehler des Modells, bei denen die Zusammenhänge selbst in der gesamten Population nicht richtig wiedergegeben würden (Modellfehler; MacCallum & Tucker, 1991).

Die Residuen liefern damit ein Kriterium zur Prüfung der im Modell aufgestellten Annahmen über die Beziehungen zwischen den Variablen: Wenn auch bei bestmöglicher Parameterwahl immer noch große Diskrepanzen zwischen empirischer und modelltheoretischer Matrix verbleiben, deutet dies an, dass sich die empirischen Zusammenhänge der Indikatorvariablen insgesamt nur schlecht über die im Modell angenommenen Beziehungen erklären lassen. Also wäre zu vermuten, dass die im Modell aufgestellten Annahmen zumindest teilweise falsch sind. Sind die Residuen dagegen klein, lassen sich die im Modell aufgestellten Annahmen mit den empirischen Daten vereinbaren.

Die Größe der Residuen $S - \hat{\Sigma}$ ist allerdings – wie die Größe der Kovarianzen selbst – von den Maßstäben, d. h. den Varianzen der Variablen abhängig. Zur praktischen Beurteilung der Passung zwischen Modell und Daten, des sog. Modell-Fits, sind daher **standardisierte Residuen** besser geeignet, die von Strukturgleichungsmodellprogrammen ausgegeben werden können. Standardisierte Residuen sind unabhängig von den Maßstäben der Variablen wie z-Werte interpretierbar, d. h. Werte betraglich > 1.96 wären auf dem 5%-Niveau signifikant. Dabei betrachtet man die Residuen jeweils einzeln, d. h. man kann anhand der standardisierten Residuen lokale Aussagen machen, inwieweit das Modell stimmt: Eine global recht gute Passung zwischen Modell und Daten kann z. B. dadurch zustande kommen, dass ein Modell alle empirischen Zusammenhänge recht gut wiedergibt, somit alle Residuen relativ klein sind. Sie kann aber auch dadurch zustande kommen, dass fast alle Zusammenhänge perfekt wiedergegeben werden, einige wenige Zusammenhänge aber nur sehr schlecht, so dass diese Residuen recht groß sind. Ein auffällig großes standardisiertes Residuum kann damit auf eine lokale Fehlspezifikation (falsche Annahme im Modell) hinweisen.

Um die Passung zwischen Modell und Daten als Ganzes inferenzstatistisch zu beurteilen, gibt es einen **χ^2-Test** der globalen Abweichungen zwischen Modell und Daten:

Wenn die Annahmen des Modells richtig sind, sollte der χ^2-Test kein signifikantes Ergebnis anzeigen, also nicht auf Abweichungen zwischen Modell und Daten hindeuten (Beibehaltung der Nullhypothese als Wunschhypothese ▶ Abschn. 15.3).

Zur zusammenfassenden quantitativen Einschätzung, wie stark Modell und Daten insgesamt voneinander abweichen bzw. wie gut ein Modell die empirischen Zusammenhänge erklärt, gibt es außerdem verschiedene **deskriptive Gütekriterien** als globale Maße der Übereinstimmung bzw. Abweichung zwischen Modell und Daten, worauf genauer beim folgenden Anwendungsbeispiel eingegangen wird.

Wenn das Modell insgesamt zufriedenstellend zu den Daten passt, so dass man davon ausgehen kann, dass es korrekt spezifiziert ist, können auch die einzelnen **Parameterschätzungen** sinnvoll inhaltlich interpretiert werden. Interessierende Parameter können anhand der geschätzten Standardfehler der Parameter auf Signifikanz getestet werden, alternativ auch durch Likelihood-Ratio- bzw. χ^2-Differenztests (vgl. Bollen, 1989; Kline, 2011). Weiterhin kann die Vorhersagequalität des Modells beurteilt werden, d. h. die erklärte Varianz der einzelnen abhängigen Variablen. Dies zeigen wir am folgenden Anwendungsbeispiel.

17.4 Anwendungsbeispiel: Ein Strukturgleichungsmodell zu Burnout

Konstrukte und Fragestellung

Burnout ist ein psychologisches Syndrom, das durch emotionale Erschöpfung, eine distanzierte und zynische Einstellung zur Arbeit (Depersonalisation) und reduzierte persönliche Leistungsfähigkeit gekennzeichnet ist (Maslach, Jackson & Leiter, 1997). Die Entstehung von Burnout wurde in zahlreichen Studien durch Arbeitsplatzmerkmale zu erklären versucht (z. B. Bakker, Demerouti & Verbeke, 2004), wobei sich insbesondere hohe Arbeitsanforderungen (z. B. Arbeitsmenge, Zeitdruck) als Prädiktoren von Burnout erwiesen haben. Trotz Identifikation relevanter Arbeitsplatzmerkmale ist noch nicht hinreichend erklärt, warum einige Personen unter bestimmten Arbeitsplatzmerkmalen Burnout entwickeln, andere dagegen nicht. Persönlichkeitsfaktoren könnten helfen, derartige interindividuelle Unterschiede zu erklären.

Philp, Egan und Kane (2012) untersuchten die Ausprägung des Perfektionismus als möglichen Prädiktor auf Persönlichkeitsebene. In seiner negativen Ausprägung ist Perfektionismus gekennzeichnet durch das Setzen unrealistisch hoher Standards, selektive Aufmerksamkeit gegenüber Fehlern und deren Übergeneralisierung, kritische Evaluation des eigenen Handelns sowie eine Tendenz zum „Alles-oder-Nichts-Denken", so dass für Personen mit hohem Perfektionismus nur vollkommener Erfolg oder vollständiges Versagen möglich sei (Hewitt & Flett, 1991). Philp et al. (2012) gehen davon aus, dass der Effekt von Perfektionismus auf Burnout durch einen bestimmten Coping-Mechanismus zum Umgang mit beruflichen Anforderungen vermittelt (mediiert) wird, der durch hohe Einsatzbereitschaft für die Arbeit über das geforderte Maß hinaus sowie mangelnde Distanzierungsfähigkeit von beruflichen Anforderungen gekennzeichnet ist. Dies wird als „übersteigertes Engagement" (Peter, 2002) oder Overcommitment (Siegrist et al., 2004) bezeichnet. Es könnte auch vermutet werden, dass hohe Arbeitsanforderungen ihrerseits zu Overcommitment beitragen und sie so neben ihrem direkten Effekt auch indirekt über Overcommitment auf Burnout wirken. Diese Beziehungen sollen in ihrer Gesamtheit anhand eines Strukturgleichungsmodells überprüft werden. Die angenommenen Beziehungen entsprechen dem schon in ◘ Abb. 17.2 dargestellten Pfadmodell.

Hypothesen

Das Strukturmodell umfasst die folgenden Hypothesen über die direkten Effekte zwischen den Konstrukten:

- H1: Hohe Arbeitsanforderungen führen zu Burnout.
- H2: Hohes Overcommitment führt zu Burnout.
- H3: Hoher Perfektionismus führt zu Overcommitment.
- H4: Hohe Arbeitsanforderungen führen zu Overcommitment.

Neben diesen vier direkten Effekten ergeben sich zwei indirekte Effekte. Perfektionismus und Arbeitsanforderungen haben jeweils einen indirekten Effekt auf Burnout, der durch Overcommitment vermittelt wird: Perfektionismus und Arbeitsanforderungen führen zu Overcommitment und dieses wiederum zu Burnout. Ein direkter Effekt von Perfektionismus auf Burnout wird nicht erwartet. Es wird davon ausgegangen, dass Perfektionismus und Arbeitsanforderungen korrelieren können.

Messung der Konstrukte

In den Messmodellen ist festgelegt, wie die Konstrukte über 10 Indikatorvariablen operationalisiert werden. Perfektionismus wird gemessen über die Variablen „Überhöhte Standards", „Übermäßige Selbstkritik" und „Alles-oder-nichts-Denken". Die Arbeitsanforderungen werden durch „Arbeitsmenge" und „Zeitdruck" erfasst. Für die Messung von Overcommitment werden die Variablen „Aufopferung", „Unfähigkeit abzuschalten" sowie „Arbeitsbezogenes Grübeln" als Indikatoren gewählt. Burnout wird durch „Erschöpfung" und „Depersonalisation"

Tabelle 17.3 Fiktive Daten der ersten drei Befragungsteilnehmer im Anwendungsbeispiel

standards	selbstkritik	alles.nichts	menge	zeitdruck	aufopferung	nicht.abschalten	gruebeln	erschoepfung	depersonalisation
39	28	36	23	28	37	28	26	30	20
19	34	20	11	5	13	18	20	12	20
31	39	49	42	32	27	25	35	39	37
...									

operationalisiert. Die Operationalisierung erfolgt über Fragebogenskalen (▶ Abschn. 8.4.4).

Empirische Daten

Eine Stichprobe von $N = 260$ Berufstätigen nahm an der Fragebogenstudie zum Thema Burnout teil (fiktive Daten). Für jede Fragebogenskala wurde jeweils der Summenwert aller Items der Skala bestimmt. Somit liegen pro Befragungsteilnehmer Werte für 10 Variablen vor. Diese Daten stehen hier durch Leerzeichen getrennt in einer einfachen Textdatei, wobei jede Spalte einer Variablen und jede Zeile einem Fall entspricht.

Im Auszug aus der Rohdatendatei in Tab. 17.3 sind die Daten der ersten 3 Befragungsteilnehmer zu sehen. Der vollständige Datensatz zu diesem Beispiel kann im ▶ Webangebot zum Buch heruntergeladen werden.

Analyse des Modells

Wir zeigen die Analyse hier exemplarisch für die Statistikprogrammumgebung R (R Development Core Team, 2015) und das dazugehörige R-Paket „lavaan“ zur Analyse von Strukturgleichungsmodellen (Rosseel, 2015, ausführlich beschrieben in Rosseel, 2012; ▶ Abschn. 17.6). R und lavaan sind als freie Software verfügbar. Wie auch bei mehreren anderen Programmen für Strukturgleichungsmodelle werden die Anweisungen zur Analyse in eine Textdatei geschrieben (Programmsyntax). Darin wird angegeben, welche Daten analysiert werden sollen, es wird das zu analysierende Modell definiert, und es können unterschiedliche Arten von Ergebnissen angefordert werden. Die Art dieser Programmsyntax wie auch die Präsentation der Ergebnisse unterscheidet sich zwischen verschiedenen Strukturgleichungsmodellprogrammen oft stark. Die statistischen Ergebnisse sind dagegen, von ungewöhnlicheren Modellen oder Analysemethoden abgesehen, bei verschiedenen Programmen meist weitgehend identisch.

In Abb. 17.8 ist die Syntax zur Strukturgleichungsmodellanalyse des Anwendungsbeispiels mit dem lavaan-Paket abgebildet. Die Syntax kann auch im ▶ Webangebot zum Buch heruntergeladen werden.

Zur Durchführung der Analyse wird zuerst die Statistikprogrammumgebung R installiert, falls noch nicht vorhanden, und aufgerufen. In R wird dann 1.) in das Verzeichnis gewechselt, in dem sich die Daten und die Analysesyntax befinden (unter deutschsprachigem R für Windows: Menü Datei → Verzeichnis wechseln), 2.) die Datei mit der Analysesyntax geöffnet (Menü Datei → Öffne Skript) und 3.) die Syntax ausgeführt (Menü Bearbeiten → Alles ausführen). Bei der ersten Benutzung wird das R-Zusatzpaket lavaan auf dem Rechner installiert (Anweisung install.packages("lavaan") zu Beginn der Syntax, setzt eine Internetverbindung voraus; alternativ kann das lavaan-Paket auf Windows- und Mac-Rechnern über das R-Menü zur Paketverwaltung installiert werden). Anschließend wird das installierte lavaan-Paket geladen, damit seine speziellen Funktionen genutzt werden können, es werden die Daten eingelesen, Mess- und Strukturmodell definiert und Ergebnisse angefordert.

Ergebnisse

Als Ergebnisse (▶ Ergebnisse der Analyse mit dem R-Paket lavaan, hier in Abb. 17.9 leicht gekürzt wiedergegeben) erhalten wir zuerst Angaben zur Analyse und zu den Modellgütekriterien. Die am Anfang aufgeführte „Minimum Function Test Statistic“ ist die χ^2-Teststatistik unseres Modells. Mit $\chi^2 = 36.292$ bei 30 Freiheitsgraden (df) und $p = .199$ weichen Modell und Daten nicht signifikant voneinander ab.

Es folgt die χ^2-Teststatistik für das Unabhängigkeitsmodell (Baseline), die in die Berechnung relativer Gütekriterien – wie z. B. des Comparative-Fit-Index (CFI) – eingeht. Der hohe Wert des CFI von .993 sowie die niedrigen Werte des Root Mean Square Error of Approximation (RMSEA) von .028 und des Standardized Root Mean Square Residual (SRMR) von .038 deuten ebenfalls darauf hin, dass das Modell korrekt ist und die Parameterschätzungen somit interpretiert werden können (▶ Modellgüte in ▶ Abschn. 17.5.4).

In den Ergebnissen (Abb. 17.9) folgen die Parameterschätzungen 1.) für die Faktorladungen (λ^x, λ^y) in

```
# Analysesyntax für das Beispiel-Strukturgleichungsmodell zu Burnout
# Analyse mit der freien Statistikumgebung R (http://www.r-project.org/)
# und dem lavaan-Paket für Strukturgleichungsmodelle (http://lavaan.org/)
# (für lavaan 0.5-18/April 2015; Aktualisierung ggf. im Webangebot zum Buch)

# R-Paket "lavaan" für Strukturgleichungsmodelle installieren
# (nur einmalig; sobald das Paket installiert ist, kann die
# install.packages()-Anweisung mit # vor der Zeile auskommentiert werden)
install.packages("lavaan")

# lavaan-Paket für Strukturgleichungsmodelle laden (bei jeder Analyse)
library("lavaan")

# Datendatei einlesen: In der ersten Zeile (header) stehen die Variablennamen
burnout.daten <- read.table("burnout_beispieldaten.dat", header=TRUE)
# Struktur der Daten anschauen: 260 Fälle, 10 Variablen
str(burnout.daten)

# Modell definieren: Zwei exogene latente Variablen
# (Prädiktor-Konstrukte) Perfektionismus und Arbeitsanforderungen,
# zwei endogene latente Variablen (Mediator-/Kriteriums-Konstrukte)
# Overcommitment und Burnout
# Die Modelldefinition steht in einfachen Anführungszeichen '.
# Das Symbol =~ bedeutet "wird gemessen durch",
# das Symbol ~ bedeutet "wird vorhergesagt durch".
burnout.modell <- '

    # Messmodelle
    perfektionismus     =~ standards + selbstkritik + alles.nichts
    anforderungen       =~ menge + zeitdruck
    overcommitment      =~ aufopferung + nicht.abschalten + gruebeln
    burnout             =~ erschoepfung + depersonalisation

    # Strukturmodell
    overcommitment      ~ perfektionismus + anforderungen
    burnout             ~ overcommitment + anforderungen
'

# Modell analysieren mit der Funktion sem()
burnout.analyse <- sem(model=burnout.modell, data=burnout.daten)

# Ergebnisse ausgeben einschliesslich Gütekriterien
# und standardisierter Parameterschätzungen
summary(burnout.analyse, fit.measures=TRUE, standardized=TRUE)
# Auf Wunsch werden noch ausführlichere Ergebnisse ausgegeben;
# bei Bedarf Kommentarzeichen vor den folgenden beiden Zeilenentfernen:
# summary(burnout.analyse,
# fit.measures=TRUE, standardized=TRUE, rsquare=TRUE, modindices=TRUE)
```

Abb. 17.8 Syntax zur Analyse mit dem R-Paket lavaan. In den mit # beginnenden Zeilen stehen Kommentare zur näheren Erläuterung der Anweisungen

lavaan (0.5–18) converged normally after 146 iterations

Number of observations	260
Estimator	ML
Minimum Function Test Statistic	36.292
Degrees of freedom	30
P-value (Chi-square)	0.199

Model test baseline model:

Minimum Function Test Statistic	963.995
Degrees of freedom	45
P-value	0.000

User model versus baseline model:

Comparative Fit Index (CFI)	0.993
Tucker-Lewis Index (TLI)	0.990

Loglikelihood and Information Criteria:

Loglikelihood user model (H_0)	–9183.562
Loglikelihood unrestricted model (H_1)	–9165.416
Number of free parameters	25
Akaike (AIC)	18417.124
Bayesian (BIC)	18506.141
Sample-size adjusted Bayesian (BIC)	18426.881

Root Mean Square Error of Approximation:

RMSEA	0.028
90 Percent Confidence Interval	0.000 0.057
P-value RMSEA <= 0.05	0.876

Standardized Root Mean Square Residual:

SRMR	0.038

Parameter estimates:

Information	Expected
Standard Errors	Standard

	Estimate	Std.err	Z-value	P(>|z|)	Std.lv	Std.all
Latent variables:						
perfektionismus =~						
standards	1.000				8.306	0.802
selbstkritik	0.929	0.087	10.628	0.000	7.715	0.741
alles.nichts	0.858	0.081	10.566	0.000	7.125	0.733
anforderungen =~						
menge	1.000				8.847	0.833
zeitdruck	0.829	0.109	7.615	0.000	7.335	0.733
overcommitment =~						
aufopferung	1.000				6.723	0.674
nicht.bschltn	1.117	0.110	10.190	0.000	7.512	0.772
gruebeln	1.095	0.104	10.501	0.000	7.361	0.838
burnout =~						
erschoepfung	1.000				7.729	0.814
depersonalstn	0.968	0.099	9.806	0.000	7.481	0.752

Abb. 17.9 Ergebnisse der Analyse mit dem R-Paket lavaan

Regressions:						
overcommitment ~						
perfektionsms	0.402	0.068	5.941	0.000	0.497	0.497
anforderungen	0.086	0.056	1.552	0.121	0.114	0.114
burnout ~						
overcommitmnt	0.523	0.087	6.042	0.000	0.455	0.455
anforderungen	0.445	0.073	6.132	0.000	0.509	0.509
Covariances:						
perfektionismus ~~						
anforderungen	15.461	5.889	2.626	0.009	0.210	0.210
Variances:						
standards	38.201	5.917			38.201	0.356
selbstkritik	48.971	6.094			48.971	0.451
alles.nichts	43.609	5.324			43.609	0.462
menge	34.499	9.644			34.499	0.306
zeitdruck	46.370	7.497			46.370	0.463
aufopferung	54.276	5.727			54.276	0.546
nicht.bschltn	38.203	4.920			38.203	0.404
gruebeln	23.017	3.906			23.017	0.298
erschoepfung	30.311	5.659			30.311	0.337
depersonalstn	43.006	6.011			43.006	0.435
perfektionsms	68.986	10.047			1.000	1.000
anforderungen	78.272	13.135			1.000	1.000
overcommitmnt	32.361	6.007			0.716	0.716
burnout	25.844	5.595			0.433	0.433

Abb. 17.9 (Fortsetzung)

den Messmodellen der latenten Variablen („Latent variables"), 2.) für die Effekte (γ, β) zwischen den latenten Variablen im Strukturmodell („Regressions"), 3.) für Kovarianzen („Covariances"), hier die Kovarianz zwischen den latenten exogenen Variablen (φ_{12}), und 4.) für Varianzen („Variances"), nämlich die Messfehlervarianzen der manifesten Variablen (θ^{δ}, θ^{ε}), die (Gesamt-)Varianzen der latenten exogenen (Prädiktor-)Variablen (φ_{11}, φ_{22}) sowie die unerklärten Varianzen der latenten endogenen (Mediator-/Kriteriums-)Variablen (ψ_{11}, ψ_{22}).

In der ersten Spalte der Ergebnisse (unter der Überschrift „Estimate") stehen unstandardisierte Schätzungen. Da wir zur Skalierung der latenten Variablen keine bestimmten Faktorladungen auf 1 festgesetzt hatten, wurde von lavaan bei jeder latenten Variable automatisch die Ladung der jeweils ersten manifesten Variable auf 1 gesetzt (also z. B. bei Perfektionismus die Ladung von Standards). Für alle nicht festgesetzten Ladungen erhalten wir in den drei folgenden Spalten eine Schätzung des Standardfehlers (Std.err), einen z-Wert (unstandardisierte Schätzung dividiert durch Standardfehler, d. h. erste durch zweite Spalte) sowie einen p-Wert für den z-Test, der prüft, ob der jeweilige Parameter ungleich null ist. Mit einer Irrtumswahrscheinlichkeit von jeweils 5 % sind hier also alle Modellparameter signifikant von null verschieden mit Ausnahme des Effekts von Arbeitsanforderungen auf Overcommitment ($p = .121$).

Es folgen 2 Spalten mit standardisierten Parameterschätzungen, zuerst für den Fall, dass nur die latenten Variablen standardisiert sind (Std.lv, nur selten benötigt), dann auf der Basis, dass alle latenten und manifesten Variablen standardisiert sind (Std.all). Standardisierte Variablen haben den Mittelwert null und die Standardabweichung eins. Bei Beziehungen zwischen Variablen (Faktorladungen, Effekte im Strukturmodell) sind daher die Angaben der letzten Spalte vom Wertebereich her interpretierbar als Korrelationskoeffizienten oder standardisierte Regressionskoeffizienten. Bei Varianzen (Fehlervarianzen der manifesten Variablen, unerklärte Varianzen der latenten endogenen Variablen) sind die Angaben der letzten Spalte interpretierbar als Anteile an der Gesamtvarianz, d. h. als Fehlervarianzanteil bei manifesten Variablen bzw. unerklärter Varianzanteil bei den latenten endogenen Variablen. Die Angabe von .433 bei Burnout bedeutet also, dass 43.3 % der Varianz der latenten Variable Burnout nicht im Modell erklärt werden können (Residualvarianz) oder umgekehrt, dass 56.7 % der Varianz im Modell erklärt werden. Durch den eindeutigen

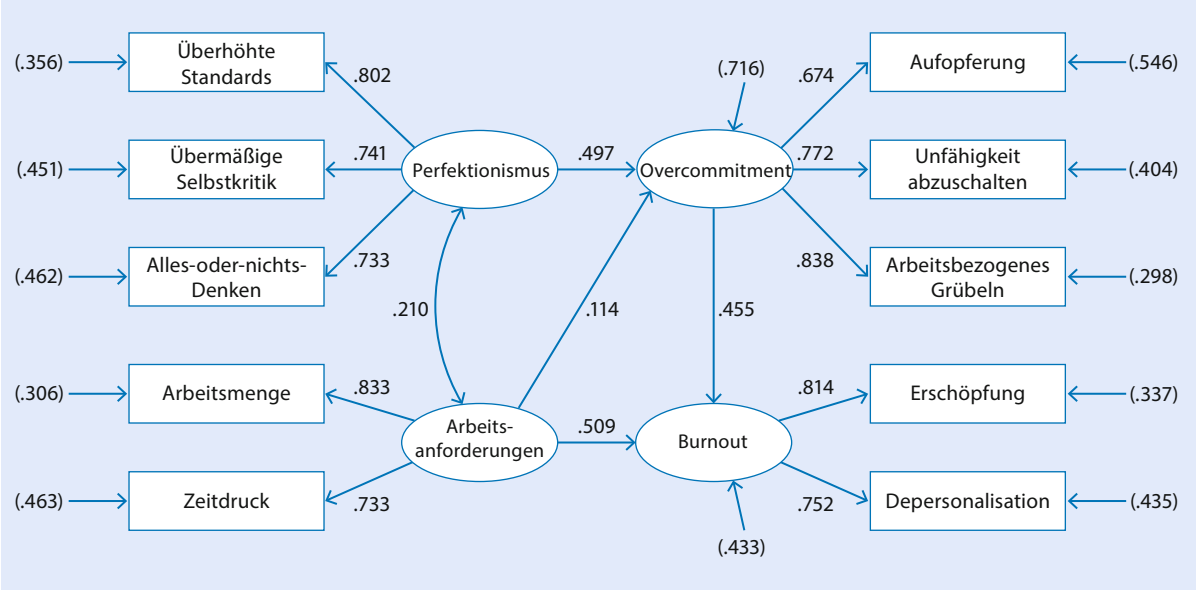


Abb. 17.10 Pfaddiagramm zum Anwendungsbeispiel mit den Parameterschätzungen auf der Basis standardisierter latenter und manifester Variablen

Wertebereich eignen sich diese Parameterschätzungen auf der Basis durchweg standardisierter Variablen meist besser zur inhaltlichen Interpretation als unstandardisierte Schätzungen, die durch unterschiedliche Maßeinheiten der Variablen betraglich beliebige Werte aufweisen können und damit schlecht vergleichbar sind. Als grafische Veranschaulichung der Ergebnisse sind diese standardisierten Parameterschätzungen in Abb. 17.10 dargestellt.

Zusammenfassend können fast alle Annahmen bestätigt werden: Das Modell passt insgesamt zu den Daten, insofern kann davon ausgegangen werden, dass sich die Konstrukte tatsächlich so operationalisieren lassen wie im Modell angenommen. Da das Modell zu den Daten passt, kann auch davon ausgegangen werden, dass keine wesentlichen empirischen Beziehungen zwischen den Variablen im Modell fehlen. Bis auf den angenommenen Effekt von Arbeitsanforderungen auf Overcommitment sind alle vermuteten Beziehungen signifikant. Perfektionismus wirkt somit nur indirekt (vermittelt über Overcommitment) auf Burnout, während sich die Arbeitsanforderungen nur direkt (jedoch nicht indirekt über Overcommitment) auf Burnout auswirken.

17.5 Praktischer Einsatz, Herausforderungen und Probleme

Strukturgleichungsmodelle bieten für quantitative Datenanalysen in den Sozialwissenschaften eine Reihe von Vorteilen: Durch die ausdrückliche Unterscheidung zwischen manifesten und latenten Variablen können Messfehlereinflüsse explizit berücksichtigt werden und oft validere Aussagen getroffen werden als bei Analysen, die nur auf manifesten Variablen beruhen. Komplexe Modelle können als Ganzes analysiert und insgesamt auf ihre Übereinstimmung mit empirischen Daten geprüft werden. Diese Komplexität bringt aber bei der statistischen Analyse und Interpretation auch Herausforderungen und potenzielle Probleme mit sich. Zur Abwägung zwischen Nutzen und Aufwand werden im Folgenden einige für die Anwendung zentrale Punkte erläutert.

17.5.1 Erklärung vs. Vorhersage von Variablen

Strukturgleichungsmodelle sind aufwendiger als Regressionsanalysen. Dieser Aufwand lohnt sich in Situationen, in denen eine **„regression for explanation“** (Maruyama, 1998) angestrebt wird. Hierbei besteht das Ziel darin, latente Kriteriumsvariablen inhaltlich auf latente Prädiktorvariablen zurückzuführen sowie zu beurteilen, welche Einflussfaktoren hierbei relevant sind und wie stark deren Einflüsse sind. Strukturgleichungsmodelle erlauben valide Aussagen über Effekte von Konstrukten, die nicht durch Messfehler beeinträchtigt sind und die getrennt werden können von Fragen der Operationalisierung der Konstrukte.

Wenig spezifischen Nutzen bieten Strukturgleichungsmodelle dagegen bei einer **„regression for prediction“**, bei der es primär um eine möglichst gute statistische Vor-

hersage einer beobachteten Kriteriumsvariablen geht, also um hohe Varianzerklärung, und nicht um die Prüfung gegebener theoretischer Annahmen. In diesem Falle kann ebenso gut eine regressionsanalytische Vorhersage auf der Ebene manifester Variablen zum Einsatz kommen.

17.5.2 Modellidentifikation

In Strukturgleichungsmodellen werden zahlreiche Parameter (Pfadkoeffizienten, Faktorladungen, Varianzen etc.) zur Modellierung vieler Hypothesen simultan so geschätzt, dass die empirischen Zusammenhänge der beobachteten Variablen insgesamt so gut wie möglich im Modell wiedergegeben werden. Dies funktioniert aber nur, wenn die empirischen Daten ausreichende Informationen für die Schätzung aller im Modell enthaltenen Parameter bereitstellen.

Strukturgleichungsmodelle beruhen üblicherweise nicht auf Rohdaten als Informationen, sondern auf der empirischen Kovarianzmatrix der Indikatorvariablen. Bei einer Anzahl von k Indikatorvariablen enthält diese (symmetrische) Kovarianzmatrix $k(k+1)/2 = s$ voneinander verschiedene Einträge als Informationen, bei 3 Indikatorvariablen also z. B. $3 \times 4/2 = 6$ Einträge (3 Varianzen, 3 Kovarianzen). Bezeichnet man die Anzahl der zu schätzenden Parameter im Modell mit t, so muss die Anzahl s der Einträge der Kovarianzmatrix groß genug sein, um alle t Modellparameter schätzen zu können. Es gibt 3 Fälle (Bollen, 1989):

1. **Genau identifiziertes Modell:** Ist $t = s$, so reichen die empirischen Informationen gerade aus, um alle Parameter schätzen zu können. Genau identifizierte Modelle können die empirischen Zusammenhänge stets perfekt erklären, dadurch sind sie aber inhaltlich auch nicht aussagekräftig.
2. **Unteridentifiziertes Modell:** Ist $t > s$, so müssten im Modell mehr Parameter geschätzt werden, als empirische Informationen vorhanden sind. Dies ist nicht möglich, solche unteridentifizierten (nicht identifizierten) Modelle können also nicht analysiert werden.
3. **Überidentifiziertes Modell:** Ist $s > t$, sind mehr empirische Informationen vorhanden, als für die Schätzung der Parameter im Modell benötigt werden. Nur in diesem Fall sind eine sinnvolle Analyse und das Testen der Übereinstimmung zwischen Modell und Daten anhand des χ^2-Testes überhaupt möglich.

Die Differenz $s - t$ zwischen der Anzahl empirischer Informationen und der Anzahl der Modellparameter bezeichnet man als die **Freiheitsgrade** (df) des Modells. Die Freiheitsgrade müssen also für sinnvolle Analysen in jedem Fall positiv sein. Je höher die Zahl der Freiheitsgrade, desto aussagekräftiger sind Tests auf Übereinstimmung zwischen Modell und Daten: Bei vielen Freiheitsgraden wird nur ein geringer Teil der empirischen Information für die Parameterschätzung verwendet, und es steht noch entsprechend viel Information zur Testung des Modells zur Verfügung.

Die globale Bedingung $s - t \geq 0$ ist für die Identifikation des gesamten Modells notwendig, aber nicht hinreichend. So kann es vorkommen, dass trotz global erfüllter Bedingung einzelne Teile des Modells nicht identifiziert sind, z. B. bei empirisch nur sehr schwachen Zusammenhängen zwischen Variablen: Betrachtet man z. B. das Messmodell eines Konstruktes für sich alleine, so ergeben sich bei $k = 3$ manifesten Indikatorvariablen $k(k+1)/2 = 6$ empirische Varianzen und Kovarianzen, aber auch 6 Modellparameter: 1 Varianz der latenten Variablen, 2 Faktorladungen (da eine zur Skalierung auf den Wert 1 festgesetzt ist) und 3 Fehlervarianzen. Bei 3 Indikatorvariablen ist das Messmodell eines einzelnen Konstruktes also genau identifiziert, bei 2 Indikatorvariablen wäre es für sich alleine unteridentifiziert, und erst bei 4 Indikatorvariablen ist es für sich alleine überidentifiziert, so dass eine sinnvolle separate Analyse dieses Messmodells möglich würde. Das Beispielmodell in Abb. 17.6 weist bei 10 Indikatorvariablen zwar global $55 - 25 = 30$ df auf, ist aber dennoch nur identifiziert, solange die beiden latenten Variablen ξ_2 und η_2 tatsächlich Zusammenhänge zu anderen latenten Variablen haben: Weder das Messmodell von ξ_2, noch das Messmodell von η_2 wären mit jeweils nur 2 Indikatorvariablen für sich alleine identifiziert.

Abhilfe bei Identifikationsproblemen schafft u. U. die Aufnahme zusätzlicher Variablen oder als teilweiser Ausweg die separate Prüfung eines für sich allein identifizierten Modellteils. Bei beschränkter Information bleibt allenfalls die Möglichkeit, das Modell mit weniger Parametern zu reformulieren: Sofern inhaltlich vertretbar, könnten z. B. Ladungen oder Fehlervarianzen zweier manifester Variablen eines Konstruktes gleichgesetzt werden. Bei Konstrukten, die nur über einzelne manifeste Variablen gemessen wurden, können Fehlervarianzen dieser Messungen nicht geschätzt werden und müssen auf null – oder auf einen aufgrund von Vorinformationen bekannten Wert – festgesetzt werden. Damit geht allerdings der Vorteil von Strukturgleichungsmodellen verloren, Messfehlereinflüsse genau abschätzen zu können, so dass die Analyse dann nur eine Pfadanalyse auf der Ebene beobachteter Variablen darstellt.

17.5.3 Schätzprobleme

Die simultane Berücksichtigung vieler Zusammenhänge ist rechenaufwendig und erfolgt durch **iterative Schätzungen**, d. h. durch Austesten von Werten mit einer schrittweisen Annäherung an eine optimale Lösung. Im Gegensatz zu einer Regressionsanalyse, bei der die Regressionsgewichte als Parameter in einem Schritt direkt aus den Daten geschätzt werden, können hierbei Schätzprobleme auftreten:

1. Es kann vorkommen, dass die iterative Schätzung nicht konvergiert, d. h. es wird beim Austesten unterschiedlicher Parameterwerte überhaupt keine optimale Lösung gefunden.
2. Es kann vorkommen, dass die Schätzung zwar konvergiert und eine vermeintlich optimale Lösung liefert, die Parameterschätzungen aber nicht sinnvoll interpretierbar sind, sog. **Heywood Cases** (Heywood, 1931). So können z. B. negative Schätzungen für Varianzen ausgegeben werden, wenn diese die Gleichungssysteme rechnerisch erfüllen, obwohl Varianzen empirisch nicht negativ sein können.

Schätzprobleme treten vergleichsweise häufig bei ML-Schätzungen auf, in kleinen Stichproben, wenn Voraussetzungen der Schätzmethode verletzt sind oder wenn das analysierte Modell auf falschen Annahmen beruht, d. h. fehlspezifiziert ist. Abhilfe kann z. B. eine andere, voraussetzungsärmere Schätzmethode bieten, eine Vereinfachung des Modells, eine schrittweise Analyse von Teilmodellen und/oder eine größere Stichprobe.

17.5.4 Ergebnisinterpretation

Die Beurteilung der Güte eines Strukturgleichungsmodells erfolgt sinnvollerweise nicht nach einem einzigen Signifikanztest oder einem Maß für Varianzerklärung wie in einer Regressionsanalyse, sondern umfasst mehrere Schritte sowie zahlreiche Tests und Koeffizienten (Abb. 17.11):

1. Ist die Schätzung **konvergiert?** Hat das Programm also nicht etwa nur vorläufige oder gar keine Werte ausgegeben, weil eine optimale Lösung nicht gefunden werden konnte?
2. Sind die ausgegebenen Parameterschätzungen **inhaltlich plausibel?** Es sollten insbesondere keine negativen Schätzungen für Varianzen auftreten, da solche Heywood Cases jede weitere Interpretation der Ergebnisse infrage stellen.

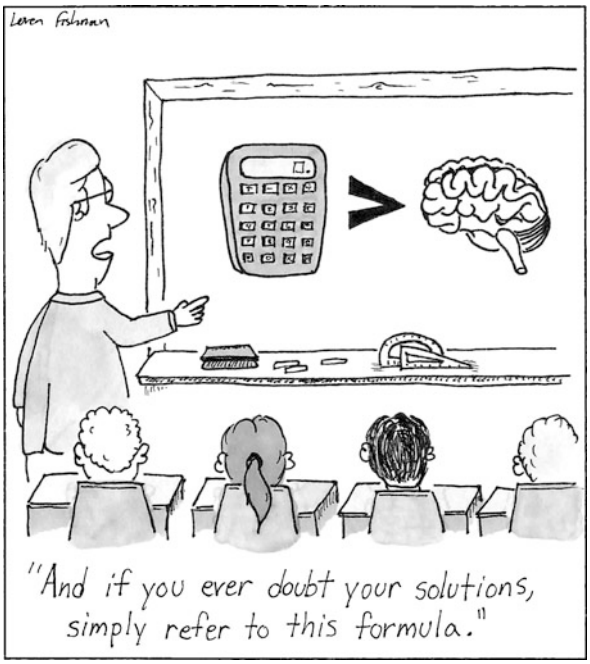


Abb. 17.11 Bei Strukturgleichungsmodellen sind die mit entsprechenden Statistikprogrammen berechneten Parameterschätzungen und verschiedenen Modellgütekriterien stets einer kritischen Interpretation zu unterziehen. (© Loren Fishman/Search ID: lfin95, Rights Available from CartoonStock.com)

3. Ist das Modell mit den Daten zu vereinbaren? Neben der Beurteilung der standardisierten Residuen steht der **χ^2-Test auf Abweichungen zwischen Modell und Daten** zur Verfügung. Wie bei jedem Signifikanztest werden allerdings mit zunehmendem Stichprobenumfang auch minimale, inhaltlich u. U. kaum relevante Abweichungen zwischen Modell und Daten signifikant. In den Sozialwissenschaften erlauben mathematische Modelle so gut wie immer nur Annäherungen an die empirische Realität, so dass man stets mit zumindest kleineren Abweichungen zwischen Modell und Daten rechnen kann und muss (MacCallum & Tucker, 1991). Gerade in großen Stichproben und bei vielen Freiheitsgraden muss daher ein signifikanter χ^2-Test nicht bedeuten, dass das Modell gravierende Fehler enthält, sondern kann auch auftreten, wenn das Modell trotz kleiner Abweichungen die Daten insgesamt recht gut beschreibt.
4. Wie gut ist das Ausmaß der Übereinstimmung zwischen den empirischen Zusammenhängen der Indikatorvariablen und den im Modell reproduzierten Zusammenhängen insgesamt? Unabhängig von der Frage einer signifikanten Abweichung zwischen Modell und Daten können **deskriptive Gütekriterien**

als Maße für den Grad der Übereinstimmung bzw. Abweichungen zwischen Modell und Daten herangezogen werden (▶ Modellgüte; ausführlich z. B. in Schermelleh-Engel et al., 2003). Bei der Beurteilung eines Modells sollten hiervon möglichst unterschiedliche Kriterien berücksichtigt werden, die verschiedene Aspekte der Güte eines Modells erfassen. Durch Eigenschaften von Modell und Daten (z. B. absolute Höhe der empirischen Zusammenhänge) können verschiedene Kriterien auch unterschiedliche Beurteilungen nahelegen.

5. Passt das Modell insgesamt nicht zufriedenstellend zu den Daten, können einerseits die Residuen Hinweise auf konkrete Fehler im Modell geben. Andererseits können Strukturgleichungsmodellprogramme teilweise selbst Hinweise auf mögliche Modellverbesserungen geben: Sog. **Modifikationsindizes** zeigen an, um wie viel sich die Anpassung zwischen empirischer und modelltheoretischer Kovarianzmatrix (im Sinne einer Verringerung des χ^2-Wertes) verbessern würde, wenn ein zusätzlicher Parameter ins Modell aufgenommen wird, z. B. ein zusätzlicher Pfadkoeffizient. Solche explorativen Hinweise können dann daraufhin geprüft werden, ob sie sinnvoll interpretierbar und mit der zugrunde gelegten Theorie verträglich sind. Modifiziert man das Modell tatsächlich, um eine bessere Passung zwischen Modell und Daten zu erhalten, muss aber bei der Ergebnisinterpretation berücksichtigt werden, dass empirische Information in die Modellstruktur eingeflossen ist, wodurch der χ^2-Test und weitere Modellgütekriterien zugunsten des Modells verzerrt werden. Idealerweise sollte das modifizierte Modell daher an einer weiteren Stichprobe kreuzvalidiert werden.
6. Statt ein Modell schrittweise zu modifizieren, bis es passt, ist es in der Praxis oft ratsam, von vornherein verschiedene **konkurrierende, aber potenziell sinnvolle Modelle** aufzustellen (z. B. mit ausschließlich indirekten Effekten vs. indirekten und direkten Effekten oder mit unkorrelierten vs. korrelierten Faktoren) und diese am selben Datensatz zu analysieren. Handelt es sich dabei um sog. **hierarchisch geschachtelte Modelle** („nested models"), bei denen ein Modell durch zusätzliche Parameter aus einem anderen hervorgeht, so können diese direkt mittels Likelihood-Ratio- bzw. χ^2-Differenztests gegeneinander getestet werden, um festzustellen, welches am besten mit den Daten zu vereinbaren ist (vgl. Bollen, 1989; Kline, 2011).

17

17.5.5 Voraussetzungen

Häufig erfolgt die Parameterschätzung in Strukturgleichungsmodellen als ML-Schätzung. Diese basiert, wie erwähnt, auf bestimmten Voraussetzungen:

1. Multivariate Normalverteilung der Indikatorvariablen,
2. Stichprobenumfang angemessener Größe sowie
3. Richtigkeit des angenommenen Modells in der Population.

In der Praxis sind Daten aber äußerst selten multivariat normalverteilt und weichen oft bereits univariat deutlich von Normalverteilung ab (Micceri, 1989). Sozialwissenschaftliche Variablen, z. B. Antwortskalen in Fragebögen, sind außerdem oft relativ grobstufig und weichen damit schon grundsätzlich von einer kontinuierlichen Normalverteilung ab. Schließlich umfassen Stichproben z. B. in der Psychologie oftmals bestenfalls einige hundert Fälle, während die mathematische Ableitung der ML-Schätzung auf asymptotisch großen Stichproben (d. h. gegen unendlich gehendem Umfang) beruht. Dennoch werden Strukturgleichungsmodelle in der Praxis oft auch dann eingesetzt, wenn diese formalen Voraussetzungen nicht erfüllt sind (für Übersichten s. Breckler, 1990; Jackson, Gillaspy & Purc-Stephenson, 2009; MacCallum & Austin, 2000; McDonald & Ho, 2002).

Angesichts der Vielfalt möglicher Modelle und unterschiedlicher empirischer Datensituationen ist es kaum möglich, allgemeingültige Aussagen darüber zu treffen, wann bestimmte Analysen trotz Voraussetzungsverletzungen noch vertretbar sind und wann nicht mehr. Nützliche Anhaltspunkte geben aber methodische Simulationsstudien, in denen gezielt Daten analysiert werden, die Voraussetzungen einer Schätzmethode verletzen (z. B. schief verteilte Daten, Modelle mit kleinen Fehlern), so dass untersucht werden kann, ob Schätzergebnisse unter diesen Umständen noch brauchbar sind oder nicht. Oftmals zeigt sich dabei, dass ML-Parameterschätzungen auch bei Nichtnormalität unverzerrt bleiben, Standardfehlerschätzungen sowie der χ^2-Test aber verzerrt werden (vgl. Curran et al., 1996; Olsson, Foss, Troye & Howell, 2000; West, Finch & Curran, 1995). Abhilfe schaffen hier Korrekturmöglichkeiten für Standardfehlerschätzungen und den Modelltest, die die Nichtnormalität von Daten einbeziehen, wie der oben (▶ Modellgüte) angesprochene **Satorra-Bentler-scaled-χ^2-Test** (Satorra & Bentler, 1994) und entsprechend angepasste, sog. robuste Standardfehlerschätzungen, die in verschiedenen Strukturgleichungsmodellprogrammen implementiert sind.

Modellgüte

Die Modellgüte (oder Modell-Fit) eines Strukturgleichungsmodells gibt an, wie gut ein aufgestelltes Modell den empirisch beobachteten Daten entspricht. Mithilfe verschiedener Modellgütekriterien soll beurteilt werden, ob die Hypothesen, auf denen das Modell beruht, insgesamt beibehalten werden können, d. h. ob das aufgestellte Modell die empirisch beobachteten Zusammenhänge gut erklären kann.
Die Bestimmung dieser Gütekriterien basiert im Wesentlichen auf den Abweichungen zwischen der **empirischen Kovarianzmatrix** S und der **modelltheoretischen Kovarianzmatrix** $\hat{\Sigma}$ (► Abschn. 17.3.1). Bei einer geringen Diskrepanz zwischen diesen beiden Matrizen besteht eine hohe Übereinstimmung zwischen Modell und Daten und damit ein guter Modell-Fit. Die Modellgüte kann inferenzstatistisch mit dem χ^2-Test geprüft werden. Ergänzend stehen verschiedene deskriptive Gütekriterien zur Verfügung, die das Ausmaß der Übereinstimmung oder Abweichung zwischen Modell und Daten quantifizieren und auf bestimmte Dateneigenschaften (z. B. Stichprobengröße) weniger sensibel reagieren als der χ^2-Test.
Im Folgenden wird eine Auswahl wichtiger Gütekriterien genauer dargestellt. Gütekriterien sind am besten für die **ML-Schätzmethode** (► Abschn. 17.3.2) untersucht, daher wird diese im Folgenden angenommen.

χ^2-Test

Der χ^2-Test ist der einzige inferenzstatistische Test zur Beurteilung der Güte von Strukturgleichungsmodellen (Schermelleh-Engel, Moosbrugger & Müller, 2003). Er prüft als Wunschhypothese die Nullhypothese, dass die aus den Parametern des Modells reproduzierte **modelltheoretische Kovarianzmatrix** der **empirischen Kovarianzmatrix der Population** entspricht. Da diese Populationsmatrizen unbekannt sind, werden als Schätzung dieser Matrizen die empirische Kovarianzmatrix S und die modelltheoretische Kovarianzmatrix $\hat{\Sigma}$ verwendet und es wird getestet, ob die Differenz der Matrizen $\boldsymbol{S - \hat{\Sigma}}$ signifikant von null verschieden ist (Bollen, 1989). Ein **nichtsignifikanter χ^2-Wert** (bezogen auf die Anzahl der Freiheitsgrade, ► Abschn. 17.5.2) führt zur Beibehaltung der Nullhypothese. Die Unterschiede zwischen S und $\hat{\Sigma}$ sind dann klein genug, um sie auf zufällige Stichprobenfluktuationen zurückführen zu können. Ein nichtsignifikanter χ^2-Test führt also zu der Aussage, dass das aufgestellte Modell zu den Daten passt.
Bei der Anwendung der ML-Schätzmethode muss die Voraussetzung **multivariat normalverteilter Indikatorvariablen** erfüllt sein. Verwendet man ML-Schätzungen trotz nicht normalverteilter Daten, so resultieren bei steigender Nichtnormalität χ^2-Werte, die zu groß sind (Curran, West & Finch, 1996), so dass zu viele wahre Modelle abgelehnt werden. Alternativ kann bei Nichtnormalität der Daten der Satorra-Bentler-scaled-χ^2-Test (Satorra & Bentler, 1994) verwendet werden, der diesen Einfluss der Nichtnormalität berücksichtigt, oder auf Schätzmethoden ohne Verteilungsvoraussetzungen ausgewichen werden (► Abschn. 17.5.5).
Außerdem hat der χ^2-Test den Nachteil, dass er vom **Stichprobenumfang beeinflusst** wird: Wenn ein Modell nicht perfekt passt, wird mit zunehmenden Stichprobenumfang der χ^2-Wert immer größer. Praktisch bedeutet das, dass in sehr großen Stichproben aufgrund der damit verbundenen hohen Teststärke bereits minimale Unterschiede zwischen S und $\hat{\Sigma}$ statistisch signifikant werden können. Ein signifikanter χ^2-Test zeigt also nur auf, dass Abweichungen zwischen dem Modell und den Daten bestehen und das Modell nicht perfekt zu den Daten passt, man kann jedoch keine Aussagen über die Größe dieser Abweichungen treffen. Daher ist die zusätzliche Betrachtung deskriptiver Gütekriterien sinnvoll, die weniger stark vom Stichprobenumfang beeinflusst werden.

Deskriptive Gütekriterien

Deskriptive Gütekriterien ermöglichen (im Gegensatz zum χ^2-Test) eine graduelle Beurteilung der Abweichung zwischen Modell und Daten. Man kann zwischen deskriptiven Kriterien unterscheiden, die allein das untersuchte Modell betrachten und als **absolute Maße** der Modellanpassung angesehen werden können, und denen, die auf einem Modellvergleich beruhen und **relative Maße** im Vergleich zu einem Baseline-Modell darstellen.

Absolute deskriptive Gütekriterien

Die absoluten deskriptiven Gütekriterien betrachten die Abweichungen zwischen empirischer Kovarianzmatrix S und modelltheoretischer Kovarianzmatrix $\hat{\Sigma}$. Im Folgenden werden zwei absolute Kriterien vorgestellt.
Root Mean Square Error of Approximation: Der **RMSEA** (Browne & Cudeck, 1992; Steiger & Lind, 1980) ist ein Gütemaß zur Bestimmung der approximativen Passung zwischen Modell und Daten bezogen auf die Stichprobengröße und die Anzahl der Freiheitsgrade. Da die Annahme, dass ein Modell in der Population exakt passt, sehr streng ist, und Modelle deshalb oftmals abgelehnt werden müssten, wurde der RMSEA als Maß für die näherungsweise Passung entwickelt. Da die Stichprobengröße im RMSEA berücksichtigt wird, ist der RMSEA von dieser relativ unabhängig (Curran, Bollen, Chen, Paxton & Kirby, 2003; Sugawara & MacCallum, 1993). Der RMSEA sollte für einen guten Modell-Fit möglichst klein sein, wobei 0 für eine perfekte Anpassung steht. Simulationsstudien legen nahe, dass Werte $\leq$ .05 oder $\leq$ .06 für einen guten Fit sprechen (Hu & Bentler, 1998, 1999). Zusätzlich kann ein Konfidenzintervall bestimmt werden, um Aussagen über die Präzision der RMSEA-Schätzung machen zu können.
Standardized Root Mean Square Residual: Der **SRMR** (Bentler, 1995; Jöreskog & Sörbom, 1981) misst die mittleren standardisierten Differenzen zwischen den Elementen der empirischen Kovarianzmatrix S und der modelltheoretischen Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}$ und kann somit als durchschnittliches Residuum zwischen standardisierten Variablen betrachtet werden. Der SRMR ist eines der wenigen deskriptiven Kriterien, das nicht auf dem χ^2-Test beruht. Aber auch der SRMR wird vom Stichprobenumfang beeinflusst: Mit steigendem Stichprobenumfang wird dieser kleiner (Heene, Hilbert, Draxler, Ziegler & Bühner, 2011). Damit das Modell gut zu den Daten passt, sollten die mittleren Differenzen möglichst klein sein, idealerweise bei null liegen. Gemäß Simulationsstudien sollen Werte $\leq$ .05 bis $\leq$ .08 eine gute Passung zwischen Modell und Daten anzeigen (Hu & Bentler, 1995, 1999).

Relative deskriptive Gütekriterien

Relative Gütekriterien prüfen, inwieweit das untersuchte Modell besser zu den Daten passt als ein denkbar schlecht passendes Vergleichsmodell (Bentler & Bonett, 1980). Als Vergleichsmodell kann z. B. das **Unabhängigkeitsmodell** verwendet werden, bei dem davon ausgegangen wird, dass alle beobachteten Variablen unkorreliert sind. Dieses Modell kann daher keinerlei Kovarianzen der Variablen erklären und passt entsprechend schlecht zu den Daten, wenn die beobachteten Variablen tatsächlich zusammenhängen. Neben dem Unabhängigkeitsmodell kann auch das **Nullmodell** als

Vergleichsmodell dienen. Das Nullmodell ist noch restriktiver als das Unabhängigkeitsmodell, da mit dem Nullmodell keine Kovarianzen und auch keine Varianzen der beobachteten Variablen erklärt werden können. Gütemaße, die auf Modellvergleichen beruhen, haben meist Werte zwischen 0 und 1, wobei größere Werte für eine bessere Modellanpassung stehen.

Der **Comparative-Fit-Index** (**CFI**; Bentler, 1990) ist ein deskriptives Gütemaß, das auf einem Vergleich des untersuchten Modells mit dem Unabhängigkeitsmodell beruht. Dabei wird bestimmt, um wie viel das untersuchte Modell die Daten besser erklären kann als das Unabhängigkeitsmodell. Von allen Gütemaßen, die auf einem Modellvergleich beruhen, scheint der CFI am wenigsten von der Stichprobengröße abhängig zu sein (vgl. Fan, Thompson & Wang, 1999; Hu & Bentler, 1999). Simulationsstudien sprechen für einen guten Fit (in Relation zum Unabhängigkeitsmodell), wenn der CFI Werte $\geq .95$ aufweist (Hu & Bentler, 1999).

Der **Goodness-of-Fit-Index** (**GFI**; Tanaka & Huba, 1984) beruht auf einem Vergleich des untersuchten Modells mit dem Nullmodell und gibt den durch das untersuchte Modell erklärten Anteil an der Gesamtvarianz und -kovarianz der Variablen an. Wenn der GFI Werte $\geq .95$ aufweist, wird im Allgemeinen von einem guten Fit ausgegangen (Marsh & Grayson, 1995). Aus dem GFI wurde der **Adjusted-Goodness-of-Fit-Index** (**AGFI**; Jöreskog & Sörbom, 1989) entwickelt, der die Anzahl der Freiheitsgrade in Relation zur Anzahl der beobachteten Variablen berücksichtigt. Dadurch fällt der AGFI für weniger komplexe Modelle mit einer geringeren Anzahl zu schätzender Parameter höher aus, als für komplexere Modelle mit einer größeren Anzahl zu schätzender Parameter. Die Werte des AGFI sind in der Regel kleiner als die des GFI, sollten für eine gute Modellanpassung jedoch nahe am GFI liegen. Der GFI und der AGFI sind stärker von der Stichprobengröße beeinflusst als der CFI: je größer die Stichprobe, desto höher die Werte des (A)GFI (vgl. Hammervold & Olsson, 2012; Hu & Bentler, 1998).

Insgesamt sollten aufgrund der Sensibilität der verschiedenen Kriterien gegenüber verschiedenen Dateneigenheiten (Stichprobengröße, Modellkomplexität etc.) immer mehrere Gütekriterien gleichzeitig betrachtet und zur Beurteilung eines Strukturgleichungsmodells herangezogen werden. Es ist zudem zu berücksichtigen, dass die hier vorgestellten Cut-Off-Werte nur eine grobe Orientierung bieten können, da diese Werte auch von dem untersuchten Modell selbst abhängig sind und sie je nach untersuchtem Modell ihre Aussagekraft verlieren können (Fan & Sivo, 2007; Nye & Drasgow, 2011). In Schermelleh-Engel et al. (2003) finden sich weitere Gütekriterien im Rahmen der Strukturgleichungsmodellierung sowie detailliertere Informationen und Formeln der verschiedenen Kriterien.

Alternativ kann man zur Parameterschätzung auf andere Methoden ausweichen, z. B. auf die erwähnte ULS-Schätzung, die keine Verteilungsannahmen benötigt. Dies kann jedoch andere Nachteile mitbringen, z. B. eine geringere Effizienz gegenüber ML-Schätzung, so dass für gleiche Genauigkeit der Parameterschätzungen (gleiche Standardfehler) größere Stichproben benötigt würden. Außerdem werden bei ULS-Schätzung unterschiedliche Maßstäbe, d. h. unterschiedliche Varianzen der beobachteten Variablen nicht berücksichtigt, was bei empirisch stark unterschiedlichen Varianzen ein Nachteil ist. Formal ohne Verteilungsvoraussetzungen auskommende Weighted-Least-Squares-Schätzungen, die dies berücksichtigen würden, benötigen dagegen sehr große Stichproben (Olsson et al., 2000).

Für Modelle mit grobstufigen Variablen bietet es sich einerseits an, polychorische statt normaler Produkt-Moment-Zusammenhänge zu berechnen, sofern man davon ausgeht, dass hinter den grobstufigen Variablen kontinuierliche, normalverteilte Variablen stehen. Alternativ gibt es Schätzmethoden, die Schwellenwerte für die einzelnen Stufen der Variable mitbestimmen, wie die WLSMV-Schätzmethode („weighted least squares, mean- and variance-adjusted“) der Programme Mplus (Muthén & Muthén, 2012) und lavaan (Rosseel, 2015), die für grobstufige Daten auch in relativ kleinen Stichproben vorteilhaft sein kann (vgl. Beauducel & Herzberg, 2006; Rhemtulla, Brosseau-Liard & Savalei, 2012).

Schließlich kann man beim Einsatz von Strukturgleichungsmodellen in der Praxis nicht im Vorhinein wissen, ob das angenommene Modell, das ja erst geprüft werden soll, richtig ist. Gravierend falsche Annahmen führen zwar meist auch schon in kleinen Stichproben zu auffällig schlechter Modellanpassung, doch ist es grundsätzlich nicht völlig auszuschließen, dass auch schon kleine, inhaltlich unbedeutende und unbemerkte Fehler im Modell zu Verzerrungen von Schätzergebnissen führen. Darüber hinaus können besonders Kombinationen aus kleinen Stichproben, niedrigen Zusammenhängen der Variablen und/oder nicht ganz korrekten Modellen bei ML-Schätzungen leicht zu Problemen und wenig verlässlichen Ergebnissen führen (Briggs & MacCallum, 2003; MacCallum, Tucker & Briggs, 2001; Ximénez, 2006, 2007)

17.5.6 Kausalaussagen

Strukturgleichungsmodelle werden oft mit dem Ziel verwendet, mittels statistischer Methoden Kausalaussagen zu prüfen.

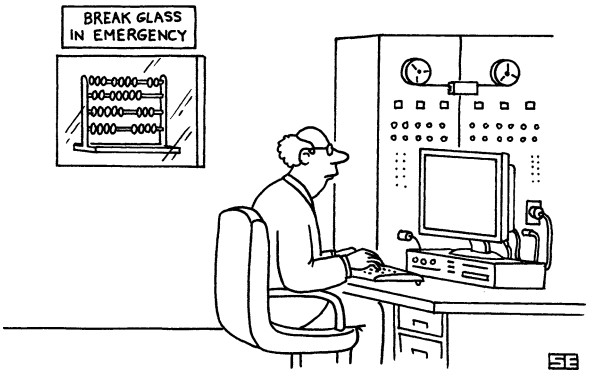
Durch das Aufstellen eines Modells mit gerichteten Beziehungen zwischen Variablen werden Annahmen über Einflussrichtungen nicht „verifiziert“, selbst dann nicht, wenn das Modell mit den Daten vereinbar ist.

Es kann sogar vorkommen, dass zwei oder mehr alternative Modelle, die inhaltlich völlig unterschiedliche Annahmen machen, statistisch trotzdem gleich gut mit den Daten zu vereinbaren sind: Aus sich inhaltlich widersprechenden Hypothesen kann sich u. U. das gleiche statistische Zusammenhangsmuster ergeben, so dass eine Entscheidung zwischen unterschiedlichen, sich inhaltlich widersprechenden Kausalmodellen anhand statistischer Kriterien allein nicht immer möglich ist (MacCallum, Wegener, Uchino & Fabrigar, 1993; Stelzl, 1986). Die manchmal für Strukturgleichungsmodelle verwendete Bezeichnung „Kausalanalyse“ kann insofern missverständlich sein.

Strukturgleichungsmodelle prüfen Kausalhypothesen grundsätzlich in gleicher Weise wie andere statistische Verfahren, indem der bei Gültigkeit einer Kausalhypothese zu erwartende statistische Zusammenhang zwischen Variablen untersucht wird. Die Ergebnisse der Parameterschätzungen in Strukturgleichungsmodellen und das Ausmaß der Abweichungen zwischen Modell und Daten können damit aber zur Prüfung im Sinne einer möglichen Widerlegung von Kausalannahmen herangezogen werden. Sie ermöglichen so in Bereichen, in denen experimentelle Forschung u. U. schwer möglich ist, zumindest konzeptuell klarere Aussagen als von vornherein ungerichtete korrelative Studien oder Untersuchungen, die nur einzelne beobachtete Variablen einbeziehen. Eine strenge Prüfung von Annahmen über Kausalrichtungen erfordert aber – von logisch unbeeinflussbaren Variablen abgesehen – zusätzlich eine zeitliche Perspektive der Datenerhebung im Rahmen von Längsschnittuntersuchungen (▶ Abschn. 12.5.4) oder durch eine zeitlich vorangegangene experimentelle Manipulation sowie die Kontrolle von Störvariablen, die mit statistischen Mitteln allein nicht immer möglich ist (zu experimentellen Untersuchungen ▶ Abschn. 12.5.3).

17.6 Software für Strukturgleichungsmodelle

Seit den 1970er-Jahren wurden verschiedene Programme für die Analyse von Strukturgleichungsmodellen entwickelt. Sie unterscheiden sich in der Art der Bedienung, d. h. darin, ob Modelle grafisch, syntaxbasiert über einzelne Pfade oder syntaxbasiert in Matrixform spezifiziert werden sowie in der Verfügbarkeit (frei/kommerziell) und in fortgeschrittenen Analysefunktionen. Bei einfachen Analysen spielt es in der Regel keine methodische Rolle, welches Programm verwendet wird, so dass man benutzen kann, was man kennt und was verfügbar ist. Ungewöhnlichere Analysen sind aber teilweise nur mit bestimmten Programmen möglich (◘ Abb. 17.12).



◘ **Abb. 17.12** Eine Berechnung von Strukturgleichungsmodellen ist in der Praxis nur mithilfe entsprechender Spezialsoftware möglich. (© Stan Eales/Search ID: sea0478, Rights Available from CartoonStock.com)

Relativ bekannt sind die folgenden Programme (alphabetisch ohne Anspruch auf Vollständigkeit; Stand der Angaben August 2015):

AMOS (Analysis of Moment Structures) erlaubt, Strukturgleichungsmodelle grafisch durch das Zeichnen von Pfaddiagrammen aufzustellen. Zu Anfang kann dies den Lernaufwand für die ersten Analysen vermindern: Man kann mit der Maus relativ schnell ein Modell zeichnen. Bei komplexen Modellen kann das grafische Vorgehen aber umständlich und arbeitsaufwendig werden. AMOS wird über SPSS lizenziert.

EQS ist ein ursprünglich syntaxbasiertes Programm, bei dem Modelle anhand der angenommenen Beziehungen zwischen einzelnen Variablen spezifiziert werden. EQS ist für Windows, Mac OS X und verschiedene Unix-Systeme erhältlich, wobei die Windows-Version inzwischen auch eine grafische Modellspezifikation erlaubt. Eine Demoversion wird angeboten, ist derzeit aber nicht direkt herunterladbar. Seite des Anbieters Multivariate Software:

- http://www.mvsoft.com/

LISREL (Linear Structural Relationships) ist eines der am längsten existierenden Programme für Strukturgleichungsmodelle. Die Modellspezifikation erfolgt syntaxbasiert in Matrixform, d. h. man legt nicht einzelne Beziehungen zwischen Variablen fest, sondern komplette Matrizen, die z. B. jeweils alle gleichartigen Pfadkoeffizienten oder alle Fehlervarianzen enthalten. Dies erfordert zu Anfang mehr Lernaufwand, kann aber gerade bei komplexen Modellen mit vielen Variablen die Arbeit vereinfachen. Alternativ können in der Syntaxvariante SIMPLIS Modelle anhand der angenommenen Beziehungen zwischen einzelnen Variablen festgelegt werden. In der Windows-Version von LISREL können Ergebnisse

anschaulich als Pfaddiagramm präsentiert werden, außerdem ist auch eine grafische Aufstellung des Modells möglich.

Für Windows gibt es eine frei verfügbare „student version“, die lediglich auf Modelle mit maximal 16 beobachteten Variablen beschränkt ist, was zum Lernen und für viele Analysen ausreicht. Diese ist direkt auf der Seite des Anbieters Scientific Software International herunterladbar:

- http://www.ssicentral.com/

Mplus ist ein syntaxbasiertes Programm, bei dem Modelle anhand der angenommenen Beziehungen zwischen einzelnen Variablen aufgestellt werden. Mplus bietet verschiedene Analysefunktionen, die in anderen Programmen nicht vorhanden sind, u. a. für nichtlineare Strukturgleichungsmodelle (Interaktions-/Moderatoreffekte zwischen quantitativen Variablen, kurvilineare Effekte).

Für Windows, Mac und Linux gibt es eine frei verfügbare Demoversion, mit der Modelle mit maximal zwei unabhängigen und sechs abhängigen Variablen analysiert werden können. Diese ist direkt von der Seite der Anbieter Muthén & Muthén herunterladbar:

- http://www.statmodel.com/

R Die Statistikprogrammumgebung R (http://www.r-project.org/) ist ein frei („open source“) verfügbares System mit sehr vielfältigen Möglichkeiten, das v. a. von Wissenschaftlerinnen und Wissenschaftlern aus der Statistik und ihren Anwendungsfeldern nichtkommerziell entwickelt wird. In vielen Disziplinen wird R zunehmend zum Standard der Datenanalyse. R ist für Windows, Mac und Linux verfügbar und grundsätzlich syntaxbasiert, wobei es für einfachere Funktionen auch grafische Bedienungsoberflächen gibt (allerdings aktuell nicht für Strukturgleichungsmodelle). Erhältlich ist R über CRAN, das Comprehensive R Archive Network:

- http://cran.r-project.org/

Für R gibt es mehrere Zusatzpakete („packages“), mit denen Strukturgleichungsmodelle analysiert werden können:

R: lavaan package Das lavaan-Zusatzpaket ist ein aktuell aktiv weiterentwickeltes Strukturgleichungsmodellpaket für R. Es benutzt eine einfache, leicht verständliche Syntax zur Definition des Modells über die einzelnen Pfade, die sich an übliche R-Konventionen anlehnt und sinnvolle Voreinstellungen (ähnlich dem kommerziellen Programm Mplus) verwendet. Auch die einführende Dokumentation ist sehr gut verständlich. Das lavaan-Paket gilt zwar noch als Beta-Software, beherrscht aber bereits komplexere Modelle (z. B. benutzerfreundlich gestaltete Gruppenvergleiche) sowie unterschiedliche Schätzmethoden (Maximum Likelihood, Generalized Least Squares, Weighted Least Squares). Das Paket ist auf CRAN erhältlich, einführende Dokumentation auf der Seite des Autors Yves Rosseel (Universität Gent, Belgien):

- http://lavaan.org/

R: OpenMx package Das OpenMx-Paket für R ist eine Neu-Implementierung des schon länger existierenden Mx-Programms für Strukturgleichungsmodelle. In OpenMx können Modelle syntaxbasiert entweder anhand einzelner angenommener Beziehungen zwischen Variablen oder anhand kompletter Parametermatrizen aufgestellt werden. OpenMx strebt umfangreiche Funktionen für Strukturgleichungsmodelle an, verwendet allerdings eine recht komplexe und nicht so eng an anderen R-Konventionen orientierte Syntax für die Definition von Modellen und die Benennung von Objekten. Die Dokumentation ist umfangreich, allerdings auch komplex. Das Paket ist derzeit direkt von der Seite des Projekts erhältlich:

- http://openmx.psyc.virginia.edu/

R: sem package Im schon länger erhältlichen sem-Zusatzpaket für R werden Strukturgleichungsmodelle anhand der angenommenen Beziehungen zwischen einzelnen Variablen spezifiziert. Das sem-Paket erlaubt eine Parameterschätzung mittels Maximum Likelihood oder Two-Stage Least Squares. Erhältlich ist es über CRAN, weitere Informationen finden sich auf der Seite des Autors John Fox (McMaster-Universität, Hamilton, Ontario, Kanada):

- http://socserv.socsci.mcmaster.ca/jfox/Misc/sem/

Lernquiz

1. Welche Vorteile bietet die Analyse von Strukturgleichungsmodellen in den Human- und Sozialwissenschaften gegenüber Methoden auf der Ebene von beobachteten Variablen, wie z. B. Regressionsanalysen?
2. Was sind manifeste und latente Variablen?
3. Was ist eine Mediatorvariable?
4. Welche Teilmodelle kann man in Strukturgleichungsmodellen unterscheiden? Welche Beziehungen werden darin festgelegt?
5. Was ist ein Skalierer? Wofür wird er benötigt?
6. Was ist die modelltheoretische Kovarianzmatrix? Welche Einträge stehen darin?
7. Welche Hypothese prüft der χ^2-Test im Zusammenhang mit Strukturgleichungsmodellen? Soll der Test für ein Modell, das zu den Daten passt, signifikant sein oder nicht?
8. Angenommen, Sie möchten die vermuteten Zusammenhänge zwischen 4 Konstrukten mit einer konfirmatorischen Faktorenanalyse untersuchen.

Zu jedem Konstrukt haben Sie Daten von zwei beobachteten Variablen erhoben. Ist dies ausreichend, damit das Modell identifiziert ist?

9. Sie prüfen einen angenommenen Effekt von Arbeitsanforderungen auf Burnout mit einem Strukturgleichungsmodell. Die Analyse der Daten von $N = 185$ Beschäftigten ergibt einen standardisierten Pfadkoeffizienten der Arbeitsanforderungen auf Burnout von .38, $p = .042$. Als Modellgütekriterien erhalten Sie $\chi^2 = 55.40$, df $= 30$, $p = .003$ sowie RMSEA $= .18$ und CFI $= .83$. Wie interpretieren Sie das Ergebnis? Besteht der angenommene Effekt?
10. Sie untersuchen einen angenommenen Effekt von Burnout bei Mitarbeitern auf deren Kündigungsneigung mithilfe eines Strukturgleichungsmodells. Sie erhalten Daten von $N = 112$ Beschäftigten. Die Analyse ergibt einen signifikanten Effekt von Burnout auf die Kündigungsneigung (standardisierter Pfadkoeffizient .32, $p = .011$). Das Modell passt offenbar gut zu den Daten ($\chi^2 = 9.32$, df $= 8$, $p = .316$). Für das Item „Ich suche gezielt nach Stellenangeboten, die für mich infrage kommen“ erhalten Sie als standardisierte Parameterschätzungen eine Faktorladung von 1.05 und eine Fehlervarianz von -.10. Wie interpretieren Sie dieses Ergebnis? Besteht der vermutete Effekt?

Übungsaufgaben

1. Installieren Sie ein Programm Ihrer Wahl zur Analyse von Strukturgleichungsmodellen. Falls an Ihrer Hochschule keines verfügbar ist, können Sie zum Beispiel R mit einem der beschriebenen Zusatzpakete verwenden oder die LISREL student version.
2. Setzen Sie das im Anwendungsbeispiel beschriebene Modell mit diesem Programm um. Schreiben Sie die entsprechenden Anweisungen für das Programm bzw. legen Sie das Modell grafisch über ein Pfaddiagramm fest. Die Rohdaten erhalten Sie im ▶ Webangebot zu diesem Buch.
3. Gehen Sie die Ergebnisse schrittweise durch und versuchen Sie, die Bedeutung der Ausgaben zu verstehen. Die Darstellung der Ergebnisse von Strukturgleichungsmodellanalysen ist sehr programmabhängig, Informationen zur Darstellung finden Sie in der jeweiligen Programmdokumentation.

Weiterführende Literatur

Die englischsprachige methodische Literatur zu Strukturgleichungsmodellen ist sehr umfangreich. Neben der spezialisierten Zeitschrift „Structural Equation Modeling“ finden sich viele Artikel in methodischen Zeitschriften der Fächer, die Strukturgleichungsmodelle anwenden, z. B. in „Psychological Methods“ und „Sociological Methods and Research“.

Eine gut lesbare und aktuelle Einführung im Umfang eines ganzen Buches bietet z. B.

- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Die erste Auflage des Buches von 1998 ist in manchen Punkten noch ausführlicher.

Ein sehr detailliertes Standardwerk, das sich mit ausführlichen Herleitungen gut zum Nachschlagen und Vertiefen eignet, ist

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Eine deutschsprachige Einführung gibt das Buch von

- Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften*. München: Oldenbourg.

Unterschiedliche Modelltypen werden verständlich und an praktischen Beispielen gut nachvollziehbar erläutert.

Übersichten zu sozialwissenschaftlichen Anwendungen von Strukturgleichungsmodellen und Empfehlungen für die Darstellung von Analyseergebnissen in Forschungsarbeiten bieten:

- Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling, 7*, 461–483.
- Breckler, S. J. (1990). Application of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107*, 260–273.
- Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*, 6–23.
- MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

- McDonald, R. P. & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64–82.
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*, 323–337.

Literatur

Bakker, A. B., Demerouti, E., & Verbeke, W. (2004). Using the job demands-resources model to predict burnout and performance. *Human Resource Management, 43*, 83–104.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Beauducel, A. & Herzberg, P. Y. (2006). On the performance of maximum likelihood vs. means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186–203.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in covariance structures. *Psychological Bulletin, 88*, 588–606.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

Breckler, S. J. (1990). Application of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107*, 260–273.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38*, 25–56.

Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230–258.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. (2003). The finite sampling properties of the RMSEA: Point estimates and confidence intervals. *Sociological Methods and Research, 32*, 208–252.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16–29.

Duncan, O. D. (1966). Path analysis. Sociological examples. *American Journal of Sociology, 72*, 1–16.

Fan, X. & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*, 509–529.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56–83.

Gelfand, L. A., Mensinger, J. L., & Tenhave, T. (2009). Mediation analysis: A retrospective snapshot of practice and more recent directions. *The Journal of General Psychology, 136*, 153–178.

Goldberger, A. S. & Duncan, O. D. (1973). *Structural equation models in the social sciences.* New York, NY: Seminar Press.

Hammervold, R. & Olsson, U. H. (2012). Testing structural equation models: The impact of error variances in the data generating process. *Quality and Quantity*, *46,* 1547–1570.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*, 319–336.

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling, 10*, 35–46.

Hewitt, P. L. & Flett, G. L. (1991). Perfectionism in the self and social contexts: Conceptualisation, assessment, and association with psychopathology. *Journal of Personality and Social Psychology, 60*, 456–470.

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London, Series A, 134*, 486–501.

Hu, L.-T. & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 76–99). London, UK: Sage.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria vs. new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*, 6–23.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*, 239–251.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*, (pp. 85–112). New York, NY: Seminar Press.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Uppsala, Sweden: University of Uppsala, Department of Statistics.

Jöreskog, K. G. & Sörbom, D. (1989). *LISREL 7: User's reference guide.* Chicago, IL: SPSS.

Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* New York, NY: Guilford.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

MacCallum, R. C. & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502–511.

MacCallum, R. C., Tucker, L. R., & Briggs, N. E. (2001). An alternative perspective on parameter estimation in factor analysis and related methods. In R. Cudeck, S. H. C. du Toit & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 39–57). Lincolnwood, IL: Scientific Software International.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*, 185–199.

Marsh, H. W. & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 177–198). Thousand Oaks, CA: Sage.

Maruyama, G. M. (1998). *Basics of structural equation modeling.* Thousand Oaks, CA: Sage.

Maslach, C., Jackson, S. E., & Leiter, M. P. (1997). Maslach burnout inventory. In C. P. Zalaquett & R. J. Wood (Eds.), *Evaluating stress: A book of resources* (pp. 191–218). Lanham, MD: Scarecrow.

McDonald, R. P. & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64–82.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.

Moosbrugger, H. (2011). Klassische Testtheorie. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 103–117). Berlin: Springer.

Muthén, L. K. & Muthén, B. O. (2012). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

Nye, C. D. & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14*, 548–570.

Olsson, U. H., Foss, T., Troye, S. V. & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*, 557–595.

Peter, R. (2002). Berufliche Gratifikationskrisen und Gesundheit. *Psychotherapeut, 47*, 386–398.

Philp, M., Egan, S., & Kane, R. (2012). Perfectionism, over commitment to work, and burnout in employees seeking workplace counseling. *Australian Journal of Psychology, 64,* 68–74.

R Development Core Team (2015). *R: A language and environment for statistical computing (version 3.2.1)* [Computer software and manual]. Wien: R Foundation for Statistical Computing. Retrieved August 19, 2015, from http://r-project.org.

Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften.* München: Oldenbourg.

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2012). How many categories is enough to treat data as continuous? A comparison of robust continuous and categorical SEM estimation methods under a range of non-ideal situations. *Psychological Methods, 17*, 354-373.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48(2),* 1–36. Retrieved November 13, 2013, from http://www.jstatsoft.org/v48/i02.

Rosseel, Y. (2015). *R package lavaan: Latent variable analysis (version 0.5-18)* [Computer software and manual]. Ghent University, Ghent, Belgium. Retrieved August 19, 2015, from http://cran.r-project.org/web/packages/lavaan/.

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23–74. Retrieved November 13, 2013, from http://www.dgps.de/fachgruppen/methoden/mpr-online/.

Siegrist, J., Starke, D., Chandola, T., Godin, I., Marmot, M., Niedhammer, I., & Peter, R. (2004). The measurement of effort-reward imbalance at work: European comparisons. *Social Science and Medicine, 58*, 1483–1499.

Steiger, J. H. & Lind, J. C. (1980). *Statistically-based tests for the number of factors.* Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City, IA. Retrieved November 13, 2013, from http://www.statpower.net/SteigerBiblio/Steiger-Lind1980.pdf.

Stelzl, I. (1986). Changing a causal hypothesis whithout changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research, 21*, 309–331.

Sugawara, H. M. & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement, 17*, 365–377.

Tanaka, J. S. & Huba, G. J. (1984). Confirmatory hierarchical factor analyses of psychological distress measures. *Journal of Personality and Social Psychology, 46*, 621–635.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation modeling with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 37–55). Thousand Oakes, CA: Sage.

Wright, S. (1918). On the nature of size factors. *Genetics, 3*, 367–374.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161–215.

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling, 13*, 587–614.

Ximénez, C. (2007). Effect of variable and subject sampling on recovery of weak factors in CFA. *Methodology, 3*, 67–80.

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*, 197–206.

Evaluationsforschung

Nicola Döring, Jürgen Bortz

18.1 Evaluationsforschung im Überblick – 977

18.2 Verschiedene Ansätze in der Evaluationsforschung – 995

18.3 Phasen der Evaluationsforschung – 1017

Literatur – 1033

N. Döring, J. Bortz (unter Mitarbeit von S. Pöschl),
Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften, Springer-Lehrbuch,
DOI 10.1007/978-3-642-41089-5_18, © Springer-Verlag Berlin Heidelberg 2016

Lernziele

- Die Evaluationsforschung von der Grundlagen- und Interventionsforschung abgrenzen können.
- Zentrale Konzepte der Evaluationsforschung definieren und erläutern können.
- Wissen, was die Evaluationsstandards beinhalten.
- Wichtige Evaluationsansätze anhand bekannter Evaluationsmodelle charakterisieren können.
- Wissen, was in den einzelnen Phasen einer Evaluationsstudie zu tun ist.

Bewertungen geben wir im Alltag ständig ab und tauschen sie mit anderen Menschen aus: Wir beurteilen Restaurants und Hotels, bewerten die Arbeit einer politischen Partei oder des kommunalen Krankenhauses, rezensieren Bücher und Filme oder diskutieren über den Nutzen eines neuen Gesetzes. Im Unterschied zu derartigen, mehr oder minder willkürlichen Einschätzungen durch Laien strebt die Evaluationsforschung eine wissenschaftlich fundierte Bewertung von Produkten, Maßnahmen oder anderen Gegenständen an. Die Evaluationsforschung („evaluation research") ist ein anwendungsbezogener und oft auftraggeberfinanzierter Teilbereich der empirischen Sozialforschung.

Welche erwünschten Wirkungen und unerwünschten Effekte konkrete – geplante oder bereits realisierte – Maßnahmen der Kriminalitätsprävention, der Verkehrssicherheit, des Daten- und Umweltschutzes, der Wirtschafts- oder Wissenschaftsförderung etc. nach sich ziehen, wird meist politisch kontrovers diskutiert. Von einer wissenschaftlichen Evaluationsstudie verspricht man sich eine unabhängige bzw. überparteiliche Bewertung, mit der dann Entscheidungen Pro bzw. Kontra die entsprechende Maßnahme überzeugend begründet werden können. Auch erwartet man sich von Evaluationsstudien Hinweise für die Verbesserung von Maßnahmen und Produkten.

Die moderne Evaluationsforschung entwickelte sich etwa ab den 1930er-Jahren in den USA (z. B. Bateman, 2000). Ihr oblag u. a. die Bewertung von Programmen, Interventionen und Maßnahmen im Bildungs- und Gesundheitswesen sowie die Entwicklung formaler Regeln und Kriterien für die wissenschaftliche Erfolgs- und Wirkungskontrolle. Eine größere Verbreitung erlangte die Evaluationsforschung im Zuge sozialreformerische Bestrebungen in den 1960er- und 1970er-Jahren. Im deutschsprachigen Raum konnte die Evaluation v. a. in der Bildungs- und Psychotherapieforschung, der Arbeitspsychologie sowie in vielen Feldern der Politikforschung erste Erfolge verzeichnen. Inzwischen ist die Bedeutung der wissenschaftlichen Evaluation stark gewachsen. Besonders, weil man in ihr eine Legitimationsgrundlage für Maßnahmen sieht, die mit öffentlichen Geldern finanziert werden (sollen) bzw. die die Allgemeinheit betreffen (für einen Geschichtsabriss der Evaluationsforschung Gollwitzer & Jäger, 2009; Widmer & De Rocchi, 2012).

In diesem Kapitel werden zunächst die wichtigsten Charakteristika der Evaluationsforschung – v. a. auch im Vergleich zur empirischen Grundlagenforschung – herausgearbeitet und zentrale Konzepte eingeführt (▶ Abschn. 18.1). Anschließend werden verschiedene Ansätze und Modelle der Evaluationsforschung voneinander abgegrenzt (▶ Abschn. 18.2). Abschließend wenden wir uns dem Ablauf einer typischen Evaluationsstudie zu und behandeln die in den einzelnen Phasen anfallenden Arbeitsschritte von der Akquise eines Evaluationsprojekts bis zum Abschlussbericht an den Auftraggeber (▶ Abschn. 18.3).

Wir stellen die Evaluationsforschung im letzten Kapitel des vorliegenden Lehrbuches vor, da zur Durchführung eigener Evaluationsstudien solide Grundkenntnisse in den Methoden der empirischen Sozialforschung notwendig sind. Darüber hinaus stellt eine Evaluationsstudie zusätzliche Anforderungen an die Forschenden: Sie müssen angesichts des engen Praxisbezugs ihrer Arbeit viel Zeit und Energie für die Moderation zwischen unterschiedlichen Interessensgruppen aufwenden, ihr Vorgehen unter ethischen Gesichtspunkten besonders kritisch durchdenken, die methodische Umsetzung der Studie den praktischen Bedingungen und Beschränkungen im Untersuchungsfeld immer wieder flexibel anpassen und eine angemessene Position im Spannungsfeld zwischen ihrer Rolle als **Dienstleistende** für den Auftraggeber einerseits und dem Anspruch **wissenschaftlicher Unabhängigkeit** andererseits finden. Im Kontext der Evaluationsforschung meint wissenschaftliche Unabhängigkeit vor allem, dass Forschende sich nicht von Auftraggebenden oder anderen Anspruchsgruppen instrumentalisieren lassen, um bestimmte Wunschbewertungen zu produzieren oder unerwünschte Befunde zu unterdrücken, sondern dass sie die verschiedenen Sichtweisen der Anspruchsgruppen aufnehmen und in einem transparenten Prozess die vorliegenden und noch zu erhebenden empirischen Informationen über die Wirklichkeit in **Bewertungen** transformieren. Dazu gehört auch eine Reflexion des jeweiligen Vorgehens hinsichtlich seiner ethischen, methodischen und theoretischen Voraussetzungen und Grenzen, die sich aus dem gewählten Evaluationsansatz ergeben.

Die sozialwissenschaftliche Evaluation ist zu unterscheiden von verwandten Konzepten, die ebenfalls dazu dienen, datenbasiert Maßnahmen und Organisationen zu bewerten und zu verbessern, etwa das betriebswirtschaftliche **Controlling** oder das betriebliche **Qualitätsmanagement**. Hierbei handelt es sich um Aktivitäten, die kontinuierlich durchgeführt werden (Evaluationsstudien kommen dagegen punktuell zum Einsatz) und stets im Auftrag des Managements erfolgen (in der Evaluati-

onsforschung gibt es auch andere Auftraggebende). Ein weiterer Hauptunterschied betrifft die Zielsetzung, die in der Evaluationsforschung über betrieblichen und betriebswirtschaftlichen Nutzen hinausgeht, teils ganz andere Evaluationsgegenstände als Unternehmen in den Blick nimmt und nicht zuletzt die Wissenschaftlichkeit von Verfahren und Ergebnissen betont.

18.1 Evaluationsforschung im Überblick

Um die Besonderheiten der Evaluationsforschung herauszuarbeiten beginnen wir mit der Abgrenzung gegenüber der Grundlagenforschung (▶ Abschn. 18.1.1) und gehen zur Klärung wichtiger Evaluationsbegriffe über (▶ Abschn. 18.1.2 bis ▶ Abschn. 18.1.7). Wir stellen verschiedene Typen von Evaluationsstudien vor (▶ Abschn. 18.1.8) und gehen darauf ein, welche besonderen Standards bei der Evaluationsforschung bei guten Studien einzuhalten sind (▶ Abschn. 18.1.9). Anschließend umreißen wir, welche spezifischen Anforderungen an die Qualifikation von Evaluierenden gestellt werden (▶ Abschn. 18.1.10). Verweise auf Fachinformationen zur Evaluationsforschung runden den Abschnitt ab (▶ Abschn. 18.1.11).

18.1.1 Abgrenzung von Grundlagen-, Interventions- und Evaluationsforschung

Grundlagen-, Interventions- und Evaluationsforschung wenden prinzipiell den gesamten Kanon sozialwissenschaftlicher Forschungsmethoden an und sind alle denselben Prinzipien der Wissenschaftlichkeit verpflichtet. Sie unterscheiden sich jedoch in punkto Theoriebildung sowie in ihren operativen Rahmenbedingungen erheblich.

Während die **Grundlagenforschung** Theorien entwickelt und überprüft, die dazu dienen, Sachverhalte zu beschreiben, zu erklären und vorherzusagen (z. B. grundlagenwissenschaftliche Theorien zu Formen und Ursachen der Depression), widmet sich die **Interventionsforschung** der Entwicklung und Überprüfung von Theorien, die auf Techniken der zielgerichteten Veränderung von Sachverhalten abzielen (z. B. technologische Theorien zur Depressionstherapie; ▶ Grundlagenwissenschaftliche und technologische Theorien). Die **Evaluationsforschung** schließlich befasst sich mit der wissenschaftlich fundierten Bewertung von Sachverhalten und insbesondere von Interventionsmaßnahmen hinsichtlich verschiedener Bewertungskriterien, etwa Effektivität, Effizienz, Akzeptanz oder Nachhaltigkeit (z. B. Beurteilung der Effizienz von Depressionstherapie A versus Depressionstherapie B). Ergebnisse der Evaluationsforschung dienen in erster Linie dazu, die evaluierte Intervention zu verbessern oder Entscheidungen über die Nutzung oder Nichtnutzung bzw. Weiterführung einer Intervention zu treffen. Im Zuge der Evaluationsforschung können aber auch Theoriemodelle zur Wirkungsweisen der Maßnahme entwickelt und geprüft werden, sog. Programmtheorien.

Grundlagenforschung – Die Grundlagenforschung widmet sich der Entwicklung und Überprüfung wissenschaftlicher Theorien.

Grundlagenwissenschaftliche Theorien – Grundlagenwissenschaftliche Theorien dienen der Beschreibung, Erklärung und Vorhersage von Sachverhalten jeglicher Art.

Interventionsforschung bzw. anwendungsorientierte Forschung – Die Interventionsforschung bzw. anwendungsorientierte Forschung widmet sich der Entwicklung und Überprüfung technologischer Theorien.

Technologische Theorien – Technologische Theorien dienen der Beschreibung, Erklärung und Vorhersage von Interventionen und geben dadurch Handlungsanweisungen.

Evaluationsforschung – Die Evaluationsforschung widmet sich der Bewertung von Maßnahmen, Programmen (Maßnahmenbündeln), aber auch von anderen Evaluationsgegenständen. Dabei wird auf technologische oder auch grundlagenwissenschaftliche Theorien zurückgegriffen. Evaluationsforschung operiert meist stärker theorieanwendend als theorieentwickelnd.

Programmtheorien – Programmtheorien erklären detailliert die Wirkungsweise einer evaluierten Maßnahme bzw. eines evaluierten Programms. Sie liegen meist nicht im Vorfeld vor, sondern werden im Zuge der Evaluationsforschung entwickelt und geprüft. Bei einer Programmtheorie handelt es sich also um eine Unterform der technologischen Theorien.

Grundlagen- und Evaluationsforschung unterscheiden sich nicht nur in ihren Zielsetzungen bei der Theoriebildung, sondern auch in ihren operativen **Rahmenbedingungen**. So ist Evaluationsforschung als anwendungsorientierte Forschung in ihren Themenstellungen nicht so frei wie die Grundlagenforschung, sondern richtet sich an den Fragestellungen, aber auch an den Zeit- und Budgetvorgaben von Auftraggebenden aus, die die Studie finanzieren (zu Unterschieden zwischen Grundlagen- und Anwendungsforschung ▶ Abschn. 1.1.3).

18.1.2 Definition der Evaluationsforschung

Wir verwenden den Begriff der **Evaluationsforschung** synonym mit **wissenschaftlicher Evaluation** (kurz: **Evaluation**), um zum Ausdruck zu bringen, dass es sich um Bewertungen handelt, die auf der Basis von Forschungsergebnissen in einem wissenschaftlichen Prozess durch sozialwissenschaftlich qualifizierte Evaluationsfachleute vorgenommen werden. Die der Bewertung zugrunde lie-

Grundlagenwissenschaftliche und technologische Theorien

Theo Herrmann (1979, Kap. 9) unterscheidet wissenschaftliche und technologische Theorien. Grundlagenwissenschaftliche Theorien beinhalten ein in sich schlüssiges Annahmengefüge über Ursachen und Wirkungen eines Sachverhaltes oder Phänomens (zu Merkmalen von Theorien ► Abschn. 2.2.10). Eine wichtige Aufgabe der empirischen Grundlagenforschung besteht darin, derartige Theorien zu entwickeln und ihre Gültigkeit zu überprüfen.

Technologische Theorien hingegen knüpfen an grundlagenwissenschaftliche Theorien an. Sie stellen die Basis für die Gewinnung von Regeln dar, mit denen die wissenschaftlichen Erkenntnisse praktisch nutzbar gemacht werden können. Ihr primäres Erkenntnisinteresse sind Formen des Handelns, mit denen etwas hervorgebracht, vermieden, verändert oder verbessert werden kann (sog. Interventionen).

Der jeweiligen instrumentellen Funktion entsprechend gelten für grundlagenwissenschaftliche und technologische Theorien unterschiedliche Bewertungsmaßstäbe:

- Eine **gute grundlagenwissenschaftliche Theorie** ist durch Kriterien wie eine präzise Terminologie, einen logisch konsistenten Informationsgehalt (Widerspruchslosigkeit), durch eine möglichst breite inhaltliche Tragweite sowie durch eine begrenzte Anzahl von Annahmen (Sparsamkeit) gekennzeichnet (► Abschn. 2.2.10).
- Eine **gute technologische Theorie** sollte wissenschaftliche Erkenntnisse in effiziente, routinisierbare Handlungsanleitungen umsetzen und Wege ihrer praktischen Nutzbarmachung aufzeigen.

Grundlagenwissenschaftliche und technologische Theorien sind für eine Wissenschaft gleichermaßen wichtig (Herrmann, 1976, S. 135, der in diesem Zusammenhang Physiker/innen als Vertreter wissenschaftlicher Theorien und Ingenieur/innen als Vertreter technologischer Theorien vergleicht).

Generell sind Evaluationsstudien in der Regel nicht in dem Maße theorieprüfend bzw. theoriebildend angelegt wie grundlagenwissenschaftliche Forschungsprojekte, vielmehr sind sie v. a. als **theorieverwendend** zu kennzeichnen. Denn der Fokus muss darauf liegen, einen wissenschaftlich fundierten Beitrag zu leisten, um die praxisbezogenen Fragestellungen der Auftraggebenden und anderer Anspruchsgruppen zu beantworten und ggf. identifizierte Praxisprobleme besser lösen zu können.

Ein kleines Beispiel soll das Gesagte verdeutlichen: Die **grundlagenwissenschaftliche Theorie zur Politikverdrossenheit** möge behaupten, Politikverdrossenheit sei ursächlich auf die überwiegend negative Berichterstattung über politische Ereignisse in den Medien bzw. insbesondere im Fernsehen zurückzuführen („Videomalaise"; Robinson, 1976). Die Grundlagenforschung würde sich nun z. B. damit befassen, was unter „Politikverdrossenheit" genau zu verstehen ist, wie mediale Stimuli geartet sind, die aversive Reaktionen auslösen, und welche Randbedingungen hierfür erfüllt sein müssen oder unter welchen Umständen negative Einstellungen zur Politik handlungsrelevant werden.

Nehmen wir einmal an, diese Grundlagenforschung käme zu dem Resultat, dass die Art, wie im Fernsehen über Politik berichtet wird, das Ausmaß der Politikverdrossenheit tatsächlich beeinflusst. Eine technologische Theorie hätte nun z. B. die Aufgabe, Annahmen darüber zu formulieren, durch welche Maßnahmen Politikverdrossenheit reduziert werden kann. Die **technologische Theorie über Strategien zur Reduzierung von Politikverdrossenheit** könnte z. B. behaupten, dass politische Nachrichten unterhaltsamer präsentiert werden müssen („Infotainment"), dass Politiker und Politikerinnen des Öfteren von ihrer menschlichen Seite gezeigt werden sollten oder dass „Good News" und „Bad News" in ihrem Verhältnis ausgewogener sein sollten. Diese Annahmen zu einer technologischen Theorie zu verdichten, wäre Aufgabe der Interventionsforschung bzw. anwendungsorientierten Forschung.

Entsprechende Maßnahmen bewertend zu überprüfen, obliegt dann der **Evaluationsforschung**. Vielleicht fände sich eine Fernsehanstalt bereit, ihre politische Berichterstattung nach den Empfehlungen der technologischen Theorie zu ändern und diese Maßnahme wissenschaftlich evaluieren zu lassen. Die Evaluation könnte sich auf einen Wirksamkeitsnachweis konzentrieren, damit wäre dann die technologische Theorie vorläufig bestätigt. Die Evaluationsstudie könnte aber auch bestrebt sein, die kausalen Wirkmechanismen zwischen einzelnen Merkmalen der modifizierten Politikberichterstattung und der Politikverdrossenheit bei unterschiedlichen Publika des Senders genauer theoretisch und empirisch zu rekonstruieren. Damit würde dann eine **Programmtheorie zu der evaluierten Maßnahme** entwickelt und überprüft. Bei Programmtheorien handelt es sich um jeweils für einen oder mehrere Anwendungsfälle (Programme als Bündel von einzelnen Interventionen) konkretisierte technologische Theorien.

Ob es überhaupt wünschenswert ist, Politikverdrossenheit zu reduzieren, oder ob man sie nicht eher steigern sollte, um damit langfristig politische Veränderungen zu forcieren, ist eine Frage der **Wertung**, die nicht wissenschaftlich, sondern nur ethisch-normativ zu begründen ist (Keuth, 1989). Die technologische Theorie präsentiert Handlungsoptionen; man kann sie zu Rate ziehen, um Politikverdrossenheit zu steigern oder zu senken, und man kann sich auch dafür entscheiden, nicht in die Politikverdrossenheit einzugreifen. Diese Entscheidungen sind von entsprechenden Entscheidungsträger/innen nach Maßgabe persönlicher Überzeugungen und Zielsetzungen zu treffen. In demokratischen Strukturen sind solche Entscheidungen eine Folge von Diskussionen und Prozessen der Konsensbildung, aber auch der Macht.

genden Daten werden in den meisten Evaluationsstudien neu erhoben. Es sind aber auch wissenschaftliche Evaluationen möglich, bei denen ausschließlich auf bereits systematisch gesammeltes Datenmaterial zurückgegriffen wird, sofern dieses eine ausreichende Datenqualität aufweist. Die Forschungsbetonung ist uns so wichtig, weil das Stichwort „Evaluation" heute für alle möglichen Bewertungen recht inflationär verwendet wird, die jedoch nicht wissenschaftlichen Standards genügen.

Es ist jedoch zu beachten, dass die Bezeichnung „Evaluationsforschung" (mit der wir die Wissenschaftlichkeit akzentuieren) auch hinterfragt und kritisiert wird (Kromrey, 2003; Lüders, 2006). Etwa in Fällen, in denen eine Evaluationsstudie mit vorhandenem Datenmaterial

auskommt und somit keinen eigenen Forschungsprozess mehr beinhaltet, sondern sich ganz auf Bewertungsvorgänge konzentriert. Auch wird moniert, dass bei Evaluationsforschung unklar sei, ob hier Forschung zum Zweck der Evaluation oder – was nämlich auch existiert – Forschung über Evaluationsprozesse gemeint ist. Evaluationsforschung im letztgenannten Sinne liegt vor, wenn sich Studien damit befassen, wie Methoden und Ansätze der Evaluation begründet, geprüft oder weiterentwickelt werden können (Widmer & De Rocchi, 2012, S. 14). Inwiefern Evaluationsmodelle (die ein bestimmtes methodisches Vorgehen vorgeben) und die Evaluationspraxis (also der Ablauf konkreter Evaluationsstudien) übereinstimmen, ist z. B. eine Fragestellung der Evaluationsforschung im Sinne einer Forschung über Evaluation.

In der Fachliteratur finden sich zahlreiche, jeweils leicht abweichende Definitionen der wissenschaftlichen Evaluation bzw. Evaluationsforschung (Mittag & Hager, 2000; Rossi & Freeman, 1993):

- Evaluation ist ein Prozess, „… bei dem nach zuvor festgelegten Zielen und explizit auf den Sachverhalt bezogenen und begründeten Kriterien ein Evaluationsgegenstand bewertet wird. Dies geschieht unter Zuhilfenahme sozialwissenschaftlicher Methoden durch Personen, welche hierfür besonders qualifiziert sind. Das Produkt eines Evaluationsprozesses besteht in einer Rückmeldung verwertbarer Ergebnisse in Form von Beschreibungen, begründeten Interpretationen und Empfehlungen an möglichst viele Beteiligte und Betroffene, um den Evaluationsgegenstand zu optimieren und zukünftiges Handeln zu unterstützen.“ (Balzer, 2005, S. 16)
- „Evaluationsforschung untersucht wissenschaftsgestützt unter Berücksichtigung geltender Standards (Zielexplikation und Standards der Evaluationsforschung) die Effektivität (Ausmaß der Zielerreichung) und Effizienz (Verhältnis von Aufwand und Nutzen) von Gegenständen.“ (Spiel, 2003, S. 883)
- „Evaluation: The systematic process of delineating, obtaining, reporting, and applying descriptive and judgemental information about some object’s merit, worth, probity, feasibility, safety, significance, or equity. The result of an evaluation process is an evaluation as product.“ (Stufflebeam & Shinkfield, 2007, S. 698)
- „Unter Evaluation wird eine wissenschaftliche Dienstleistung verstanden, die sich mit der systematischen und transparenten Bewertung eines Gegenstandes befasst.“ (Widmer & De Rocchi, 2012, S. 11)

Die Definitionen unterscheiden sich v. a. darin, a) ob und wie sie zentrale Aspekte wie z. B. den Evaluationsgegenstand oder die Evaluationskriterien durch Beispiele ausführen und b) welche weniger zentralen Aspekte sie noch anführen (z. B. Qualifikation der Evaluierenden). Wir schlagen eine Definition vor, die wichtige Elemente jedes Evaluationsforschungsprojektes benennt und exemplarisch illustriert.

Evaluationsforschung bzw. wissenschaftliche Evaluation – Die Evaluationsforschung („evaluation research“) bzw. wissenschaftliche Evaluation („evaluation“) nutzt sozialwissenschaftliche Methoden, um einen **Evaluationsgegenstand** (z. B. ein Produkt oder eine Maßnahme) unter Berücksichtigung der relevanten **Anspruchsgruppen** (z. B. Patienten, Angehörige, Produktenwickler, Evaluationsauftraggeber) anhand bestimmter **Evaluationskriterien** (z. B. Akzeptanz, Wirksamkeit, Effizienz, Nachhaltigkeit) und Maßgaben zu ihren Ausprägungen zu **bewerten**. Die durch den Prozess der Evaluationsforschung im Ergebnis erlangte Bewertung soll in der Praxis unterschiedliche **Evaluationsfunktionen** erfüllen (z. B. Legitimations- oder Optimierungsfunktion), was eine aktive **Evaluationsnutzung**, d. h. ein praktisches Aufgreifen der laufenden und/oder abschließenden Evaluationsergebnisse verlangt (z. B. unterstützt durch schriftliche Reports, mündliche Berichte und/oder Workshops des Evaluationsteams für die relevanten Anspruchsgruppen).
Die Evaluationsforschung ist den üblichen Prinzipien der Wissenschaftlichkeit sowie der Wissenschafts- und Forschungsethik verpflichtet, die für die besonderen Bedingungen der Evaluationsforschung zusätzlich in verbindliche **Evaluationsstandards** als Gütestandards für Evaluationsstudien gefasst wurden. Die Evaluationsforschung stellt über sozialwissenschaftliche Methodenkompetenz hinaus besondere **Anforderungen an die Qualifikation der Evaluierenden**.

18.1.3 Der Evaluationsgegenstand

Während man in der Grundlagenforschung vom Untersuchungsgegenstand bzw. Forschungsobjekt („research object“) spricht, ist in der Evaluationsforschung analog vom **Evaluationsgegenstand** bzw. **Evaluationsobjekt** oder auch vom **Evaluandum** („evaluation object“, „evaluand“) die Rede. Das Spektrum der Evaluationsgegenstände ist dabei sehr breit und kann kaum trennscharf und erschöpfend dargestellt werden. Wissenschaftlich evaluiert werden z. B. (Gollwitzer & Jäger, 2009; Wottawa & Thierau, 2003):

- Bauvorhaben (z. B. Bewertung unterschiedlicher Konzepte für ein neues kommunales Sportzentrum; Bewertung unterschiedlicher Standorte für einen neuen Flughafen hinsichtlich der Belastung der Anwohner durch Fluglärm),
- Evaluationsprojekte (Bewertung der Qualität einer Evaluationsstudie im Sinne einer Metaevaluation, ► Abschn. 18.3.5 „Durchführung einer Metaevaluation“),
- Forschung (z. B. Bewertung der wissenschaftlichen Leistungen unterschiedlicher universitärer und außeruniversitärer Forschungseinrichtungen),
- Gesetze und Gesetzesänderungen (z. B. Auswirkungen der Abschaffung der Todesstrafe auf die Kriminalitätsrate; Wirksamkeit von Anti-Terror-Gesetzen),

- Infrastruktureinrichtungen (z. B. Bewertung von Sicherheit und Zuverlässigkeit des öffentlichen Personennahverkehrs),
- Institutionen und öffentliche Einrichtungen (z. B. Vergleich von Privathochschulen und staatlichen Hochschulen),
- Interventionsmaßnahmen (z. B. Wirkung von bestimmten psychotherapeutischen, pädagogischen, oder medizinischen Maßnahmen gegen die Volkskrankheit Depression),
- Methoden (z. B. Vergleich der Tauglichkeit verschiedener Methoden zur Förderung der kindlichen Kreativität),
- Personen (z. B. Evaluation des Personals einer Organisation),
- Politikfelder (z. B. Bewertung des Erfolgs der Familienpolitik unterschiedlicher europäischer Länder),
- Präventionsmaßnahmen (z. B. Wirksamkeit von Kampagnen für gesundheitliche Vorsorgeuntersuchungen),
- Produkte (z. B. vergleichende Analysen der Wirkung verschiedener Psychopharmaka; Gesundheitsschäden durch verschiedene Holzschutzmittel; Usability verschiedener technischer Geräte),
- Programme (z. B. Evaluation eines Maßnahmenpakets zur Förderung der Gleichstellung von Menschen mit Handicaps),
- Projekte (z. B. Bewertung eines schulischen Pilotprojekts gegen Mobbing),
- Strategien (z. B. Zukunftsfähigkeit einer Unternehmensstrategie),
- Systeme (z. B. Auswirkungen verschiedener Unternehmensstrukturen auf die Zufriedenheit des Personals; Bewertung nationaler Schulsysteme durch internationale Bildungsstudien wie PISA) oder
- Unterricht und akademische Lehre (z. B. Evaluation von Studiengängen; Abb. 18.1).

Die Beispiele verdeutlichen die unterschiedlichen disziplinären Bezüge der Evaluationsforschung, denn evaluiert wird in der Psychologie, Soziologie und Medizin ebenso wie z. B. in der Politik-, Kommunikations-, Pflege-, Sport-, Erziehungs- oder Wirtschaftswissenschaft. Nicht nur in sozial-, auch in geistes- und technikwissenschaftlichen Fächern wird Evaluationsforschung betrieben, denn Museen werden ebenso evaluiert wie Serviceroboter (zum wissenschaftlichen Fächerkanon ▶ Abschn. 1.1.2).

Wenn wir im Folgenden zuweilen vereinfachend von „Maßnahmen“ sprechen, dann steht dieser Begriff stellvertretend für das gesamte Spektrum möglicher Evaluationsobjekte. Da die Evaluation von Maßnahmen, Interventionen bzw. Programmen in der Praxis besonders wichtig ist, wird im Kontext der sozialwissenschaftlichen Evaluationsforschung gern auch direkt von **Programmevaluation** („program evaluation“) gesprochen. Diese umfasst die Bewertung sämtlicher Maßnahmen, die eine zielgerichtete und planvolle Verbesserung bestehender Zustände beabsichtigen (Intervention) oder das Eintreten eines unerwünschter Zustandes verhindern wollen (Prävention). Unter dieses breite Verständnis von „Programmen“ fallen dann viele der o. g. Evaluationsgegenstände wie Interventionen, Techniken/Methoden, Projekte oder auch Bildungsmaßnahmen aller Art.

Evaluationsgegenstand oder Evaluationsobjekt – Als Evaluationsgegenstand oder Evaluationsobjekt („evaluation object“, „evaluand“) bezeichnet man den Untersuchungsgegenstand, auf den sich eine wissenschaftliche Evaluation bezieht. Ein besonders typischer Evaluationsgegenstand sind einzelne **Interventionsmaßnahmen** (z. B. psychotherapeutische, pädagogische oder medizinische Verfahren) bzw. größer angelegte **Programme** (z. B. Weiterbildungs-, Frauenförder-, Integrationsprogramme), die auf bestimmte individuelle und kollektive Veränderungen abzielen.

Evaluationsgegenstände sind in der Regel keine natürlich vorgefundenen Entitäten (eine Ausnahme bildet die Evaluation von Personal), sondern es sind durch Personen oder Institutionen (Praktiker) künstlich hergestellte Sachverhalte, mit denen Veränderungen erzielt werden sollen. Entsprechende Maßnahmen oder Programme basieren somit auf einer mehr oder minder explizierten **technologischen Theorie** (▶ Abschn. 18.1.1). Einfache Programmtheorien beschreiben als sog. **logische Modelle** relativ schematisch die Ablauflogik eines Programms anhand typischer Programmelemente (Century, Rudnick, & Freeman, 2010): Dementsprechend basiert ein Programm auf bestimmten Inputs und Aktivitäten, die zu Outputs (Produkte und Dienstleistungen), Outcomes (Wirkungen bei den Mitgliedern der Zielgruppe) und schließlich auch zu Impacts (langfristigen und überindividuellen Wirkungen z. B. auf regionaler oder gesellschaftlicher Ebene) führen sollen. Andere Programmtheorien arbeiten genauer aus, wie das Programm von den Verantwortlichen durchzuführen ist **(Aktionstheorie)** und wie es dann auf die Zielgruppen wirken soll (**Veränderungstheorie**, Chen, 1990, 2005; zur theorieorientierten Evaluation ▶ Abschn. 18.2.3). Während manche Evaluationsstudien gemäß ihrem Evaluationsauftrag darauf abzielen, die Wirkungsweise einer Maßnahme genau zu rekonstruieren und dazu **Programmtheorien über den Evaluationsgegenstand explizieren und prüfen**, zielen andere darauf ab, nur die Wirksamkeit einer Maßnahme nachzuweisen. Dabei wird dann die Maßnahme im Gesamtpaket als Ursachenfaktor betrachtet und im Zeitverlauf oder im Kontrast mit anderen Maßnahmen untersucht. Bei einer solchen sog. Black-Box-Evaluation werden die Wirkmechanismen des Evaluationsgegenstandes nicht genauer unter die Lupe genommen.

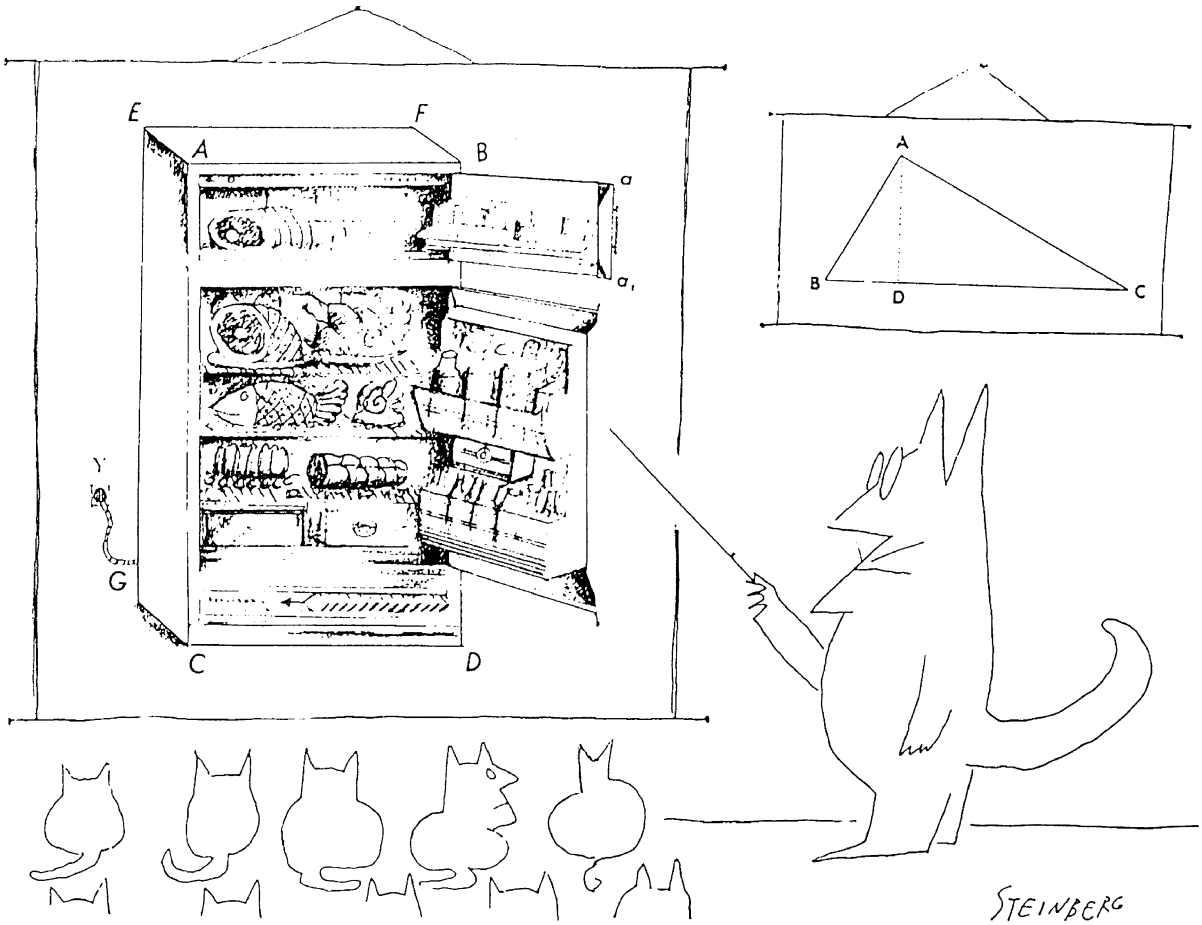


Abb. 18.1 Unterricht und Training sind wichtige Anwendungsbereiche der Evaluationsforschung. (© The Saul Steinberg Foundation/VG Bild-Kunst, Bonn 2015)

18.1.4 Die Anspruchsgruppen

Da die Evaluationsforschung auf eine wissenschaftlich fundierte Bewertung des jeweiligen Evaluationsgegenstandes hinausläuft, die oft als Entscheidungsgrundlage dient und weitreichende praktische Konsequenzen haben kann, ist es wichtig, die Sichtweisen der **Betroffenen** und **Beteiligten** bzw. der **Anspruchsgruppen** („stakeholder“) zu berücksichtigen. Für den Grad der Beteiligung der Stakeholder gibt es unterschiedliche Stufen der Partizipation beginnend mit Studien, die durch das wissenschaftliche Evaluationsteam geleitet und durchgeführt werden und die Anspruchsgruppen (bzw. Vertreter/innen derselben) vor allem als Experten und Informanten hinzuziehen über Studien, in denen den Anspruchsgruppen über die gesamte Projektlaufzeit hinweg eine beratende Rolle zukommt, bis zu partizipativen Studien, bei denen die Forschenden mit Vertreter/innen von Anspruchsgruppen als gleichberechtigten Forschungspartner/innen über alle Phasen des Prozesses hinweg kooperieren (dabei werden Vertreter/innen von Anspruchsgruppen teilweise auch auf Stellen im Evaluationsprojekt beschäftigt). Welche Form der Einbeziehung der Stakeholder gewählt wird, ist anhand des Forschungsproblems sowie der Kontextbedingungen zu entscheiden. Dass die Einbeziehung möglichst aller relevanten Anspruchsgruppen für gute Evaluationen notwendig ist, ergibt sich aus vier Gründen:

1. **Erkenntnistheoretisch** ist ein umfassendes Verständnis des Evaluationsgegenstandes, seiner Wirkungsweise und Effekte nur möglich, wenn man das Verhalten und Erleben der Beteiligten und Betroffenen umfassend berücksichtigt. Viele Evaluationskriterien setzen eine Erhebung der Bedürfnisse, Erwartungen, Erfahrungen, Vorkenntnisse, Veränderungen etc. der Zielgruppen einer Maßnahme voraus (z. B. Zufriedenheit mit einem Produkt, Lernerfolg durch ein Training).
2. **Forschungsethisch** muss sichergestellt werden, dass Betroffene und Beteiligte durch die Evaluationsforschung nicht ungewollt beeinträchtigt oder geschädigt

werden (zur Forschungsethik ▶ Abschn. 4.1; zur Forschungsethik im Kontext der Evaluationsforschung ▶ Abschn. 18.1.9 „Die Evaluationsstandards"). Dazu gehört im Zusammenhang mit Evaluationsmaßnahmen z. B. eine faire Berücksichtigung kritischer Minderheitenmeinungen oder auch eine rechtzeitige Information über Evaluationsergebnisse mit der Möglichkeit zur Korrektur von Missverständnissen im Abschlussbericht.

3. Die Einbeziehung von Stakeholdern ist ein wichtiger Faktor für die Nutzung von Evaluationsergebnissen. Wenn Stakeholder an der Evaluationsstudie aktiv beteiligt werden, den Forschungsprozess somit besser verstehen und den Ergebnissen stärker vertrauen, kann dies dazu beitragen, dass Evaluationen wirklich ihren intendierten Einfluss im Interventionsfeld entfalten und nicht ungenutzt in der Schublade liegen bleiben (zur **Evaluationsnutzung** ▶ Abschn. 18.1.7).
4. Schließlich existieren noch Evaluationsansätze, die auf der Basis ihrer Wertvorstellungen (v. a. Menschen- und Gesellschaftsbild) davon ausgehen, dass Betroffene ein Evaluationsprojekt von Anfang an maßgeblich aktiv mitgestalten und durch den Evaluationsprozess ihr Wissen und ihre Kompetenzen im Sinne von **Emanzipation** und **Empowerment** stärken sollen (zu stakeholderorientierten Evaluationsmodellen ▶ Abschn. 18.2.4 „Stakeholderorientierte Evaluationsmodelle").

Jedes Evaluationsforschungsprojekt ist nicht zuletzt gemäß den als Gütekriterien fungierenden Evaluationsstandards (▶ Abschn. 18.1.9) verpflichtet, die relevanten Stakeholder zu identifizieren und – je nach Fragestellung und Evaluationsansatz – angemessen einzubeziehen. Wenn Betroffene und Beteiligte an einer reaktiven Datenerhebung teilnehmen (z. B. an Interviews und Fragebogenerhebungen), bezeichnet man sie auch als **Evaluationsrespondenten** („evaluation respondents").

Es ist zu beachten, dass Stakeholder nicht nur die **unmittelbaren Zielgruppen** einer Maßnahme sind (z. B. Kunden, Patienten, Teilnehmer), sondern auch **indirekt Betroffene** (z. B. Angehörige, Freunde und Kinder der Zielgruppen). Zu den Stakeholdern zählen auch **Benachteiligte**, also jene Personengruppen, die durch die Maßnahme direkt oder indirekt Nachteile erleiden – etwa weil sie nicht zur angesprochenen Zielgruppe gehören oder weil Fördermaßnahmen, die dem Evaluationsgegenstand zugutekommen, gleichzeitig eine Förderung anderer Maßnahmen verhindern.

Daneben sind die an der Entwicklung und Optimierung des Evaluationsgegenstandes **Beteiligten** wichtige Stakeholder (z. B. Personal der betroffenen Einrichtungen, Produkt- und Maßnahmenentwickler, Maßnahmenverantwortliche; zu klientenorientierten Evaluationsmodellen ▶ Abschn. 18.2.4).

Beispiel: In einer Seniorenwohnanlage soll ein Pilotprojekt zum Einsatz von Gesundheitsassistenzrobotern durchgeführt und evaluiert werden. Die Roboter sollen das Pflegepersonal von Routinetätigkeiten entlasten (z. B. demente Personen in die Teeküche begleiten), die Bewohner/innen geistig, emotional und körperlich aktivieren (z. B. gymnastische und kognitive Übungen anleiten; Spielfunktionen anbieten) und ihnen bei der selbstständigen Alltagsbewältigung helfen (z. B. an die Medikamenteneinnahme und an ausreichendes Trinken erinnern). Welche Funktionen ein Assistenzroboter im Einzelnen ausführen soll, wie er zu gestalten und einzusetzen ist, kann aus Sicht unterschiedlicher Stakeholder (z. B. Pflegepersonal, Bewohnerinnen und Bewohner, Angehörige, Physiotherapeuten, Ergotherapeuten, Ärzte, Träger der Wohnanlage, Krankenkassen, Roboterentwickler) ganz unterschiedlich bewertet werden. Die Herausforderung besteht also nicht nur darin, alle relevanten Anspruchsgruppen zu erkennen und ihre jeweiligen Sichtweisen zu erfassen, sondern auch darin, mit divergierenden Positionen angemessen umzugehen, etwa wenn Krankenkassen die Maßnahme primär aus der Perspektive einer möglichen Kosten- und Personaleinsparung betrachten, während Bewohner/innen und Angehörige v. a. eine Erhöhung der Betreuungsqualität am Herzen liegt, während wiederum die Robotikforschenden auf die Grenzen bzw. Kosten der technischen Realisierbarkeit bestimmter Anforderungen hinweisen.

Evaluationsforschung wird teilweise durch Forschungsförderer oder aus den Haushaltsmitteln von Hochschulen finanziert. Oft jedoch handelt es sich um **Auftragsforschung** durch Personen oder Einrichtungen, die den Evaluationsgegenstand selbst entwickeln, verbessern, vermarkten oder auf der Grundlage der Evaluationsergebnisse Entscheidungen treffen wollen. Innerhalb der relevanten Stakeholder haben die **Auftraggebenden** der Evaluationsstudie eine zentrale Position, da sie das Forschungsprojekt finanzieren, die Forschungsfrage (mit-)definieren, die Ergebnisse abnehmen – und dann möglichst auch praktisch nutzen wollen bzw. sollen. Weitere wichtige Stakeholder sind **Entscheidungsträger** in Wirtschaft, Verwaltung und Politik, die über die Weiterführung, Einstellung oder Modifikation einer Maßnahme auf der Basis von Evaluationsergebnissen befinden, wobei diese in bestimmten Kontexten identisch mit den Auftraggebenden sein können. Eine enge Zusammenarbeit und häufige Rücksprache mit den Auftraggebenden und EntscheidungsträgerInnen – sowie generell allen intendierten Evaluationsnutzenden – ist deswegen empfehlenswert, um den Evaluationsauftrag bestmöglich zu erfüllen.

Eine zu enge Ausrichtung am Auftraggeber kann jedoch mit Blick auf die gewünschte Unabhängigkeit und Überparteilichkeit einer wissenschaftlichen Evaluation zum Problem werden. Etwa wenn sich ein Evaluationsteam unter mehr oder minder subtilem Druck sieht, dem Auftraggeber wunschgemäße bzw. positiv gefärbte Ergebnisse vorlegen zu müssen. Dieser Versuchung zu widerstehen gehört zu den Fähigkeiten entsprechend **qualifizierter Evaluationsfachleute**, die ihre Arbeit an den etablierten **Evaluationsstandards** (▶ Abschn. 18.1.9) orientieren. Ein wichtiges Korrektiv gegenüber unsachgemäßer Einflussnahme des Auftraggebers kann eine von Anfang an installierte **Metaevaluation** sein, etwa in Form einer Begleitung des Evaluationsprojektes durch einen externen wissenschaftlichen Beirat ▶ Abschn. 18.3.5 „Durchführung von Metaevaluation"). Die **Publikation** von Evaluationsstudien ist ein weiteres Instrument der Qualitätssicherung, da die Angemessenheit des methodischen Vorgehens sowie die Ausgewogenheit der Ergebnisdarstellung und -interpretation dann fachlicher Kritik der Scientific Community und anderer Interessierter unterzogen werden können.

Anspruchsgruppen bzw. Stakeholdern – Im Kontext der Evaluationsforschung sind mit **Anspruchsgruppen** bzw. **Stakeholdern** („stakeholder") alle direkt und indirekt von der Nutzung und Wirkung des Evaluationsgegenstandes **Betroffenen** („program beneficiary" bzw. **Zielgruppe**: Kunden, Patienten, Teilnehmer sowie deren Angehörige, Freunde etc.) sowie alle an der Entwicklung, Umsetzung und Optimierung des Evaluationsgegenstandes **Beteiligten** („program developer/provider": Produkt- und Maßnahmenentwickler, Maßnahmenverantwortliche, Dienstleistungspersonal, politische Entscheidungsträger etc.) gemeint.

Evaluationsklienten – Unter Evaluationsklienten („evaluation client") versteht man Beteiligte und/oder Betroffene, die eine Evaluation in Auftrag geben bzw. denen das Evaluationsteam Dienstleistungen bietet.

Evaluationsrespondenten – Unter Evaluationsrespondenten („evaluation respondent") versteht man Betroffene und Beteiligte, die an Datenerhebungen im Zuge einer Evaluationsstudie teilnehmen (z. B. mittels Fragebogen, Interview).

Im Rahmen der Evaluationsforschung gilt es, für die jeweilige Evaluationsfrage die **relevanten Stakeholder** zu identifizieren und deren Anliegen angemessen in den Forschungsprozess einzubeziehen. Besonders wichtige Stakeholder sind in der Regel die Auftraggebenden der Evaluation, die Maßnahmenverantwortlichen sowie die Zielgruppen der Maßnahme.

18.1.5 Die Evaluationskriterien

Ziel der Evaluationsforschung ist die wissenschaftlich fundierte **Bewertung** des Evaluationsgegenstandes. Dazu müssen klare und messbare Bewertungskriterien definiert, entsprechende Daten erhoben und interpretiert werden (◘ Abb. 18.2). Die **Auswahl und Festlegung der Evaluationskriterien** bestimmt maßgeblich das spätere Evaluationsergebnis. Kriterien sind unbrauchbar, wenn sie sich nicht empirisch erfassen lassen oder wenn sie inhaltlich nicht zur Lösung des Evaluationsproblems beitragen.



◘ **Abb. 18.2** Voraussetzung einer Evaluation ist die präzise Definition von Erfolgskriterien. (Mick Stevens/The New Yorker Collection/The Cartoon Bank; © Condé Nast)

Es ist zu beachten, dass neben der Definition der Evaluationskriterien auch noch die jeweiligen **Bewertungsstandards** festzulegen sind: Welche Ausprägungen des Evaluationskriteriums sollen als positives, ausreichendes oder negatives Ergebnis gewertet werden? Standards können z. B. als **Mindest-, Regel- oder Maximalstandards** gefasst werden. Sie können empirisch relativ zu einem Ausgangs- oder anderen Vergleichszustand formuliert sein (z. B. Symptomreduktion im Vorher-Nachher-Vergleich um 10 %) oder auch als absolute Werte bzw. Idealnormen festgelegt werden (z. B. Symptomfreiheit aller Patienten als Behandlungsziel; zu Vergleichskriterien Gollwitzer & Jäger, 2009, S. 48). Werden die abhängigen Variablen in einer Evaluationsstudie mit psychologischen Tests erhoben, so sind individuelle oder Gruppenvergleiche mit Normwerten möglich (▶ Abschn. 10.4). Bei **normorientierten Bewertungsmaßstäben** geht es um den Vergleich mit anderen (z. B. besser als die Vergleichsgruppe abschneiden), bei **kriteriumsorientierten Bewertungsmaßstäben** um den Vergleich mit einem festgelegten Wert (z. B. ein vordefiniertes Kriterium für gute Leistung erreichen).

> ***Bewertungsstandards* im Sinne von geforderten Merkmalsausprägungen des Evaluationsgegenstandes hinsichtlich der Evaluationskriterien sind Grundlage der Bewertung und bestimmen somit das inhaltliche Ergebnis einer Evaluation.**

Beispiel zu Bewertungsstandards: „Die evaluierte Lehrveranstaltung wurde mit der Schulnote 2.3 und somit überdurchschnittlich positiv von den Studierenden bewertet, ein Verbesserungsbedarf besteht hier nicht" (Be-

wertungsstandard ist hier die Durchschnittsnote von vergleichbaren Lehrveranstaltungsbewertungen).

Im Unterschied dazu sind die *Evaluationsstandards* Gütekriterien zur Bewertung der Qualität von Evaluationsstudien (► Abschn. 18.1.9).

Beispiel zu Evaluationsstandards: „Die Evaluationsstudie hat ihre Bewertungskriterien und Bewertungsstandards nicht vollständig offengelegt und mit Bezug zu den implizierten Werten begründet, sie verletzt somit den Nützlichkeitsstandard und ist in ihrer Aussagekraft eingeschränkt".

Beispiel: Die Qualität einer universitären Lehrveranstaltung soll evaluiert werden. Dazu wird die Gesamtbewertung der Lehrveranstaltung durch Studierende auf einer Schulnotenskala als ein Evaluationskriterium neben anderen (z. B. Bewertung durch Fachkollegen) ausgewählt. Interpretierbar ist das Ergebnis der Studierendenbefragung im Sinne einer Evaluation indessen nur, wenn auch die Bewertungsstandards klar sind (Döring, 2005): Bei welchen Durchschnittsnoten soll die Lehrveranstaltung aus Sicht der Evaluation als gut, bei welchen Noten als verbesserungswürdig eingeordnet werden? Die Festlegung der Bewertungsstandards ist dabei oft differenziert vorzunehmen. So ist eine Studierendenbewertung mit Note 3.0 anders einzuordnen, wenn es sich um eine thematisch unbeliebte, aber fachlich wichtige Pflichtvorlesung handelt oder um ein wahlobligatorisches Seminarangebot mit beliebten Themen. Bewertungsstandards sind teils kriterienorientiert festzulegen (z. B. bei Noten zwischen 1 und 3 wird kein Verbesserungsbedarf festgestellt), teils aber auch normorientiert durch entsprechende Vergleichsdaten (z. B. Durchschnittsnoten desselben Veranstaltungstyps an anderen Hochschulen). Normorientierte Bewertungsstandards (z. B. die 10 % der Lehrveranstaltungen mit den besten Noten werden als „exzellent", die 10 % mit den schlechtesten Noten als „verbesserungsbedürftig" eingestuft) können problematisch sein, da an einem Fachbereich oder in einem Studiengang, in dem insgesamt gute Lehre angeboten wird, auch die unteren 10 % gute Leistungen darstellen.

Je nach der aktuellen Entwicklungsphase des Evaluationsgegenstandes und in Abhängigkeit von der Funktion der Evaluation, beziehen sich die Evaluationskriterien auf drei verschiedene Aspekte der Maßnahme:

- **Kriterien zur Bewertung des Konzeptes einer Maßnahme**: Für die fragliche Maßnahme liegt ein Konzept vor und die Evaluationsstudie soll als **Konzeptevaluation** im Vorfeld einer möglichen Realisierung die Umsetzbarkeit bzw. Durchführbarkeit in der Praxis evaluieren. Dies kann Kriterien wie erwartete Kosten, erwarteter Nutzen, technische Realisierbarkeit, Personalaufwand, Bedarf in der Zielgruppe, Größe der Zielgruppe, erwartete Resonanz in der Öffentlichkeit etc. betreffen.
- **Kriterien zur Bewertung der Durchführung einer Maßnahme**: Die fragliche Maßnahme wird realisiert und die Merkmale der Umsetzung sollen im Zuge einer **Prozess- oder Implementierungsevaluation** („process evaluation", „implementation evaluation") bewertet werden, z. B. mit Blick auf möglichen Verbesserungsbedarf. Dies kann Kriterien der Intensität und Art der Nutzung sowie Nutzungsbarrieren, der Zusammensetzung und Größe der tatsächlich erreichten Zielgruppen, der Zufriedenheit bzw. Unzufriedenheit der Zielgruppen, des Zeit- und Kostenaufwandes für die Maßnahmenanbieter, unerwartete Probleme und Konflikte etc. umfassen.
- **Kriterien zur Bewertung der Ergebnisse einer Maßnahme**: Die fragliche Maßnahme wird bzw. wurde durchgeführt und ihre unmittelbaren Leistungen/Ergebnisse (output) sowie ihre kurzfristigen Wirkungen (outcome) und langfristigen und überindividuellen Folgen (impact) sollen im Rahmen einer **Ergebnis- bzw. Produktevaluation** („product evaluation") bewertet werden. Dies betrifft Kriterien der Wirksamkeit, Effizienz, Zufriedenheit, Akzeptanz, Sicherheit, Nützlichkeit, Nachhaltigkeit etc., aber auch mögliche unerwünschte Wirkungen. Bei der Ergebnisevaluation wird besonders häufig die **Wirksamkeit** bzw. **Effektivität** („effectiveness") als Kriterium dafür betrachtet, ob und in welchem Ausmaß die evaluierte Intervention in der Lage ist, den gewünschten Effekt unter Realbedingungen kausal zu verursachen (z. B. Lernfortschritt, Heilerfolg, Umsatzsteigerung; zur wirksamkeitsorientierten Evaluation ► Abschn. 18.2.1). Ergänzend wird mit dem Kriterium der **Effizienz** („efficiency") betrachtet, in welcher Relation Kosten und Nutzen der Maßnahme stehen (zur effizienzorientierten Evaluation ► Abschn. 18.2.1).

  Wirksamkeit und Effizienz einer Maßnahme lassen sich im Hinblick auf Output sowie Outcome untersuchen. Als **Output** werden dabei gewöhnlich die „zählbaren" Dienstleistungen und Produkte bezeichnet, die von der evaluierten Maßnahme hervorgebracht wurden. Der **Outcome** dagegen umfasst sämtliche Wirkungen, die durch den Output ausgelöst werden, wobei meist primär Wirkungen bei der Zielgruppe von Interesse sind. Der Output einer Maßnahme ist definitionsgemäß in der Regel sehr viel besser messbar als der Outcome; der Outcome ist dafür inhaltlich meist relevanter.

  Beispiel: Im Zuge der Lehrevaluation einer Universität kann die Anzahl ihrer Absolventinnen und Absolventen (mit Bachelorabschluss, Masterabschluss,

Promotion, Habilitation) als Output-Variable erfasst werden. Entsprechende Daten sind der ohnehin geführten Hochschulstatistik zu entnehmen. Eine Universität, die eine nennenswerte Zahl erfolgreicher Absolventen hervorbringt, wäre somit effektiv. Wenn sie dies mit vergleichsweise geringem Budget erreicht, wäre sie aus Output-Perspektive zudem effizient. Als Outcome-Variablen sind zudem die im Studium erworbenen Kompetenzen der Absolventen sowie ihr späterer Berufserfolg in den Blick zu nehmen. Entsprechende Outcome-Variablen wären mit relativ großem Aufwand zusätzlich zu erfassen (z. B. mittels fachspezifischer Kompetenztests sowie repräsentativer Absolventenbefragungen über den weiteren Berufs- und Karriereweg etc.). Auf Outcome-Ebene wäre die betreffende Universität effektiv, wenn sie Absolventen hervorbringt, die nicht nur über den formalen Abschluss verfügen (Output), sondern auch entsprechende Kompetenzen und Berufserfolg nachweisen können, die kausal auf die universitäre Ausbildung zurückführbar sind. Die Kriterien können dabei auch in einem Spannungsverhältnis stehen, so könnte höhere Output-Effizienz (viele Absolventen bei kleinem Budget) mit geringerer Outcome-Effektivität zusammenhängen (die Absolventen verfügen über weniger Kompetenzen). Noch schwieriger zu erfassen und kausal auf die Maßnahme zurückzuführen ist der **Impact**: So könnte die erfolgreiche Lehrstrategie einer Hochschule, die kausal vermehrt zu gut qualifizierten Absolventen führt, langfristig weitere übergeordnete (d. h. nicht nur die Absolventen als Zielgruppe der Lehre direkt betreffende) Positivfolgen nach sich ziehen (z. B. Reputationsgewinn für die Hochschule, höhere Bewerberzahlen bei den Studienanfängern, höheres Aufkommen von Spenden durch Ehemalige für ihre Hochschule).

Die konkreten Bewertungskriterien für eine Evaluationsstudie – etwa eine Ergebnisevaluation – lassen sich auf unterschiedliche Weise festlegen, wobei auch eine Kombination der Vorgehensweisen zur Kriterienfestlegung möglich ist:

- Die durch den Evaluationsgegenstand bzw. die jeweilige Maßnahme ausdrücklich angestrebten **Zielsetzungen** werden als Erfolgskriterien betrachtet („goal-based evaluation"). Dies setzt voraus, dass die zu evaluierende Maßnahme mit klaren Zielsetzungen versehen ist bzw. dass sich differenzierte Zielsetzungen ableiten lassen. Aus dem Projektmanagement ist z. B. die Faustregel bekannt, gute Ziele sollten **SMART** (spezifisch, messbar, akzeptiert, realistisch und terminiert) sein (Doran, 1981)
- Nicht selten werden die Interventionsziele und zugehörigen Bewertungskriterien durch die **Auftraggebenden** der Evaluationsstudie festgelegt, die z. B. aus ökonomischer Perspektive wissen möchten, ob die Maßnahme Kriterien der Wirtschaftlichkeit genügt bzw. wie sich die Kosten einer Maßnahme senken lassen. Evaluierende sollten Mängel in der Zielexplikation (z. B. zu vage formulierte Ziele) erkennen und Auftraggebenden ggf. bei der Zielexplikation behilflich sein. Wenn hierbei die subjektive Wertigkeit alternativer Ziele zur Diskussion steht, kann u. a. auf Techniken der Entscheidungsanalyse zurückgegriffen werden (Keeney, 1992).
- Zielsetzungen und Bewertungskriterien können und sollen zudem aus den Bedürfnissen, Erwartungen und Anforderungen aller **Stakeholder** der Maßnahme (nicht nur der Auftraggebenden bzw. Klienten der Evaluation) abgeleitet werden (z. B. Befragung der Zielgruppe der Maßnahme zu ihren Bedürfnissen und Erwartungen).
- Es existieren – je nach Art des Evaluationsgegenstandes – auch etablierte **Qualitätsstandards**, die zur Beurteilung herangezogen werden können. So kann eine neu entwickelte Software nach ihrer Usability, ein veränderter Arbeitsplatz nach seiner Ergonomie, eine Großveranstaltung nach ihrer Umweltverträglichkeit, eine politische Fördermaßnahme nach ihrer Geschlechtergerechtigkeit oder eine Bildungseinrichtung nach ihrer Barrierefreiheit beurteilt werden. Für die entsprechenden Kriterien liegen Theoriemodelle und Erhebungsverfahren vor.
- Relevante Bewertungskriterien für Evaluationsgegenstände und ihre intendierten sowie nicht-intendierten Wirkungen sind durch eine Recherche der **Fachliteratur** (theoretische Rahmenmodelle, vergleichbare Evaluationsstudien), anhand in Interviews erhobener **Urteile von Fachexperten** oder mittels systematischer Beobachtung und Analyse der **Praxis** (z. B. Orientierung an den Merkmalen international erfolgreicher Konkurrenzmaßnahmen durch Benchmarking) zu generieren.
- Neben den intendierten Wirkungen (gemäß offiziellen Zielsetzungen der Maßnahme, Erwartungen der Auftraggebenden etc.) sollten in einer umfassenden Evaluationsstudie auch **nicht-intendierte Wirkungen** erfasst werden. Dabei kann es sich um positive sowie negative Effekte handeln, die bei der Konzeption der Maßnahme ursprünglich nicht beabsichtigt waren. Um diesen auf die Spur zu kommen, dürfen sich Evaluierende von den offiziellen Maßnahmenzielen nicht blenden und von den Erwartungen der Auftraggebenden nicht vereinnahmen lassen. Sie müssen über un-

erwartete Wirkprozesse und Effekte nachdenken und v. a. auch sozialwissenschaftliche Forschungsmethoden einsetzen, die deren Aufdeckung erlauben (dies sind v. a. nicht-strukturierte **qualitative Methoden der Datenerhebung** wie z. B. ethnografische Feldforschung ► Abschn. 10.1.2 oder qualitative Interviews ► Abschn. 10.2.2).

Zudem wird die Methode der **zielfreien Evaluation** („goal-free evaluation"; Scriven, 1974; Stufflebeam & Shinkfield, 2007, S. 374ff.) vorgeschlagen, um nicht-intendierte Wirkungen einer Maßnahme aufzuspüren. Bei der zielfreien Evaluation werden die Zielsetzungen einer Maßnahme dem Evaluationsteam gar nicht mitgeteilt, das somit den Auftrag hat, alle möglichen beobachtbaren Effekte der Maßnahme zu erkunden und zu erfassen. Kennzeichen der zielfreien Evaluation ist die **fehlende Kenntnis der Evaluierenden über die offiziellen Programmziele**. Die Bewertung der zielfrei gefundenen Effekte erfolgt nachträglich anhand der Bedürfnisse der relevanten Stakeholder oder allgemein konsensfähiger Qualitätskriterien und Qualitätsstandards. Zielorientierte und zielfreie Evaluationsmethoden können sinnvoll kombiniert werden, sei es sequenziell (indem zunächst zielfrei und anschließend zielorientiert evaluiert wird) oder parallel (indem ein Teil des Teams zielorientiert, ein anderer Teil zielfrei vorgeht). Zu beachten ist, dass die Bezeichnung „zielfrei" insofern missverständlich ist, als es nicht darum geht, ziellos vorzugehen, sondern sich gedanklich von den offiziellen Programmzielen als Bewertungsrichtlinie zu lösen.

Beispiel: In einem strukturschwachen Wohngebiet wird ein kommunales Internetcafé eingerichtet mit dem Ziel, dass die stark von Arbeitslosigkeit betroffene lokale Bevölkerung über die kostenfreien Internetzugänge verstärkt nach Arbeitsplätzen sucht und elektronische Bewerbungen versendet. Tatsächlich stellt sich im Zuge der zielfreien Evaluation heraus, dass im Internetcafé in erster Linie Online-Chats und Internetspiele genutzt werden und dass sich das Café zu einem Jugendtreffpunkt entwickelt hat. Während eine ausschließlich zielorientierte Evaluation eine negative Bilanz ziehen würde (die bereit gestellten Internetrechner werden kaum zur Arbeitsplatzsuche genutzt), kann die zielfreie Evaluation eine positive Bilanz bieten (die soziale Integration im Kiez wird gefördert). Ob und inwiefern nicht-intendierte Effekte als Positiv- und/oder Negativwirkungen eingestuft und wie bedeutsam sie im Vergleich zu den ursprünglich avisierten Zielen eingeordnet werden, bedarf wiederum der Begründung und Diskussion zwischen Evaluierenden, Auftraggebenden, Beteiligten und Betroffenen sowie fachwissenschaftlicher und breiter Öffentlichkeit.

Es sei jedoch angemerkt, dass eine zielfreie Evaluation in Reinform in den meisten Kontexten praktisch nicht durchführbar erscheint: Ist etwa im Beispiel wirklich zu erwarten, dass die zielfrei Evaluierenden sich keine Gedanken über die Intentionen der Einrichtung des Internetcafés machen und diese nicht mehr oder weniger genau erraten können? Wie viele relevante Kontextfaktoren muss und kann man vor dem zielfrei operierenden Evaluationsteam verbergen, um das zu verhindern? Dürfen die Evaluierenden dann überhaupt mit der Zielgruppe oder mit dem Personal des Cafés sprechen, denn daraus könnten sie Rückschlüsse auf die offizielle Zielsetzung der Maßnahme ziehen? Muss man vor einer Vor-Ort-Beobachtung durch ein zielfrei operierendes Evaluationsteam alle möglicherweise verräterischen Hinweise (Prospekte, Poster, Flyer über die Maßnahme) entfernen? Diese Fragen sollen die Grenzen des zielfreien Ansatzes verdeutlichen. Wenn also die Programmziele in vielen Fällen kaum vor den Evaluierenden verborgen werden können, so ist und bleibt der Hinweis umso wichtiger, eine Evaluationsstudie – neben der Nutzung offizieller Programmziele als Erfolgskriterien (zielorientierter Ansatz) – zusätzlich so anzulegen, dass auch unerwartete Negativ- oder Positivwirkungen entdeckt werden können und somit eine ganzheitliche Bewertung der Maßnahme möglich wird. Empfehlenswert ist also in der Praxis ein **zielfrei ausgerichtetes Vorgehen trotz Kenntnis der Programmziele**. Damit wird dann zwar keine zielfreie Evaluation im engeren Sinne realisiert (die Unkenntnis der Programmziele verlangt), aber die Intention dieses Ansatzes aufgegriffen. Möglicherweise ist eine zielfreie Evaluation durch bewusste gedankliche Distanzierung der Evaluierenden von den – ihnen im Detail offengelegten – Interventionszielen sogar erfolgversprechender als ein Studiendesign, das krampfhaft versucht, die Evaluierenden „dumm" zu halten.

Evaluationskriterien – Evaluationskriterien („evaluation criteria") beziehen sich auf das Konzept, die Durchführung sowie die Ergebnisse einer Maßnahme bzw. eines Evaluationsobjektes und können ganz unterschiedliche Bewertungsaspekte ansprechen (z. B. ökonomische, ökologische, psychologische, soziale, ethische Kriterien).

Die Festlegung der relevanten Kriterien erfolgt anhand der Zielsetzungen der Maßnahme, der Vorgaben der Auftraggebenden, der Erwartungen weiterer relevanter Stakeholder, anhand allgemein konsensfähiger Qualitätsstandards, auf der Basis von Hinweisen aus der Fachliteratur und Fachcommunity und/oder aus der Praxis.

Ergänzend zur **zielorientierten Evaluation** anhand vordefinierter Erfolgskriterien, die sich aus den offiziellen Programmzielen ableiten („goal-based evaluation"), ist es oft sinnvoll, auch für **nicht-intendierte Wirkungen** einer Maßnahme offen zu sein und diese möglichst umfassend zu erheben. Hierzu eignen sich qualitative Methoden besonders gut. Die nicht-intendierten Effekte werden dann nachträglich anhand unterschiedlicher Kriterien bewertet. Bei der spezifischen Methode der

zielfreien Evaluation („goal-free evaluation") werden die offiziellen Maßnahmenziele den Evaluierenden gar nicht mitgeteilt, um so die Wahrscheinlichkeit zu erhöhen, dass sie auch nicht-intendierte Wirkungen aufdecken. Alternativ kann auch so vorgegangen werden, dass die Evaluierenden sich bewusst von den ihnen mitgeteilten offiziellen Maßnahmenzielen lösen und andere Wirkungen und Effekte erkunden.
Von besonders großer praktischer Bedeutung ist die zielorientierte Evaluation der Ergebnisse einer Maßnahme im Hinblick auf ihre **Effektivität** (Wirksamkeit, Ausmaß der Zielerreichung) sowie ihre **Effizienz** (Kosten für die Zielerreichung bzw. Kosten-Nutzen-Bilanz).

18.1.6 Die Evaluationsfunktionen

Die im Zuge einer Evaluation erarbeitete Bewertung des Evaluationsgegenstandes anhand unterschiedlicher Evaluationskriterien soll in der Praxis unterschiedlichen Funktionen, Zwecken bzw. Zielen dienen. Die **Evaluationsfunktionen** (bzw. **Evaluationszwecke** oder **Evaluationsziele**) werden in der Literatur unterschiedlich gegliedert (Stockmann, 2002, S. 3ff.; Widmer & De Rocchi, 2012, S. 27f.) und hängen eng miteinander zusammen. Wir unterscheiden pragmatisch 5 Funktionen:

1. **Erkenntnisfunktion:** Evaluationsforschung trägt definitionsgemäß dazu bei, wissenschaftliche Erkenntnisse über die Eigenschaften und Wirkungen von Evaluationsgegenständen zu sammeln. Diese Erkenntnisse erweitern – je nach ihrer Vermittlung – das Wissen der Beteiligten und Betroffenen sowie der Fachöffentlichkeit und zuweilen sogar der breiten Öffentlichkeit. Um die Erkenntnisfunktion zu erfüllen, müssen bei einer konkreten Evaluationsstudie die relevanten Evaluationsfragen konkretisiert werden und untersuchbar sein.
2. **Lern- und Dialogfunktion**: Im Zuge der Evaluation vollziehen die Beteiligten Lernprozesse und treten verschiedene Anspruchsgruppen miteinander in Dialog. Eine Evaluationsstudie kann die Funktion haben, einen solchen Dialog über den Evaluationsgegenstand zu fördern und fruchtbar zu gestalten. Damit eine Evaluationsstudie die Lern- und Dialogfunktion erfüllt, muss das Evaluationsteam besonderes Augenmerk darauf legen, relevante Stakeholder aktiv in den Evaluationsprozess einzubeziehen und zwischen divergierenden Positionen unterschiedlicher Stakeholdergruppen zu moderieren.
3. **Optimierungsfunktion:** Über die reine Wissenserweiterung und damit verbundene Lern- und Dialogprozesse hinaus soll Evaluation oft Erkenntnisse liefern, die zur zielgerichteten Verbesserung des Evaluationsgegenstandes beitragen können: „evaluation's most important purpose is not to prove, but to improve" (Stufflebeam, 2003, S. 4). Damit eine Evaluationsstudie die Optimierungsfunktion erfüllt, muss das Evaluationsteam sehr konkrete, detaillierte und umsetzbare Verbesserungsvorschläge erarbeiten und diese fortlaufend (formative Evaluation; ▶ Abschn. 18.1.8) oder am Ende gebündelt (summative Evaluation) an die Maßnahmenverantwortlichen zurückmelden.
4. **Entscheidungsfunktion:** Evaluationsergebnisse bilden nicht selten die Basis für Wahlentscheidungen zwischen verschiedenen Evaluationsgegenständen (soll Maßnahme A oder B umgesetzt werden?) oder auch für Entscheidungen über die Implementierung oder Weiterführung einer Maßnahme (lohnt es sich, die Maßnahme fortzusetzen oder sollte sie abgebrochen werden?). Damit eine Evaluationsstudie die Entscheidungsfunktion erfüllt, muss das Evaluationsteam die Evaluationsergebnisse zu klaren Praxisempfehlungen verdichten. Die Entscheidung selbst wird indessen nicht vom Evaluationsteam getroffen, es liefert im Zuge der unabhängigen wissenschaftlichen Bewertung des Evaluationsgegenstandes allenfalls eine **Entscheidungshilfe**.
5. **Legitimationsfunktion:** Sowohl die Durchführung von Evaluationsforschung als auch ihre Befunde sollen dazu beitragen, die Entwicklung und Durchführung einer Intervention zu kontrollieren und nach außen zu legitimieren. Vor allem bei öffentlich geförderten Programmen und Maßnahmen dient regelmäßige Evaluation dazu, über die sachgerechte Verwendung öffentlicher Gelder Rechenschaft abzulegen. Damit eine Evaluationsstudie die Legitimationsfunktion erfüllt, muss sie angesichts möglicher Zweifel oder Kritik an der Maßnahme und ihren Kosten, deren Merkmale und Wirkung sehr transparent darstellen und in überzeugender Weise unparteilich bzw. überparteilich bewerten.

Evaluationsfunktionen – Im Zuge der Evaluationsforschung wird eine wissenschaftlich fundierte Bewertung des Evaluationsgegenstandes erarbeitet, die wiederum dazu dient, übergeordnete praxisbezogene Funktionen zu erfüllen, die sich als **Erkenntnis-, Lern- und Dialog-, Optimierungs-, Entscheidungs-** und **Legitimationsfunktionen** fassen lassen.

Pseudoevaluation – Zuweilen werden Evaluationen aber auch für vordefinierte Zwecke instrumentalisiert bzw. missbraucht: Bei einer solchen Pseudoevaluation wird gar nicht ergebnisoffen eine wissenschaftliche Bewertung des Evaluationsgegenstandes vorgenommen, sondern das Ergebnis strategisch beeinflusst, manipuliert oder ignoriert.

Neben den o. g. Evaluationsfunktionen, die mit dem wissenschaftlichen Selbstverständnis harmonieren und Ergebnisoffenheit beinhalten, wird Evaluation teilweise auch für fragwürdige Zwecke **strategisch instrumentalisiert** und dadurch zur **Pseudoevaluation** (Stufflebeam & Shinkfield, 2007, S. 145ff.). So nutzen manche Auf-

traggebende eine Evaluation nur zur Selbstdarstellung, indem sie Mitarbeiter- oder Kundenbefragungen durchführen und somit Mitarbeiter- bzw. Kundenorientierung demonstrieren, ohne jedoch die dabei gewonnenen Daten jemals systematisch auszuwerten, geschweige denn aus den Befunden praktische Konsequenzen zu ziehen. Evaluation kann zudem als Durchsetzungshilfe eingesetzt werden, sei es von Befürwortern oder Gegnern einer Intervention, indem sie gezielt Einfluss nehmen auf die Durchführung und Ergebnisse der Evaluation (z. B. durch Vermeidung heikler Fragen im Evaluationsinstrument, durch Befragung vorselektierter Informanten und/oder durch Unterdrückung negativer Befunde im Evaluationsbericht). Schließlich kann bei heiklen Interventionen die Evaluation auch für eine Verantwortungsdelegation der Entscheidungsträger an die Evaluierenden genutzt werden oder auch eine Methode sein, um praktisches Handeln zu verzögern. Auftraggebende mit dominanten Instrumentalisierungsabsichten sind – wenn möglich – im Vorfeld zu identifizieren und zu meiden (für ein Beispiel ► Abschn. 18.3.1 „Festlegung und Prüfung der Evaluationsziele“).

Doch strategische Ziele verfolgen nicht nur Auftraggebende, sondern teilweise auch Evaluierende, etwa wenn sie überzogene Erwartungen von Auftraggebern (z. B. hinsichtlich globaler Wirksamkeitsnachweise) nicht als empirisch undurchführbar zurückweisen, um einen lukrativen Evaluationsauftrag nicht zu verlieren (Uhl, 1997). Oder wenn sie Evaluationsergebnisse vorauseilend im Sinne des Auftraggebers „gewichten“, um sich dessen Gunst und Folgeaufträge zu sichern.

18.1.7 Die Evaluationsnutzung

Damit das Evaluationsergebnis sachgerecht im Sinne der zu erfüllenden Evaluationsfunktionen genutzt werden kann – etwa hinsichtlich der Optimierung der Intervention oder einer finalen Entscheidung über den Einsatz der Intervention – sollte die spätere praxisbezogene Evaluationsnutzung möglichst von Anfang an geplant und vom Evaluationsteam auch aktiv unterstützt werden.

Verantwortung im Hinblick auf die Evaluationsnutzung zu übernehmen bedeutet für die Evaluierenden zunächst, ihre Ergebnisse verständlich darzustellen und sich vor eindeutigen Bewertungen und daraus abgeleiteten **Praxisempfehlungen** nicht zu scheuen. Anders als in der Grundlagenforschung, in der zu Recht vor überinterpretierten Ergebnissen zu warnen ist, kann die Evaluationsforschung ihrer **Ratgeberpflicht** nur nachkommen, wenn sie ihre Ergebnisse zu klaren Empfehlungen verdichtet. Derart eindeutige Empfehlungen werden umso leichter fallen, je besser dafür Sorge getragen wurde, dass die Evaluationsstudie nach Kriterien der Wissenschaftlichkeit und unter Einhaltung der Evaluationsstandards von und mit Fachleuten geplant und umgesetzt wurde.

Evaluationsergebnisse und darauf aufbauende Praxisempfehlungen in einem Abschlussbericht darzustellen, reicht für eine erfolgreiche Nutzung der Evaluationsergebnisse meist nicht aus. Denn oft kann nicht davon ausgegangen werden, dass Evaluationsergebnisse auch wirklich rezipiert, verstanden und akzeptiert werden und dann auch noch zur Ableitung und Umsetzung der richtigen Konsequenzen führen, wie etwa Hense und Mandl (2003) am Beispiel der Unterrichtsverbesserung durch Evaluationsrückmeldungen zeigen. Die Förderung der Evaluationsnutzung umfasst deswegen weitere Aktivitäten wie **Beratung**, **Coaching**, **Workshops**, **Weiterbildungskurse** etc., in denen Bedeutung und praktische Konsequenzen der Evaluationsergebnisse diskutiert und konzipiert werden (zur nutzungsfokussierten Evaluation ► Abschn. 18.2.4 „Klientenorientierte Evaluationsmodelle“ und zur Organisation der Evaluationsnutzung in der Abschlussphase eines Evaluationsprojektes ► Abschn. 18.3.5). Bei Fremdevaluation und externer Evaluation (zu diesen und weiteren Evaluationstypen ► Abschn. 18.1.8) sind definitionsgemäß außenstehende Evaluationsfachleute beteiligt, die dann auch als überparteiliche Vermittler und Moderatoren die Evaluationsnutzung unterstützen können. Bei Selbstevaluation und interner Evaluation sind die Evaluierenden gleichzeitig Maßnahmenbeteiligte und können in Loyalitätskonflikte geraten, wenn sie auf der Basis der Evaluationsdaten eine Negativbilanz ziehen oder unpopuläre Konsequenzen empfehlen. Eine gewisse Absicherung für alle Evaluierenden – und insbesondere diejenigen im Kontext der Selbstevaluation – bieten verbindliche Mandate und Richtlinien (zum Evaluationsvertrag ► Abschn. 18.3.2; zu Evaluationsstandards ► Abschn. 18.1.9).

Welche Faktoren konkret dazu beitragen, dass Evaluationsergebnisse tatsächlich in der Praxis genutzt werden, und wie diese Nutzungsprozesse im Einzelnen aussehen, wird seit den 1970er-Jahren intensiv untersucht (Miller & Campbell, 2006). Dabei stellte sich heraus, dass eine gezielte Evaluationsnutzung in dem Sinne, dass die empirischen Ergebnisse der Evaluationsstudie zu unmittelbaren Praxisveränderungen führen, nicht alle Folgen der Evaluation abdeckt. So können auch nichtintendierte Nutzungsweisen von Evaluationsergebnissen vorkommen oder Effekte erzielt werden, die nicht aus den Ergebnissen der Evaluation, sondern aus dem Evaluationsprozess selbst resultieren. Statt von Evaluationsnutzung („evaluation utilization“) wird umfassender vom **Evaluationseinfluss** („evaluation influence“) gesprochen, um die vielfältigen kurz- und langfristigen Wirkungen der Durchführung von Evaluationsforschung auf das jeweilige Praxisfeld anzusprechen (Beywl, Bestvater, & Friedrich, 2011; Coryn, Noakes, Westine, & Schröter, 2011).

Evaluationsnutzung – Mit der Evaluationsnutzung („evaluation utilization") ist die Verwendung der Ergebnisse einer wissenschaftlichen Evaluation durch Verantwortliche in der Praxis gemeint. Damit die Evaluationsergebnisse zu sachgerechten Entscheidungen (z. B. Weiterführung einer Maßnahme oder nicht) und sinnvollen Veränderungen (z. B. spezifische Verbesserungen einer Maßnahme) führen, ist die Abgabe eines Evaluationsberichtes mit entsprechenden Empfehlungen meist nicht ausreichend. Stattdessen sollten die Bedeutung und Konsequenzen der Evaluationsergebnisse mit den entsprechenden Stakeholdern diskutiert und Umsetzungsprozesse organisiert werden (z. B. durch Workshops, Weiterbildungsmaßnahmen, Einrichtung von Arbeitsgruppen).
Insbesondere wenn eine Evaluation die Funktion hat, die Maßnahme fortlaufend zu verbessern, sind Bemühungen um die Evaluationsnutzung nicht nur am Projektende sinnvoll, sondern häufige Rückkopplungen von Evaluationsergebnissen an die für die Entwicklung, Realisierung und Umsetzung des Evaluationsgegenstandes Zuständigen notwendig, etwa im Sinne eines Regelkreissystems.

18.1.8 Wichtige Typen von Evaluationsstudien

Evaluationsstudien werden nach ihren Kennzeichen in unterschiedliche Typen eingeteilt, wobei Mischformen möglich sind. Während die Typen sich eher auf **organisatorische Aspekte** der Studie beziehen, beinhalten die verschiedenen Evaluationsansätze unterschiedliche methodologische und theoretische Modelle des Evaluationsgeschehens (▶ Abschn. 18.2).

Ein erstes Einteilungskriterium für unterschiedliche Typen von Evaluationsstudien ist die **Verteilung der Rollen** zwischen Evaluationsteam und Evaluierten:

- **Fremd- und Selbstevaluation**: Der klassische Fall der Evaluation ist die Fremdevaluation. Dabei sind die Evaluierenden nicht an der untersuchten Maßnahme beteiligt oder von ihr betroffen, d. h. sie sind keine Stakeholder. Anders gesagt: Evaluierende und Evaluierte sind unterschiedliche Personen. Damit ist u. a. der Vorteil verbunden, dass Evaluierende als Fachleute sozialwissenschaftliche Methodenkompetenz in das Praxisfeld mitbringen und ihnen zudem als Außenstehenden ein neutralerer und unvoreingenommenerer Blick auf die Maßnahme zugeschrieben wird als den Maßnahmenbeteiligten. Fremdevaluatoren können aus derselben Institution stammen, die auch für die Maßnahme verantwortlich ist (interne Evaluation) oder von außerhalb der Institution kommen (externe Evaluation).
  Bei der Selbstevaluation sind die Maßnahmenbeteiligten bzw. Praktiker gleichzeitig die Evaluierenden: Sie treffen selbst Entscheidungen darüber, ob, wann und wie ihre Maßnahme unter welchen Fragestellungen evaluiert werden soll (Evaluationssteuerung), sind selbst an der Erhebung und Auswertung der Daten beteiligt (Evaluationsdurchführung) und sind selbst die Hauptnutzer der Evaluationsergebnisse (Evaluationsnutzung). Dabei ist zu beachten, dass die Datensammlung bei der Selbstevaluation sich keineswegs auf Selbsteinschätzungen beschränkt, sondern z. B. die Befragung und Beobachtung von Teilnehmern, Kunden, Klienten oder Expertenurteile beinhaltet. Der Vorteil der Selbstevaluation liegt darin, dass von einer Evaluation, die von den Praktikern selbst organisiert wird, ein besonders enger Praxisbezug und somit eine hohe praktische Verwertbarkeit der Ergebnisse zu erwarten sind. Gleichzeitig hat Selbstevaluation den Nachteil, dass sie teilweise als weniger „streng" wahrgenommen wird als Fremdevaluation. Sie steht eher im Verdacht, gewollt oder ungewollt verzerrte oder geschönte Ergebnisse zu produzieren. Die Installation von Metaevaluation (d. h. Evaluationsexperten begleiten und bewerten Selbstevaluationsaktivitäten; ▶ Abschn. 18.3.5) kann hier ein mögliches Regulativ sein. Selbstevaluation verlangt entsprechende Evaluationskompetenz, die bei den Praktikern oft zunächst aufgebaut werden muss und zusätzliche Arbeit bedeutet. Im deutschsprachigen Raum hat Selbstevaluation seit den 1980er-Jahren eine Tradition im Feld der sozialen Arbeit (Dusenbury, Brannigan, Falco, & Hansen, 2003; Gutknecht-Gmeiner, Hense, Kihm, Lüth, & Steckhan, 2011). Große Bedeutung hat Selbstevaluation zudem im Bildungsbereich, wo einzelne Lernende wie Lehrende oder auch ganze Schulen und Hochschulen systematische Selbstevaluation dazu nutzen können, ihre Bildungsaktivitäten zu verbessern (zur Selbstevaluation in der Lehre Beywl et al., 2011; Hense, 2006 sowie das Informationsportal http://www.selbstevaluation.de). In der angloamerikanischen Literatur werden Evaluationsansätze, die den Praktikern viel Verantwortung bei der Steuerung und Durchführung von Evaluationen überlassen, u. a. als partizipative Evaluation oder Empowermentevaluation konzeptualisiert (zu stakeholderorientierten Evaluationsmodellen ▶ Abschn. 18.2.4).
  Bei einer Fremdevaluation, bei der die Evaluierenden nicht direkt an der Maßnahme beteiligt sind, wird typischerweise ihre **institutionelle Zugehörigkeit** als weiteres Unterscheidungskriterium zur Typeneinteilung von Evaluationsstudien herangezogen:
  - **Interne und externe Fremdevaluation:** Bei der internen Fremdevaluation stammen die Evaluierenden aus derselben Institution, die auch für die Maßnahme verantwortlich ist. Dies ist z. B. der Fall, wenn Fachleute aus dem hochschuleigenen Didaktik- und Evaluationszentrum die Evaluation der Lehrveranstaltungen an ihrer eigenen Hochschule durchführen. Bei einer externen Fremdevaluation stammen die Evaluierenden dagegen

aus einer anderen Institution als der, an welcher die zu evaluierende Maßnahme angesiedelt ist. Der Vorteil der internen Evaluation besteht darin, dass die Evaluierenden ständig vor Ort sind und Zugriff auf eine große Fülle an Informationen und Kontakten haben, während externe Evaluierende, die im Rahmen der Evaluationsstudie – wenn überhaupt – nur wenige Besuche vor Ort abstatten können, kaum in der Lage sind, sich ein derart detailliertes Bild zu machen. Allerdings haben externe Evaluierende wiederum den Vorteil, dass sie feldfremde, neue Kenntnisse und Erfahrungen einbringen und gerade in kontroversen Fragen eine neutralere Haltung einnehmen können. Interne Evaluierende geraten eher in den Verdacht einseitiger Parteilichkeit oder können in Loyalitätskonflikte verstrickt werden. Ob interne oder externe Evaluationen kostspieliger sind, hängt vom Einzelfall ab. Bei internen Evaluationen sind die verdeckten Kosten (z. B. Zeitaufwand von bereits beschäftigten Mitarbeitenden einer Organisation) zu beachten.

Bei der Selbstevaluation entfällt die Differenzierung in intern und extern, denn durch die Personalunion von Evaluierten und Evaluierenden erfolgt die Selbstevaluation generell innerhalb ein und derselben Institution, also intern. Zu beachten ist daher: Jede Selbstevaluation ist intern, aber nicht jede interne Evaluation ist Selbstevaluation. Deswegen ist es problematisch, die Begriffe interne Evaluation und Selbstevaluation synonym zu verwenden, wie es in der deutsch- und englischsprachigen Fachliteratur jedoch immer wieder vorkommt.

Neben der institutionellen Verortung der Evaluationstätigkeiten bieten die **Evaluationsfunktionen** (► Abschn. 18.1.6) ein weiteres Unterscheidungskriterium für Typen von Evaluationsstudien:

- **Formative und summative Evaluation:** Die gestaltende bzw. formative Evaluation („formative evaluation“) erfüllt v. a. eine Optimierungsfunktion: Es geht darum, die Maßnahme zu verbessern. Dementsprechend sollen die Evaluationsergebnisse bei einer formativen Evaluation primär von den Maßnahmenbeteiligten genutzt werden. In der Regel ist dabei eine kontinuierliche Rückmeldung von Evaluationsergebnissen in die Praxis hilfreich, so dass in mehreren Rückkopplungsschleifen Verbesserungen vorgenommen und geprüft werden können. Qualitative Methoden spielen im Kontext der formativen Evaluation eine wichtige Rolle (z. B. werden Leitfadeninterviews eingesetzt, in denen einzelne Maßnahmennutzende Vor- und Nachteile der Maßnahme detailliert beschreiben und konkrete Änderungswünsche formulieren). Demgegenüber hat die bilanzierende bzw. summative Evaluation („summative evaluation“) v. a. eine Kontroll- und Legitimationsfunktion: Es geht darum, die Maßnahme zusammenfassend zu bewerten und auf diese Weise Rechenschaft abzulegen. Zuweilen erfüllen die Ergebnisse einer summativen Evaluation auch eine Entscheidungsfunktion: Soll die Maßnahme beendet oder durch eine bessere ersetzt werden? Quantitative Methoden spielen im Kontext der summativen Evaluation eine wichtige Rolle (z. B. soll eine möglichst repräsentative Stichprobe der Maßnahmennutzenden ihre Zufriedenheit mit der Maßnahme auf einer Schulnotenskala angeben). Die Differenzierung zwischen bilanzierend-summativer und gestaltend-formativer Evaluation geht auf den bekannten britisch-amerikanischen Evaluationsforscher **Michael Scriven** zurück (Scriven, 1967). Formative wie summative Evaluationen können jeweils als Selbstevaluationen sowie als interne oder externe Fremdevaluationen durchgeführt werden.

Eine andere Einteilungsmöglichkeit von Evaluationsstudien ergibt sich aus den Fragen, welche **Aspekte des Evaluationsgegenstandes** zu welchen **Zeitpunkten** untersucht werden:

- **Konzept-, Prozess- und Ergebnisevaluation**: Die Evaluation kann sich auf das Konzept der Maßnahme, auf die Durchführung (Implementierung) und/oder auf die Ergebnisse konzentrieren. Bei vielen Evaluationsstudien handelt es sich um Ergebnisevaluationen, die die prinzipielle Wirksamkeit der Maßnahme prüfen, also der Frage nachgehen, ob die Maßnahme die gewünschten Veränderungen hervorbringt (**Wirksamkeitsevaluation;** ► Abschn. 18.2.1 „Wirksamkeitsorientierte Evaluationsmodelle“). Andere Evaluationsstudien verfolgen das Ziel, die Wirkprozesse einer Maßnahme zu prüfen, was eine theoretische Modellierung der kausalen Relationen zwischen Elementen der Maßnahme sowie der Maßnahme und den Effekten voraussetzt (**theorieorientierte Wirkungsevaluation;** ► Abschn. 18.2.3). Für eine Wirkungsevaluation reicht es nicht aus, die Ergebnisse der Maßnahme zu betrachten, vielmehr muss detailliert auch die Durchführung der Maßnahme untersucht werden. Eine Prozess-, Durchführungs- bzw. Implementationsevaluation geht etwa der Frage nach, wie regelkonform eine Maßnahme umgesetzt wird (z. B. orientieren sich Therapie- oder Rehabilitationsprogramme in der Praxis wirklich genau an den Vorgaben des jeweiligen theoretischen Ansatzes, dem sie nominell folgen?), was u. a. eine Evaluation des für die Maßnahmendurchführung zuständigen Personals beinhaltet (z. B. Welche Qualifikationen bringen

die in der Maßnahme tätigen Therapeuten mit, wie gut sind sie im Detail mit den Prinzipien des offiziell verfolgten Therapieansatzes vertraut, inwiefern zeigen sie in der Praxis Verhaltensweisen, die der Ethik oder Methodik des Ansatzes widersprechen?). Personalevaluation ist politisch besonders heikel und trifft nicht selten auf große Widerstände und Ängste bei den Evaluierten, was von professionellen Evaluierenden entsprechend zu berücksichtigen ist, um eine vertrauensvolle Zusammenarbeit zu ermöglichen (Stufflebeam & Shinkfield, 2007, S. 115ff.). Um einen Evaluationsgegenstand ganzheitlich bzw. systemisch zu untersuchen, kann es sinnvoll sein, Konzept-, Prozess- und Ergebnisevaluation kombiniert zu realisieren, wie es z. B. das CIPP-Modell der Evaluation vorschlägt (zur **systemischen Evaluation** ▶ Abschn. 18.2.2).

- **Prospektive, begleitende und retrospektive Evaluation:** Auf der Zeitdimension wird differenziert, ob die Evaluation vor der Durchführung der Maßnahme (prospektive, Ex-ante-Evaluation), während der Durchführung der Maßnahme (begleitende, Ongoing-Evaluation) oder erst nach dem Ende der Maßnahme (retrospektive, Ex-post-Evaluation) erfolgt. Nicht selten wird Evaluationsbedarf erst im Nachhinein deutlich, etwa wenn Geldgeber für eine Fortsetzung der Maßnahme anzuwerben sind und dafür rückblickend der Wert der bereits beendeten Maßnahme festgestellt werden soll (z. B. anhand einer Sekundärauswertung von Dokumenten über die Durchführung und Ergebnisse der Maßnahme oder eine retrospektive Befragung der Beteiligten und Betroffenen). Viele Evaluationsstudien werden begleitend durchgeführt, denn während der Durchführung der Maßnahme bestehen die umfangreichsten Möglichkeiten der Datengewinnung. Allerdings ist eine begleitende Evaluation auch kostenintensiv, v. a., wenn es sich um eine zeitlich weit gestreckte Maßnahme handelt. Für eine genauere Darstellung speziell der prospektiven Evaluation s. Mittag und Hager (2000) sowie Gollwitzer und Jäger (2009, Kap. 7).

Nicht zuletzt findet sich auch die Differenzierung nach den drei großen **wissenschaftstheoretischen Paradigmen** (▶ Kap. 2) der empirischen Sozialforschung in der Evaluationsforschung wieder:

- **Quantitative, qualitative und Mixed-Methods-Evaluation:** Analog der wissenschaftstheoretischen Einteilung der Grundlagenforschung in einen quantitativ-statistischen und einen qualitativ-interpretativen Ansatz (▶ Kap. 2) werden auch Evaluationsstudien zuweilen entsprechend als quantitative versus qualitative Evaluation etikettiert. In diesem Sinne ist eine „qualitative Evaluationsstudie“ dadurch gekennzeichnet, dass sie einem qualitativen Wissenschaftsverständnis folgend qualitative Methoden einsetzt (zum Überblick siehe Flick, 2006; Shaw, 1999; für die Nutzung der dokumentarischen Methode innerhalb der Evaluationsforschung: Bohnsack & Nentwig-Gesemann, 2010; ▶ Abschn. 12.1.1). Die Differenzierung zwischen quantitativen und qualitativen Evaluationsstudien ist in der Praxis jedoch von **untergeordneter Bedeutung,** da die meisten Evaluationsstudien sowohl quantitative als auch qualitative Daten erheben und auswerten, also einen Methodenmix nutzen. Die meisten theoretischen Evaluationsmodelle (▶ Abschn. 18.2.3) sind für qualitative wie quantitative Methoden gleichermaßen ausgelegt oder empfehlen ausdrücklich einen Methodenmix. Dennoch sollte nicht jede Evaluationsstudie mit Methodenmix gleich als „Mixed-Methods-Evaluation“ etikettiert werden. Denn der Mixed-Methods-Ansatz (Creswell & Plano Clark, 2010; Greene, Caracelli, & Graham, 1989; ▶ Abschn. 7.1) verlangt nicht nur die Erhebung unterschiedlicher Datentypen, sondern v. a. eine systematische und wissenschaftstheoretisch reflektierte Integration qualitativer und quantitativer Methoden und Daten über den gesamten Forschungsprozess hinweg (z. B. Kuckartz & Busch, 2012; ▶ Abschn. 2.4). Eine Evaluationsstudie, die qualitative und quantitative Daten jeweils separat auswertet und interpretiert, folgt somit nicht dem Mixed-Methods-Paradigma.

18.1.9 Die Evaluationsstandards

Evaluationsstandards sind Vorgaben für gute wissenschaftliche Evaluationen, die bei allen Typen von Evaluationsstudien erfüllt werden sollten. Gemäß den Evaluationsstandards der Gesellschaft für Evaluation e. V. (DeGEval), die aus dem einschlägigen amerikanischen Standardset abgeleitet sind (Widmer, 2000, Kap. 5 sowie das Handbuch der Evaluationsstandards, Joint Committee on Standards for Educational Evaluation, 2006), haben **gute Evaluationsstudien** nach aktuellem Diskussionsstand vier grundlegende Eigenschaften aufzuweisen (die Definitionen orientieren sich überwiegend im Wortlaut an Gesellschaft für Evaluation e.V., 2008b):

- **Nützlichkeit** („utility“): Die Nützlichkeitsstandards sollen sicherstellen, dass sich die Evaluation an den geklärten Evaluationszwecken sowie am Informationsbedarf der vorgesehenen Nutzer und Nutzerinnen ausrichtet und die der Bewertung zugrunde liegenden Werte transparent macht. Die Interessen und Bedürfnisse der Beteiligten und Betroffenen sind dabei zu berücksichtigen. Die Evaluatorinnen und Evaluato-

ren sollen vertrauenswürdig und kompetent bzw. einschlägig qualifiziert sein. Evaluationsberichte sollen klar und verständlich sein, rechtzeitig vorliegen und ihre Nutzung soll gefördert werden.

- **Durchführbarkeit** („feasability"): Die Durchführbarkeitsstandards sollen sicherstellen, dass eine Evaluation realistisch, gut durchdacht, diplomatisch und kostenbewusst ausgeführt wird. Evaluationen sollten so ausgelegt sein, dass Störungen minimalisiert und die erforderlichen Informationen beschafft werden können. Durch diplomatisches Vorgehen soll die Akzeptanz der Stakeholder für den Evaluationsprozess sowie dessen Ergebnisse gefördert werden. Die für eine Evaluation eingesetzten finanziellen Mittel sollten in vernünftiger Relation zum Nutzen der Studie stehen.
- **Fairness** („propriety"): Die Fairnessstandards sollen sicherstellen, dass in einer Evaluation respektvoll und fair mit den betroffenen Personen und Gruppen umgegangen wird. Vereinbarungen zwischen den Vertragspartnern einer Evaluationsstudie sollten schriftlich festgehalten werden. Die individuellen Rechte der in die Evaluation einbezogenen Personen sind stets zu schützen. Die Stärken und Schwächen des evaluierten Programms sollten vollständig und fair dargestellt werden. Evaluationsergebnisse müssen allen an der Evaluation beteiligten Personen zugänglich gemacht werden.
- **Genauigkeit** („accuracy"): Die Genauigkeitsstandards sollen sicherstellen, dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervorbringt und vermittelt. Das zu evaluierende Programm muss korrekt beschrieben werden. Die genutzten Informationsquellen müssen reliabel und valide sein. Sowohl quantitative als auch qualitative Informationen sollen angemessen und systematisch aufgearbeitet und interpretiert werden. Die Evaluationsarbeit ist exakt zu dokumentieren.

Den vier Qualitätsdimensionen sind insgesamt 25 Einzelstandards zugeordnet. Nicht in jeder Evaluationsstudie lassen sich alle Einzelstandards realisieren, da diese teilweise auch in Konkurrenz zueinander stehen. Die Evaluationsstandards greifen – speziell auf den Bereich der Evaluationsforschung zugeschnitten und etwas anders gewichtet – die vier generellen Qualitätskriterien in der Wissenschaft auf (▶ Abschn. 3.1), nämlich inhaltliche Relevanz (Nützlichkeitsstandards), methodische Strenge (Genauigkeitsstandards, die z. B. die Testgütekriterien Objektivität, Reliabilität und Validität von Erhebungsinstrumenten ansprechen), ethische Strenge (Fairnessstandards, die u. a. auf Anonymität und Datenschutz sowie auf die Fairness von Testverfahren abheben; ▶ Abschn. 10.4.1) und Präsentationsqualität (z. B. Verständlichkeit von Ergebnisberichten, was unter die Nützlichkeitsstandards fällt). Da Evaluationsforschung oft Auftragsforschung ist, kommt der vertragskonformen Durchführung besondere Bedeutung zu (Durchführbarkeitsstandards). Evaluationsstandards sollen nicht zuletzt dabei helfen, in der Praxis sowohl die Interessen von Auftraggebenden und anderen Stakeholder einerseits als auch die Interessen von Fremd- bzw. Selbstevaluierenden andererseits zu wahren.

Bereits vor Projektstart ist zu analysieren, wie die Einhaltung der Evaluationsstandards während des gesamten Evaluationsprojekts sichergestellt und bei Bedarf nach Abschluss der Evaluation auch zusammenfassend bewertet werden soll. Im Falle von **Selbstevaluationen** sind weitere Empfehlungen zu beachten, die u. a. für gesicherte Rahmenbedingungen der – typischerweise abhängig beschäftigten – Selbstevaluatoren sorgen sollen (Müller-Kohlenberg & Beywl, 2003). Für **Metaevaluationen** (▶ Abschn. 18.3.5 „Durchführung einer Metaevaluation") sind ebenfalls spezifische Richtlinien zu beachten. Auch liegen teilweise eigene Standards für Evaluationsforschung in bestimmten Bereichen vor (z. B. Standards der Personalevaluation oder Standards der gesundheitsökonomischen Evaluation).

18.1.10 Anforderungen an Evaluierende

Damit Evaluationsforschung tatsächlich gemäß wissenschaftlichen Gütekriterien respektive den international anerkannten Evaluationsstandards durchgeführt wird, müssen die Evaluierenden entsprechende Kenntnisse und Erfahrungen mitbringen. Über Sach- und Methodenkompetenzen hinaus sind im Feld der Evaluationsforschung viel stärker als in der Grundlagenforschung auch persönliche und soziale Kompetenzen gefragt, etwa um bei Interessenskonflikten zwischen verschiedenen Anspruchsgruppen vermitteln oder Beeinflussungsversuchen und politischem Druck widerstehen und wissenschaftliche Unabhängigkeit wahren zu können (über Probleme des Selbstbildes von Evaluatoren berichtet Alkin, 1990).

Entscheidend für den Erfolg eines Evaluationsprojektes ist das gesamte **Evaluationsteam**, dessen Größe und Zusammensetzung von Art und Umfang des avisierten Projektes abhängt. Das Spektrum reicht von **Single-Evaluator-Projekten** bis zu arbeitsteiligen Evaluationsteams mit einem Dutzend oder mehr Mitgliedern (Stufflebeam & Shinkfield, 2007, S. 42ff.). Bei größeren Projekten wird meist eine Arbeitsteilung zwischen Projektmanagement (Projektleiter/in) und Evaluationsdurchführung vorgenommen, wobei aufseiten der Forschung erfahrene Evaluationsexpert/innen insbesondere für die Untersu-

chungsplanung und Interpretation der Ergebnisse notwendig sind, während einzelne Schritte der Datenerhebung und Datenauswertung auch von geschulten Hilfskräften übernommen werden können (Tätigkeiten als Beobachter, Interviewer, Kodierer etc.). Zur Qualitätssicherung sollte möglichst von vorne herein eine Metaevaluation (Widmer, 1996) in das Evaluationsprojekt integriert werden, d. h., eine Überprüfung der Evaluationsarbeit durch externe Evaluationsexperten bzw. einen wissenschaftlichen Beirat. Große Evaluationsprojekte kommen heute kaum ohne spezifische Informations- und Kommunikationstechnologie aus (Aufbau von Datenbanken und Intranets, Beschaffung und Wartung von Spezialsoftware etc.), weshalb teilweise auch IT-Fachleute in einem Evaluationsteam mitarbeiten.

Während sich die Evaluationsstandards auf die Qualität von Evaluationsstudien beziehen, sind die von der American Evaluation Association (AEA) herausgegebenen **Guiding Principles for Evaluators** (AEA, 2004; http://www.eval.org/publications/guidingprinciples.asp) auf die Evaluierenden und ihr Selbstverständnis zugeschnitten und fordern

- „systematic inquiry“: die systematische, datenbasierte Untersuchung des zu evaluierenden Gegenstands,
- „competence“: den Beleg ausreichender Fachkompetenz zur Durchführung von Evaluationsstudien,
- „integrity/honesty“: die Gewährleistung eines fairen und integren Evaluationsprozesses,
- „respect for people“: den angemessenen Respekt gegenüber den Persönlichkeitsrechten aller Beteiligten,
- „responsibility for general and public welfare“: ein allgemeines Verantwortungsgefühl für das durch Programme beeinflusste öffentliche Gemeinwohl.

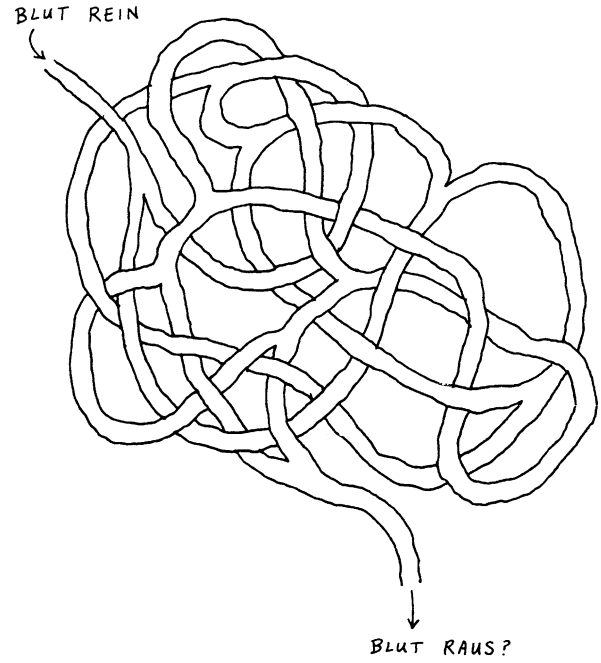
Auftraggebenden wird empfohlen, nur Evaluierende zu engagieren, die sich diesen Prinzipien verpflichtet haben und entsprechende Kompetenzen nachweisen können. Dementsprechend ist es umgekehrt von Seiten der Evaluatorinnen und Evaluatoren wichtig, ihre Forschungsleitlinien transparent zu machen.

Die Gesellschaft für Evaluation e. V. DeGEval hat zentrale **Anforderungen an Evaluierende** formuliert und daraus auch Empfehlungen für die Ausbildung abgeleitet (Heiner, 1988; Mowbray, Holter, Teague, & Bybee, 2003; Rädiker & Niklesz, 2012). Dementsprechend werden vier zentrale Kompetenzfelder unterschieden; zudem wird von Evaluierenden ein Mindestmaß an Praxiserfahrung im jeweiligen Evaluationsbereich gefordert (DeGEval, 2004):

1. **Kenntnisse über Theorie und Geschichte der Evaluationsforschung**: Evaluierende müssen die zentralen Konzepte der Evaluationsforschung kennen (z. B. Definition der Evaluation, Evaluationsfunktionen, Evaluationsnutzung), über die geschichtliche Entwicklung der Evaluationsforschung auf dem Laufenden (z. B. nordamerikanische vs. europäische Evaluationstraditionen) und mit unterschiedlichen Ansätzen und Modellen der Evaluation sowie mit den Evaluationsstandards vertraut sein.
2. **Methodenkompetenzen**: Evaluierende müssen über solide Kenntnisse der Methoden empirischer Sozialforschung verfügen und dabei speziell auch die Besonderheiten der Anwendungs- und Evaluationsforschung kennen: Sie müssen sich auskennen mit Untersuchungsplanung, Datenerhebung, qualitativer und quantitativer Datenanalyse inklusive entsprechender Analysesoftware sowie Ergebnisinterpretation und Berichterstattung. Neben der sozialwissenschaftlichen Methodenkompetenz benötigen sie auch Methodenkompetenz im Bereich des Projektmanagements (z. B. Zeit- und Kostenplanung und entsprechende Kontrolle).
3. **Kenntnisse über Evaluationsgegenstände und Evaluationskontexte**: Eine gute Evaluationsstudie muss detailliert auf den jeweiligen Evaluationsgegenstand abgestimmt sein, dementsprechend ist eine solide Wissensbasis notwendig. Angesichts der Vielfalt der Evaluationsgegenstände ist generelles Überblickswissen kaum möglich, aber in zumindest zwei inhaltlichen Evaluationsbereichen (z. B. Evaluation von Schulen, von psychotherapeutischen Interventionen oder von technischen Innovationen) sollten Evaluierende Kenntnisse aufweisen. Zudem sind Kenntnisse über die organisationalen (z. B. Strukturen von und Abläufe in Organisationen, Bedingungen für Organisationswandel) und rechtlichen (z. B. Rechte und Pflichten von Organisationsmitgliedern, juristische Grenzen von Veränderungsanliegen) Rahmenbedingungen von Evaluationen erforderlich. So können Evaluierende keine praktisch nützlichen Empfehlungen aus den Evaluationsbefunden ableiten, wenn sie sich über die faktischen und rechtlichen Möglichkeiten und Grenzen von Veränderungen (z. B. in Schulen) nicht im Klaren sind. Wissenslücken der Evaluierenden hinsichtlich des Evaluationsgegenstandes und bei den Kontextbedingungen fallen Betroffenen und Beteiligten der Evaluation meist sehr negativ auf und beeinträchtigen die Glaubwürdigkeit der Evaluierenden.
4. **Soziale und personale Kompetenzen**: Evaluierende müssen über Sozialkompetenz (z. B. Einfühlungsvermögen), über kommunikative Kompetenz (z. B. verständliches Schreiben) sowie über kooperative Kompetenz (z. B. Moderation von Arbeitsgruppen) verfügen. Darüber hinaus ist als persönliche Kompetenz die Selbstmanagementkompetenz gefragt (z. B. eigene Arbeitsorganisation, Selbstmotivation). Nicht zuletzt sind in dem komplexen und dynamischen

Arbeitsfeld der Evaluation Lern- und Problemlösungskompetenzen unabdingbar (z. B. Fähigkeit zur Anpassung von Untersuchungsplänen an veränderte Feldbedingungen; Bereitschaft und Fähigkeit zum Erwerb neuer interkultureller Kompetenzen). Gerade der Aspekt der sozialen Kompetenz ist nicht zu unterschätzen. So werden Evaluationsbemühungen durchaus nicht von allen Anspruchsgruppen begrüßt, sondern stoßen nicht selten auf Vorbehalte, Ängste und massive Widerstände. Hier sind Feingefühl, diplomatisches Geschick, Überzeugungskraft und Durchhaltevermögen gefragt, um konstruktive Arbeitsbedingungen zu schaffen und zu erhalten.

5. **Praxiserfahrungen in der Evaluationsforschung**: Es werden mindestens zwei sechswöchige Praktika in unterschiedlichen Evaluationsfeldern verlangt, die durch Fachleute mit mehrjähriger Evaluationserfahrung betreut werden sollen. Wichtig ist dabei jeweils eine praktische Beteiligung an allen Phasen der Evaluationsstudie (▶ Abschn. 18.3), da in den einzelnen Phasen teilweise sehr unterschiedliche Kenntnisse und Fähigkeiten einzubringen sind (z. B. Entwicklung des Evaluationskonzeptes; schriftliche und mündliche Präsentation der Evaluationsergebnisse; ◘ Abb. 18.3).



◘ **Abb. 18.3** Genau wie die Intervention sollte auch die Evaluation von Experten durchgeführt werden. (Aus Poskitt & Appleby, 1993. Mit freundlicher Genehmigung von Steven Appleby)

18.1.11 Fachinformationen zur Evaluationsforschung

Die Evaluationsforschung hat in den letzten Jahren einen immensen Boom erlebt und ist dementsprechend mittlerweile sehr gut institutionalisiert: Es existieren weltweit wissenschaftliche Fachgesellschaften zur Evaluationsforschung, ebenso Fachzeitschriften und einschlägige Lehr- und Handbücher. Zudem haben sich als Handreichungen für die Praxis sog. Evaluationschecklisten eingebürgert.

Wissenschaftliche Fachgesellschaften

In den disziplinären Fachgesellschaften existieren oft eigene Methoden- und Evaluationsfachgruppen, etwa die Fachgruppe „Methoden und Evaluation" in der Deutschen Gesellschaft für Psychologie. Daneben haben sich aber auch eigene **Evaluationsfachgesellschaften** mit informativen Internetpräsenzen formiert:

- American Evaluation Association, AEA (http://www.eval.org)
- Australasien Evaluation Society, AES (http://www.aes.asn.au)
- Canadian Evaluation Society, CES (http://www.evaluationcanada.ca)
- European Evaluation Society, EES (http://www.europeanevaluation.org)
- Gesellschaft für Evaluation e. V., DeGEval (http://www.degeval.de)
- Schweizerische Evaluationsgesellschaft, SEVAL (http://www.seval.ch)
- UK Evaluation Society, UKES (http://www.evaluation.org.uk)

Wissenschaftliche Fachzeitschriften

Evaluationsstudien werden vereinzelt in den disziplinären Fachzeitschriften publiziert. Einschlägiger sind die dezidierten **Evaluationsfachzeitschriften**, von denen hier einige in alphabetischer Reihenfolge mit ihren Verlagen genannt seien:

- Advances in Program Evaluation (Emerald)
- American Journal of Evaluation (Sage)
- Canadian Journal of Program Evaluation (Canadian Evaluation Society)
- Evaluation and Program Planning (Elsevier)
- Evaluation and the Health Professions (Sage)

- Evaluation Review – A Journal of Applied Social Research (Sage)
- Evaluation – The International Journal of Theory, Research and Practice (Sage)
- International Journal of Educational Research (Elsevier)
- Journal of Evaluation in Clinical Practice (Wiley)
- Journal of Multidisciplinary Evaluation, JMDE (Online Journal, http://survey.ate.wmich.edu/jmde/)
- Journal of Personnel Evaluation in Education (Springer)
- New Directions for Evaluation (Wiley)
- Klinische Diagnostik und Evaluation (V and R Unipress)
- LeGes – Gesetzgebung and Evaluation (Mitteilungsblatt der Schweizerischen Gesellschaft für Gesetzgebung, SGG, und der Schweizerischen Evaluationsgesellschaft, SEVAL: http://www.seval.ch/de/le_ges/)
- Practical Assessment, Research and Evaluation, PARE (Online-Journal, http://pareonline.net/)
- Studies in Educational Evaluation (Elsevier)
- Zeitschrift für Evaluation (Waxmann).

Einschlägige Lehr- und Handbücher

Eine Auswahl deutsch- und englischsprachiger Lehr- und Handbücher ist unter „Weiterführende Literatur“ aufgeführt.

Evaluationschecklisten

Zur Erleichterung und Professionalisierung der praktischen Arbeit in der Evaluationsforschung hat der Evaluationsforscher Michael Scriven die **Checklistenmethode** eingeführt (eine Sammlung von Evaluationschecklisten bietet die Website des Evaluationszentrums der Universität Michigan: http://www.wmich.edu/evalctr/checklists/). Als Gedächtnisstütze liefert die Checkliste eine geordnete – und oft auch chronologische – Auflistung wichtiger Fragen, Kriterien oder Arbeitsschritte der Evaluationspraxis, implizit enthält sie dabei auch jeweils ein theoretisches Modell (bei chronologischen Listen ein Phasenmodell) der Evaluationsarbeit (Scriven, 2007).

Social Media

Für den Austausch und die Diskussion von wissenschaftlichen Informationen gewinnen Internet und speziell Social Media an Bedeutung (► Abschn. 13.4.1). Das Angebotsspektrum ist hier sehr heterogen und kann naturgemäß nur exemplarisch beleuchtet werden.

- So bündelt z. B. die englischsprachige Plattform **EvalCentral** (http://www.evalcentral.com) die Beiträge unterschiedlicher Weblogs, die von Evaluationsfachleuten geführt werden.
- Das **Eval-Wiki**: Glossar der Evaluation ist ein redaktionell durch das Unternehmen Univation – Institut für Evaluation Dr. Beywl & Associates betreutes deutschsprachiges Wiki, in dem Begriffsdefinitionen aus dem Bereich der Evaluationsforschung mit Quellenangaben zusammengetragen werden (http://www.eval-wiki.org/glossar/Eval-Wiki:_Glossar_der_Evaluation).
- **Evalinfo** ist ein Wiki zur Evaluation von Informationssystemen, das vom Deutschen Institut für Pädagogische Forschung, DIPF, organisiert wird und zur Mitarbeit einlädt (http://evalinfo.dipf.de/evalinfo). Es bietet u. a. Checklisten und Fragebögen, die bei der Durchführung eigener Evaluationsstudien helfen können.
- Ein weiteres nützliches **Wörterbuch der Evaluation** findet sich unter http://www.evoluation.de (verfasst von Jan Hense).

Die führende deutschsprachige **Mailingliste** mit rund 700 Mitgliedern ist die seit 1997 von Wolfgang Beywl in Kooperation mit der Gesellschaft für Evaluation e. V. DeGeEval administrierte Liste **forum-evaluation** (https://lists.uni-koeln.de/mailman/listinfo/forum-evaluation). Über diese Liste werden vorwiegend Veranstaltungshinweise, Stellenausschreibungen aber auch Projektausschreibungen aus dem Bereich der Evaluationsforschung verbreitet. Demgegenüber finden über die Mailingliste **EVALTALK** der American Evaluation Association, AEA, (http://www.eval.org/resources/listservs.asp) häufiger auch fachliche Diskussionen statt.

18.2 Verschiedene Ansätze in der Evaluationsforschung

Innerhalb der Evaluationsforschung haben sich in den letzten Jahrzehnten unterschiedliche theoretische Ansätze herausgebildet, die sich in ihren Zielsetzungen, ihrem Evaluationsverständnis und ihren jeweiligen Methoden unterscheiden (Miller, 2010). Für **jeden Ansatz** liegen meist mehrere verschiedene **Evaluationsmodelle** vor, die den wissenschaftstheoretischen Hintergrund und das methodische Vorgehen präzisieren. Dabei sind weder die Bündelung der Evaluationsmodelle zu übergeordneten Evaluationsansätzen noch deren Etikettierungen in der Literatur einheitlich (z. B. Soellner, 2010: 1. ergebnisorientierte Modelle, 2. Prozessmodelle, 3. Nutzenmodelle, 4. bewertungsorientierte Modelle, 5. Programmtheoriemodelle; Stufflebeam & Shinkfield, 2007: 1. questions- and methods-oriented evaluation approaches, 2. improvement- and accountability-oriented evaluation approaches, 3. social agenda and advocacy ap-

proaches, 4. eclectic approaches; ; Widmer & De Rocchi, 2012: 1. Effektivitätsmodelle, 2. ökonomische Modelle, 3. akteursorientierte Modelle).

Die folgende Darstellung der Ansätze erhebt keinen Anspruch auf Vollständigkeit, sondern soll durch didaktische Vereinfachung die Charakteristika einiger wichtiger Evaluationsansätze im Kontrast zueinander verdeutlichen und an ausgewählten Modellen und Studienbeispielen illustrieren:

- **Ergebnisorientierter Evaluationsansatz:** Er ist in der Praxis sehr verbreitet und konzentriert sich meist auf zwei Merkmale der zu evaluierenden Maßnahme: ihre Wirksamkeit (Effektivität) sowie das Verhältnis von Nutzen und Kosten (Effizienz).
- **Systemischer Evaluationsansatz:** Dieser geht einen Schritt weiter und betrachtet den Evaluationsgegenstand umfassender, also untersucht nicht nur die Ergebnisse der Maßnahme, sondern auch die Eingangs- und Kontextbedingungen sowie die Art und Weise der Umsetzung der Maßnahme. Eine systemische Evaluation liefert also potenziell ein recht umfassendes Bild der Maßnahme.
- **Theorieorientierter Evaluationsansatz:** Er ist noch anspruchsvoller und analysiert im Detail die Ursache-Wirkungs-Mechanismen, die innerhalb einer Maßnahme greifen und für die unterschiedlichen Effekte einer Maßnahme verantwortlich sind.
- **Akteursorientierter Evaluationsansatz:** Dieser ist schließlich dadurch gekennzeichnet, dass seine Forschungsaktivitäten in den Dienst einzelner Stakeholder-Gruppen gestellt werden (klientenorientierte Modelle) oder alle Stakeholder gleichermaßen aktiv einbeziehen sollen bis hin zur Steuerung der Evaluation durch die Stakeholder selbst (stakeholderorientierte Modelle).

Die im Bereich der Evaluationsforschung vorliegenden theoretischen Ansätze und Modelle haben **präskriptiven Charakter**, d. h. sie geben methodologisch und inhaltlich begründet mit unterschiedlichen Akzentsetzungen vor, wie eine aussagekräftige Evaluationsstudie zu realisieren ist. Dabei machen verschiedene Ansätze teilweise unterschiedliche oder gar gegensätzliche Vorgaben etwa zur Rolle der Evaluierenden, zur Bedeutung von Programmtheorien oder zur Wahl geeigneter Untersuchungsdesigns und Datenerhebungsmethoden. Teilweise herrscht regelrechter Streit um das „richtige" Vorgehen zwischen den verschiedenen Ansätzen. Abgesehen von theoretischen Diskursen, in denen die Modelle argumentativ aneinander gemessen werden, lassen sich die verschiedenen theoretischen Ansätze in der Evaluationsforschung auch einer empirischen Analyse unterziehen, die auf eine **Deskription der Evaluationspraxis** hinausläuft (etwa anhand von Dokumentenanalysen, Interviews mit Evaluierenden, Beobachtungen der Evaluationspraxis etc.). Hierbei soll die Frage geklärt werden, welche Rolle die einzelnen theoretischen Ansätze überhaupt in der Evaluationspraxis spielen: Orientieren sich die Evaluationsforschenden in ihrer täglichen Arbeit wirklich an den Vorgaben eines Evaluationsmodells? Wie theoriekonform fällt die Evaluationspraxis in den einzelnen Ansätzen aus? Und sind Evaluationsstudien, die streng einem bestimmten theoretischen Evaluationsmodell folgen, besonders ertragreich? In den letzten Jahren mehren sich empirische Analysen, die die theoretischen Evaluationsmodelle und die Evaluationspraxis miteinander in Beziehung setzen (Hense, 2010; Christie, 2003). Dabei zeigt sich im Groben, dass **Evaluationsforschende oft eklektisch vorgehen**, d. h. Elemente aus unterschiedlichen Evaluationsansätzen miteinander kombinieren, anstatt nur einem bestimmten Modell streng zu folgen (Gesellschaft für Evaluation e. V. [DeGEval], 2008a). Ähnliche Befunde sind aus anderen Praxisfeldern bekannt. So zeigt die Psychotherapieforschung, dass trotz teilweise erbittertem Streit zwischen Vertretern unterschiedlicher theoretischer Schulen die Praktiker nicht selten Konzepte und Techniken aus verschiedenen Modellen (z. B. psychoanalytisch und kognitiv-verhaltenstheoretisch) flexibel kombinieren (Christie, 2007; Soellner, Hapkemeyer, & Scheibner, 2010; einen guten Einblick in Psychotherapieintegration in Forschungs und Praxis bietet auch die Zeitschrift *Journal of Psychotherapy Integration* der American Psychological Association).

Ein weiteres Ergebnis der Forschung zur **Theorie-Praxis-Diskrepanz in der Evaluation** (Kromrey, 2003) bezieht sich darauf, dass zuweilen gerade jene Aspekte, die ein bestimmtes theoretisches Evaluationsmodell charakterisieren, in der Praxis gar nicht prominent realisiert werden: So etwa, wenn ausgerechnet als Empowerment-Evaluationen etikettierte Studien die Zielgruppen einer Maßnahme kaum einbeziehen und deren Empowerment gar nicht erfassen (Spiel, Gradinger, & Lüftenegger, 2010) oder wenn theoriegeleitete Evaluationen bei der Formulierung von Evaluationsfragen und bei der Planung und Durchführung der Studie gar nicht erkennbar theoriegeleitet vorgehen (Johnson et al., 2009). Ein gewisses Spannungsverhältnis zwischen Theorie und Praxis eines Handlungsfeldes ist üblich und spricht nicht gegen die Bedeutung von Theoriebildung, sondern für eine fortwährende sorgfältige Reflexion, empirische Prüfung und Weiterentwicklung der Theoriemodelle – sowie auch der Praxistätigkeit. Denn auf welchem Wege eine festgestellte Theorie-Praxis-Diskrepanz sinnvoll aufzulösen ist, muss im Einzelfall begründet werden: Möglicherweise ist das jeweilige Evaluationsmodell zu realitätsfremd und muss der Evaluationspraxis angepasst werden, möglicher-

Vorschlag für einen umfassenden Theorierahmen der Evaluation

Welche Teiltheorien ein umfassender Theorierahmen der Evaluation gemäß Shadish, Cook und Leviton (1991, Kap. 2) idealerweise haben sollte, wird hier in der von Jan Hense (2013) verfassten Kurzform widergegeben (http://www.evoluation.de/node/view/151 [04.11.2013]):

1. **Theorie des sozialen Wandels** durch soziale Programme
   - **Problemstellung**: Wie können soziale Programme zum Lösen sozialer Probleme beitragen und wie kann man Programme dabei verbessern?
   - **Wissensbasis**:
     1. Beschreibung der internen Struktur von Programmen, ihren Funktionen und ihren Wirkweisen.
     2. Beschreibung der externen Kontexte, die Programme formen und einschränken.
     3. Beschreibung, wie sozialer Wandel allgemein zustande kommt, wie Programme Veränderungen herbeiführen und wie diese Veränderungen zum allgemeinen sozialen Wandel beitragen.
2. **Theorie der Wissensgenerierung** durch Evaluation
   - **Problemstellung:** Wie wird in Evaluationen Wissen generiert und was ist daran besonders?
   - **Wissensbasis:**
     1. Ontologische Aussagen zur Natur von „Wirklichkeit".
     2. Epistemologische Aussagen zu Natur, Quellen und Grenzen von Wissen.
     3. Methodologische Aussagen zu Verfahren und Methoden der Wissensgenerierung.
3. **Theorie von Werthaltungen** in sozialen Programmen
   - **Problemstellung**: Werthaltungen sind allgegenwärtig in sozialen Programmen. Wie können in Evaluationen diesbezügliche Probleme expliziert und offen behandelt werden, und wie können die Werteimplikationen von Programmen sensibel analysiert werden?
   - **Wissensbasis**:
     1. Metatheoretische Aussagen zu Natur und Rechtfertigung von Werthaltungen.
     2. Präskriptive Aussagen, die bestimmte Werthaltungen und Werte präferieren.
     3. Deskriptive Aussagen, die Werthaltungen beschreiben ohne dabei eine als die beste zu markieren.
4. **Theorie der Nutzung** von Evaluationsergebnissen
   - **Problemstellung**: Wie können Evaluationsergebnisse hervorgebracht werden, die zur Lösung sozialer Probleme beitragen?
   - **Wissensbasis**:
     1. Beschreibung möglicher Arten von Nutzung.
     2. Einlassungen zum zeitlichen Rahmen, in dem Nutzung auftritt.
     3. Erklärungen, wie Nutzung in unterschiedlichen Umständen durch Evaluatoren gefördert werden kann.
5. **Theorie der Evaluationspraxis**
   - **Problemstellung**: Praktizierende Evaluatoren benötigen pragmatische Konzepte als Orientierung und generelle Strategien bei ihrer Arbeit; zu deren Umsetzung brauchen sie praxistaugliche Methoden, die innerhalb der praktischen Beschränkungen funktionieren, unter denen sie arbeiten.
   - **Wissensbasis**:
     1. Aussagen dazu, wann evaluiert werden sollte.
     2. Aussagen dazu, was der Zweck einer Evaluation sein sollte.
     3. Aussagen dazu, welche Rollen Evaluatoren einnehmen sollten.
     4. Aussagen dazu, welche Art von Fragen gestellt werden sollten.
     5. Aussagen dazu, welches Design verwendet werden sollte.
     6. Aussagen dazu, was zur Förderung der Nutzung von Ergebnissen getan werden sollte.

weise müssen aber auch die Evaluationspraktiker besser geschult werden, um einen theoretischen Ansatz modellkonformer in ihrer Forschungstätigkeit umsetzen zu können.

Ein Grund für die Theorie-Praxis-Diskrepanzen in der Evaluationsforschung mag auch darin liegen, dass die einzelnen Evaluationsmodelle jeweils keine vollständigen Theorie der Evaluation bieten, sondern Einzelaspekte des Evaluationsgeschehens hervorheben, zu anderen Aspekten aber keine genauen Aussagen treffen. Welche Komponenten eine vollständige Evaluationstheorie idealerweise haben sollte, wird von Shadish, Cook und Leviton (1991) ausführlich erläutert (► Vorschlag für einen umfassenden Theorierahmen der Evaluation). Die Konstruktion eines solchen Theorierahmens steht allerdings noch aus.

18.2.1 Der ergebnisorientierte Evaluationsansatz

Der ergebnisorientierte Evaluationsansatz konzentriert sich auf einige wenige ergebnisbezogene Merkmale des Evaluationsgegenstandes. Der Fokus liegt also auf einer konkreten **kriteriumsbezogenen Fragestellung** (z. B. nach der Akzeptanz einer Maßnahme, nach der Usability eines technischen Gerätes). Diese wird im ergebnisorientierten Ansatz mit einzelnen **spezifischen Datenerhebungsmethoden** (z. B. Bewertung auf einer Schulnotenskala, Usabilitytest) beantwortet. Der ergebnisorientierte Ansatz wird in der Literatur auch als „questions-and methods-oriented evaluation" (Stufflebeam & Shinkfield, 2007) bezeichnet.

Es ist zu beachten, dass innerhalb jedes Evaluationsansatzes große Unterschiede in der Aussagekraft der Ergebnisse bestehen, je nachdem, welches Modell in welcher Weise umgesetzt wird. So kann eine ergebnisorientierte Evaluation summativ erfolgen (z. B. vergleichende Bewertung der Usability unterschiedlicher Zeitmanagement-Apps für Smartphones) oder formativ angelegt sein (etwa wenn Ergebnisse von Usabilitytests in mehreren Entwicklungszyklen zur Optimierung des Produktes genutzt werden). Auch hat ein positives Evaluationsergebnis im ergebnisorientierten Ansatz eine ganz unterschiedliche Geltung, je nachdem, ob es sich um eine One-Shot-Case-Study handelt (d. h. eine vorexperimentelle Studie ohne Kontrollgruppe und ohne Messwiederholung) oder ob Kontrollgruppen, Vorher-Nachher-Messungen oder gar Langzeitbeobachtungen des bzw. der Zielkriterien einbezogen werden. So mag es beeindruckend sein, wenn eine Evaluation anhand einer aussagekräftigen Stichprobe der Zielgruppe zeigt, dass 95 % der Nutzenden die evaluierte Maßnahme weiterempfehlen würden oder sie auf einer Schulnotenskala mit der Note 1.2 bewerten. Ob sie sie jedoch langfristig im Alltag einsetzen würden, ob eine Alternativmaßnahme nicht hinsichtlich Kosten-Nutzen-Bilanz viel vorteilhafter wäre oder ob wichtige Stakeholder, die andere Einschätzungen haben, unberücksichtigt geblieben sind, bleibt offen.

In dem großen Spektrum der Evaluationskriterien (► Abschn. 18.1.5) haben Effektivität und Effizienz eine herausgehobene Position. Als Varianten der ergebnisorientierten Evaluation werden deswegen im Folgenden wirksamkeits- und effizienzorientierte Ansätze genauer betrachtet.

Wirksamkeitsorientierte Evaluationsmodelle

Mit der **Effektivität** oder Wirksamkeit ist gemeint, ob und in welchem Ausmaß eine Intervention die angestrebten Zielsetzungen bzw. Wirkungen erreicht. Dabei unterscheiden wird die unmittelbare Leistung bzw. Produktivität einer Maßnahme **(Output),** die Wirksamkeit für die Mitglieder der Zielgruppe **(Outcome)** sowie die langfristigen Folgen auf überindividueller Ebene (**Impact**; ► Abschn. 18.1.5). In der Politikwissenschaft ist für die Wirksamkeitsevaluation die Bezeichnung **Impact Evaluation** gängig (Khandker, Koolwal, & Samad, 2009).

Je geringer die Differenz zwischen dem erreichten Zielzustand und dem angestrebten Ziel, umso höher die Wirksamkeit bzw. Effektivität einer Maßnahme.

Bei einer wirksamkeitsorientierten Evaluation wird zuweilen noch differenziert, ob es sich um die **grundsätzliche Wirksamkeit** handelt, die unter kontrollierten bzw. optimalen Bedingungen nachweisbar ist **(Efficacy)**, oder ob die **praktische Wirksamkeit** gemeint ist, die sich unter Realbedingungen im Feld zeigt **(Effectiveness)**. So lässt sich z. B. in der Medizin abgrenzen, ob ein Wirkstoff erfolgreich im Labor getestet wurde oder in klinischen Studien an echten Patientinnen und Patienten.

Der Nachweis, dass eine Maßnahme a) überhaupt die intendierten Wirkungen entfaltet (Nachweis eines **statistisch überzufälligen Effektes**, der kausal auf die Maßnahme zurückzuführen ist) und dies b) in einem praktisch bedeutsam starken Maße tut (praktisch bedeutsame **Effektgröße**; ► Abschn. 14.2), ist intuitiv nachvollziehbar ein zentrales Evaluationsanliegen, v. a. auch der summativen Evaluation. In der Wirksamkeitsevaluation spielen quantitative Methoden eine große Rolle. Die APA-Datenbank PsycInfo bietet zu „effectiveness evaluation“ als „Keyword“ 14 543 Artikel in referierten Fachzeitschriften, was die große Bedeutung dieses Ansatzes in der Psychologie unterstreicht (Stand: November 2013). Die meisten dieser Publikationen beziehen sich auf die Evaluation der Wirksamkeit von therapeutischen Interventionen sowie von Unterricht.

Es ist zu beachten, dass die Wirksamkeit nicht mit dem Wert einer Maßnahme gleichzusetzen ist, sondern nur eine Bewertungskomponente darstellt. In der Evaluationsforschung wird der intrinsische Wert eines Evaluationsgegenstands auch als dessen **Güte (Merit)** bezeichnet und von der zweckbezogenen Verwendbarkeit bzw. **Tauglichkeit (Worth)** sowie von der praktischen **Bedeutsamkeit** (Significance) abgegrenzt.

Beispiel: Ein computerbasiertes Training kann gemäß didaktischen Kriterien eine hohe Güte besitzen („merit“). Dennoch kann es in einem bestimmten Anwendungskontext mangelnde Tauglichkeit („worth“) aufweisen, etwa wenn das Training nicht auf die konkrete Zielgruppe abgestimmt ist (z. B. weil sich die Zielgruppe durch geringe Computeraffinität oder sogar Computerangst auszeichnet). Aber selbst wenn das computerbasierte Training hohe Güte und bei der intendierten Zielgruppe auch hohe Tauglichkeit zeigt, kann es trotzdem von geringer praktischer Bedeutsamkeit sein (z. B. wenn es inhaltlich ein Spezialthema behandelt, das nur in wenigen Kontexten relevant ist und/oder Gefahr läuft, bald veraltet zu sein; http://www.evoluation.de: Güte [04.11.2013]).

Für Hinweise zur korrekten Verwendung des ausdifferenzierten Fachvokabulars der Evaluationsforschung sei u. a. auf die bereits in ► Abschn. 18.1.11 erwähnten deutschsprachigen Onlinewörterbücher verwiesen (http://www.evoluation.de; http://www.eval-wiki.org/glossar).

Bei Wirksamkeitsevaluationen ist der wissenschaftliche **Nachweis der Kausalität** zentral: Eine Maßnahme kann nur Wirksamkeit für sich beanspruchen, wenn man belegen kann, dass die beobachteten positiven Effekte

nicht auch ohne die Maßnahme zustande gekommen wären. In der Grundlagenforschung gilt das Experiment als „Goldstandard" für einen wissenschaftlichen Kausalitätsnachweis (► Abschn. 7.6.1). Randomisierte Kontrollgruppenstudien spielen auch in der Wirksamkeitsevaluation eine wichtige Rolle, etwa bei Medikamententests oder bei der Überprüfung der Wirksamkeit von Lehr- oder Therapiemethoden. Relativ häufig kommen bei Wirkungsevaluationen auch quasi-experimentelle Feldstudien zum Einsatz. Die geringste Aussagekraft haben nicht- bzw. vorexperimentelle Studien, etwa die alleinige Messung abhängiger Variablen nach einer Maßnahme oder auch ein Vorher-Nachher-Vergleich ohne Kontrollgruppe.

In der Wirkungsevaluation wird im Zuge der Professionalisierung der Evaluationsforschung in jüngster Zeit auf einen Kausalitätsnachweis mit hoher interner Validität (► Abschn. 3.2.2) immer größerer Wert gelegt (► Wirksamkeit einer Heilstollenbehandlung: Radon oder Kurschatten?). Vor allem in der Entwicklungszusammenarbeit werden entsprechende Wirkungsevaluationen auch als **Rigorous Impact Evaluation** bezeichnet (Reade, 2008). Sollen etwa Programme zur Armutsbekämpfung in ländlichen Regionen Afrikas evaluiert werden, so ist für einen intern validen Kausalitätsnachweis zu fordern, dass eine Stichprobe vergleichbarer Dörfer gezogen wird und Experimental- und Kontrollgruppe randomisiert zugewiesen werden, so dass die Entwicklung der durch die Maßnahme geförderten Dörfer einerseits mit der ohne Maßnahme stattfindenden Entwicklung in vergleichbaren Dörfern andererseits systematisch kontrastiert werden kann. Gerade in Evaluationsfeldern, in denen es um die existenziellen Lebensbedingungen von Menschen geht, ist es aus ethischer Sicht wichtig, Kausalität genau zu prüfen. Andernfalls läuft man Gefahr, dass knappe Mittel immer wieder falsch eingesetzt werden, wirkungslose Programme stattfinden und somit vielen Menschen eine wirksame Unterstützung vorenthalten wird (Clemens & Demombynes, 2011). Randomisierte Kontrollgruppendesigns sind in der Entwicklungsforschung vielfältig anwendbar und durchaus nicht auf Laborexperimente beschränkt (White, 2013). Sie lassen sich auch mit qualitativen und partizipativen Evaluationsansätzen kombinieren (Karlan, 2009).

Bei der Konzeption einer wirksamkeitsorientierten Evaluationsstudie ist die **Abstimmung von Maßnahme und Wirkung** von zentraler Bedeutung. Wittmann (1988, 1990) macht zu Recht auf ein Problem aufmerksam, das häufig Ursache für das Misslingen von wirksamkeitsorientierten Evaluationsstudien ist: Die Maßnahme und die Erfassung ihrer Wirkung (d. h. die Evaluationskriterien, ► Abschn. 18.1.5) sind nicht genügend aufeinander abgestimmt bzw. nicht symmetrisch. **Symmetrie** wäre gegeben, wenn das Wirkkriterium (die abhängige Variable) genau das erfasst, worauf die Maßnahme (die unabhängige Variable) Einfluss nehmen soll.

Wenn z. B. eine neue Trainingsmethode Weitsprungleistungen verbessern soll, wäre die Sprungweite in Zentimetern nichts anderes als ein symmetrisches Kriterium. Bei komplexeren Maßnahmen kann es – insbesondere bei einer nachlässigen Zielexplikation – dazu kommen, dass die Wirkkriterien zu diffus sind bzw. „neben" dem eigentlichen Ziel der Maßnahme liegen, was die Chancen einer erfolgreichen Evaluation mindert. Wittmann (1990) unterscheidet in diesem Zusammenhang in Anlehnung an Brunswik (1955) vier Formen der Asymmetrie:

- **Fall 1:** Das Kriterium erfasst Sachverhalte, die von der Maßnahme nicht beeinflusst werden (Beispiel: Die Maßnahme „Rollstuhlgerechte Stadt" wird über den Tablettenkonsum der Rollstuhlfahrer evaluiert).
- **Fall 2:** Das Kriterium ist gegenüber einer breitgefächerten Maßnahme zu spezifisch (Beispiel: Eine Maßnahme zur Förderung des Breitensports wird über die Anzahl der Mitgliedschaften in Sportvereinen evaluiert).
- **Fall 3:** Das Kriterium ist gegenüber einer spezifischen Maßnahme zu breit angelegt (Beispiel: Ein gezieltes Training „Logisches Schlussfolgern" wird über allgemeine Intelligenzindikatoren evaluiert).
- **Fall 4:** Die Schnittmenge zwischen einer breitgefächerten Maßnahme und einem breitgefächerten Kriterium ist zu klein (Beispiel: Die Wirkung eines Grippemittels gegen Husten, Schnupfen, Heiserkeit, Fieber, Gelenkschmerzen etc. wird über einen kompletten Gesundheitsstatus inklusive Belastungs-EKG, Blutzucker, Cholesterinwerte, Blutsenkung, Harnprobe etc. geprüft).

Die bewusst etwas überzeichneten Beispiele für Asymmetrie sollen auf mögliche Fehlerquellen bei der Auswahl und Operationalisierung der Evaluations- bzw. Wirkkriterien aufmerksam machen. Planungsfehler dieser Art entstehen z. B., wenn das Evaluationsteam Kriterien messen will, die den Auftraggeber nicht interessieren (Fall 1), wenn die Auftraggeberinteressen sehr einseitig sind (Fall 2), wenn der Evaluator über das Ziel der Maßnahme hinausgehende Eigeninteressen hat (Fall 3) oder wenn der Auftraggeber (bzw. die Interventoren) und der Evaluator nicht genügend über den Sinn der Maßnahme kommuniziert haben und sicherheitshalber lieber zu viel als zu wenig beeinflussen bzw. kontrollieren wollen (Fall 4).

Die Vorgabe einer praktisch bedeutsamen **Effektgröße** (► Abschn. 14.2) fällt bei **einfachen Wirkhypothesen** mit einem eindeutig operationalisierten Wirkkriterium (z. B. Anzahl der Diktatfehler, der Krankheitstage, der Unfälle) leichter als bei einer komplexen **multivariaten**

Wirksamkeit einer Heilstollenbehandlung: Radon oder Kurschatten?

In einem Kurort befindet sich in einem stillgelegten Bergwerk ein radonhaltiger Heilstollen, der u. a. von Patientinnen und Patienten mit **Morbus Bechterew** oder anderen schweren rheumatischen Krankheiten aufgesucht wird. Ohne Frage hat diese ärztliche Maßnahme das Ziel, die Lebensqualität der Erkrankten zu steigern. Die ärztliche Leitung dieser Einrichtung beschließt, die Heilstollenbehandlung zu evaluieren und verteilt hierfür nach Abschluss der Kur Fragebögen an die Patienten, in denen nach dem Wohlbefinden im Allgemeinen und nach den rheumatischen Beschwerden im Besonderen gefragt wird. Diese als **Eingruppen-Posttest-Design** (One-Shot-Case-Study) durchgeführte Evaluation ist kritisierbar, denn das Ergebnis der Befragung ist keineswegs zwingend auf die ärztlichen Maßnahmen zurückzuführen. Die schwache interne Validität (► Abschn. 3.2.2) ergibt sich unmittelbar daraus, dass der Gesundheitszustand der Patienten vor der Behandlung nicht systematisch erhoben wurde, d. h., es fehlt ein Bezugsrahmen, aus dem heraus die ex post erhobenen Daten interpretiert werden können. Die externe Validität (► Abschn. 3.2.2) wäre kritisierbar, wenn die Stichprobe der untersuchten Patienten in Bezug auf bestimmte therapierelevante Merkmale verzerrt wäre (zu viele alte Patienten, zu wenig Kassenpatienten etc.). Die Aussagekraft der Dateninterpretation wird durch eine zusätzliche Befragung zu Befinden und Beschwerden vor der Behandlung deutlich erhöht. Aber auch dieses sog. **Eingruppen-Pretest-Posttest-Design** führt noch nicht zu der gewünschten Eindeutigkeit der Ergebnisse. Nehmen wir einmal an, die Befragung nach der Behandlung hätte statistisch signifikante und deutlich bessere Gesundheitswerte ergeben als die Befragung vor der Behandlung. Könnte man hieraus folgern, die radonhaltige Stollenkur sei die Ursache für das verbesserte Befinden? Die folgenden Alternativerklärungen belegen, dass auch dieser Untersuchungstyp noch immer keine zufriedenstellende interne Validität aufweist:

- Zu nennen wären zunächst weitere **zwischenzeitliche Einflüsse**, die unabhängig von der Stollenbehandlung wirksam gewesen sein können: Die Kur findet in einer landschaftlich schönen Gegend statt, man lernt nette Leute kennen, man findet Gelegenheit, sich mit anderen Patienten auszutauschen, die Ernährungsgewohnheiten werden umgestellt, man hat mehr körperliche Bewegung etc. Dies alles könnten auch Ursachen für die am Kurende registrierte Verbesserung des Gesundheitszustandes sein.
- Eine weitere Alternativerklärung ist mit der sog. **instrumentellen Reaktivität** gegeben. Durch den Pretest bzw. die Befragung vor der Behandlung kann eine gedankliche Auseinandersetzung mit dem eigenen Befinden angeregt werden, die bei einer erneuten Befragung zu einer anderen Einschätzung der eigenen Lebenssituation führt. Mögliche Unterschiede im Wohlbefinden vor und nach der Kur hätten in diesem Falle weder etwas mit den eigentlich interessierenden ärztlichen Maßnahmen noch mit den o. g. zeitlichen Einflüssen zu tun.
- Eine dritte Erklärung lässt sich aus dem untersuchten Merkmal selbst bzw. dessen **natürlicher Variabilität** ableiten. Geht man davon aus, dass das Befinden von Patienten mit Morbus Bechterew in starkem Maße durch die Intensität ihrer Schmerzen beeinträchtigt wird, ist zu bedenken, dass intensive Schmerzzustände bei dieser Krankheit schubweise auftreten, d. h., das Befinden dieser Patientinnen und Patienten unterliegt starken Schwankungen. Ein verbessertes Befinden nach der Kur könnte also schlicht auf die Tatsache zurückgeführt werden, dass die Betroffenen zu Beginn der Kur unter einem akuten Schub litten, der in der Zeit der Nachbefragung ohnehin, d. h. auch ohne ärztliche Behandlung, abgeklungen wäre (zu dem Effekt der Regression zur Mitte ► Abschn. 12.5.4).

Die genannten Probleme der internen Validität könnten durch eine Erweiterung der Studie um **mehrere Vortestmessungen** teilweise behoben werden. Denn wenn die abhängigen Variablen über längere Zeit wiederholt gemessen werden (wobei dann zwischenzeitliche Einflüsse, instrumentelle Reaktivität und natürliche Variabilität greifen), kann von diesen Schwankungen zwischen den Messzeitpunkten eine dezidierte Maßnahmenwirkung abgegrenzt werden. Nämlich wenn bei dem Messzeitpunkt nach der Maßnahme plötzlich deutlich abweichende Werte erscheinen, die über die sonstigen Schwankungen hinausgehen. Bei Designs mit einer einzigen Untersuchungsgruppe und mehr als einem Messzeitpunkt stellt die **Zeit die unabhängige Variable** dar.

Noch höher wird die interne Validität, wenn neben der Zeit noch die **Maßnahme als zweite unabhängige Variable** variiert, d. h. neben der zu evaluierenden Maßnahme auch mindestens eine Kontrollgruppe untersucht wird: **Prä-Post-Kontrollgruppen-Design**. Dazu müssten mindestens zwei hinsichtlich ihrer Erkrankung an Morbus Bechterew und sonstiger relevanter Hintergrundvariablen (z. B. Alter) vergleichbare Patientengruppen gebildet werden (z. B. durch Randomisierung oder Matching, ► Abschn. 7.6). Während die Untersuchungsgruppe die zu evaluierende Heilstollenbehandlung durchläuft, wird die Kontrollgruppe möglichst ähnlich behandelt, nur das Radon entfällt (z. B. Aufenthalt gleicher Länge mit vergleichbarer Betreuungsintensität in einem Luftkurort). Zeitliche Einflüsse während der Maßnahme (z. B. neue soziale Kontakte, Ernährungsumstellung), instrumentelle Reaktivität durch die Vorhermessung und natürliche Variabilität wirken in Untersuchungs- und Kontrollgruppe gleichermaßen. Zeigen sich im Vergleich jedoch bei der Untersuchungsgruppe deutlich stärkere Verbesserungen als in der Kontrollgruppe kann dies mit großer Sicherheit auf die Wirksamkeit speziell des Heilstollens bzw. des Radons zurückgeführt werden (hohe interne Validität).

Wirkhypothese, bei der die Wirksamkeit einer Maßnahme über mehrere Wirkkriterien zu erfassen ist. So sind z. B. die meisten psychotherapeutischen Interventionsprogramme nicht darauf ausgerichtet, nur ein einziges Symptom zu kurieren; die Kontrolle der Therapiewirkung erstreckt sich in der Regel auf verschiedene Störungen, deren Beseitigung oder Veränderung über jeweils spezifische Instrumente zu registrieren ist. Aber auch hier sollte man so gut wie möglich von Vorstellungen über eine „irgendwie" geartete positive Wirkung Abstand nehmen und sich bemühen, den erwarteten Wirkprozess in Bezug auf alle Wirkkriterien möglichst genau zu beschreiben. Ebenso wie in der Grundlagenforschung ist aufgrund des Teststärkeproblems die Wirksamkeit einer Maßnah-

me nicht allein an der statistischen Signifikanz, sondern an der Effektgröße und ihrer Interpretation hinsichtlich praktischer oder klinischer Bedeutsamkeit festzumachen (► Kap. 14).

Die Wirkung einer Maßnahme bezeichnen wir in der Tradition der quantitativen Sozialforschung als **abhängige Variable** und die Maßnahme selbst als **unabhängige Variable**. Um den Terminus „Variable" rechtfertigen zu können, muss die unabhängige Variable mindestens 2 Stufen aufweisen (z. B. Experimental- und Kontrollgruppe), wobei die zu evaluierende Maßnahme eine Stufe der unabhängigen Variablen darstellt. Den Ausprägungen der abhängigen Variablen für die einzelnen Stufen der unabhängigen Variablen (oder ggf. auch dem Zusammenhang zwischen unabhängiger und abhängiger Variable) ist dann die relative Wirkung der Maßnahme zu entnehmen. Man beachte, dass nicht-experimentelle Evaluationsstudien, bei denen die unabhängige „Variable" nur aus der zu bewertenden Maßnahme besteht, wenig aussagekräftig sind, weil die auf der abhängigen Variablen festgestellten Merkmalsausprägungen nicht zwingend auf die Maßnahmen zurückgeführt werden können (► Wirksamkeit einer Heilstollenbehandlung: Radon oder Kurschatten?).

Bevor man experimentell oder quasi-experimentell untersucht, **wie** sich bestimmte unabhängige Variablen auf die abhängigen Variablen auswirken, muss sichergestellt werden, **ob** die unabhängigen Variablen überhaupt in vorgesehener Weise in der Stichprobe realisiert sind. Angenommen es soll untersucht werden, ob Sparmaßnahmen in Schulen (Kürzung des Papieretats um 40 %) eine höhere Akzeptanz (abhängige Variable) erreichen, wenn sie mit ökonomischen oder mit ökologischen Argumenten begründet werden (unabhängige Variable). In zehn zufällig ausgewählten Schulen eines Bundeslandes werden zeitgleich mit der Etatkürzung Rundbriefe an die Lehrerkollegien verschickt, die entweder mit den Worten „Unsere Schule schützt die Umwelt …" oder „Unsere Schule spart Steuergelder …" beginnen.

Nachdem das Sparprogramm ein Jahr gelaufen ist, werden alle Schülerinnen und Schüler sowie alle Lehrkräfte um eine Einschätzung gebeten. Dabei stellt sich heraus, dass der Papiermangel auf das schärfste kritisiert und als Zumutung und Ärgernis eingestuft wird. Die Landesregierung nimmt dieses Ergebnis äußerst missbilligend zur Kenntnis. Auch die Presse spart nicht mit Schülerschelte und beklagt die Verantwortungslosigkeit der jungen Generation, die weder bereit sei, für das Gemeinwohl noch für die Umwelt Opfer zu bringen.

Erst als sich einige Schüler in einem Leserbrief an die Presse wenden und mitteilen, sie hätten den Sinn des Sparprogramms nicht verstanden, beginnt sich der Unmut gegen die Evaluatoren zu wenden, die es offensichtlich versäumt hatten nachzuprüfen, ob die intendierte künstliche Bedingungsvariation (Sparen für den Fiskus vs. Sparen für die Umwelt) in der Praxis umgesetzt wurde. Wie sich später herausstellte, wurde dies in der Tat versäumt, denn die ohnehin mit Umläufen und Rundbriefen überhäuften Lehrkräfte hatten die Mitteilung meist nur oberflächlich gelesen und erst recht nicht an ihre Schüler weitergegeben.

Wo die Intervention nicht greift, kann auch die Evaluation nichts ausrichten. Deswegen sollte man empirisch prüfen, ob die im Untersuchungsplan vorgesehenen Stufen der unabhängigen Variablen in der Praxis tatsächlich realisiert sind. Diese Kontrolle nennt man im grundlagenwissenschaftlichen Experiment meist Manipulation Check, in der Evaluationsforschung spricht man davon, die **Umsetzungskonformität** („compliance"), die **Programmtreue** („program fidelity", „program integrity") bzw. – das ist im Feld der Programmevaluation der gängigste Begriff – die **Implementierungstreue** („implementation fidelity") der Maßnahme im Zuge einer Implementationskontrolle zu prüfen (Heiner, 1996; Kirkhart, 2000; Wholey, 1979). Eine sorgfältige Implementierungskontrolle sorgt dafür, dass die Maßnahme nicht als völlige „Black Box" betrachtet wird, sondern die zentralen Abläufe und Aktivitäten zumindest in den wichtigsten Zügen bekannt und empirisch gesichert sind (für eine genauere Analyse der Wirkungsweise einer Maßnahme ist der theorieorientierte Ansatz zuständig, ► Abschn. 18.2.3).

Im Beispiel hätte man während der Maßnahme nachfragen müssen, was Schüler und Lehrer überhaupt von dem Sparprogramm und seiner Zielsetzung wissen. Bei Überprüfungen der Implementierungstreue (z. B. wiederholte Befragung der Maßnahmenbeteiligten, um die Qualität ihrer Arbeit zu bewerten) darf jedoch nicht übersehen werden, dass diese ihrerseits einen Eingriff in das Geschehen darstellen (z. B. **instrumentelle Reaktivität**) und dadurch den Charakter von Störvariablen haben können. Die Programmtreue ist nicht nur gefährdet, wenn – wie im obigen Beispiel – die Maßnahme unvollständig realisiert wird (z. B. Zielgruppen nicht ausreichend informiert oder nicht gut betreut werden) und deswegen nicht wirken kann. Zu Verzerrungen kommt es in der Praxis nicht selten aus dem umgekehrten Grund, dass eine evaluierte Maßnahme sozusagen „viel zu gut" implementiert wird, weil mit der Evaluation eine intensive, positiv empfundene Betreuung der Zielgruppe verbunden ist und somit überschießend positive Wirkungen auftreten. In der Grundlagenforschung ist diese Reaktion der Probanden auf die Tatsache des Untersuchtwerdens (im Unterschied zur Reaktion auf den interessierenden Stimulus) als **Reaktivität der experimentellen bzw. Untersuchungssituation** oder als „Hawthorne-Effekt" bekannt und stellt eine **Bedrohung der Konstruktvalidität** und damit der **internen Validität** dar (► Abschn. 3.2.2). In der Evaluationsfor-

schung hat man ebenfalls genau zu prüfen, inwiefern Evaluationstätigkeiten rund um die untersuchte Maßnahme es letztlich verhindern, dass der Effekt der puren Maßnahme feststellbar ist.

Beispiel: Die Wirksamkeit eines handygestützten Fitnessprogramms für Seniorinnen und Senioren soll geprüft werden. Das Programm stellt individuelle Trainingspläne zusammen, zeichnet anhand von Schrittzähler und GPS absolvierte Spaziergänge und Dauerläufe auf und gibt Rückmeldungen zum Trainingserfolg. Tatsächlich erweist sich das Programm als ausgesprochen erfolgreich, da die an der Studie teilnehmenden Hochbetagten regelmäßig mit Begeisterung das Programm nutzen und große Fortschritte erzielen. Doch nachhaltig ist dieser Erfolg leider nicht. Nach Abschluss der Evaluation stellen die meisten Teilnehmenden die Nutzung des Programms rasch wieder ein. Was ihre Verhaltensänderung bewirkt hatte, war nämlich gar nicht das Trainingsprogramm als solches gewesen. Vielmehr hatten die Besuche des Evaluationsteams, die eine willkommene Abwechslung in die Alltagsroutine brachten und Gelegenheit für interessante Gespräche boten, die Fitnessmotivation hervorgebracht.

Soweit realisierbar, haben quantitative Evaluationsstudien mindestens **zweifach gestufte unabhängige Variablen** mit einer Experimental- und einer **Kontrollgruppe** (► Evaluation der Wirksamkeit eines Computerlernspiels für Vorschulkinder zum Brandschutz). Wenn man nämlich die Maßnahme inklusive der Evaluationsbetreuung mit einer Gruppe vergleichbarer Personen kontrastiert, die nur die Evaluationsbetreuung erhält, kann man die Wirkung der Maßnahme von der Wirkung der Evaluation (und anderer Einflussfaktoren) abgrenzen. Alternativ zur Untersuchung einer Kontrollgruppe lässt sich die relative Wirkung einer Maßnahme auch über folgende Varianten abschätzen:

- Vergleich mehrerer Maßnahmen (z. B. vergleichende Evaluierung verschiedener Instruktionen zur Handhabung eines Computerprogramms),
- mehrfache Anwendung der Maßnahme (z. B. Spendenaufrufe zu verschiedenen Jahreszeiten, um kumulative Einflüsse auf die Hilfsbereitschaft zu erkennen),
- künstliche Variation der Intensität der Maßnahme (z. B. Wirkungsvergleich bei einem unterschiedlich dosierten Medikament),
- natürliche Variation der Intensität der Maßnahme (z. B. vergleichende Werbewirkungsanalysen bei Personen mit unterschiedlich häufigen Werbekontakten),
- Vergleiche mit Normen (z. B. Vergleich des Zigarettenkonsums nach einer Antirauchkampagne mit statistischen Durchschnittswerten).

Evaluation der Wirksamkeit eines Computerlernspiels für Vorschulkinder zum Brandschutz

Als Beispiel für eine wirksamkeitsorientierte Evaluation wird die Studie von Morrongiello, Schwebel, Bell, Stewart, & Davis (2012) steckbriefartig dargestellt.

Evaluationsziel: Die Wirksamkeit eines Computerlernspiels für Vorschulkinder zum Brandschutz soll geprüft werden. Das Evaluationskriterium der Wirksamkeit wird definiert als brandschutzbezogenes Handlungswissen der Vorschulkinder (3.5–6 Jahre).

Untersuchungsdesign: Randomisierte Vorher-Nachher-Kontrollgruppenstudie. Zwischen Vorher- und Nachher-Messung lagen drei Wochen, in denen das Spiel zu Hause genutzt werden konnte. Die Kontrollgruppe erhielt statt des Brandschutzspiels ein vergleichbares Computerspiel zur Prävention von Unfällen mit Hunden.

Stichprobe: Selbstselektionsstichprobe von Familien mit Vorschulkindern, die bislang noch nicht direkt von einem Brand betroffen waren und auch noch nicht an einem Brandschutzprogramm teilgenommen hatten.

Instrumente: Um den kindlichen Probanden gerecht zu werden und handlungsbezogenes Wissen angemessen zu erfassen, wurde anstelle einer Befragung ein Test durchgeführt, bei dem die Kinder das Verhalten der Personen in verschiedenen brandbezogenen Situationen mithilfe von Puppen in einer Puppenstube durchspielen sollten.

Untersuchungsdurchführung: Zusammen mit einem Elternteil besuchten die Kinder zweimal im Abstand von drei Wochen das Forschungslabor. Bei der Vorher-Messung und bei der Nachher-Messung führt das Kind den brandschutzbezogenen Puppenstubentest durch. Die Kinder wurden per Zufall der Experimental- oder der Kontrollgruppe zugeordnet. Im Anschluss an die Vorher-Messung erhielt jeweils das Elternteil jedes Experimentalgruppenkindes das Brandschutzcomputerspiel samt Instruktion, das Elternteil jedes Kontrollgruppenkindes ein Computerspiel zum Schutz vor Unfällen mit Hunden. Die Eltern sollten ein standardisiertes Tagebuch zur Nutzung des Computerspiels im häuslichen Umfeld führen.

Ergebnisse: Es zeigte sich, dass brandschutzbezogenes Handlungswissen in der Behandlungsgruppe statistisch signifikant und mit mittlerer Effektgröße zunahm ($\boldsymbol{d = 0.46}$), sich in der Kontrollgruppe dagegen nur schwach veränderte ($d = 0.11$). Dem Computerlernspiel ist also gemäß der Klassifikation von Cohen eine Wirkung mittlerer Stärke zuzuschreiben, was aufgrund der potenziell lebensrettenden Wirkung entsprechenden Wissensgewinns als praktisch sehr bedeutsam aufzufassen ist (zur Interpretation von Effektgrößen ► Abschn. 14.2).

Bewertung: Bei der Studie handelt es sich offenbar um einen methodisch sauberen Wirksamkeitsnachweis.

Limitationen der wirkungsorientierten Evaluation: Nicht-intendierte negative Wirkungen werden nicht untersucht (Anregung zum Spielen mit Feuer, Alpträume, Ängste) und es erfolgt keine systemische Betrachtung des Evaluationsgegenstandes (z. B. keine Analyse der Möglichkeiten und Grenzen der Verbreitung des Spiels in unterschiedlichen Kinderpopulationen).

Mit der **Operationalisierung der abhängigen Variablen** legen wir fest, wie die Wirkung der Maßnahme erfasst werden soll. Hierbei ist darauf zu achten, dass die Operationalisierung der Maßnahmenwirkung nicht nur für die Experimentalgruppe, sondern auch für die Kontrollgruppe oder ggf. weitere Gruppen mit anderen Vergleichsmaßnahmen sinnvoll angelegt sein muss.

Eine gelungene Operationalisierung setzt eine sorgfältige Explikation der Ziele voraus, die mit der Maßnahme angestrebt werden, und leitet daraus die Evaluationskriterien (► Abschn. 18.1.5) ab. In dieser wichtigen Planungsphase muss festgelegt werden, anhand welcher Daten die (relative) Maßnahmenwirkung erfasst werden soll und wie diese Daten zu erheben sind. Über die in den Human- und Sozialwissenschaften üblichen Datenerhebungstechniken wird ausführlich in Kap. 10 berichtet, so dass wir uns hier mit einer kurzen Aufzählung begnügen können:

- **Beobachtung** (Beispiel: Wie verändern sich Art und Anzahl der Unfälle in einem Betrieb nach Einführung einer neuen Arbeitsschutzkleidung?),
- **Interview** (Beispiel: Auf welche Probleme stoßen Lehrkräfte unterschiedlicher Schultypen und Fächer beim Einsatz von Computern im Unterricht?),
- **Fragebogen** (Beispiel: Wie ändern sich Umfrageergebnisse zum Thema „Wirtschaftliche Zukunftsperspektiven“ bei einem Regierungswechsel?),
- **Psychologischer Test** (Beispiel: Wie verändert ein kognitives Training bei Kindern deren Leistungen in einem kognitiven Fähigkeitstest? Wie stark kann eine musiktherapiegestützte Intervention die Symptome von Demenz – gemessen mit einem klinischen Demenztest – mildern?),
- **Physiologische Messung** (Beispiel: Wie stark sinkt der Kortisolspiegel nach einem Antistresstraining? Wie schnell und wie stark lässt sich erhöhter Blutdruck durch ein Sportprogramm senken?)
- **Dokumentenanalyse** (Beispiel: Wie verändert sich die Sprachkompetenz von Jugendlichen – erhoben anhand von Schulaufsätzen – im historischen Wandel?)

Effizienzorientierte Evaluationsmodelle

Lässt sich die Wirksamkeit einer Maßnahme kausal nachweisen, so ist ergänzend oft noch die Frage nach der Effizienz relevant: Wie **aufwändig** und **teuer** ist es, durch die Maßnahme den gewünschten Effekt zu generieren? Bei der Effizient werden also **Kosten und Nutzen** in Relation gesetzt. Im Bereich gesundheitsbezogener Interventionen ist der Nachweis der Wirksamkeit (klinische Evaluation) notwendig für die Zulassung der Therapie, während der Nachweis der Effizienz entscheidend ist für die Kostenübernahme durch die Krankenkassen (ökonomische Evaluation). Es geht bei der effizienzorientierten bzw. ökonomischen Evaluation somit um die **Wirtschaftlichkeit**: Wenn z. B. durch Kontrollgruppenstudien gezeigt werden kann, dass eine neue Unterrichtstechnik die schulischen Deutschleistungen im Durchschnitt um eine Note verbessert, dass durch die Einführung neuer Arbeitsschutzmaßnahmen die Zahl der Arbeitsunfälle um 10 % zurückgeht oder dass sich die Aufklärungsrate für Einbruchsdelikte nach einer kriminalpolizeilichen Schulungsmaßnahme um 3 % verbessert, so sind dies sicherlich Belege einer hohen Effektivität der jeweiligen Maßnahme. Aber sind die genannten Maßnahmen auch effizient, d. h. steht der zu betreibende Aufwand in vertretbarem Verhältnis zum Ergebnis? Oder sollten die notwendigen Mittel für die Maßnahme angesichts knapper Ressourcen nicht besser anderweitig (z. B. für eine alternative Maßnahme mit derselben Zielsetzung oder für Programme mit anderen Zielen) verwendet werden?

Eine Antwort auf derartige Fragen setzt voraus, dass man den **materiellen sowie ideellen Wert eines Effekts** näher beziffern kann, dass man also Vorstellungen darüber hat, wie viel der Unterschied um eine Einheit der Schulnotenskala, die Verhinderung eines Arbeitsunfalls oder die Aufklärung eines Einbruchs „wert“ sind. Das berechtigte Interesse der Auftraggebenden, vom Evaluationsteam zu erfahren, ob die evaluierte Maßnahme nicht nur effektiv, sondern auch effizient ist, stellt Evaluationsstudien vor das Problem, zum einen die Kosten einer Maßnahme umfassend zu ermitteln, zum anderen aber auch den materiellen und ideellen Nutzen der Maßnahme bzw. ihrer Wirkungen plausibel und nachvollziehbar zu quantifizieren. Nach der Bestimmung und Quantifizierung von Kosten und Nutzen müssen beide Größen zur Ableitung der Effizienz noch sinnvoll in Bezug zueinander gesetzt und interpretiert werden.

- **Kostenanalyse** („cost analysis“): Die Quantifizierung der Kosten von Evaluationsobjekten ist ein notwendiger Bestandteil jeder ökonomischen Evaluation und erfolgt in der Regel monetär, d. h. die Kosten werden in **Geldwerteinheiten** erfasst. Entsprechende Kostenmodelle erfordern – insbesondere bei größeren Maßnahmen und Programmen – betriebs- und volkswirtschaftliches Fachwissen zur **Kosten- und Leistungsrechnung**. Bei den Kosten handelt es sich um den Wert der für die Maßnahme verbrauchten Ressourcen (Weiß & Preuschoff, 2004). Dies umfasst neben den Ausgaben (Geldzahlungen für Beschaffung und Nutzung von Ressourcen) auch die Kosten für bereits verfügbare Ressourcen (z. B. bereits vorhandene Gebäude und eingestellte Mitarbeitende). Typische Kostenarten sind z. B. Planungskosten, Personalkosten, Kosten für Fremdleistungen, Material- bzw. Verbrauchskosten sowie Gebäude- und Gerätekosten bzw. Betriebsmittel. Nicht zu vergessen sind auch die **Op-**

portunitätskosten, also die Kosten, die durch die mit der Nutzung einer Maßnahme verbundene gleichzeitige Nichtnutzung anderer vorhandener Möglichkeiten entstehen (sog. Schatten- oder Alternativkosten). Beispiel: Die Kosten für ein Studium berechnen sich nicht nur über die Kosten für den Studienplatz über die gesamte Studienzeit hinweg (Räume, Personal, Material, Geräte), sondern im Sinne der Opportunitätskosten auch darüber, welcher Verdienstausfall während der Studienzeit entsteht, da die Studierenden zur selben Zeit nicht voll erwerbstätig sind (Weiß & Preuschoff, 2004). Dabei wird mit markt- oder branchenüblichen Durchschnittswerten (Durchschnittsverdienst einer Person entsprechender Altersgruppe und Qualifikation) kalkuliert. In der Praxis können Kostenanalysen z. B. dadurch erschwert sein, dass diese sensiblen Daten (z. B. Personalkosten, Entwicklungskosten einer neuen Technologie) nicht detailliert für das Evaluationsteam zugänglich sind – entsprechende Vereinbarungen zur Bereitstellung von Daten sind möglichst vor Beginn der Evaluationsarbeit zu treffen (z. B. Festlegungen im Evaluationsvertrag, ► Abschn. 18.3.2 „Abschluss eines Evaluationsvertrags“).

Sind die Kosten eines Evaluationsgegenstandes identifiziert und monetär erfasst, gilt es, auch den Nutzen zu messen und beide Größen in Relation zueinander zu bringen. Neben der Kosten-Nutzen-Analyse im engeren Sinne sind in der effizienzorientierten bzw. ökonomischen Evaluation auch Kosten-Effektivitäts-Analysen, Kosten-Nutzwert-Analysen sowie Kosten-Nutzen-Analysen verbreitet, die im Folgenden anhand vereinfachter Beispiele illustriert werden.

- **Kosten-Effektivitäts-Analyse** bzw. Kosten-Wirksamkeits-Analyse („cost effectiveness analysis“, **CEA**): Bei dieser einfachsten Form der Effizienzanalyse werden die im Rahmen der Kostenanalyse identifizierten anfallenden Kosten an einem **natürlichen Wirksamkeitsparameter** relativiert (z. B. Kosten pro erfolgreich behandelter Person, Kosten pro Veränderungseinheit; Gollwitzer & Jäger, 2009, S. 105). Angenommen drei medizinisch sinnvolle Methoden der Gewichtsreduktion für Patienten mit Adipositas sind hinsichtlich ihrer Effizienz zu vergleichen. Dazu werden zunächst die Kosten der jeweiligen Maßnahmen bestimmt. Die Kosten werden dann an der (z. B. binnen eines Vierteljahrs) erreichten Gewichtsabnahme in kg relativiert. Ist eine Maßnahme, die pro Person 1 000 Euro kostet und eine Gewichtsreduktion von 15 kg im Durchschnitt liefert, effizienter als eine Maßnahme, die 600 Euro kostet und eine Gewichtsreduktion von 8 kg erzeugt? Die Kosten-Effektivitäts-Analyse gibt Aufschluss: Bei der ersten Maßnahme kostet ein Kilogramm Gewichtsreduktion rund 67 Euro, bei der zweiten Maßnahme 75 Euro. Die CEA hat den Vorteil, dass als Nutzen einfach die in einem festgelegten Zeitraum erzielte Wirkung in einer natürlichen Einheit verwendet werden kann (z. B. Reduktion der Angst auf einer Angstskala, Reduktion des Alkoholkonsums in Liter, Verbesserung der Schulleistung auf der Notenskala etc.). Der Nachteil des Verfahrens liegt darin, dass die berechneten Kosteneffektivitäten nur vergleichbar sind zwischen Maßnahmen, deren Effekte in derselben Einheit gemessen wurden. Die Kosteneffektivität als solche muss zudem interpretiert werden, etwa vergleichend zwischen Alternativmaßnahmen oder in Anlehnung an ein vordefiniertes Kriterium (z. B. als effizient werden unter Berücksichtigung des vorhandenen Budgets nur Maßnahmen mit Maximalkosten von 50 Euro pro Person betrachtet, die den gewünschten Effekt erbringen).
- **Kosten-Nutzwert-Analyse** („cost utility analysis“, **CUA**): Bei dieser Form der Effizienzanalyse werden die Kosten an einem neu **konstruierten Nutzwertparameter** („utility“) relativiert. Dadurch ist es möglich, mehrere Effekte einer Maßnahme jeweils mit Prioritäten zu gewichten und zu einem Gesamtnutzwert zu vereinen. So wäre der Wert einer Gewichtsreduktionsmaßnahme für Patienten mit Adipositas nicht nur an den verlorenen Kilogramm Körpergewicht zu bemessen, sondern z. B. auch an dem Zugewinn an subjektivem Wohlbefinden oder an sozialen Kontakten. Gemäß der **multiattributiven Nutzentheorie** (MAUT) lässt sich der Gesamtnutzen bestimmen, indem man die einzelnen Effekte einer Maßnahme jeweils mit ihrem Nutzwert gewichtet und aufaddiert. Dabei ist zu beachten, dass die Zuschreibung von Nutzwerten sich je nach Stakeholder-Gruppe unterscheiden kann, da z. B. Patienten, Ärzte, Krankenkassen, Arbeitgeber, soziale Gemeinschaft etc. oft unterschiedliche Bewertungsprioritäten haben. Für die Operationalisierung und Integration von Nutzwerten gibt es unterschiedliche Ansätze. Um z. B. Nutzenwerte von Experten zu erheben und zu integrieren kann die **Delphi-Methode** eingesetzt werden (► Abschn. 10.3.3). Bei der Erfassung der Nutzenwerte von Betroffenen kann die Methode der **Planungszelle** (Dienel, 1978) weiterhelfen, bei der eine Auswahl der Betroffenen in mehreren Diskussionsrunden unter Anleitung erfahrener Moderatoren eine Gruppenentscheidung erarbeitet. Ausführliche Informationen zu der Frage, wie ein gemeinsames Zielsystem bzw. eine gemeinsame Nutzenfunktion über verschiedene Zielattribute generiert

werden kann, findet man in der Literatur zur **Entscheidungstheorie** (Jungermann, Pfister, & Fischer, 2005; Wright, 1993).

Bei der Evaluation von Objekten interessiert häufig die Frage, in welchem Ausmaß einzelne Objektmerkmale den Gesamtnutzen der Objekte beeinflussen. Ein Veranstalter von Seekreuzfahrten möchte z. B. in Erfahrung bringen, in welcher Weise die Merkmale

- Qualität der Reiseleitung (hochqualifiziert, akzeptabel, unqualifiziert),
- Preis (€ 3 000,-, € 4 000,-, € 5 000,-) und
- Kabinentyp (Innenkabine, Außenkabine)

die Akzeptanz von Kreuzfahrten beeinflussen.

Aus diesen 3 Merkmalen mit zweimal 3 und einmal 2 Ausprägungen lassen sich insgesamt $3 \times 3 \times 2 = 18$ verschiedene Kreuzfahrtvarianten kombinieren, die von potenziellen Kunden oder Experten in eine Rangreihe zu bringen (oder allgemein: zu bewerten) sind (zur Bildung von Rangreihen ▶ Abschn. 8.4.3). Aufgabe des **Conjoint Measurements** (auch Verbundmessung genannt) ist es nun, den $3 + 3 + 2 = 8$ Merkmalsausprägungen Teilnutzenwerte zuzuordnen, deren objektspezifische Addition zu Gesamtnutzenwerten führt, die die Rangwerte der Objekte bestmöglich reproduzieren (ausführlichere Informationen zum Conjoint Measurement findet man z. B. bei Backhaus, Erickson, Plinke, & Weiber, 1994; ausführlichere Informationen zum genannten Beispiel bei Tscheulin, 1991).

Es ist zu beachten, dass zwischen der Intensität der Wirkung und der Ausprägung des Nutzwertes unterschiedliche Beziehungen bzw. **Nutzenfunktionen** bestehen können. Im einfachsten Fall besteht ein monotoner oder linearer Zusammenhang, d. h. mit zunehmender Wirkung steigt auch der Nutzen (z. B. Anzahl geheilter Patienten). Der Zusammenhang kann aber auch nichtmonoton bzw. nichtlinear – z. B. umgekehrt u-förmig (d. h. quadratisch) – sein, etwa wenn sehr geringe und sehr große Effekte beide wenig Nutzen, mittlere Effekte jedoch hohen Nutzen haben (z. B. Gewichtsreduktion durch ein Programm für Patienten mit Adipositas; Gollwitzer & Jäger, 2009, S. 102). Von besonderem Vorteil sind Nutzenfunktionen, wenn eine Maßnahmenwirkung sinnvollerweise nur multivariat operationalisiert werden kann und die einzelnen Kriterien sowohl positive als auch negative Wirkungen erfassen (Beispiel: Ein gutes Konditionstraining baut Kalorien ab, stärkt die Muskeln, stabilisiert den Kreislauf, es ist jedoch gleichzeitig auch arbeits- und zeitaufwändig). In diesen Fällen wird für jedes mit dem Ziel der Maßnahme verbundene Kriterium eine Nutzenfunktion definiert, deren Kombination über den gesamten Nutzen der Maßnahme informiert (sog. multiattributive Nutzenfunktion; Eisenführ & Weber, 1993; Jungermann, et al., 2005, Kap. 4).

In verschiedenen Forschungsbereichen haben sich zudem bestimmte, theoretisch eingeführte Nutzwertparameter etabliert. So wird der Nutzen von Maßnahmen im Gesundheitsbereich im Rahmen gesundheitsökonomischer Evaluationen oft mit dem Parameter **qualitätskorrigiertes Lebensjahr** („quality-adjusted life year“, **QALY**) gefasst, in den als multiplikativer Index (▶ Abschn. 8.7) eingeht, wie sich eine Maßnahme auf die **Lebenszeit** (LZ; in Jahren) sowie die **Lebensqualität** (LQ; Faktor 1: perfekte Gesundheit bis Faktor 0: Tod) auswirkt. Eine Maßnahme mit dem Nutzwert 1 QALY verlängert also die Lebenszeit um ein Jahr bei voller Gesundheit, während ein Nutzwert von 0.5 QALY besagt, dass die Lebenszeit um ein Jahr bei stark eingeschränkter Gesundheit (z. B. Bettlägerigkeit) verlängert wird ($1\ \text{LZ} \times 0.5\ \text{LQ} = 0.5$ QALY) bzw. um ein halbes Jahr bei voller Gesundheit ($0.5\ \text{LZ} \times 1\ \text{LQ} = 0.5$ QALY). Angenommen einem neuen Medikament zur Behandlung von Lungenkrebs lässt sich ein Nutzwert von 0.5 QALY zuschreiben. Im Rahmen der Kosten-Nutzwert-Analyse wird nun ermittelt, in welchem Verhältnis die Kosten für das Medikament pro Person zum Nutzwert stehen bzw. wie viel 1 QALY kostet. Wenn die Behandlung mit dem Lungenkrebsmedikament 5 000 Euro pro Patient kostet für 0.5 QALY, entstehen somit rechnerisch Kosten in Höhe von 10 000 Euro pro Patient für 1 QALY.

Kosten pro QALY stellen eine Entscheidungsgrundlage im Gesundheitssystem dar, wobei die anzulegenden **Schwellenwerte** kontrovers diskutiert werden: Welche Kosten pro QALY sollte die soziale Gemeinschaft tragen, wenn man berücksichtigt, dass angesichts knapper Ressourcen die Finanzierung einer sehr teuren Maßnahme gleichzeitig bedeutet, dass andere Maßnahmen nicht realisierbar sind und somit die QALY anderer Patientengruppen indirekt reduziert wird. Zudem ist zu betonen, dass der QALY als Nutzwert nicht den Gesamtnutzen einer Maßnahme abdeckt, z. B. die **soziale Bedeutung** einer Maßnahme ignoriert. So kann mit palliativmedizinischen Maßnahmen meist weder die Lebenszeit noch die Lebensqualität deutlich gesteigert werden, trotzdem messen wir der medizinischen Betreuung von Sterbenden nicht zuletzt aus ethischen Gründen große Bedeutung bei, so dass Kosten pro QALY nicht die einzige Entscheidungsgrundlage bilden dürfen. Im Vergleich zur einfachen Kosteneffektivität gemessen in natürlichen Wirksamkeitsparametern (z. B. eine Knieoperation mindert die Schmerzen auf einer Schmerzskala um 10 Punkte für 5 000 Euro; eine Augenoperation bessert die Sehstärke um eine Stufe für 500 Eu-

ro) hat die Verwendung von Nutzwerten den Vorteil, dass Kosteneffektivitäten über unterschiedliche Maßnahmen direkt **vergleichbar** sind. Somit kann dann rechnerisch angegeben werden, welche Kosten pro QALY jeweils durch eine bestimmte Knie- oder Augenoperation, einen Herzschrittmacher, eine Physio- oder Chemotherapie anfallen.

- **Kosten-Nutzen-Analyse** („cost benefit analysis“, **CBA**): Bei dieser theoretisch anspruchsvollsten Form der Effizienzanalyse werden die Kosten an dem ebenfalls **monetär gemessenen Nutzen** relativiert. Voraussetzung ist, dass für alle Maßnahmenwirkungen jeweils der Nutzen erfasst und sinnvoll in Geldwerteinheiten umgerechnet werden kann. Dies ist theoretisch und methodisch besonders anspruchsvoll bei Wirkungen, für die keine marktüblichen Preise bekannt sind (z. B. monetärer Wert eines Menschenlebens, einer vom Aussterben bedrohten Tierart oder einer guten Allgemeinbildung der Bevölkerung). Für eine Kosten-Nutzen-Analyse des Hochschulstudiums im Rahmen bildungsökonomischer Evaluationen müsste z. B. – je nach Fragestellung der Evaluation – auf der Nutzenseite nicht nur erfasst werden, welche Einkommensgewinne durch die verbesserte Qualifikation im weiteren Erwerbsleben erzielt werden, sondern auch welcher weitere Nutzen sich durch die verbesserte Bildung für die Individuen (z. B. Gesundheitsverhalten) und die Gesellschaft (z. B. wirtschaftliche Entwicklung, politische Partizipation) ergibt.
  Eine Möglichkeit, den monetären Nutzen eines bestimmten Effekts zu operationalisieren besteht in der Erfassung der **Zahlungsbereitschaft** („willingness to pay“, **WTP**). Man fragt dementsprechend die Stakeholder, wie viel Geld zu bezahlen sie bereit wären, um einen bestimmten positiven Effekt zu erhalten (z. B. Reduktion von Rückenschmerzen, Zugewinn an Sprachkompetenz, Verbesserung der Wasserqualität des örtlichen Sees). Umgekehrt kann auch erfasst werden, wie viel Geld man ihnen zahlen müsste, damit sie einen negativen Effekt – also z. B. gesundheitliche Einschränkungen oder Umweltbelastungen – akzeptieren („willigness to accept“, **WTA**).
  Die Grenzen des Ansatzes bestehen u. a. darin, dass die WTP **einkommensabhängig** ist und dass die subjektive Zahlungsbereitschaft wiederum je nach Stakeholder-Gruppe stark variieren dürfte. Sollen die Kosten einer Umweltschutzmaßnahme zur Verbesserung der Wasserqualität eines Sees im Rahmen einer umweltökonomischen Evaluation am monetären Nutzen relativiert werden, so ergeben sich für die Kosten-Nutzen-Analyse unterschiedliche Nutzenwerte, je nachdem, ob man die Zahlungsbereitschaft für gute Wasserqualität bei den lokalen Familien mit Kleinkindern, bei Schwimmern, bei Anglern oder bei Arbeitssuchenden erfasst. Neben der Zahlungsbereitschaft von Laien dürfte zur Einschätzung des ökologischen und volkswirtschaftlichen Wertes der Wasserqualität auch das **Urteil von Fachleuten** (z. B. Umweltexperten) wichtig sein.
  Die Ermittlung des finanziellen Wertes einer staatlichen oder betrieblichen Maßnahme zählt zu den Aufgaben der Volks- und Betriebswirtschaftslehre. Wenn z. B. gefragt wird, ob der Staat mehr Geld in die Kinder-, Jugend- oder Erwachsenenbildung investieren sollte, wäre eine Antwort auf diese Frage v. a. von einer **volkswirtschaftlichen Kosten-Nutzen-Analyse** der alternativen Maßnahmen abhängig zu machen. Ähnliches gilt z. B. für Maßnahmen im Gesundheitswesen (z. B. Methadonprogramm), im Strafvollzug (mehr psychologische Betreuung?) oder im Verkehrswesen (mehr Subventionen für den Güterverkehr auf Straße, Wasser oder Schiene?). Innerbetriebliche Maßnahmen wie z. B. eine neue Arbeitszeitenregelung, höhere Investitionen in Forschung und Entwicklung oder die Umstellung eines Produktionszweiges auf Automatisierung, sind ohne **betriebswirtschaftliche Gewinn-Verlust-Rechnungen** bzw. ohne betriebswirtschaftliches Controlling nicht effizienzorientiert evaluierbar.
  Eine ökonomische Evaluation, die eine „harte“ Kosten-Nutzen-Analyse beinhaltet, kann – ebenso wie eine Wirksamkeitsüberprüfung oder jede andere Form der Evaluierung – bei den Maßnahmenverantwortlichen mitunter starke Ängste wecken. Schließlich könnte sich die Maßnahme als viel zu kostspielig bzw. ineffizient erweisen und auf der Basis dieser Bilanz nicht mehr legitimieren lassen. Es existieren jedoch Methoden einer **Ex-ante Kosten-Nutzen-Analyse**, mit deren Hilfe sich vorab errechnen lässt, welche Effektgröße eine Maßnahme mindestens haben muss, damit sich die Investitionen lohnen bzw. der monetäre Nutzen die Kosten übersteigt (Wittmann, Nübling, & Schmidt, 2002, S. 46ff.). Dabei scheint der monetäre Nutzen von Interventionen im Gesundheitsbereich häufig unterschätzt zu werden, da ab mittleren Effektgrößen entsprechende Maßnahmen offenbar eine gute Kosten-Nutzen-Bilanz aufweisen.

Die ökonomische Evaluation erfordert neben sozialwissenschaftlichen v. a. wirtschaftswissenschaftliche Methoden und ist inhaltlich stark ausdifferenziert. Große Bedeutung haben u. a. **bildungsökonomische Evaluationen** („education economics evaluation“; Weiß & Preuschoff, 2004), **gesundheitsökonomische Evaluationen** („health economics evaluation“; z. B. Drummond, Sculpher, Torrance, & O’Brien, 2005; Schöffski & Schulenburg, 2000)

Zur ökonomischen Evaluation von Selbsthilfegruppen

Auf der Basis eines Übersichtsartikels zum ökonomischen Nutzen von Selbsthilfegruppen (Engelhardt, Trojan, & Nickel, 2009) wird im Folgenden die effizienzorientierte Evaluation am Beispiel von Kosten-Nutzen-Analysen kursorisch vorgestellt. Selbsthilfegruppen können positive Wirkungen auf drei Ebenen erzielen: 1. für die Teilnehmenden selbst (z. B. emotionale, informationelle und praktische soziale Unterstützung), 2. für Angehörige und andere Betroffene (z. B. Beratung, Vermittlung von Hilfen) sowie 3. für die Gesellschaft insgesamt (z. B. Aufklärung über Krankheitsbilder und soziale Probleme). Die Wirksamkeit von Selbsthilfegruppen, v. a. für die Teilnehmenden selbst, ist gut belegt.

Somit scheint es gerechtfertigt, dass Selbsthilfeeinrichtungen staatlich gefördert werden, dass z. B. ein kommunales Selbsthilfezentrum von der Stadt regelmäßig Zuschüsse erhält, um die Kosten für Räumlichkeiten, Telefon, Personal etc. zu decken. Angesichts staatlicher Förderung kann die Frage aufkommen, ob es sich für die Gemeinschaft nachweislich lohnt, Selbsthilfeinitiativen mit knappen öffentlichen Mitteln zu fördern. Damit ist die Frage nach der Effizienz aufgeworfen, die sich durch ökonomische Evaluationsstudien beantworten lässt.

Soll die Effizienz durch eine Kosten-Nutzen-Analyse bewertet werden, müssen zunächst die Kosten erfasst werden. Ist die Stadt die Auftraggeberin der Evaluationsstudie und will die Frage beantwortet wissen, ob ihre Förderung von kommunalen Selbsthilfeinitiativen effizient ist, so bildet das Förderbudget direkt die Kosten ab. Wie aber lässt sich der Nutzen der vielfältigen Wirkungen der vor Ort geförderten Selbsthilfeinitiativen in Geldwerteinheiten überführen? Eine Möglichkeit besteht darin, den Nutzen für die Kosten-Nutzen-Analyse über die Zahlungsbereitschaft zu operationalisieren. Dazu wäre zu ermitteln, wie viel Geld relevante Stakeholder zu zahlen bereit wären, um den Erhalt der Selbsthilfeeinrichtung für ein weiteres Jahr zu gewährleisten. Eine entsprechende ökonomische Evaluation wurde für die Göttinger Kontakt- und Informationsstelle (KIBIS; http://www.selbsthilfe-goettingen.de) durchgeführt (Wilkens, 2002). Dazu wurde jeweils eine Stichprobe der Einwohner/innen von Göttingen sowie eine Stichprobe der Selbsthilfegruppenteilnehmenden zu ihrer **Zahlungsbereitschaft** befragt. Die Angaben wurden dann auf die Gesamtbevölkerung Göttingens (Zahlungsbereitschaft: 3.41 Mio DM) sowie die Gesamtheit der Selbsthilfegruppenteilnehmenden (Zahlungsbereitschaft: 131 175 DM) hochgerechnet. Es ergab sich rechnerisch eine Gesamt-Zahlungsbereitschaft der relevanten Stakeholder von rund **1.77 Mio Euro** (bzw. damals 3.54 Mio. DM) pro Jahr. Dem so ermittelten monetären Nutzen der KIBIS standen die monetären Kosten in Höhe der städtischen Fördersumme von rund **70 000 Euro** (bzw. 142 500 DM) im Jahr 2000 gegenüber. Es wurde also der mehr als 25-fache monetäre Nutzen im Verhältnis zu den Kosten generiert, was als sehr hohe Effizienz einzustufen ist.

Ein anderer Ansatz besteht darin, den Nutzen der geförderten Selbsthilfegruppen für die Kosten-Nutzen-Analyse nicht anhand des Outcomes (soziale Unterstützung, öffentliche Aufklärung etc.), sondern anhand des **Outputs** (geleistete Arbeit in den Selbsthilfegruppen durch aktive Mitwirkung der Teilnehmenden) zu berechnen. Hierfür wird erfasst, wie viele Selbsthilfegruppen das geförderte Zentrum nutzen, wie oft sie sich treffen, wie viele Teilnehmende mitwirken etc. In der Summe lassen sich als Output-Maß die von den Selbsthilfegruppenteilnehmenden unentgeltlich geleisteten Arbeitsstunden angeben. Diese wiederum sind anhand von Durchschnittslöhnen in Geldwerte umzurechnen. Eine entsprechende Kosten-Nutzen-Analyse der Selbsthilfeinitiativen in München kam zu dem Ergebnis, dass den monetären Kosten in Höhe eines städtischen Zuschusses von **50 Euro** (bzw. damals 100 DM) als monetärer Nutzen der Wert der unbezahlten, auf das Problem bezogenen Arbeit in Höhe von rund **165 Euro** (bzw. damals 331 DM) gegenüber stand (Kandler, 1995). Es wurde nach diesem Berechnungsansatz also der dreifache monetäre Nutzen generiert, was als hohe Effizienz einzuordnen ist.

und neuerdings verstärkt auch **umweltökonomische Evaluationen** („ecological economics evaluation"). Ein Studienbeispiel zur ökonomischen Evaluation von Selbsthilfegruppen ist in ▶ Zur ökonomischen Evaluation von Selbsthilfegruppen zu finden. Da von den Ergebnissen ökonomischer wie sonstiger Evaluationsstudien weitreichende politische Entscheidungen abhängig sein können, muss das Evaluationsteam die wissenschaftliche Unabhängigkeit gegenüber Anspruchsgruppen verteidigen, die bestimmte Wunschergebnisse sehen möchten. Teilweise werden spezifische Evaluationsstandards für Teilbereiche der ökonomischen Evaluation vorgeschlagen (z. B. Guidelines zur gesundheitsökonomischen Evaluation: Walter & Zehetmayr, 2006). In der APA-Datenbank PsycInfo finden sich zu „efficiency evaluation" sowie „economic evaluation" als „Keywords" insgesamt 305 Artikel in referierten Fachzeitschriften (Stand: November 2013).

18.2.2 Der systemische Evaluationsansatz

In Abgrenzung zu Evaluationen, die sich auf einzelne Merkmale einer Maßnahme – v. a. Ergebnismerkmale der Effektivität und Effizienz – konzentrieren, sind unter dem systemischen Ansatz hier solche Evaluationsmodelle angesprochen, die eine Maßnahme ganzheitlicher betrachten: Der systemische Ansatz untersucht den Evaluationsgegenstand zum einen im Hinblick auf **Prozesse** seiner Konzeption, Umsetzung, Nutzung und Wirkung und zum anderen auch im Hinblick auf die **Umfeldbedingungen**, etwa situative, kulturelle, historische oder sonstige Kontextfaktoren.

Ein Vertreter des systemischen Ansatzes ist z. B. das **Multi-Facetten-Modell der Evaluation** („multifaceted treatment evaluation model"; Nelson & Steele, 2006), das neben Effektivität (als Outcome Evaluation) und Effizienz (als Economic Evaluation) speziell die prospektiven

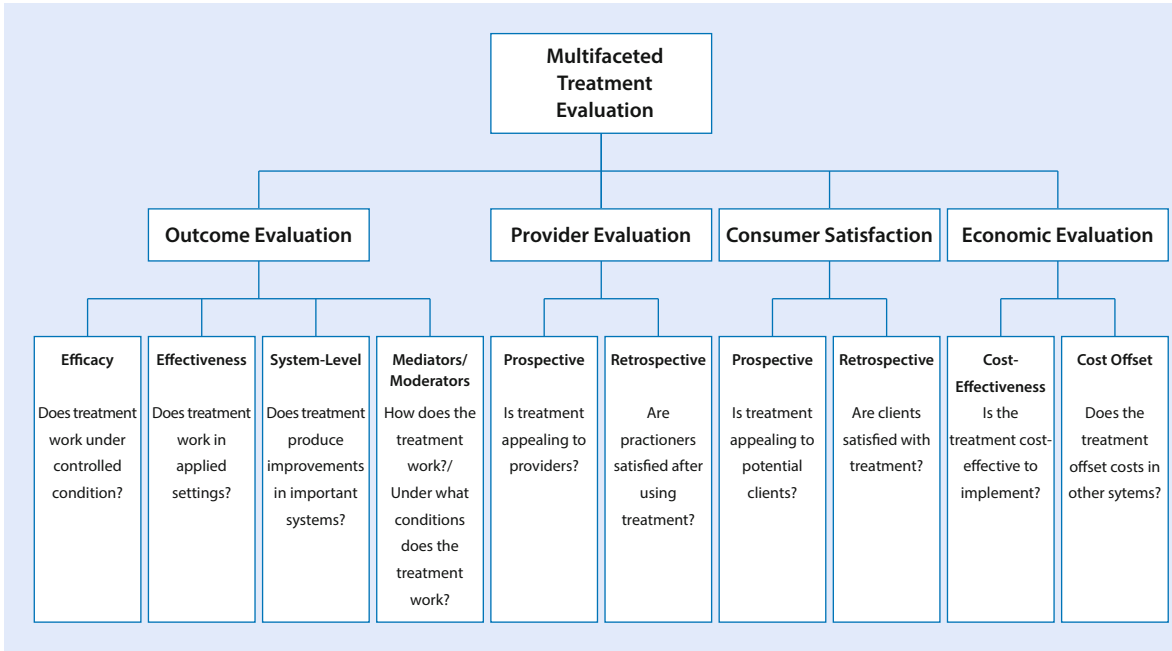


Abb. 18.4 Das Multi-Facetten-Modell der Treatmentevaluation. (Aus Nelson & Steele, 2006, Figure 1. Copyright © 2014 by the American Psychological Association. Adapted with permission. The use of APA information does not imply endorsement by APA)

und retrospektiven Bewertungen der Maßnahmenanbieter sowie Maßnahmezielgruppen hinsichtlich Akzeptanz und Zufriedenheit einbezieht (Abb. 18.4).

Das prominenteste Evaluationsmodell im systemischen Ansatz ist das bereits in den 1960er-Jahren entwickelte **CIPP-Evaluationsmodell** von **Daniel Stufflebeam** (Stufflebeam et al., 1971), das aus den vier Komponenten **C**ontext-, **I**nput-, **P**rocess- und **P**roduct-Evaluation besteht und von ihm selbst als „social systems approach to evaluation" (Stufflebeam & Shinkfield, 2007, S. 347) sowie als „improvement/accountability approach" (Stufflebeam & Shinkfield, 2007, S. 325) eingeordnet wird. Die ganzheitliche Perspektive soll dazu dienen, möglichst praxistaugliche Evaluationsergebnisse abzuleiten. Im CIPP-Modell wird mit vielfältigen qualitativen und quantitativen Datenerhebungsmethoden gearbeitet und in den einzelnen Teilkomponenten werden diverse Evaluationskriterien angelegt. Zielsetzung einer wissenschaftlichen Evaluation nach dem CIPP-Modell ist es, anhand der Evaluationsergebnisse eindeutige und objektivierbare Bewertungen und Empfehlungen zu einer Maßnahme abzugeben, die auf **explizierten Werten** basieren (z. B. Menschenrechte, Verfassung). Eine relativistische Position, nach der die Maßnahme je nach den Wirklichkeitskonstruktionen unterschiedlicher Stakeholder oder unterschiedlicher Evaluationsteams sowohl als wertvoll als auch wertlos gelten kann, wird abgelehnt (Stufflebeam & Shinkfield, 2007, S. 331).

Die Definitionen, Methoden und Zielsetzungen der 4 Teilkomponenten des CIPP-Modells sind in Tab. 18.1 kursorisch zusammengefasst. Zur Durchführung von Evaluationen nach dem CIPP-Modell liegt eine umfassende Evaluationscheckliste von Daniel Stufflebeam vor (http://www.wmich.edu/evalctr/checklists). Ein Studienbeispiel zur Umsetzung des CIPP-Modells ist ▶ Studienbeispiel „CIPP-Evaluation der Deutschen Schülerakademie" zu entnehmen.

Die APA-Literaturdatenbank PsycINFO weist 9 referierte Journal-Artikel für das Keyword „evaluation" und den Suchbegriff „CIPP" im Volltext nach (Stand: November 2013). Es ist zu beachten, dass eine wissenschaftliche CIPP-Evaluation nicht zwingend alle vier Teilevaluationen umfassend durchlaufen muss, sondern sich – je nach Evaluationsauftrag – auch auf einzelne Teil-Evaluationen konzentrieren kann (für ein entsprechendes Beispiel einer CIPP-Evaluation mit Fokus auf Produktevaluation und die Kriterien Wirksamkeit und Nachhaltigkeit Stufflebeam & Shinkfield, 2007, S. 352ff.).

18.2.3 Der theorieorientierte Evaluationsansatz

Bei der sog. theorieorientierten oder theoriebasierten Evaluation wird über die Frage der reinen Kausalität (die in der wirksamkeitsorientierten Evaluation untersucht wird) hinausgegangen (Giel, 2013): Man versucht, die Wir-

Tabelle 18.1 Überblick über die vier Komponenten des CIPP-Modells. (Mod. nach Stufflebeam & Shinkfield, 2007, S. 328, S. 335. This material is reproduced with permission of John Wiley & Sons, Inc.)

	Context-Evaluation	Input-Evaluation	Process-Evaluation	Product-Evaluation
Definition	Analyse der Kontextbedingungen des Evaluationsgegenstandes bzw. der Maßnahme, v. a. der **Zielgruppen und ihrer Bedürfnisse**, aber auch anderer Stakeholder und ihrer Anforderungen	Analyse der **Lösungsstrategie der Maßnahme** (inklusive möglicher Alternativmaßnahmen) sowie der **vorhandenen Ressourcen**, mit denen eine bestimmte Veränderung im Feld erreicht werden soll	Analyse der **Prozesse im Rahmen der tatsächlichen Umsetzung der Maßnahme**, etwa Umfang und wahrgenommene Qualität der Maßnahmennutzung durch die Zielgruppen, Umsetzungsprobleme aus Sicht des Personals etc.	Analyse von Umfang und Art der **Wirkungen der Maßnahme**, was sowohl intendierte Wirkungen als auch unerwartete positive wie negative Wirkungen umfasst
Methoden (Auswahl)	**Bedarfsanalyse** durch schriftliche Befragung, Fokusgruppendiskussionen, Experteninterviews, Dokumentenanalyse etc.	**Analyse der Maßnahmenstrategie und der vorhanden Ressourcen** durch Beobachtung, Dokumentenanalyse, Experteninterviews; Benchmarking: Lernen von erfolgreichen Vergleichsprojekten etc.	**Monitoring der Maßnahmenumsetzung** mit qualitativen und quantitativen Indikatoren, Befragung, Beobachtung, nonreaktive Datenerhebung etc.	**Erfassung der Maßnahmenwirkungen** durch Beobachtung, Befragung, Tests, nonreaktive Datenerhebung, physiologische Messungen etc. in experimentellen, quasi-experimentellen und nicht-experimentellen Designs
Zielsetzung bei formativer Evaluation	Entscheidung darüber, welche Ziele mit welcher Priorität eine geplante Maßnahme verfolgen soll bzw. ob und wie Zielsetzungen einer laufenden Maßnahme zu verändern sind	Entscheidung darüber, welche konkrete Maßnahme zur Zielerreichung zu wählen ist (Realisierbarkeit, Erfolgsaussichten) bzw. ob und wie ihr Konzept im Rahmen einer laufenden Maßnahme zu verändern ist	Entscheidung darüber, wie die Umsetzung der Maßnahme zu gestalten ist bzw. ob und wie sie im Rahmen einer laufenden Maßnahme zu verändern ist	Entscheidung darüber, ob und wie eine laufende Maßnahme hinsichtlich ihrer Wirkungen modifiziert, fortgesetzt oder beendet werden soll
Zielsetzung bei summativer Evaluation	Bewertung einer durchgeführten Maßnahme dahingehend, wie gut ihre Zielsetzung auf den Bedarf der Zielgruppe(n) sowie die Anforderungen anderer Stakeholder zugeschnitten ist	Bewertung einer durchgeführten Maßnahme dahingehend, wie gut ihr Ansatz im konkreten Fall zur Problemlösung geeignet ist, v. a. im Vergleich zu möglichen Alternativmaßnahmen mit derselben Zielsetzung	Bewertung einer durchgeführten Maßnahme dahingehend, wie gut sie im Alltagsbetrieb aus Sicht der relevanten Stakeholder umgesetzt wird bzw. wie gut sie implementiert ist	Bewertung einer durchgeführten Maßnahme dahingehend, wie gut ihre Resultate sind – in Relation zu ihren eigenen Zielen, ihren Kosten, übergeordneten Werten oder alternativen Maßnahmen

kungsmechanismen im Detail zu rekonstruieren: **Welche** Interventionskomponenten funktionieren für **welche** Zielgruppe unter **welchen** Umständen **wie** und **warum**? Es geht also darum, im Detail die **Wirkprozesse** des Evaluationsgegenstandes theoretisch zu modellieren. Die theorieorientierte Wirkungsevaluation ist das Gegenmodell zu **Black-Box-Evaluationen**, bei denen nur die Ergebnisse einer Maßnahme untersucht werden (Wirksamkeit) und die Kausalprozesse als solche (Wirkung) im Dunkeln bleiben. Bei der theorieorientierten Evaluation wird somit die Erkenntnisfunktion der Evaluation besonders stark betont und Evaluationsforschung sowohl als theoriebildend als auch theorieprüfend angelegt. Gemäß der These „Es gibt nichts Praktischeres als eine gute Theorie“ („There is nothing as practical as a good theory“, Lewin, 1951, S. 196) soll theorieorientierte Evaluation ihren Nutzen darin entfalten, dass sie effektivere Veränderungsmöglichkeiten aufzeigt, wenn die Wirkprozesse im Detail besser verstanden werden.

Bei der theorieorientierten Evaluation versucht man, die in einem vorgegebenen Interventionsprogramm oder Produkt enthaltenen impliziten Wirk- bzw. Veränderungstheorien zu explizieren (Mackenzie & Blamey, 2005) und/oder passende grundlagenwissenschaftliche Theorien auf die Interventions- und dann auch Evaluationsfragestellung anzuwenden. Ein Theoriemodell, das die Wirkungsweise einer Maßnahme detailliert beschreibt und erklärt, wird auch als **Programmtheorie** bezeichnet; wissenschaftstheoretisch handelt es sich dabei um technologische Theorien (► Abschn. 18.1.1). Im Vergleich zur populären Wirksamkeitsevaluation (► Abschn. 18.2.1) kommt die theorieorientierte Wirkungsevaluation viel seltener zum Einsatz.

Innerhalb des theorieorientierten Ansatzes sind **unterschiedliche Evaluationsmodelle** zu unterscheiden, von denen die theoriegeleitete und die realistische Evaluation im Folgenden skizziert werden. Weitere theorieorientierte Modelle sind die die theoriebasierte Evaluation im

Tabelle 18.2 Evaluation der Deutschen Schülerakademie DSA nach dem CIPP-Modell. (Mod. nach Neber & Heller, 2002, S. 216. Adapted with permission from *European Journal of Psychological Assessment*, Vol. 18(3), pp. 214–228. © 2002 Hogrefe & Huber Publishers, www.hogrefe.com. DOI: 10.1027//1015-5759.18.3.214)

Evaluationstyp	Evaluationsfunktion	Evaluationsfragestellungen (Auswahl)	Methoden	Evaluationsrespondenten	Hauptergebnisse (verkürzte Darstellung)
Inputevaluation	deskriptiv	Welche Eingangsvoraussetzungen haben die Akademieteilnehmenden?	– Objektive und subjektive Selbstauskunftsinstrumente	$N = 252$ DSA-Teilnehmende $N = 70$ abgelehnte DSA-BewerberInnen	Die Teilnehmenden zeigen überdurchschnittliche Begabung und Motivation. Fazit: Die Teilnehmenden gehören tatsächlich zur Zielgruppe der Maßnahme.
	evaluativ	Sind die Akademieteilnehmenden hochbegabt?	– Vergleich von Teilnehmenden und Abgelehnten mit Standards für Hochbegabung – Vergleich von Teilnehmenden mit Abgelehnten		Die Teilnehmenden – sowie die Abgelehnten – lassen sich als hochbegabt einordnen. Fazit: Die Abgelehnten könnten ebenfalls in die Maßnahme aufgenommen werden.
Prozessevaluation	deskriptiv	Welche Merkmale haben die Kurse der Schülerakademie?	– Selbstauskunftsinstrumente von Teilnehmenden und Lehrenden	$N = 252$ DSA-Teilnehmende $N = 22$ DSA-Lehrende	Die Kurse sind geprägt von einer Vielfalt von Lehr-/Lernmethoden, selbstgesteuertes und kooperatives Lernen spielen eine wichtige Rolle.
	evaluativ	Entsprechen die Kursmerkmale den Anforderungen hochbegabter Lernender?	– Vergleich mit Standards der Begabtenförderung – Übereinstimmung zwischen Teilnehmenden- und Lehrendenbewertungen der Kurse		Die Lehrenden berichteten, dass sie ihre Unterrichtsziele erreichten. Die Teilnehmenden bewerteten die Kurse teilweise als weniger anspruchsvoll als erwartet. Fazit: Das Leistungsniveau der Kurse könnte noch gesteigert werden, das didaktische Konzept scheint angemessen.
Produktevaluation	deskriptiv	Welche kurzfristigen Effekte zeigt die Schülerakademie hinsichtlich ausgewählter kognitiver, motivationaler und sozialer Merkmale der Teilnehmenden?	– Vorher-Nachher-Messung – Fragebögen	$N = 252$ DSA-Teilnehmende $N = 174$ Einschätzungen von DSA-Teilnehmenden durch ihre regulären LehrerInnen	Es zeigte sich im Vorher-Nachher-Vergleich eine Zunahme domänenspezifischer Selbstwirksamkeitsüberzeugungen und selbstgesteuerter Lernstrategien sowie ein Rückgang der Vorliebe für wettbewerbsorientiertes Lernen.
	evaluativ	Trägt die Teilnahme an der Schülerakademie dazu bei, Kompetenz in exzellente Leistung zu überführen?	– Bewertung der Effekte vor dem Hintergrund von Theorien		Domänenspezifische Selbstwirksamkeitsüberzeugungen und selbstgesteuerte Lernstrategien sowie geringe Orientierung an konkurrenzorientiertem Lernen werden theoriebasiert als Katalysator- bzw. Mediatorvariablen verstanden, die dazu beitragen, Kompetenzen in exzellente Leistung zu überführen. Fazit: Die DSA erreicht die intendierten Effekte im Sinne einer Beeinflussung relevanter Katalysatorvariablen.
	explanativ	Welche Merkmale der Teilnehmenden und welche Merkmale der Kurse tragen zur Wirkung der Schülerakademie bei?	– Regressionsanalyse mit Input- und Process-Variablen als Prädiktoren und Product-Variablen als Kriterien		Die Effekte ließen sich v. a. vorhersagen durch individuelle Verbesserungsmöglichkeiten (d. h. geringere Eingangsvoraussetzungen der Teilnehmenden) sowie durch didaktische Merkmale der Kurse (v. a. selbstorganisiertes und kooperatives Lernen). Fazit: Die DSA erreicht die intendierten Effekte durch bestimmte Merkmale der Kurse.

Studienbeispiel

CIPP-Evaluation der Deutschen Schülerakademie

Die Deutsche Schülerakademie (DSA; http://www.deutsche-schuelerakademie.de) ist ein 1988 gestartetes und seit 1993 auch vom Bund mitfinanziertes Programm zur Förderung besonders leistungsfähiger und motivierter Schülerinnen und Schüler der 10. bis 13. Jahrgangsstufe des Gymnasiums. Das außerschulische Programm besteht aus dem einmaligen Besuch einer 16-tägigen Sommerakademie mit fachlichem Kursangebot sowie überfachlichen Aktivitäten.

Da das Programm vom Bundesministerium für Bildung und Forschung mitfinanziert wurde und wird, gab es eine wissenschaftliche Evaluation in Auftrag, um den Nutzen der Schülerakademie bewerten zu lassen. Die Evaluation wurde nach dem CIPP-Modell umgesetzt (Neber & Heller, 2002). In Abweichung von der Terminologie von Stufflebeam (► Abschn. 18.2.2) wird die Analyse der Maßnahmenteilnehmer hier nicht als Context-, sondern als Input-Evaluation etikettiert. Die Evaluationsfunktionen werden im Unterschied zu den in ► Abschn. 18.1.6 vorgestellten 5 Hauptfunktionen hier in 3 Aspekte eingeteilt: deskriptiv, evaluativ und explanativ (◘ Tab. 18.2).

engeren Sinne (theory-based evaluation, Birckmayer & Weiss, 2000) sowie die theoriebasierte Stakeholder-Evaluation (theory-based stakeholder evaluation; Hansen & Vedung, 2010), die den theoriebasierten mit dem partizipativen Evaluationsmodell verknüpft. Dabei wird das Ziel verfolgt, für die relevanten Stakeholder-Gruppen zu rekonstruieren, welche Theorien über die Wirkungsweise der untersuchten Intervention diese jeweils haben.

Die theoriegeleitete Evaluation („theory-driven evalution")

Im Rahmen der theoriegeleiteten Evaluation werden **Programmtheorien in 2 Untertheorien** eingeteilt (Chen, 1990, S. 17ff., 2005, S. 29; Widmer & De Rocchi, 2012, S. 65ff.; ► Studienbeispiel: „Theoriegeleitete Planung und Evaluation von Gesundheitskampagnen"):

- Das **Veränderungsmodell** („change model") erklärt, in welcher Weise die Maßnahme unter welchen Bedingungen bei den Zielgruppen die erwünschten – sowie auch unerwünschte – Wirkungen hervorbringen kann.

  Beispiel: Prostitutionskunden sollen zur Verwendung von Kondomen animiert werden, um sich selbst, ihre Partner und Partnerinnen sowie Prostituierte vor sexuell übertragbaren Infektionen zu schützen. Dazu soll eine Aufklärungskampagne eingesetzt und evaluiert werden. Das Veränderungsmodell müsste nun konkret auf theoretischer Basis vorgeben, welche Botschaften einer Kampagne (z. B. positive Rollenmodelle, angstauslösende Szenarien) die gewünschte Verhaltensänderung bei der Zielgruppe auf welchem Wege (z. B. Aufmerksamkeit für das Problem, erhöhtes Risikobewusstsein, Akzeptanz von Kondomen als Lösungsansatz) und unter welchen Bedingungen (z. B. Verfügbarkeit von Kondomen) bewirken können. Dabei sollten auch unerwünschte Veränderungen mitbedacht werden (z. B. könnte erhöhtes Risikobewusstsein dazu führen, dass Freier anstelle vermehrter Kondomverwendung sich verstärkt besonders jungen oder semiprofessionellen Prostituierten zuwenden).
- Das **Aktionsmodell** („action model") erklärt, wie die Maßnahme abläuft, d. h. welche Organisationen und Akteure im Hinblick auf welche Zielgruppen was genau unter welchen Bedingungen tun müssen, damit die Maßnahme realisiert wird. Im Hinblick auf die Aufklärungskampagne betrifft

Studienbeispiel

Theoriegeleitete Planung und Evaluation von Gesundheitskampagnen

Eine wichtige Maßnahme zur Prävention von Krankheiten stellen öffentliche Gesundheitskampagnen dar, die die Bevölkerung z. B. zu gesünderer Ernährung, mehr Bewegung, weniger Alkoholkonsum etc. bringen sollen. Doch wie sind derartige Kampagnenbotschaften im Sinne eines **Veränderungsmodells** am besten zu formulieren, damit sie bei möglichst großen Teilen der Zielgruppe tatsächlich zu Verhaltensänderungen führen? Ist es z. B. angeraten, mit möglichst schockierenden Text- und Bildbotschaften zu arbeiten, um die Bevölkerung aufzurütteln? Oder sind angstauslösende Botschaften gerade kontraindiziert, weil das Publikum sich abwendet und diese möglichst schnell „verdrängt"?

Es liegt auf der Hand, dass effektive Kampagnenbotschaften nicht einfach „aus dem Bauch heraus" entwickelt werden können, sondern sich auf Theorien über Einstellungen und Einstellungsänderungen sowie deren Verknüpfung mit Verhaltensänderungen speisen sollten. Ein entsprechender Theoriehintergrund ist nicht nur für die Gestaltung der Intervention, sondern gleichzeitig für deren Evaluation notwendig. Denn wenn ein theoriebasiertes Modell der Verhaltensänderung vorliegt, das z. B. mehrere kausal miteinander verknüpfte Schritte beschreibt – von der Aufmerksamkeit für die Kampagne über das Hinterfragen eigener Einstellungen, den kurz- und langfristigen Einstellungswandel bis zur kurz- und langfristigen Verhaltensänderung etc. – dann sind damit gleichzeitig wichtige Indikatoren zur Messung der Wirkungsweise und Wirksamkeit der Kampagne gegeben, die in einer theoriebasierten Evaluation aufgegriffen werden können (Slater, 2006). So wurde im Rahmen einer theoriebasierten formativen Evaluationsstudie eine HIV/Aids-Präventionskampagne für Jugendliche in Äthiopien entwickelt, wobei mit Theorien der Risikokommunikation gearbeitet wurde, um zielgruppenorientierte Botschaften zu konzipieren (Cho & Witte, 2005).

Es ist zu beachten, dass mit dem Veränderungsmodell nur ein Teilbereich der Programmtheorie abgedeckt ist und im Sinne einer theoriegeleiteten Evaluation auch das theoretische **Aktionsmodell** aufzustellen und zu prüfen wäre.

das z. B. Fragen der medialen Aufbereitung, der Orte und Zeitpunkte der Verbreitung der Kampagnenbotschaften, damit die Zielgruppe erreicht wird.

Die realistische Evaluation

Die realistische Evaluation („realistic evaluation"; Hense & Mandl, 2011; Pawson & Tilley, 1997) widmet sich theoretischen Erklärungen von Maßnahmenwirkungen unter Realbedingungen, d. h. der Ansatz betont die Bedeutsamkeit konkreter **Kontextbedingungen im Feld**, die darüber entscheiden, ob und wie bestimmte Wirkprinzipien greifen. Die realistische Evaluation distanziert sich damit von (quasi-)experimenteller wirkungsorientierter Evaluation, die die Gesamtwirkung einer Maßnahme nachweisen will (▶ Studienbeispiel: „Realistische Evaluation der Wirkungen von Videoüberwachung in Parkhäusern"). Im Rahmen der realistischen Evaluation werden Hypothesen in der Weise aufgestellt, dass damit kleine Kausalmodelle zu Einzelkomponenten der Intervention prüfbar sind, in denen die Verursachungsprozesse zwischen Kontext, Wirkmechanismus und Ergebnismuster differenziert modelliert sind (**CMO-Modelle**: Context, Mechanism, Outcome; Pawson & Tilley, 1997).

18.2.4 Der akteursorientierte Evaluationsansatz

Der akteursorientierte Evaluationsansatz zeichnet sich dadurch aus, dass er die Interessen der Beteiligten und Betroffenen in den Mittelpunkt rückt, sie z. B. bei der Nutzung der Evaluationsergebnisse unterstützt oder maßgeblich in die Durchführung der Evaluation einbezieht. Innerhalb des akteursorientierten Ansatzes lassen sich 2 Gruppen von Evaluationsmodellen unterscheiden: Die **klientenorientierten Modelle**, die auf spezielle Akteursgruppen zugeschnitten sind, und die **stakeholderorientierten Modelle**, die möglichst alle Beteiligten und Betroffenen gleichermaßen mit ihren Interessen einbeziehen (Widmer & De Rocchi, 2012, S. 79ff.).

Klientenorientierte Evaluationsmodelle

Wenn Evaluationsforschung als Dienstleistung verstanden wird, dann lässt sich fragen, wem denn von den Evaluierenden v. a. gedient werden soll. Vordergründig ist dies der Auftraggeber oder Sponsor einer Evaluationsstudie. Man spricht von **managementorientierter Evaluation**, wenn die Forschungstätigkeit v. a. auf die Interessen der **Auftraggebenden** zugeschnitten wird. Dies

Studienbeispiel

Realistische Evaluation der Wirkungen von Videoüberwachung in Parkhäusern

Die Installation von Überwachungskameras in Parkhäusern stellt eine Maßnahme dar, die die Kriminalität (z. B. Autoeinbrüche und Autodiebstähle) reduzieren soll. Aber welche Wirkungen kommen hier im Einzelnen wodurch zustande?

- Werden durch die Videoüberwachung vermehrt Autodiebe überführt (dazu müssen die Kameras so platziert sein, dass Diebe identifizierbar sind, zudem müssen die Ressourcen vorhanden sein, um die Videoaufzeichnungen fortlaufend durchzuführen und bedarfsweise auszuwerten etc.)?
- Werden durch die Videoüberwachung Autodiebe abgeschreckt (dazu muss die Videoüberwachung eindeutig erkennbar und von potenziellen Dieben als Risiko wahrgenommen werden)?
- Werden die Autobesitzer durch die Videoüberwachung vorsichtiger und lassen weniger Wertgegenstände im Wagen (dazu muss ihnen die Videoüberwachung bewusst sein und als Indikator erhöhter Diebstahlgefahr wahrgenommen werden)?
- Werden videoüberwachte Parkhäuser von Parkenden bevorzugt und sind deswegen stärker frequentiert, so dass letztlich das erhöhte Publikumsaufkommen (und nicht die Videoüberwachung selbst) potenzielle Diebe abschreckt?

Der Forschungsprozess im Zuge der realistischen Evaluation ist zirkulär: Zunächst wird ein Satz von CMO-Konfigurationen ähnlich den vier oben angeführten Kausalfragestellungen formuliert, dann werden qualitative und quantitative Daten erhoben (z. B. Interviews mit Polizei und Parkhauskunden, quantitative Daten zu Nutzungs- und Kriminalitätsraten in videoüberwachten und nicht-videoüberwachten Parkhäusern etc.). Der gesamte Satz von CMO-Konfigurationen wird durch die Daten geprüft, wobei qualitative und quantitative Analysetechniken zum Einsatz kommen. Das Gesamtergebnis wird interpretiert, was auf eine Modifikation der CMO-Konfigurationen und einen neuen Forschungszyklus hinauslaufen kann.

Das von Pawson und Tilley (1997, S. 75ff.; 2004) selbst zur Erläuterung der realistischen Evaluation genutzte Beispiel der Bewertung von Videoüberwachung in Parkhäusern lässt erahnen, wie aufwändig eine realistische Evaluation ist, die das Generieren, Prüfen und Verbessern eines ganzen Satzes detaillierter CMO-Konfigurationen verlangt. Der Ansatz ist vermutlich eher für kleinere Maßnahmen als für groß angelegte Programme geeignet. Die kleinteiligen Ergebnisse einer realistischen Evaluation, die die Komplexität von Interventionsprozessen und ihren Wirkungen abzubilden versuchen, können im Zuge der Evaluationsnutzung zu Problemen führen, etwa wenn Auftraggebende möglichst einfache und eindeutige Ergebnisse erwarten („Ist Videoüberwachung nun generell erfolgreich oder nicht?"). Der praktische Nutzen überprüfter CMO-Konfigurationen erschließt sich jedoch, wenn die Detailinformationen genutzt werden, um für konkrete reale Settings die Eignung einer Interventionsvariante zu verbessern oder einzuschätzen (z. B. lohnt sich Videoüberwachung auch in einem wenig frequentierten Parkhaus, wenn zudem nur wenig Ressourcen für die Auswertung der Videobilder zur Verfügung stehen?).

kann in der Praxis z. B. bedeuten, dass eine Maßnahme primär unter dem Aspekt möglicher Kosteneinsparungen evaluiert wird (Widmer & De Rocchi, 2012, S. 82).

Demgegenüber versteht sich die fürsprechende bzw. **anwaltschaftliche Evaluation** („advocacy evaluation") als Dienstleister v. a. für die – oft sozial unterprivilegierten – Adressaten einer sozialen Maßnahme (Greene, 1997). Unter der erkenntnistheoretischen Prämisse, dass Forschende niemals eine völlig neutrale Position einnehmen können, sondern – oft unreflektiert – durch ihre Vorannahmen, Menschenbilder und Werthaltungen ohnehin bestimmte Standpunkte beziehen, wird die Forderung abgeleitet, bewusst einen solchen Standpunkt zu wählen, der v. a. die Interessen von sozial und gesellschaftlich benachteiligten **Interventionsnutzenden** bzw. Zielgruppen sozialer Maßnahmen (z. B. Langzeitarbeitslose, Drogenkonsumenten, Prostituierte, Straßenkinder etc.) wahrt. Dieser Standpunkt basiert auf Werten der Fairness, Gerechtigkeit und Gleichberechtigung. Kontrovers diskutiert wird jedoch die konkrete Umsetzbarkeit sowie die Vereinbarkeit anwaltschaftlichen Vorgehens mit dem Anspruch wissenschaftlicher Unabhängigkeit (Widmer & De Rocchi, 2012, S. 83ff.). So kann das Evaluationsteam in Dilemmata geraten, wenn verschiedene Stakeholder-Gruppen unterschiedliche Ansprüche gegenüber einer Maßnahme geltend machen: Können und sollen anwaltschaftliche Evaluierende dann immer Partei für die jeweils weniger privilegierte Stakeholder-Gruppe ergreifen?

Die **nutzungsfokussierte Evaluation** („utilization-focused evaluation") schließlich versteht sich in erster Linie als Dienstleister für all diejenigen Stakeholder, die als **Evaluationsnutzende** praktische Konsequenzen aus den Evaluationsergebnissen ziehen sollen (Patton, 2008). Ein weit verbreitetes Problem in der Evaluationsforschung besteht nämlich darin, dass Evaluationsergebnisse völlig konsequenzenlos „in der Schublade" verschwinden (Fleischer & Christie, 2009, S. 166). Die spätere sachgerechte Nutzung von Evaluationsergebnissen sollte also, so die Prämisse der nutzungsfokussierten Evaluation, von Anfang an im Evaluationsprozess ausdrücklich berücksichtigt werden. Die damit verbundenen Anforderungen und Anstrengungen unterscheiden sich sehr stark in Abhängigkeit von der Evaluationsfunktion: Soll eine Evaluationsstudie in erster Linie Entscheidungs- oder Legitimationsfunktionen erfüllen, so mag das im Evaluationsbericht dargestellte Votum des Evaluationsteams, dass die betreffende Maßnahme sich als erfolgreich erwiesen hat und eine Weiterführung empfehlenswert ist, vollkommen ausreichend sein. Anders sieht es aus, wenn die Evaluationsstudie v. a. eine Optimierungsfunktion erfüllen soll und ihre ausdifferenzierten Ergebnisse somit zum Anlass zu nehmen sind, Teile der Maßnahme und ihrer Umsetzung zu modifizieren. Entsprechende evaluationsgestützte Veränderungen in der Praxis – auch gegen diverse Widerstände – erfolgreich und nachhaltig durchzusetzen, ist keine Selbstverständlichkeit, weshalb die Dienstleistungsfunktion in der nutzungsfokussierten Evaluation sich bewusst bis auf die Phase der Umsetzung der Evaluationsergebnisse erstreckt. Arbeitsmittel der nutzungsfokussierten Evaluation sind z. B. moderierte Workshops, in denen die Evaluierenden gemeinsam mit den Evaluationsnutzern die Befunde diskutieren und daraus abzuleitende Konsequenzen besprechen und planen (▶ Abschn. 18.3.5 „Organisation der Evolutionsnutzung").

Das nutzungsfokussierte Evaluationsmodell basiert auf der philosophischen Position des Pragmatismus und arbeitet eklektisch mit Elementen unterschiedlicher anderer Evaluationsmodelle. Es ist ausdrücklich offen für alle quantitativen und qualitativen Forschungsmethoden. Nutzungsfokussierte Evaluation wird von Patton (2008, S. 37) folgendermaßen definiert:

> » Utilization-focused program evaluation is evaluation done for and with specific intended primary users for specific, intended uses. Utilization-focused evaluation begins with the premise that evaluation should be judged by their utility and actual use; therefore, evaluators should facilitate the evaluation process and design any evaluation with careful consideration for how everything that is done, from beginning to end, will affect use. Use concerns how real people in the real world apply evaluation findings and experience the evaluation process. Therefore, the focus in utilization-focused evaluation is on *intended use by intended users*.

Der Schlüssel der nutzungsfokussierten Evaluation liegt zunächst in der Klärung der **intendierten Nutzenden der Evaluationsergebnisse**. Dabei ist zu beachten, dass intendierte Nutzerinnen und Nutzer von Evaluationsergebnissen (z. B. politische oder organisationale Entscheidungsträger, Programmentwickler) oft andere Stakeholder sind als die Zielgruppen einer Maßnahme. Des Weiteren sind die **intendierten Nutzungsweisen der Evaluationsergebnisse** (was genau soll auf ihrer Basis getan werden?) detailliert festzulegen (▶ Studienbeispiel: „Nutzungsfokussierte Evaluation eines Hospizes"). Dabei sind unterschiedliche Typen von Nutzungsweisen zu differenzieren (Patton, 2008, S. 112):

- **Instrumentelle Nutzung**: Die Evaluationsergebnisse werden z. B. für Optimierungs-, Kontroll- und Entscheidungsfunktionen genutzt, etwa in Form konkreter Veränderungen der Maßnahme und/oder ihrer Durchführung.
- **Konzeptuelle Nutzung**: Die Evaluationsergebnisse werden im Sinne der Erkenntnisfunktion genutzt und tragen zum besseren Verständnis der Maßnahme auf

Studienbeispiel

Nutzungsfokussierte Evaluation eines Hospizes

Die Hospizbewegung hat das Ziel, Sterbenden und ihren Angehörigen eine ganzheitliche Begleitung anzubieten und setzt sich dafür ein, Sterben und Tod nicht gesellschaftlich auszugrenzen, sondern als Bestandteil des Lebens anzuerkennen und würdevoll zu gestalten. Es existieren vielfältige ambulante, teilstationäre und stationäre Hospizdienste unterschiedlicher Träger.

Ein US-amerikanisches Hospiz gab eine Evaluationsstudie in Auftrag, um die Stärken und Schwächen ihrer Dienstleistungen zu erfassen und Maßnahmen zur Weiterentwicklung und Verbesserung zu konzipieren. Das Evaluationsteam entschied sich angesichts dieses Evaluationsauftrags für eine nutzungsfokussierte Evaluation (Hiatt, Stelle, Muslow, & Scott, 2007): In enger Zusammenarbeit mit Leitung und Personal des Hospizes sollten die Bewertungen unterschiedlicher Stakeholder erfasst und mit Blick auf die Entwicklung konkreter und umsetzbarer Verbesserungsmaßnahmen ausgewertet werden. Mithilfe von Einzelinterviews und Fokusgruppendiskussionen wurden die Perspektiven von insgesamt 72 Vertreterinnen und Vertretern sieben unterschiedlicher Stakeholder-Gruppen erfasst: 1. aktuelle Patienten, 2. aktuelle Angehörige, 3. hinterbliebene Angehörige früherer Patienten, 4. ehrenamtliche Hospizmitarbeiter, 5. hauptamtliche Hospizmitarbeiter, 6. Ärzte, 7. Einwohner im Einzugsgebiet des Hospizes.

Die Äußerungen aller Evaluationsrespondenten wurden im Zuge der qualitativen Auswertung der Interviewdaten in drei thematische Blöcke gegliedert: 1. Stärken des Hospizes, 2. Schwächen des Hospizes, 3. Entwicklungspotenziale des Hospizes. Die Befunde wurden zusammen mit der Hospizleitung als primärer Evaluationsnutzerin und mit Blick auf die Konzeption von umsetzbaren Verbesserungsmaßnahmen diskutiert.

Als ein Problemfeld wurde z. B. identifiziert, dass eine hohe Personalfluktuation vorliegt (v. a. bei den Ehrenamtlichen) und somit teilweise Kompetenzdefizite auftreten (z. B. im Umgang mit Sterbenden religiöser Minderheiten). Als Lösungsansatz wurde ein Mentoringprogramm vorgeschlagen, das neue Teammitglieder jeweils mit erfahrenen Mitarbeitern zusammenbringt. Als ein weiteres Problemfeld erwies sich, dass die Dienste des Hospizes im Einzugsgebiet zu wenig bekannt oder mit einem negativen Image verknüpft waren – hier wurden als Gegenmaßnahmen Informationsveranstaltungen und Workshops in Pflegeheimen, öffentlichen Bibliotheken etc. vorgeschlagen, in denen die Hospizarbeit durch Personal sowie durch Angehörige von Sterbenden bzw. Verstorbenen vorgestellt werden soll.

Seiten der Stakeholder bei. Losgelöst von der konkreten Maßnahme können die Evaluationsergebnisse auch zur gesellschaftlichen Aufklärung über bestimmte Probleme beitragen und öffentliche Debatten beeinflussen.

- **Prozessbezogene Nutzung**: Allein die Teilnahme am Evaluationsprozess führt (unabhängig von dessen Ergebnissen) zu instrumentellen und/oder konzeptuellen Veränderungen, etwa wenn Stakeholder anlässlich ihrer Involvierung in den Evaluationsprozess plötzlich intensiver über die Zielsetzungen und Probleme der Maßnahme nachdenken, untereinander in Dialog treten oder wenn allein die regelmäßige Messung von bestimmten Kennwerten der Maßnahme das Verhalten der Maßnahmenbeteiligten beeinflusst.

Stakeholderorientierte Evaluationsmodelle

Während sich die klientenorientierten Evaluationsmodelle jeweils primär in den Dienst einzelner Stakeholdergruppen stellen, verfolgen stakeholderorientierte Modelle das Ziel, alle Beteiligten und Betroffenen gleichermaßen einzubeziehen. Dabei unterscheiden sich die einzelnen Modelle teilweise beträchtlich dahingehend, wie die vielfältigen und auch widerstreitenden Stakeholder-Interessen angemessen zu berücksichtigen sind und welche Rolle (z. B. vermittelnd, moderierend) die Evaluierenden in diesem Prozess einnehmen.

Da das Einbeziehen aller relevanten Anspruchsgruppen gemäß Evaluationsstandards gefordert wird und somit ein universelles Kennzeichen von Evaluationsstudien sein sollte, sind Studien gemäß stakeholderorientierter Evaluationsmodelle v. a. hinsichtlich der Intensität und Akzentsetzung der Stakeholder-Involvierung abzugrenzen, insbesondere dahingehend, dass es ihr Anliegen ist, gesellschaftlich marginalisierten Gruppen Mitbestimmung zu ermöglichen. Widmer und De Rocchi (2012, S. 85ff.) unterscheiden vier verschiedene stakeholderorientierte Evaluationsmodelle:

1. **Responsive Evaluation** („responsive evaluation“; „responsive/client-centered evaluation“) läuft auf eine qualitative Fallstudie hinaus, bei der die Sichtweisen der Beteiligten und Betroffenen sowie der Evaluierenden über den Evaluationsgegenstand auf der Basis von teilnehmenden Beobachtungen, qualitativen Interviews und Dokumentenanalysen etc. zusammengetragen werden (Stake, 1975, 2004).
2. **Demokratische Evaluation** („democratic evaluation“) trägt ebenfalls die Sichtweisen aller Beteiligten und Betroffenen im Rahmen einer Fallstudie zusammen und lässt sie dabei auch kontrollieren, ob und wie ihre Äußerungen im Evaluationsbericht auftauchen. Das Gesamtbild der Maßnahme soll sich also aus den Stimmen von Vertretern möglichst aller Stakeholder-Gruppen zusammensetzen, wobei die Evaluierenden selbst kein Urteil abgeben (MacDonald, 1976).
3. **Partizipative Evaluation** („participatory evaluation“) bezieht die Beteiligten und Betroffenen nicht nur als Informanten, sondern auch als **Forschungspartner** in das Evaluationsprojekt mit ein. Dies geschieht etwa durch die Mitwirkung an der Ausformulierung von Evaluationsfragen, an der Planung des Untersuchungs-Designs oder an der Datenauswertung. Auf diese Weise soll zum einen erreicht werden, dass die Evaluation die Interessen aller Be-

teiligten berücksichtigt und somit praxisrelevantere Ergebnisse liefert. Zum anderen wird der aktiven Mitwirkung selbst ein positiver Nutzen zugeschrieben, indem sich die Beteiligten stärker mit der Maßnahme und ihrer Verbesserung identifizieren (► Abschn. 18.1.6 zur Dialog- und Lernfunktion der Evaluation). Typische Einsatzfelder der partizipativen Evaluation sind beispielsweise die Entwicklungsforschung und die Gesundheitsforschung: Maßnahmen und Projekte der Armutsbekämpfung oder der Prävention sexuell übertragbarer Infektionen werden als besonders erfolgversprechend eingeschätzt, wenn sie partnerschaftlich zusammen mit den jeweiligen Zielgruppen entwickelt, erforscht und bewertet und auf dieser Basis bei Bedarf verbessert werden (z. B. Zukoski & Luluquisen, 2002; Chambers, 2009; Chouinard, 2013; ► Studienbeispiel: „Partizipative Interventions- und Evaluationsforschung zu einer Gesundheitskampagne gegen Übergewicht"). In partizipativen Evaluationsmodellen dominieren qualitative Forschungsmethoden und es zeigen sich Parallelen zum Ansatz der **Aktionsforschung** (Lewin, 1946; Whyte, 1991). Partizipative Evaluationsmodelle existieren in diversen Varianten, etwa das partizipative Evaluationsmodell im engeren Sinne („participatory evaluation"; Cousins & Earl, 1992; Earl & Cousins, 1995), das deliberativ-demokratische Evaluationsmodell („deliberative democratic evalution"; House & Howe, 1998) oder das konstruktivistische Evaluationsmodell („constructivist/fourth generation evaluation"; Guba & Lincoln, 1989). Es liegt auch Spezialliteratur zur partizipativen Evaluationsarbeit mit speziellen Stakeholder-Gruppen wie z. B. Kindern und Jugendlichen vor (McCabe & Horsley, 2008).

Die APA-Literaturdatenbank PsycINFO weist 59 referierte Journal-Artikel für das Keyword „participatory evaluation" nach (Stand: November 2013). Partizipative Evaluation setzt voraus, dass Vertreter/innen der relevanten Stakeholder-Gruppen sich entsprechend intensiv und dauerhaft engagieren (teilweise werden sie als Mitforschende auf Projektstellen oder auf Honorarbasis beschäftigt), denn der Zeit- und Arbeitsaufwand für alle Beteiligten ist hoch. Zudem muss eine große Kommunikations- und Konfliktbereitschaft mitgebracht werden für die Vermittlung und Aushandlung der unterschiedlichen Positionen. Da der Evaluationsprozess im partizipativen Ansatz nicht „von oben herab" durch das Evaluationsteam und den Auftraggeber bestimmt, sondern partnerschaftlich an den Bedürfnissen und Ideen der Betroffenen und Beteiligten ausgerichtet werden soll, entsteht operativ das Problem, dass bei Abschluss des Evaluationsvertrages meist noch kein vollständiger Zeit- und Kostenplan vorgelegt werden kann bzw. dieser im Laufe des Projekts immer wieder verändert werden muss. Um von den Vorteilen des partizipativen Evaluationsmodells zu profitieren und gleichzeitig dessen Herausforderungen und Probleme (Gregory, 2000) beherrschbar zu halten, wird in großen Evaluationsprojekten mitunter so vorgegangen, dass nur Teilfragen gemäß einem partizipativen Modell bearbeitet werden und in anderen Projektteilen bzw. Projektphasen ein geringerer Grad an Partizipation realisiert wird. Es liegen Messinstrumente vor, um den Grad der Stakeholder-Partizipation innerhalb von Evaluationsstudien zu erfassen, z. B. das „Participatory Evaluation Measurement Instrument" (PEMI; Daigneault, Jacob, & Trembly, 2012). Partizipative Evaluationsansätze haben sich in einer Reihe von Studien als erfolgreich erwiesen und erleben in den letzten Jahren deutlichen Zulauf. Sie sollten jedoch nicht als Patentlösung für jegliches Evaluationsproblem begriffen werden. Vielmehr muss auch in der Evaluationsforschung das methodische Vorgehen der jeweiligen Fragestellung und den Kontextbedingungen entsprechen. So wird in der Evaluations-Community der Alleinvertretungsanspruch eines Ansatzes meist zurückgewiesen und auf die Fruchtbarkeit von Methodenpluralität verwiesen (Datta, 2013).

4. **Empowerment-Evaluation** („empowerment evaluation") geht noch einen Schritt weiter als partizipative Evaluation: Die Stakeholder werden nicht nur zu Forschungspartnern der Evaluierenden, sondern sollen die Evaluation gänzlich selbst in die Hand nehmen. Dadurch sollen v. a. auch sozial unterprivilegierte Zielgruppen sozialer Maßnahmen in eine Position der Selbstermächtigung („empowerment") kommen, die es ihnen im Zuge der Evaluation und darüber hinaus durch Kompetenzgewinn erlaubt, aktiver für eine Verbesserung ihrer Lebensverhältnisse einzutreten. Dem Evaluationsteam fällt dabei nur noch eine beratende Rolle zu (Fetterman, 1994, 2001). Indem bei der Empowerment-Evaluation die Evaluierten weitgehend selbstständig die Evaluation organisieren sollen, ähnelt das Vorgehen dem einer Selbstevaluation (► Abschn. 18.1.8). Dass das theoretische Modell der Empowerment-Evaluation in der Praxis ganz unterschiedlich umgesetzt wird und dabei nicht selten die Zielgruppen gar nicht eingebunden werden (sondern nur die Maßnahmenverantwortlichen) und somit auch kein Zuwachs an Empowerment bei den Zielgruppen gefördert wird, zeigt die bereits angesprochene Forschung zum Verhältnis von Theorie und Praxis der Evaluation. Eine Auswertung der Forschungsliteratur zeigte (Spiel et al., 2010), dass von insgesamt 47 zwischen 1994 und 2005 publizierten Empowerment-

Studienbeispiel

Partizipative Interventions- und Evaluationsforschung zu einer Gesundheitskampagne gegen Übergewicht

Medienkampagnen sind ein bewährtes Mittel der Gesundheitsprävention. Zur Vorbeugung und Bekämpfung von Übergewicht und daraus resultierenden Krankheiten sollte eine entsprechende Kampagne für die mexikanisch-amerikanische Bevölkerung in Texas entwickelt, umgesetzt und evaluiert werden (Reininger et al., 2010). Denn diese Bevölkerungsgruppe ist besonders stark von Übergewicht betroffen. Um die Kampagne und deren Evaluation optimal an den Erwartungen und Bedürfnissen der Zielgruppe auszurichten, wurde ein **partizipativer Forschungsansatz** gewählt („community based participatory research", CBPR): Das universitäre Forschungsteam arbeitete mit einer Gruppe von Vertreter/innen der Zielgruppe von Anfang an zusammen. Dabei handelte es sich um fachbezogene Praxispartner (z. B. vor Ort tätige Ernährungsberater/innen, Gesundheitseinrichtungen, Ärzte), um für die Bevölkerung zentrale Organisationen (z. B. lokale Religionsgemeinschaften) sowie um Einwohner/innen – ihre Gesamtzahl wuchs im Laufe der Studie auf rund 100 Forschungspartner.

Auf Anregung der Forschungspartner sollte in der ersten Projektphase erkundet werden, welche **Gesundheitsanliegen rund um Übergewicht** die mexikanisch-amerikanische Bevölkerung hat und welche Medien sie nutzt. Dazu wurden amtliche Statistiken zu Rate gezogen und Fokusgruppendiskussionen durchgeführt. Die Daten wurden von den Wissenschaftler/innen zusammen mit den Forschungspartnern interpretiert. Auf dieser Basis wurde beschlossen, **spanischsprachige Kampagnen-Botschaften** zu kreieren, die auf gesunde Ernährung und Bewegung abzielen und mit Rollenmodellen arbeiten. Theoretische Grundlage war die sozial-kognitive Lerntheorie („Lernen am Modell"). Gemeinsam wurden spanischsprachige Botschaften für die Kampagne „Tu Salud ¡Sí Cuenta!" („Deine Gesundheit zählt!") entworfen und produziert. Durch die enge Zusammenarbeit mit vielen Forschungspartnern in der Community konnten eine ausreichende Zahl geeigneter Rollenmodelle rekrutiert und in den Botschaften auch spezifische, vor Ort kursierende Gesundheitsmythen (z. B. über die Entstehung von Diabetes) aufgegriffen werden. Die Kampagnen-Botschaften wurden teils über gedruckte Newsletter in der Gemeinde verbreitet, vor allem aber im **Frühstücksprogramm des beliebtesten lokalen TV-Senders** ausgestrahlt. Der Weg zur Kooperation mit dem Fernsehsender war durch eine gemeinsame Anfrage der Wissenschaftler/innen mit den einhundert Community-Vertretern geebnet worden.

Die Forschungspartner aus der Zielgruppe wirkten nicht nur an der Analyse der Ausgangslage, der Entwicklung der Kampagne und deren Umsetzung mit, sondern waren auch in die Planung und die Durchführung der Evaluation einbezogen. Im Sinne des CIPP-Modells wurden eine Prozess-Evaluation und eine Produkt-Evaluation (wirksamkeitsorientierte Evaluation) angestrebt. Für die **wirksamkeitsorientierte Evaluation** schlug das Forschungsteam ein quasi-experimentelles Kontrollgruppendesign vor, d. h. der lokalen Gemeinde, in der die Kampagne laufen sollte, wurde eine vergleichbare Gemeinde ohne Kampagne als Kontrollgruppe gegenübergestellt. Verglichen werden sollte dann z. B. das Ernährungs- und Bewegungsverhalten der Bevölkerung. Als Ziel wurde eine langfristige Verhaltensänderung dahingehend definiert, täglich fünf Portionen Obst oder Gemüse zu essen und sich mindestens 30 Minuten zu bewegen (v. a. Spazierengehen).

Im Rahmen der **Prozess-Evaluation** sollte geprüft werden, wie stark sich die Kampagnen-Botschaften verbreiten und wie sie von der Zielgruppe aufgenommen werden. Dazu wurde z. B. gemessen, wie viel Sendezeit im Fernsehen die Kampagnen-Botschaften im Laufe eines Jahres einnahmen (rund acht Stunden pro Jahr bzw. durchschnittlich zwei Minuten täglich im Frühstücksprogramm und 1.5 Minuten in Nachrichtensendungen). Zudem wurden sechs Fokusgruppen-Diskussionen durchgeführt, um Einschätzungen zu den Kampagnen-Botschaften zu erheben. Dabei stellte sich heraus, dass die Rollenmodelle sehr gut ankamen (allerdings nur, sofern sie korrektes Spanisch sprachen, ein Sprachmix aus Englisch und Spanisch wurde nicht akzeptiert) und dass mehr konkrete Verhaltenstipps (Rezepte, Sportübungen) gewünscht wurden. Die Ergebnisse der Prozess-Evaluation wurden mit den Forschungspartnern diskutiert und führten zu einer Modifikation der Kampagnen-Botschaften.

Im Zuge der engen Kooperation zwischen Wissenschaft und Community-Vertretern bei der Interventions- und Evaluationsforschung zur Prävention von Übergewicht wurden über die Medienkampagne hinaus vor allem lokale **Umweltfaktoren als entscheidend für das Gesundheitsverhalten** herauskristallisiert. Das betrifft das Anlegen geeigneter **Spazierwege** in der Nachbarschaft, die Durchführung von angeleiteten **Walking-Gruppen** oder den Aufbau eines lokalen **Wochenmarktes** mit bezahlbarem Obst- und Gemüseangebot. Entsprechende Praxisprojekte wurden im Gefolge der Studie zur Medienkampagne in Kooperation zwischen Community und Universität in Angriff genommen.

Das Studienbeispiel verdeutlicht die Potenziale des partizipativen Vorgehens für die Interventions- und Evaluationsforschung. Zu beachten ist jedoch der hohe zeitliche und organisatorische Aufwand, um den regelmäßigen Kontakt und die forschungspraktische Zusammenarbeit mit bis zu einhundert Forschungspartner/innen zu organisieren. Dies beinhaltet teilweise auch die Qualifizierung der Forschungspartner für bestimmte Schritte der Erhebung, Auswertung und Interpretation empirischer Daten. Auch wird deutlich, dass derartige partizipative Projekte von vorne herein mit langen Laufzeiten geplant werden sollten, um auch aufwändige Veränderungen anzustoßen und zu etablieren. Offen war zum Publikationszeitpunkt (Reininger et al., 2010) noch die Frage, ob im Sinne einer wirksamkeitsorientierten Evaluation durch Kontrollgruppenvergleich nachweisbar sein wird, dass die Kampagne samt den Begleitmaßnahmen (Spazierwege, Wochenmarkt) das Gesundheitsverhalten der lokalen Bevölkerung kausal beeinflusst im Sinne eines mittleren oder starken Effektes in Richtung auf das Zielverhalten.

Evaluationen nur eine einzige alle 10 Prinzipien des Empowerment-Ansatzes gemäß Wandersman et al. (2005) erfüllte. Ob diese Theorie-Praxis-Diskrepanz darauf zurückzuführen ist, dass in manchen Evaluationskontexten aus praktischen Gründen gar nicht alle Empowerment-Prinzipien umsetzbar sind und somit das Modell anzupassen ist, oder ob die Evaluierenden mehr Schulung und Anleitung benötigen, um sich enger an das vorliegende Empowerment-Modell halten zu können, ist Gegenstand weiterer Klärungen.

Analog sind auch die anderen Evaluationsansätze gefordert, Widersprüche zwischen ihren theoretischen Vorgaben und der praktischen Umsetzung aufzudecken und angemessen zu behandeln.

18.3 Phasen der Evaluationsforschung

Nach der theoretischen Ausrichtung einer Evaluationsstudie wenden wir uns nun ihrem **zeitlichen Ablauf** zu. Generell folgt eine wissenschaftliche Evaluationsstudie einerseits den typischen Phasen des empirischen Forschungsprozesses (Teil II „Anwendung" des vorliegenden Buches) und kann andererseits aus Projektmanagementperspektive in Projektphasen eingeteilt werden. Im vorliegenden Abschnitt stellen wir ein Phasenmodell für Evaluationsforschungsprojekte vor, das beide Sichtweisen kombiniert (Tab. 18.3; siehe den Beitrag von Döring, 2009, auf den sich dieses Teilkapitel stützt). Der Phasenablauf ist dabei nicht starr und linear aufzufassen, sondern er beinhaltet in der Praxis nicht selten auch Bearbeitungszyklen, z. B. das Rückspringen auf frühere Phasen zur Korrektur vorausgegangener Entscheidungen oder zur Verfeinerung von Vorarbeiten.

18.3.1 Initiierungsphase der Evaluation

In der Initiierungsphase geht es aus Sicht der Evaluierenden darum, potenzielle Evaluationsprojekte zu akquirieren, indem Evaluationsmöglichkeiten identifiziert und Evaluationsziele kritisch beleuchtet werden. Da Evaluation meist Auftragsforschung ist, müssen sich Evaluierende darum bemühen, entsprechende Aufträge zu erhalten. Umgekehrt geht es für Auftraggebende darum, für ihr jeweiliges Evaluationsanliegen ein geeignetes Evaluationsteam zu finden (für Empfehlungen für Auftraggebende Gesellschaft für Evaluation e. V. e.V. [DeGEval], o. J.).

Identifikation von Evaluationsmöglichkeiten

Es existieren im Wesentlichen 4 Wege für Evaluatorinnen und Evaluatoren, um an mögliche Evaluationsprojekte zu kommen und als Dienstleistende für Auftraggebende tätig zu werden (Stufflebeam & Shinkfield, 2007, S. 453ff.):

- **Projektausschreibungen**: Öffentliche sowie private/gewerbliche Ausschreibungen (RFP: „request for proposal") für Evaluationsprojekte skizzieren meist das Evaluationsobjekt, Evaluationsziele sowie Merkmale der Evaluationsstudie (z. B. Zeitrahmen, Budget, bevorzugte Forschungsmethoden; Bussmann, 1995, S. 25). Die Anforderungsliste der Auftraggebenden wird auch als **Pflichtenheft** bezeichnet. Bieter sind eingeladen, Angebote für Evaluationsstudien einzureichen (Evaluationskonzept, Zeitplan, Kostenplan etc.) sowie ihre Expertise nachzuweisen (z. B. Belege früherer Studien, Firmenprofil, Lebenslauf etc.). Üblicherweise wird über die Angebote in Abhängigkeit von inhaltlicher Qualität und Kostenvoranschlag entschieden. Öffentliche Evaluationsausschreibungen finden sich u. a. im Deutschen Ausschreibungsblatt (online unter http://www.deutsches-ausschreibungsblatt.de). Ausschreibungen unterscheiden sich hinsichtlich ihres Umfangs und ihres Detaillierungsgrades erheblich. Oftmals ist ausdrücklich vorgesehen, dass das eingereichte Evaluationskonzept im Falle eines Zuschlags später gemeinsam mit dem Auftraggeber verfeinert oder modifiziert werden kann.
- **Projektaufträge**: Viele Evaluatoren und Evaluatorinnen arbeiten nicht selbstständig, sondern in Unternehmen oder Evaluationsinstituten. Dabei sind sie teilweise in die Projektakquise involviert, teilweise bekommen sie Evaluationsaufträge zugeteilt. Der Evaluationsauftrag kann sich dabei auf einen unternehmensinternen Gegenstand beziehen (interne Evaluation: z. B. Evaluation des unternehmensinternen Weiterbildungssystems) oder beim Kunden angesiedelt sein (externe Evaluation).
- **Projektanfragen**: Manchmal suchen sich Auftraggeber auch gezielt Evaluatoren aus und bieten ihnen Projekte an. Dabei handelt es sich typischerweise um besonders namhafte Evaluatoren oder um Evaluatoren, mit denen das Unternehmen bereits früher zusammengearbeitet hat. Die Chancen, Projektanfragen zu erhalten, erhöhen sich für Evaluatorinnen und Evaluatoren in Abhängigkeit von Art und Anzahl erfolgreich durchgeführter früherer Evaluationsprojekte sowie entsprechender Öffentlichkeitsarbeit (z. B. öffentliche Vorträge, Seminare, Internetpräsenz).
- **Evaluator-initiierte Projekte**: Auf der Basis ihrer Fachexpertise und Praxiserfahrung in einem bestimmten Evaluationsbereich können Evaluatorinnen und Evaluatoren auch selbst Projektideen entwickeln. Dabei kann es sich einerseits um Projekte handeln, die aus Evaluatorperspektive v. a. von wissenschaftlichem Interesse sind, publizierbare Ergebnisse versprechen und der fachlichen Profilierung dienen. Derartige Evaluationsforschungsprojekte können in kleinem Rahmen teilweise auf Selbstkostenbasis umgesetzt werden (z. B. informelle Kooperation mit Praxispartnern). Geht es jedoch um umfangreichere Projekte und finanzierte Projektaufträge, so müssen finanzkräftige potenzielle Auftraggebende von der Evaluationsidee überzeugt werden.

Tabelle 18.3 Ein Phasenmodell für Evaluationsforschungsprojekte

Projektphasen der Evaluation	Arbeitsschritte
1. Initiierung	1.1 Identifikation von Evaluationsmöglichkeiten 1.2 Festlegung und Prüfung der Evaluationsziele
2. Konzeption	2.1 Exploration des Evaluationsbedingungen 2.2 Entwicklung des Evaluationskonzeptes 2.3 Einbeziehung der Evaluationsstandards 2.4 Vorbereitung des Projektmanagements 2.5 Abschluss des Evaluationsvertrags
3. Planung	3.1 Formulierung detaillierter Evaluationsfragen und -hypothesen 3.2 Untersuchungsplanung 3.3 Projektplanung
4. Realisierung	4.1 Erhebung von Evaluationsdaten 4.2 Analyse von Evaluationsdaten 4.3 Interpretation von Evaluationsdaten 4.4 Projektcontrolling
5. Abschluss	5.1 Abgabe der Evaluationsberichte 5.2 Organisation der Evaluationsnutzung 5.3 Durchführung einer Metaevaluation 5.4 Publikation der Evaluationsstudie

Festlegung und Prüfung der Evaluationsziele

Nicht jede Evaluationsmöglichkeit ist es wert, weiterverfolgt zu werden. Erfährt man etwa sehr spät von einer Projektausschreibung, so ist oft nicht genug Zeit vorhanden, um ein konkurrenzfähiges Angebot auszuarbeiten. Vorsicht ist auch geboten bei Ausschreibungen, deren Evaluationsbereich relativ weit von der eigenen Fachexpertise entfernt ist. Denn dies verlangt einerseits deutlichen Mehraufwand bei der Einarbeitung und außerdem wird der Zuschlag vom Qualifikationsprofil des Bieters abhängig gemacht (zu Evaluationsbereichen innerhalb der Psychologie: Moosbrugger & Schweizer, 2002). Fachliche Nähe der Evaluierenden zum Evaluationsgegenstand erleichtert und verbessert nicht nur das theoretische und methodische Arbeiten, sondern geht in der Regel auch mit höherer Glaubwürdigkeit und Autorität gegenüber den Stakeholdern einher sowie mit dem Zugriff auf einschlägiges soziales Kapital (z. B. Expertenkontakte, Kollegennetzwerke). Nicht nur Projektausschreibungen, sondern auch externe Projektaufträge müssen kritisch überprüft werden, etwa dahingehend, ob auf Evaluatorseite genügend personelle Ressourcen für ein unvorhergesehen angebotenes Projekt bereitgestellt werden können.

Neben den operativen Restriktionen (Zeit, Personal, Fachexpertise) ist bei der Entscheidung für ein Evaluationsprojekt die **Problemdefinition bzw. Zielsetzung** von herausragender Bedeutung für den Projekterfolg. Man unterscheidet zwischen den **Zielen einer Intervention** und den **Zielen, Zwecken oder Funktionen der Evaluation** (► Abschn. 18.1.6). Für potenzielle Evaluatorinnen und Evaluatoren einer Intervention gilt es im Vorfeld einzuschätzen, wie erfolgversprechend eine Transformation von Grobzielen („goals“) zu praktikablen Feinzielen bzw. Zielzuständen („objectives“ in Kooperation mit Auftraggebern und weiteren Stakeholdern erscheint. Denn ohne eine klare Problemdefinition oder einen klaren Evaluationsauftrag kann Evaluationsforschung meist nicht erfolgreich betrieben werden (Ausnahme: zielfreie Evaluation, ► Abschn. 18.1.5). Mit den Feinzielen der Evaluation sind auch die **zentralen Evaluationsfragen** festgelegt (z. B. Soll die Qualität der Durchführung einer Maßnahme bewertet werden, ihre Wirksamkeit oder ihre Kosten-Nutzen-Bilanz?), die im späteren Verlauf noch detaillierter ausgearbeitet werden (► Abschn. 18.3.3 „Formulierung detaillierter Evolutionsfragen und -hypothesen“).

Wichtig ist es auch, im Vorfeld auszuloten, ob tatsächlich ein ernsthaftes Evaluationsinteresse besteht, oder nicht vielmehr eine **Pseudoevaluation** erwartet wird (► Abschn. 18.1.6). Ein authentisches Fallbeispiel soll die Umstände einer Pseudoevaluation illustrieren (Stufflebeam & Shinkfield, 2007, S. 152f.): Ein Forschungszentrum wird seit Jahren regelmäßig mit sehr positiven Ergebnissen hinsichtlich der Bedeutung, Qualität und Menge der produzierten Forschungsergebnisse extern evaluiert. Der neue Direktor des Zentrums bespricht mit dem Evaluationsteam, dass die vorliegenden positiven Evaluationen zwar die Moral der Beschäftigten steigerten, aber keinerlei Optimierungsideen lieferten. Deswegen sei es

im Sinne der Qualitätssteigerung erforderlich, dass die Evaluation sich nun verstärkt darauf konzentriere, Schwächen aufzudecken, aus denen sich dann Verbesserungsmaßnahmen ableiten lassen. Dementsprechend erarbeitet das Evaluationsteam eine umfassende Aufstellung von Schwächen der Einrichtung und der dort durchgeführten Forschungsprogramme. Als das Evaluationsteam seine Ergebnisse präsentiert, stellt sich heraus, dass nicht nur die Beschäftigten des Forschungszentrums, sondern auch die Geldgeber anwesend und über die Negativbilanz entsetzt sind. Das Evaluationsteam muss erkennen, dass es durch den einseitigen Evaluationsauftrag von dem neuen Direktor instrumentalisiert wurde, dessen Interesse darin bestand, die Arbeit seines Vorgängers zu diskreditieren und laufende Forschungsprogramme durch andere seiner Wahl zu ersetzen. Um Auftraggeber mit Instrumentalisierungsabsichten zu erkennen, können Hintergrundinformationen aus dem Praxisfeld und Kollegenkontakte hilfreich sein.

Abgesehen von **Fachkompetenz** und **Interesse** am Inhaltsbereich sind in der Evaluationsforschung **Wertfragen** oft virulenter als in der thematisch selbst bestimmten und praxisferneren Grundlagenforschung. Wenn Evaluierende z. B. daran mitarbeiten, Methoden des militärischen Schießtrainings oder Kampagnen radikaler Parteien hinsichtlich ihrer Stärken und Schwächen zu bewerten, damit sie dann verbessert werden können, nehmen sie damit Einfluss auf die Gestaltung unserer Lebensbedingungen und müssen dies mit ihren Werten vereinbaren können. Wertfragen im Bereich ethisch kontroverser Praxisfelder und Interventionen werden teilweise durch Institutionen beantwortet, in denen Evaluatorinnen und Evaluatoren tätig sind (z. B. Leitbilder von Forschungseinrichtungen, die z. B. eine Kooperation mit der Rüstungsindustrie verbieten), meist sind sie jedoch eine Frage der persönlichen Haltung.

18.3.2 Konzeptionsphase der Evaluation

Wurden mögliche Evaluationsprojekte identifiziert, deren Zielsetzungen machbar und verantwortbar erscheinen, so müssen im nächsten Schritt die Evaluationsbedingungen genauer exploriert und das Evaluationskonzept bzw. Projektangebot entwickelt werden. Dabei ist es notwendig sicherzustellen, dass die gängigen Evaluationsstandards von Anfang an und über den weiteren Evaluationsprozess hinweg eingehalten, das Projektmanagement vorbereitet, und verbindliche Vereinbarungen zwischen Evaluierenden und Auftraggebenden getroffen werden.

Exploration der Evaluationsbedingungen

▪ Evaluationsgegenstand und Stakeholder

Ein Evaluationskonzept kann nur dann adäquat vorgeschlagen werden, wenn die Merkmale der Intervention und ihrer Zielgruppen, die Bedingungen am Evaluationsort sowie die Erwartungen des Auftraggebers möglichst genau bekannt sind. Insbesondere wenn ein konkurrenzfähiges Angebot für eine Projektausschreibung erarbeitet werden soll, ist es angeraten, Hintergrundinformationen zu recherchieren. Abgesehen von publizierten Materialien sind v. a. informelle Gespräche wertvoll und notwendig, um die „eigentlichen" Intentionen der Auftraggebenden oder absehbare Probleme und Konflikte zu eruieren. Inhaltliche Vorklärungen sollten v. a. folgende Aspekte betreffen: Identifikation der relevanten Stakeholder (z. B. Entwickler, Umsetzer und Nutzer der Intervention) und ihrer jeweiligen Fragen, Bedürfnisse und Vorbehalte sowie die Möglichkeiten der Datenerhebung im Feld (z. B. Orte für Beobachtungen, Zeitpunkte für Befragungen, Zugriff auf Dokumente etc.). Hilfreich kann es auch sein, sich im Vorfeld genau zu überlegen, welche Personen, Gruppen oder Institutionen durch die zu untersuchende Intervention und/oder durch die Evaluation möglicherweise beeinträchtigt werden, um hier entsprechend Rücksprache zu halten oder Vorkehrungen zu treffen.

▪ Evaluationsteam

Frühzeitige Überlegungen zur Zusammensetzung und Rekrutierbarkeit des Evaluationsteams sind notwendig, da Angaben zum Team für das Projektangebot benötigt werden und potenzielle Teammitglieder zur Vorbereitung des Projektes beitragen können. Je nachdem, ob potenzielle Teammitglieder extern rekrutiert werden (z. B. Evaluationsexperten einer fremden Universität für die Metaevaluation) oder von der eigenen Organisation stammen (z. B. Mitarbeiter einer Evaluationsagentur) sind Vorabsprachen mit den Betroffenen sowie mit der Leitungsebene geboten (zu Anforderungen an Evaluierende ► Abschn. 18.1.10).

Entwicklung des Evaluationskonzeptes

Das Evaluationskonzept („evaluation design") muss darlegen, wie die Evaluationsziele des Auftraggebers im vorgegebenen zeitlichen und finanziellen Rahmen und mit hoher wissenschaftlicher Qualität erreicht werden und die Evaluationsfragen beantwortet werden sollen. Das Evaluationskonzept ist das Kernstück eines Projektangebotes. Bei der Ausarbeitung des Evaluationskonzeptes kann man sich an erfolgreichen Konzepten aus anderen Projekten des fraglichen Evaluationsbereiches orientieren (Zugriff über Publikationen, Internet, Kollegennetzwerke). Zudem

existieren Checklisten, die die einzelnen Arbeitsschritte der Evaluationskonzeptentwicklung beschreiben, etwa die „Evaluation Design Checklist“ (Stufflebeam & Shinkfield, 2007, S. 493ff.).

Das Evaluationskonzept soll die Auftraggeberin bzw. den Auftraggeber überzeugen. Darüber hinaus muss es im Falle eines Projektzuschlages die Projektabwicklung begleiten und die Projektdarstellung unterstützen. Wichtig ist es deswegen, das Konzept sachgerecht und anschaulich zu charakterisieren. Hierfür bietet sich ein Rückgriff auf die etablierten Evaluationsmodelle an (► Abschn. 18.2).

Das Evaluationskonzept muss gemäß dem Phasenmodell der Evaluation Aussagen zu folgenden Aspekten treffen (◘ Tab. 18.1; Stufflebeam & Shinkfield, 2007, S. 493ff.):

1. Evaluationsziele (Grobziele und Feinziele),
2. Forschungsfragen und/oder Hypothesen, ggf. auch Theoriebildung,
3. Untersuchungsplanung (mit Bezug zu Evaluationsmodellen und Evaluationsstandards),
4. Projektmanagement (Zeitplan, Kostenplan, Projektorganisation),
5. Datenerhebung und Datenanalyse,
6. Art und Umfang der Evaluationsberichte,
7. Maßnahmen zur Evaluationsnutzung,
8. Metaevaluation.

Da das Evaluationskonzept vor Vertragsabschluss erstellt wird, ist auf Seiten der Evaluierenden hier Vorarbeit zu leisten, die sich möglicherweise nicht auszahlt. Dementsprechend können Angaben im Evaluationskonzept – etwa zur Untersuchungsplanung – zunächst nur kursorisch erfolgen. Eine detaillierte Ausarbeitung erfolgt dann erst in der Planungsphase nach Vertragsabschluss.

Einzelne Evaluationsfachleute sowie Evaluationsagenturen bzw. -institute haben sich in der Regel auf die Umsetzung eines bestimmten Evaluationsmodells spezialisiert, so dass sie dieses Modell dann auf die jeweilige Konzeptentwicklung anwenden und somit ihren Arbeitsaufwand reduzieren (Selbstvorstellungen von Evaluationseinrichtungen und den von ihnen präferierten Evaluationsmodellen finden sich z. B. regelmäßig in der Zeitschrift für Evaluation). Ein solches arbeitsökonomisches Vorgehen ist jedoch nur gerechtfertigt, wenn das Evaluationsmodell im konkreten Fall wirklich geeignet ist, um die Evaluationsfragen zu beantworten.

Einbeziehung der Evaluationsstandards

Die Evaluationsstandards (► Abschn. 18.1.9) sind in allen Phasen der Evaluation zu berücksichtigen. Im vorliegenden Modell werden die Evaluationsstandards hier in der Konzeptionsphase angesprochen, um zu verdeutlichen, dass Standards der ethischen und methodischen Strenge von Anfang an zu bedenken sind.

Bereits im Vorfeld einer Evaluationsstudie ist zu klären, wie die Einhaltung der Evaluationsstandards während des Evaluationsprojektes sichergestellt und bei Bedarf nach Abschluss der Evaluation auch zusammenfassend bewertet werden soll (z. B. in Form einer Metaevaluation, ► Abschn. 18.3.5 „Durchführung einer Metaevaluation“). Im Falle von **Selbstevaluationen** sind zusätzlich zu den vier allgemeinen Qualitätsnormen zusätzliche **Muss-Standards** einzuhalten, die u. a. für gesicherte Rahmenbedingungen und einen Schutz der – typischerweise abhängig beschäftigten – Selbstevaluatoren sorgen sollen (Deutsche Gesellschaft für Evaluation e. V. (DeGEval), 2004; Müller-Kohlenberg & Beywl, 2003). Es ist empfehlenswert im Evaluationskonzept ausdrücklich darauf einzugehen, an welchen Standards sich das geplante Evaluationsprojekt vordringlich orientieren soll. So steht die Anforderung der Sparsamkeit des Vorgehens in Konkurrenz mit der Anforderung der Genauigkeit der Ergebnisse. Für eine ausführliche Checkliste dazu, wie Evaluationsstandards in allen Phasen der Evaluation berücksichtigt werden können und sollen Widmer (2005).

Vorbereitung des Projektmanagements

Im Zuge der Konzeption einer Evaluationsstudie geht es nicht nur um eine Grobplanung des wissenschaftlichen Vorgehens, sondern auch um eine Grobplanung der operativen Umsetzung im Rahmen des Projektmanagements. Das sog. „magische Dreieck“ des Projektmanagements besteht aus

- **Zeit** der Projektbearbeitung bis zum Schlusstermin,
- **Kosten** für alle benötigten Ressourcen und
- **Leistungen** (Quantität und Qualität gemäß Projektzielen).

Alle 3 Komponenten sind wechselseitig voneinander abhängig. Soll etwa die Projektzeit verkürzt werden, so werden sich die Leistungen verringern und/oder die Kosten (z. B. durch zusätzliches Personal) erhöhen. In der Konzeptionsphase erfolgt eine Projektgrobplanung v. a. hinsichtlich Zeit und Kosten. In der Planungsphase wird dann eine Feinplanung vorgenommen.

Bei ausgeschriebenen Evaluationsprojekten ist die Projektlaufzeit meist vorgegeben, während die Projektkosten in der Konzeptionsphase abzuschätzen sind. Zur Vorbereitung des Zeitmanagements im Evaluationsprojekt wird in der Regel in der Konzeptionsphase bereits ein grober **Zeitplan** mit den wichtigsten Zwischenzielen bzw. **Meilensteinen** („milestones“) entwickelt, der dem Projektangebot beigelegt wird. In den Milestones sollten die **Deliverables** (Arbeitsergebnisse) genau definiert werden: Was und in welcher Form erhält der Auftraggeber wann

als Ergebnis (schriftlicher Bericht; mündliche Präsentation mit Handout; Tabellen mit statischen Ergebnissen; Original-Datensatz etc.). Bezüglich des Zeitmanagements ist insbesondere auch eine Verzahnung des Interventionsprozesses mit dem Evaluationsforschungsprozess notwendig, etwa für die Terminierung von Vorher-Messungen sowie zur Sicherstellung rechtzeitiger Ergebnisrückmeldungen.

Aus Sicht des Auftraggebers sollte der Aufwand für die Evaluation in vernünftiger Relation zum Aufwand der Intervention stehen. Bei der Entwicklung von Interventionen sind laut Literatur ca. 5 % des Projektbudgets für die Selbstevaluation zu veranschlagen; für externe Evaluationen werden Ausgaben in Höhe von rund 10–15 % der Kosten für die Intervention angegeben (Bundesamt für Gesundheit [BAG], 1997, S. 24). In der Praxis fallen die Evaluationsbudgets oft deutlich kleiner aus. Die Evaluationsbranche ist gehalten, den (auch monetären) Wert ihrer Arbeit für die Auftraggebenden stärker zu verdeutlichen, um dem zuweilen anzutreffenden Vorurteil entgegenzuwirken, Evaluation sei eine eigentlich überflüssige Pflichtübung oder eine Art Luxus.

Das **Evaluationsbudget** umfasst Personalkosten, Betriebskosten (z. B. Spesen, Sekretariatskosten), Sachkosten (z. B. Druck- und Versandkosten, Übersetzungskosten) sowie – wenn notwendig – Investitionskosten (z. B. Apparate, EDV). Typischerweise müssen Bieter bei öffentlichen Ausschreibungen Fixkostenangebote unterbreiten. Um die für die Forschungsleistung notwendigen Kosten abzuschätzen, ist anhand der Evaluationsziele und des Evaluationskonzeptes ein **Ressourcen- und Kostenplan** zu erstellen. Dazu werden zunächst die im Projekt anfallenden Aufgaben identifiziert (Projektstrukturplan). Für die einzelnen Aufgaben wird dann angegeben, welche Ressourcen sie erfordern (Personal, Geräte etc.), woraus dann die Kosten abgeleitet werden. Das Projektmanagement ist bei größeren Evaluationsprojekten neben den Forschungsaufgaben als eigene Aufgabe einzubeziehen (z. B. Kosten für die Beschäftigung einer Projektleiterin oder eines Projektleiters, Reisekosten etc.). Für die Erstellung von Kostenplänen existieren im Projektmanagement entsprechende Methoden und Tools (Kuster et al., 2006, S. 124), zudem liegen spezialisierte Checklisten zur Erstellung von Kostenplänen für Evaluationsprojekte vor. Dabei wird dann auch nach unterschiedlichen Finanzierungsmodellen differenziert (Stufflebeam & Shinkfield, 2007, S. 522ff.).

In das Konzept eines Evaluationsforschungsprojektes sind auch Angaben zur **Projektorganisation** aufzunehmen. Dies betrifft die Zusammenstellung des Evaluationsteams, die Kompetenzen der Teammitglieder sowie besondere Rollen (z. B. Projektleitung, Teilprojektleiter) und Gremien (z. B. Stakeholder-Beirat).

Für das **Kommunikationsmanagement** ist zu klären, wann, wie oft und über welche Kommunikationskanäle mit welchen Zuständigkeiten sich die einzelnen Stakeholder im Verlauf der Evaluation untereinander und mit dem Evaluationsteam austauschen und aktiv am Evaluationsprozess beteiligen. Weitergehende Fragen des projektübergreifenden Wissensmanagements in Evaluationsprojekten werden von Hense und Mandl (2011) diskutiert.

Abschluss des Evaluationsvertrags

Erhält ein Projektangebot den Zuschlag oder nimmt man einen Projektauftrag an, so folgt der Abschluss des **Evaluationsvertrags** („evaluation contract"), der die Rechte und Pflichten von Auftraggeber/in und Auftragnehmer/in fixiert. Insbesondere bei umfangreichen Projekten sind die rechtlichen Bedingungen genau zu prüfen, etwa die Zahlungsmodalitäten, Möglichkeiten des Vertragsrücktritts oder Konsequenzen bei Nichteinhaltung einzelner vertraglicher Regelungen (für eine Checkliste zur Vorbereitung von Projektverträgen Kuster, et al., 2006). Sofern sich während eines Evaluationsprojektes bedeutsame Änderungen im Evaluationsmandat ergeben, sollten diese ebenfalls vertraglich festgehalten werden. Bei internen Evaluationen oder selbst finanzierten Evaluationskooperationen sollte ebenfalls ein Vertrag oder eine Evaluationsvereinbarung schriftlich abgeschlossen werden, um Verbindlichkeit und Transparenz für alle Beteiligten sicherzustellen. Nicht juristisch bindend, aber selbstverpflichtend ist eine gemeinsame **Absichtserklärung** („memorandum of understanding"), in welcher die Konditionen der Zusammenarbeit festgehalten werden.

Aus Sicht der Evaluierenden ist es besonders wichtig zu prüfen, dass sie ihre vertraglich vereinbarten Leistungen im vorgesehenen Umfang termingerecht erbringen können. Umgekehrt müssen sie auch darauf achten, dass die Pflichten des Auftraggebers spezifiziert sind, dies betrifft etwa die zeitgerechte Lieferung bestimmter Informationen, den Zugang zu Dokumenten und Erhebungsorten, Unterstützung bei der Rekrutierung von Befragungspersonen, ggf. Bereitstellung von Arbeitsplätzen im Feld etc. (für eine Checkliste zur Erstellung von Evaluationsverträgen und Evaluationsvereinbarungen Stufflebeam & Shinkfield, 2007, S. 547f.).

18.3.3 Planungsphase der Evaluation

Nachdem der Evaluationsvertrag abgeschlossen wurde, stehen die Detailplanung und konkrete Vorbereitung der Evaluationsstudie an. Dies umfasst die Formulierung von detaillierten Evaluationsfragen bzw. -hypothesen, die Untersuchungsplanung inklusive Untersuchungsdesign, Stichprobenplänen und Instrumenten sowie die Projekt-

planung. Im Sinne der Evaluationsstandards der Nützlichkeit und Durchführbarkeit ist in der Planungsphase wiederum ein enger Kontakt zu den Stakeholdern notwendig, ebenso eine enge Rückkopplung mit den Abläufen im Praxisfeld. Andernfalls besteht die Gefahr, dass „am grünen Tisch" ein Studienentwurf ausgearbeitet wird, der rein organisatorisch nicht umsetzbar ist und/oder am Widerstand derjenigen Stakeholder scheitert, die nicht von Anfang an genügend einbezogen wurden (Goldfried, 1991).

Formulierung detaillierter Evaluationsfragen und -hypothesen

Bereits in der Initiations- und Konzeptionsphase werden die Evaluationsziele und zentralen Evaluationsfragen festgelegt. In der Planungsphase sind Evaluationsfragen weiter zu **konkretisieren** und v. a. auch **einzugrenzen**. Evaluationsfragen können sich – in Abhängigkeit von der Entwicklungsphase der evaluierten Intervention – auf alle Komponenten des CIPP-Modells von Stufflebeam (Stufflebeam & Shinkfield, 2007, ► Abschn. 18.2.2) beziehen. Dabei ist es in der Regel kaum möglich, alle Bereiche abzudecken, sondern es gilt, eine Auswahl der für die Evaluationsziele wichtigsten Einzelfragen bzw. Hypothesen zu treffen. Das systemische CIPP-Modell wird hier herangezogen, weil es sich gut zur Dimensionsanalyse (► Abschn. 18.2.2) des Evaluationsgegenstandes und zur Generierung und Schärfung von Evaluationsfragen und -hypothesen eignet. Entsprechende Fragen sind jedoch auch in Evaluationen auf der Basis anderer Evaluationsmodelle relevant bzw. entsprechende Modelle bieten weitere Heuristiken zur Formulierung von Evaluationsfragen und -hypothesen.

Context

- Welche Personengruppen haben Bedarf an einer geplanten (oder bereits realisierten) Intervention?
- Welche vergleichbaren Maßnahmen existieren bereits?
- Welche Bedürfnisse und Besonderheiten haben die Adressaten der Intervention?
- Wie lassen sich Mittel zur kurz- und langfristigen Finanzierung der Intervention beschaffen?

Input

- Wie gut entspricht das Konzept der Intervention den Bedürfnissen und Besonderheiten der Zielgruppen?
- Welche Alleinstellungsmerkmale hat das Konzept im Vergleich zu alternativen Interventionen?
- Wie kostenintensiv wird die Umsetzung des Konzeptes?
- Wie berücksichtigt das Konzept mögliche Widerstände gegen seine Umsetzung?
- Wie viel Personal steht zur Umsetzung der Maßnahme zu welchen Zeitpunkten zur Verfügung und welche Qualifikationen bringt es mit?

Process

- Welche Komponenten der Intervention werden in welchem Maße tatsächlich realisiert?
- In welchen Situationen bzw. an welchen Orten wird die Intervention wie intensiv genutzt?
- Welche Probleme treten bei der Umsetzung der Intervention auf?
- Wie kann die Intervention verbessert werden?

Product

- Wie groß ist die Nutzergruppe der Intervention? In welchem Verhältnis steht die aktuelle Nutzergruppe zur Zielgruppe? **(Impact) –** Spezieller: Ist die Intervention Menschen mit körperlichen Beeinträchtigungen zugänglich? **(Barrierefreiheit) –** Ist die Intervention von Menschen mit unterschiedlichem kulturellem Hintergrund gleichermaßen nutzbar bzw. auf andere Länder übertragbar? **(Internationalisierung)**
- Erreicht die Intervention die intendierten Wirkungen? In welchem Umfang werden die intendierten Wirkungen erzielt? **(Effektivität) –** Spezieller: Gibt es differenzielle Wirkungen in unterschiedlichen Gruppen? Profitieren Frauen und Männer gleichermaßen von der Intervention oder wird eine Gruppe benachteiligt? **(Gender Mainstreaming)**
- In welchem Verhältnis stehen Kosten und Nutzen der Intervention? Erreicht die Intervention erwünschte Ziele kostengünstiger als andere Interventionen? **(Effizienz)**
- Welche Vorkehrungen werden getroffen, um die Intervention an anderen Orten/für andere Zielgruppen anzubieten? **(Portabilität)**
- Welche Vorkehrungen werden getroffen, um die Intervention langfristig oder dauerhaft zu erhalten? Welchen Status hat die Intervention z. B. 2 Jahre nach Abschluss des ersten Evaluationsprojektes? **(Nachhaltigkeit)**

Die Formulierung der detaillierten Evaluationsfragen sollte auf die vom Auftraggeber gestellten Anforderungen zugeschnitten sein, wobei Evaluierende jedoch im Sinne der Evaluationsstandards auch dafür Verantwortung tragen, dass vom Auftraggeber ggf. vernachlässigte Aspekte dennoch in die Studie aufgenommen werden (z. B. Kosten-Nutzen-Analyse), damit eine ausgewogene und fundierte Qualitätsbewertung möglich ist.

Ein Teil der Evaluationsfragen lässt sich auch als **gerichtete Hypothesen** fassen, da implizit von der Wirkung der Intervention ausgegangen wird (z. B. statt der offenen

Frage „Erreicht die Intervention die intendierten Wirkungen?“ die gerichtete Hypothese „Die intendierte Wirkung X wird mit der Intervention in stärkerem Maße erreicht als ohne Intervention bzw. als mit einer alternativen Intervention.“).

In der Grundlagenforschung werden die untersuchten Fragestellungen und Hypothesen in der Regel aus einer umfassenden Theoriebildung abgeleitet. In der Evaluationsforschung werden Evaluationsfragen und -hypothesen dagegen oft ohne umfangreichen Theoriehintergrund eher pragmatisch – orientiert am Informationsbedarf von Auftraggebern und anderen Stakeholdern – formuliert. Es existieren jedoch auch Evaluationsansätze mit stärkerem theoretischen Erklärungsanspruch (zur theorieorientierten Evaluation ► Abschn. 18.2.3).

Im Sinne der Nützlichkeit einer Evaluationsstudie sollte bei jeder Evaluationsfrage nicht nur sichergestellt werden, dass entsprechende Antworten im Praxisfeld auf „Interesse“ stoßen, sondern auch hinterfragt und konkret geplant werden, wer die gewonnenen Ergebnisse in welcher Weise wozu nutzen kann und soll. Wird die intendierte Nutzung der Evaluationsergebnisse nicht ihrerseits von Anfang an mitbedacht und genau geplant, so wird die Evaluationsstudie ihren eigentlichen Zweck vermutlich verfehlen (zur nutzungsfokussierten Evaluation ► Abschn. 18.2.4 „Klientenorientierte Evaluationsmodelle“).

Untersuchungsplanung

Die Hauptelemente der Untersuchungsplanung sind Untersuchungsdesign, Stichprobenplan sowie Auswahl oder Entwicklung der Instrumente.

■ Untersuchungsdesign der Evaluationsstudie

Zur Kennzeichnung eines Untersuchungsdesigns sind eine Reihe von Kriterien heranzuziehen (► Kap. 7). Anhand des Untersuchungsortes sind **Labor- und Feldstudien** zu differenzieren, wobei die meisten Evaluationsstudien Feldstudien im Praxisfeld sind (z. B. in Unternehmen, in Bildungseinrichtungen, in Krankenhäusern etc.). Gemäß der Art der zu erhebenden Daten sowie der Datenauswertung wird zwischen einem **quantitativ-statistischen** und einem **qualitativ-interpretativen Ansatz** unterschieden, wobei in der Praxis eine Kombination von quantitativen und qualitativen Methoden gängig ist **(Methoden-Mix, Mehrmethodenansatz)**.

Es ist zudem zwischen **Querschnittstudien** (ohne Messwiederholung) und Längsschnittstudien (mit Messwiederholungen) zu unterscheiden. In Evaluationsstudien sind Querschnittstudien als **Treatmentgruppe-Kontrollgruppe/n-Vergleiche** einschlägig, etwa um herauszufinden, wie sich eine mit der Intervention behandelte Gruppe von einer nichtbehandelten Gruppe und/oder einer alternativ (bzw. herkömmlich) behandelten Gruppe unterscheidet. **Längsschnittstudien** sind dagegen geeignet, um Veränderungen nachzuvollziehen. Sie sind als **Vorher-Nachher-Vergleiche** in vielen Evaluationsstudien realisiert. Besonders aussagekräftig ist eine Kombination beider Ansätze, indem sowohl Treatment- als auch Kontrollgruppe/n vorher und nachher untersucht werden. Ein solches Design kann zusätzlich aufgewertet werden, indem mit zeitlichem Abstand noch ein dritter Messzeitpunkt einbezogen wird (Follow-Up-Erhebung), der mittel- oder langfristige Wirkungen erfasst. Ein **Vortest-Nachtest-Follow-Up-Vergleichsgruppen-Plan** ist somit im Sinne hoher Erklärungskraft besonders empfehlenswert (Hager, Patry, & Brezing, 2000, S. 183).

Während in **echten experimentellen Studien** die Untersuchungsteilnehmenden per Zufall den Untersuchungs- und Kontrollbedingungen zugeordnet werden (Randomisierung), wird in **quasi-experimentellen Studien** mit vorgefundenen oder willkürlich eingeteilten Gruppen gearbeitet, wobei diese zumindest hinsichtlich zentraler sozialstatistischer sowie interventionsrelevanter Variablen vergleichbar sein sollten. Will man z. B. den Effekt des Notebookeinsatzes in Schulen evaluieren (Schaumburg, 2001), so wird man typischerweise eine Klasse mit Notebooks ausstatten (Treatmentgruppe) und eine andere nicht (Kontrollgruppe), um dann Unterschiede zu erfassen. Hier müssen also die **vorgefundenen Klassenverbände** als Untersuchungsgruppen dienen, weil zum Zweck der Evaluation kaum eine im Sinne des Erkenntnisgewinns optimale zufallsgesteuerte Neubildung von Schulklassen praktikabel wäre. Ein weiteres Evaluationsdesign zur Überprüfung der Wirksamkeit einer Intervention ist der **Ist-Soll-Vergleich,** dabei wird die behandelte Gruppe anhand vorher festgelegter Soll-Werte eingeschätzt. So werden im Rahmen der PISA-Studie die Kompetenzen von 15-jährigen Schülerinnen und Schülern (ca. 10. Klassenstufe) erfasst und im Sinne einer normorientierten Bewertung an internationalen Vergleichswerten relativiert. Auf dieser Basis wird u. a. die Qualität nationaler Schulsysteme diskutiert.

Um die Wirksamkeit einer Intervention überzeugend nachweisen zu können (Kriterium der **internen Validität**; ► Abschn. 3.2.2), ist es wichtig, dass die im Vorher-Nachher-, Treatment-Kontrollgruppen- oder Ist-Soll-Vergleich gefundenen Effekte (Ausprägungen der **abhängigen Variablen**) mit hoher Sicherheit kausal auf die Intervention **(unabhängige Variable)** zurückzuführen sind. Dies ist v. a. bei experimentellen Kontrollgruppenstudien der Fall. Die interne Validität ist gering, wenn die Effekte durch andere Faktoren wie z. B. Umweltbedingungen oder bereits vorliegende Unterschiede zwischen den Untersuchungsgruppen erklärbar sind. Nichtexperimentelle Studien wie z. B. die PISA-Studie, die Kompetenzun-

terschiede zwischen Schülern verschiedener Länder aufzeigt, erlaubt keine eindeutigen kausalen Rückschlüsse. Denn die Schülergruppen der verschiedenen Länder unterscheiden sich nicht nur in ihrer Beschulung, sondern auch in zahlreichen anderen Erziehungs- und Sozialisationsfaktoren, die nur teilweise durch Kontrollvariablen erfasst und herausgerechnet werden können. Bei quasiexperimentellen Studien ist mangels randomisierter Gruppeneinteilung die interne Validität stark eingeschränkt, sofern die relevanten Störvariablen nicht erkannt und kontrolliert werden (z. B. Berücksichtigung von Kontrollvariablen wie z. B. Computererfahrung, Schulleistung, Lehrermotivation etc. im Notebook-Klassen-Beispiel). Dementsprechend müssen gute Evaluationsstudien diesen Problemen schon bei der Designentwicklung mit entsprechenden Methoden entgegenwirken. Außerdem müssen die Erkenntnisgrenzen im Ergebnisbericht klar benannt werden, um ungesicherten übergeneralisierten Kausalaussagen entgegenzuwirken.

Ein wichtiges Evaluationsziel besteht häufig darin, nicht nur die intendierten Wirkungen der Intervention nachzuweisen, sondern auch nichtintendierte Wirkungen. Dementsprechend breit muss das Spektrum der erfassten abhängigen Variablen sein. Insbesondere beim Einsatz standardisierter Methoden besteht die Gefahr, dass nichtantizipierte Effekte übersehen werden. Zum Aufspüren nichtintendierter Maßnahmenwirkungen sind deswegen qualitative Methoden meist unverzichtbar. Das Evaluationsdesign sollte mit den Auftraggebenden abgestimmt werden. Im Rahmen stakeholderorientierter Evaluationsmodelle (▶ Abschn. 18.2.4) wird auch eine Abstimmung des Designs mit weiteren Beteiligten und Betroffenen empfohlen (z. B. im Rahmen einer Sitzung des Stakeholder-Beirates), was aber oft an praktische Grenzen stößt. Zudem sollte antizipiert werden, wie das Design ggf. zu modifizieren ist, falls sich auf Seiten der Intervention Veränderungen ergeben. Ein gutes Evaluationsdesign sollte sich nicht nur durch hohe interne Validität, sondern auch durch **Robustheit** auszeichnen, d. h. so ausgelegt sein, dass selbst unter suboptimalen Datenerhebungsbedingungen (z. B. wegen unvorhergesehener Veränderungen der Feldbedingungen oder Wegfall einzelner Erhebungsmöglichkeiten) noch nützliche Resultate erzielt werden können.

Stichprobenplan der Evaluationsstudie

Die Datenerhebung im Rahmen einer Evaluationsstudie kann sich auf unterschiedliche Merkmalsträger beziehen (z. B. Personen, Situationen, Objekte, Ereignisse etc.). Personen sind bei psychologischen Interventionen die zentralen Merkmalsträger. Bei der Auswahl und Zusammensetzung der **Evaluationsstichprobe** sind Interventionspopulation und Interventionsstichprobe wichtige Bezugsgrößen.

Die Gesamtheit aller Personen, an die sich eine Intervention richtet, ist die **Interventionspopulation**. In der Regel wird eine Intervention nicht von der gesamten Zielgruppe, sondern nur von einer Teilgruppe genutzt, die als **Interventionsstichprobe** bezeichnet wird. Die Höhe der **Ausschöpfungsquote** der Intervention (Größe der Interventionsstichprobe relativiert am Umfang der Zielgruppe) kann dabei ein Qualitätskriterium der Maßnahme sein. Ein weiterer Erfolgsindikator einer Maßnahme ist die **Ausschöpfungsqualität der Intervention** (AQ) mit einem Idealwert von 100 (Rossi & Freeman, 1993):

$$\mathrm{AQ} = 100 \cdot \left[\frac{\text{Anzahl erreichte Zielobjekte}}{\text{Anzahl aller Zielobjekte}} - \frac{\text{Anzahl unbefugte Teilnehmer}}{\text{Anzahl aller Teilnehmer}} \right] \tag{18.1}$$

Je größer die Anzahl der „unbefugten" Teilnehmer, umso geringer die Ausschöpfungsqualität. „Unbefugte" Teilnehmer können Personen sein, die eine Maßnahme nutzen oder nutzen müssen, ohne sie wirklich zu benötigen (z. B. an einer unternehmensinternen Trainingsmaßnahme nehmen Mitarbeiter teil, die die Inhalte bereits beherrschen oder die die Inhalte in ihrem Tätigkeitsfeld gar nicht benötigen). Wenn z. B. von 1 000 Wohnungslosen 600 an einem Winterhilfsprogramm teilnehmen und außerdem 200 Personen, die der Zielgruppendefinition nicht genügen, resultiert hieraus eine Ausschöpfungsqualität von

$$100 \cdot \left[\frac{600}{1\,000} - \frac{200}{800} \right] = 35.$$

Ein Beispiel für eine negative Ausschöpfungsqualität ist die Kinderserie „Sesamstraße", die ursprünglich für Kinder mit Lernschwächen vorgesehen war, aber überwiegend von „unbefugten" Kindern genutzt wurde (Cook et al., 1975).

Voraussetzungen für eine hohe Ausschöpfungsqualität der Intervention sind v. a. eine praktikable, trennscharfe Zielgruppendefinition sowie eine gut durchdachte Strategie zur zielgruppengerechten Umsetzung der Maßnahme. Die Evaluationsstichprobe basiert oft auf der Interventionsstichprobe: D. h., von allen Teilnehmenden an der Maßnahme (Interventionsstichprobe) wird eine Auswahl (Evaluationsstichprobe) untersucht. Man beachte, dass eine schlechte Ausschöpfungsqualität der Interventions- und somit auch der Evaluationsstichprobe die externe Validität der Evaluationsstudie erheblich verringert. Die Ausschöpfungsqualität ist natürlich auch unter finanziellen Gesichtspunkten für Evaluationen nicht unerheblich.

Erreicht eine Maßnahme auch „unbefugte" Personen, sind dies letztlich Fehlinvestitionen, die auf Fehler bei der Implementierung der Maßnahme schließen lassen. Dass Fehlinvestitionen dieser Art nicht immer negativ zu Buche schlagen, belegt das oben erwähnte Sesamstraßen-Beispiel, bei dem sich die Tatsache, dass das Programm auch von Kindern ohne Lernschwächen gesehen wird, nicht auf die Kosten der Programmerstellung auswirkte.

In der Terminologie der medizinischen Epidemiologie (einführend hierzu Bortz & Lienert, 2003, S. 237ff.) kann die Stichprobenausschöpfung der Intervention auch über folgende Kennziffern charakterisiert werden:

- **Falsch-positiver Wert:** Dieser Wert entspricht der Anzahl der unbefugten Personen, die an der Maßnahme teilnahmen. Diese Zahl wird an der Anzahl aller Unbefugten relativiert.
- **Falsch-negativer Wert:** Dieser Wert entspricht der Anzahl der befugten Personen, die an der Maßnahme nicht teilnahmen. Diese Zahl wird an der Anzahl aller Befugten relativiert.
- **Sensitivität:** Dieser Wert entspricht der Anzahl der befugten Personen, die auch an der Maßnahme teilnahmen. Er wird relativiert an der Anzahl aller befugten Personen.
- **Spezifität:** Dieser Wert entspricht der Anzahl der unbefugten Personen, die tatsächlich auch nicht an der Maßnahme teilnahmen. Sie wird an der Anzahl aller unbefugten Personen relativiert.

Im oben genannten Beispiel (Winterhilfsprogramm) ergibt sich ein falsch-negativer Wert von $400/1\,000 = 0.4$ und eine Sensitivität von $600/1\,000 = 0.6$. Über den falsch-positiven Wert und über die Spezifität können im Beispiel keine Aussagen gemacht werden, weil die Prävalenzrate (Anteil aller befugten Personen an der Gesamtpopulation) unbekannt ist.

Wichtig ist in diesem Zusammenhang der Hinweis, dass sich Ausfälle oder Teilnahmeverweigerungen (Noncompliance) in (randomisierten) Experimentalkontrollgruppenuntersuchungen negativ auf die Teststärke der Hypothesentests auswirken (zur Teststärke ▶ Abschn. 14.1). Wie Verweigerungsraten das Studiendesign und kovariate Kontrollvariablen die Teststärke beeinflussen, wird von Jo (2002) untersucht. Im Mittelpunkt der Überlegungen steht ein Vergleich von **ITT** („intent to treat") und **CACE** („complier average causal effect"). Im ITT-Ansatz werden eine randomisierte Experimental- und Kontrollgruppe verglichen, ohne zu überprüfen, ob die Untersuchungsteilnehmer der Experimentalgruppe das Treatment bzw. die Maßnahme tatsächlich erhalten haben. Wenn z. B. eine Gesundheitsbehörde die Gesamteffektivität einer Impfkampagne gegen Grippe in Erfahrung bringen möchte, werden alle Personen, für die die Impfung vorgesehen war (Experimentalgruppe), verglichen mit allen Personen, für die die Impfung nicht vorgesehen war. Im Unterschied hierzu werden im CACE-Ansatz nur diejenigen Experimentalgruppenuntersuchungsteilnehmer mit der Kontrollgruppe verglichen, die das Treatment tatsächlich auch erhalten haben. CACE ist ITT vorzuziehen, wenn es um den Nachweis der Wirksamkeit einer Behandlung (Medikamente, medizinische oder psychologische Techniken, Therapien etc.) geht.

Selten wird man die regulären Maßnahmenteilnehmenden vollständig in die Evaluation einbeziehen (Vollerhebung), sondern sich auf Auswahlen (Stichproben) beschränken. Für die Zusammenstellung von Evaluationsstichproben ist es unerlässlich, Personen aus der Interventionsstichprobe zu rekrutieren. Eine Evaluationsstichprobe lässt sich nur in Ausnahmefällen als **echte Zufallsstichprobe** aus der Grundgesamtheit aller interessierenden Objekte realisieren, stattdessen sind nichtprobabilistische bzw. **willkürliche Stichproben** gängig (z. B. Ad-hoc-Stichprobe bei einer Veranstaltung, Teilnehmerrekrutierung im Schneeball-Prinzip, Zusammenstellung einer Quotenstichprobe von Freiwilligen; ▶ Kap. 9). Willkürliche Stichproben sind leichter realisierbar, haben aber den Nachteil, dass aufgrund von Verzerrungen ihre statistische Repräsentativität fraglich bzw. unklar ist. Bei der Planung von Evaluationsstichproben ist jeweils zu berücksichtigen, dass die Zielgruppe der Intervention möglichst ohne systematische Verzerrungen (z. B. Ausschluss ganzer Stakeholder-Gruppen, Vernachlässigung von Wenignutzern etc.) erfasst wird. Bei der Evaluation des Bildschirmtext-Feldversuchs in Deutschland wurden damals mehrere Stichproben kombiniert, um Auswahlverzerrungen zu reduzieren (Kromrey, 1987).

Um nicht-intendierte Wirkungen der Intervention zu erfassen, kann es sinnvoll sein, auch die Gruppe der unbefugt Teilnehmenden genauer zu beleuchten. Zudem werden für Evaluationsstichproben auch Mitglieder der Interventionszielgruppe benötigt, die nicht (**non-user**) oder nicht mehr (**drop out**) an der Maßnahme teilnehmen, da diese Gruppen Hinweise auf Zugangsbarrieren oder Nutzungsprobleme liefern können. Mitglieder der Zielgruppe, die nicht an der Maßnahme teilnehmen, werden zudem für Kontrollgruppendesigns benötigt. Abgesehen von den Maßnahmenteilnehmern sind insbesondere für Evaluationen in frühen Phasen der Interventionsentwicklung auch diejenigen Personen relevante Informanten, die an der Konzeption und Umsetzung der Intervention mitwirken (z. B. Lehrer, Trainer, Psychotherapeuten).

Für aussagekräftige inferenzstatistische Analysen ist jeweils von minimalen bzw. **optimalen Stichprobengrößen** auszugehen, die umso größer angelegt sein müssen, je kleiner der erwartete Effekt ist (zum optimalen Stichpro-

benumfang ▶ Abschn. 14.3). Um etwa die Wirksamkeit einer auf Jugendliche zugeschnittenen Primärpräventionsmaßnahme im Bereich illegaler Drogen statistisch nachzuweisen, wären – aufgrund der geringen Verbreitung des Konsums illegaler Drogen – Stichprobengrößen im 6-stelligen Bereich notwendig (Uhl, 1997). Wie mit statistischen Teststärkeproblemen bei geringen Stichprobenumfängen umzugehen ist, wird in ▶ Abschn. 14.3 erläutert.

In qualitativen Evaluationsstudien werden verhältnismäßig wenige Informantinnen und Informanten sehr ausführlich untersucht (z. B. durch Feldbeobachtung oder Leitfaden-Interview). Anstelle von zufälligen oder zufalls-ähnlichen Auswahlverfahren sind hier *theoretische Stichprobenkonstruktionen* einschlägig, bei denen bewusst besonders interessante oder kontrastierende Fälle herausgesucht werden (zu verschiedenen Stichprobentypen ▶ Kap. 9).

▪ Instrumente der Evaluationsstudie

Die für die Evaluationsfragen und -hypothesen relevanten Merkmale (Eigenschaften der Intervention und der Interventionsumsetzung, Kosten und Nutzen der Maßnahme, positive und negative Nebenwirkungen etc.) gilt es angemessen zu definieren und durch passende quantitative und/oder qualitative Datenerhebungsmethoden zu erfassen. Als Datenerhebungsmethoden kommen in der Evaluationsforschung sämtliche **sozialwissenschaftlichen Methoden** zum Einsatz, das heißt v. a. Beobachtung inklusive der Analyse von Verhaltensspuren (nonreaktive Methoden), mündliche und schriftliche Befragung, Testung, physiologische Messungen und Dokumentenanalyse (▶ Kap. 10). Für diese Datenerhebungsmethoden sind jeweils konkrete **Instrumente** notwendig, seien es Beobachtungssysteme, Interviewleitfäden, Fragebögen, Tests, Checklisten oder Kategoriensysteme.

Sofern möglich, sollten in quantitativen Evaluationsstudien bzw. Studienteilen bereits erprobte und hinsichtlich ihrer wissenschaftlichen Gütemerkmale (Objektivität, Reliabilität, Validität) geprüfte Datenerhebungsinstrumente als Evaluationsinstrumente **ausgewählt** werden. Liegen diese nicht vor, so ist eine **Modifikation** vorhandener Instrumente angeraten. Die **Neukonstruktion** von standardisierten Erhebungsinstrumenten erlaubt eine optimale Anpassung an die Evaluationsfragen, ist jedoch recht aufwändig – insbesondere, da die Qualität der neu entwickelten Instrumente erst im Rahmen von Pretests ermittelt und gesichert werden muss (zu **den Standards von psychologischen und pädagogischen Testverfahren** ▶ American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2002). Abgesehen von der Forschungsökonomie ist eine Verwendung etablierter Instrumente oft auch deswegen zu präferieren, weil Evaluationsbefunde dann mit bereits publizierten Ergebnissen vergleichbar sind. Damit die Datenerhebung im Feld nicht störend wirkt, ist auf eine **ökonomische Abwicklung** besonderen Wert zu legen (z. B. durch kurze Instrumente und komfortable Bearbeitung).

Neben den allgemeinen sozialwissenschaftlichen Datenerhebungsmethoden kommen in Evaluationsstudien teilweise auch spezifische Methoden und Instrumente zum Einsatz, etwa wirtschaftswissenschaftliche Methoden der **Kosten-Nutzen-Rechnung** zur Effizienzprüfung ▶ Abschn. 18.2.1 „Effizienzorientierte Evaluationsmodelle“) oder das **Benchmarking-Verfahren** für einen systematischen Vergleich der Merkmale verschiedener Interventionen und der Identifikation von Best-Practice-Beispielen, an denen man sich orientieren kann (für den Einsatz von Benchmarking im Rahmen der Evaluation von Hochschulen Ziegele & Hener, 2004).

Projektplanung

Während bei der Untersuchungsplanung der angestrebte wissenschaftliche Erkenntnisgewinn bzw. die Evaluationsziele im Vordergrund stehen, konzentriert sich die Projektplanung auf die operative Umsetzung und v. a. auf die Zeit-, Kosten- und Arbeitsplanung. Die Grobpläne aus der Konzeptionsphase werden verfeinert und je nach Projektbedarf durch zusätzliche Pläne ergänzt (z. B. Qualifizierungsplan, Berichtswesenplan, Keßler & Winkelhofer, 2004).

18.3.4 Realisierungsphase der Evaluation

Auf der Basis der Ergebnisse der Planungsphase werden in der Realisierungs- oder Durchführungsphase die für die Evaluationsziele notwendigen Evaluationsdaten erhoben, analysiert und interpretiert, zudem wird der operative Projektablauf durch ein Projektcontrolling begleitet. Je nach Umfang und Dauer einer Intervention sowie der aktuellen Phase im Interventionslebenszyklus werden Erhebungs-, Analyse- und Interpretationsschritte mehrfach nacheinander oder auch parallel stattfinden, wobei es im Sinne eines **iterativen Vorgehens** oft Rücksprünge in die Planungsphase gibt.

Aus dem Satz der Evaluationsstandards sind bei der Datenerhebung v. a. Standards der Durchführbarkeit und Fairness relevant, während bei der Datenauswertung und Dateninterpretation der Schwerpunkt auf Standards der Genauigkeit liegt (Gesellschaft für Evaluation e. V. [DeGEval], 2008b).

Erhebung von Evaluationsdaten

Im Vergleich zur Grundlagenforschung handelt es sich bei der Evaluationsforschung deutlich häufiger um Feldforschung, so dass **Störungen der Praxis** im Zuge der Da-

tenerhebung zu vermeiden sind (z. B. durch ökonomische Instrumente, günstige Auswahl von Erhebungszeitpunkten, nonreaktive Datenerhebung etc.). Zudem sind Fragen des **Datenschutzes** besonders zu beachten, weil im Feld oftmals die Anonymität nicht gewährleistet werden kann und zudem Beteiligte vor möglichen negativen Konsequenzen geschützt werden sollen. Nicht zuletzt sind die bereichsspezifischen **sozialen Normen** und **rechtlichen Regelungen** (z. B. Arbeitsrecht) beim Agieren des Evaluationsteams im Feld zu berücksichtigen. Die geringe Kontrolle der Feldbedingungen gegenüber Laborbedingungen erfordert es auch, dass Evaluationsforschende darauf vorbereitet sind, ihr Design und ihre Instrumente bei Bedarf kurzfristig zu modifizieren, um veränderten Praxisbedingungen gerecht zu werden. Entsprechende Entscheidungen sind dabei mit dem Auftraggeber und anderen Stakeholdern abzustimmen.

Gerade bei großen Projekten, vielen verschiedenen Erhebungsorten und Erhebungszeitpunkten ist eine vollständige Dokumentation der Bedingungen der Datenerhebung besonders wichtig. Weiterhin ist zu beachten, dass sich die Datensammlung – je nachdem, ob eine Ergebnis-, Prozess- und oder Konzeptevaluation ansteht – eben nicht immer nur auf Wirkungen der Intervention bezieht, sondern Umsetzung und Konzeption der Intervention. Zur Unterstützung der Erhebung von Evaluationsdaten liegen Handreichungen und Checklisten vor, die verschiedene Datenquellen systematisieren (Stufflebeam & Shinkfield, 2007, S. 569ff.).

Aus Kostengründen wird in manchen Evaluationsstudien auf eine Primärdatenerhebung ganz oder teilweise verzichtet und stattdessen eine **Sekundäranalyse** von vorhandenem Datenmaterial durchgeführt (z. B. Controlling- und Marketingdaten eines Unternehmens).

Analyse von Evaluationsdaten

Je nach Art der erhobenen quantitativen und/oder qualitativen Daten sowie der interessierenden Evaluationsfragen sind jeweils **adäquate quantitative und/oder qualitative Datenauswertungsmethoden** zu wählen und korrekt anzuwenden (► Kap. 12). Im Falle einer formativen bzw. einer begleitenden Evaluation sind häufige Zwischenauswertungen zu erstellen.

Evaluationsspezifische Besonderheiten im Vergleich zur Grundlagenforschung ergeben sich im Zusammenhang mit einer sachgerechten Datenanalyse anhand der formulierten Fragestellungen und Hypothesen kaum (Stufflebeam & Shinkfield, 2007). Besonders wichtig ist es allerdings, den praktischen Nutzwert im Hinblick auf die Evaluationsziele bei der Datenanalyse immer in den Vordergrund zu rücken. Für umfassende und komplexe inferenzstatistische Datenanalysen werden sich Auftraggeber und andere Evaluationsnutzer allenfalls dann interessieren, wenn diese Antworten auf die drängenden Evaluationsfragen liefern. Um den Nutzwert der Datenanalyse zu steigern empfiehlt es sich, Zwischenergebnisse mit den intendierten Evaluationsnutzern zu diskutieren und von ihnen ggf. noch Anregungen für zusätzliche Datenanalysen entgegen zu nehmen.

Angesichts der Ausgangsfragen und der Datenlage ist bei Evaluationsstudien jeweils kritisch zu überdenken, ob überhaupt **inferenzstatistische Methoden** indiziert sind. Dies ist nur dann der Fall, wenn a priori theoriebasierte Hypothesen formuliert wurden, während zur Beantwortung von vielen Evaluationsfragen deskriptive Befunde ausreichen. Ebenso erübrigen sich Signifikanztests, wenn einerseits alle Teilnehmer einer Maßnahme untersucht wurden (Vollerhebung) oder wenn nur sehr wenige Untersuchungsteilnehmende gewonnen werden konnten und somit die statistischen Voraussetzungen für entsprechende Analysen nicht erfüllt sind.

Zu bedenken ist auch, dass sich in vielen Evaluationsstudien vermutlich nur schwache Effekte (zu Effektgrößen ► Abschn. 14.2) nachweisen lassen, die bei geringer Stichprobengröße mangels Teststärke gar nicht statistisch signifikant werden können (zur Teststärke ► Abschn. 14.1). Um hier zu gesicherten Befunden zu gelangen, empfiehlt es sich, anstelle der statistischen Signifikanz z. B. die **Musterläufigkeit** der Ergebnisse zu beachten: Angenommen eine psychologische Onlineberatung wird in einem Vorher-Nachher-Design überprüft und es werden 15 Variablen (Operationalisierungen der Evaluationskriterien) einbezogen, wobei sich keine einzige Differenz als statistisch signifikant erweist. Die Schlussfolgerung, dass die Maßnahme wirkungslos war, könnte voreilig sein, wenn die Evaluationsstichprobe z. B. $n = 300$ Personen umfasst. Denn ein schwacher Effekt hätte für einen Signifikanznachweis die 3- oder 4fache Zahl an Probanden erfordert. Betrachtet man jedoch das Ergebnismuster und stellt fest, dass sich deskriptiv bei 11 von 15 Variablen Differenzen im Sinne einer Verbesserung gezeigt haben, so könnte dies eine ermutigendere Bilanz rechtfertigen. Erst recht, wenn diese quantitativen Indikatoren der Wirksamkeit mit anderen Befunden übereinstimmen (z. B. Ergebnisse von qualitativen Interviews mit Beratern und Ratsuchenden, die über wahrgenommene Wirkungen berichten).

Interpretation von Evaluationsdaten

Bei der formativen Evaluation zielt die Interpretation der Zwischenergebnisse darauf ab, konkrete Empfehlungen für die Veränderung der Maßnahme oder ihrer Umsetzung abzuleiten. Bei der summativen Evaluation entspricht die Interpretation des empirischen Gesamtergebnisses der abschließenden Bewertung der Maßnahme hinsichtlich der betrachteten Evaluationskriterien und zugehörigen Be-

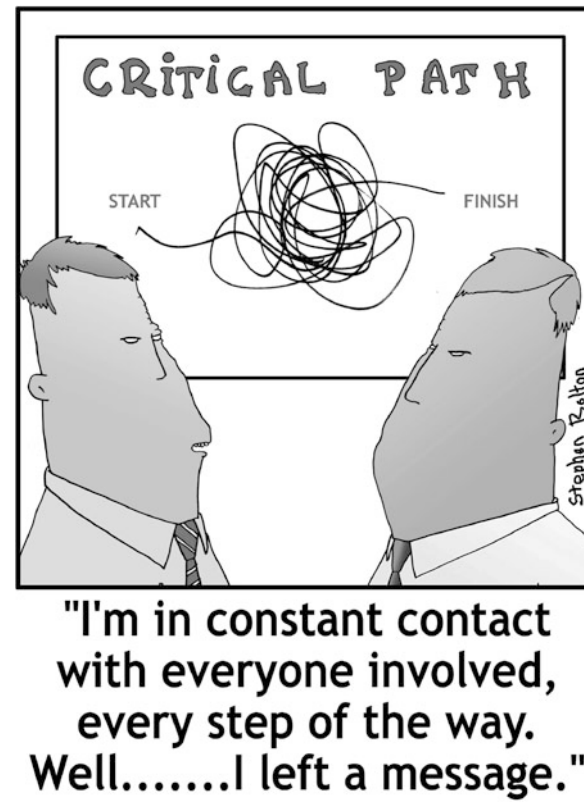
wertungsstandards (z. B. Grad der erreichten Effizienz, Nachhaltigkeit, ▶ Abschn. 18.1.5).

Während man in der Grundlagenforschung eine Studie durchaus mit divergierenden Befunden und offenen weiteren Forschungsfragen beenden kann, erfüllt eine Evaluationsstudie nur dann ihren Zweck, wenn sie die Fülle der Daten letztlich zu einem konsistenten Gesamtergebnis zusammenfasst. Auf dem Weg zu diesem Ergebnis mögen zunächst divergierende Befunde offen gelegt, reflektiert und mit den Stakeholdern diskutiert werden – aber letztlich mit dem Ziel einer wissenschaftlich abgesicherten Synthese durch das Evaluationsteam und **Ableitung von Empfehlungen**. Da es bei der Evaluation um Bewertung geht, sollten Evaluatorinnen und Evaluatoren am Ende auch einer solchen bewertenden Gesamtinterpretation nicht aus dem Weg gehen, so der vehemente Appell von Stufflebeam und Shinkfield (2007, S. 588):

» Evaluators who seek and present firm conclusions often face opposition and criticism and sometimes are wrong. However, if they ground their conclusions in systematic analysis and synthesis of a sufficient set of appropriate evidence and report appropriate caveats, we think they will be correct far more often than those who fail to practice systematic, conclusion-oriented evaluation and will be instrumental in helping their audiences make sound decisions and improve programs. Systematic evaluation is not and never will be an exact science, but it is an invaluable guide to progress. Evaluators who are steadfastly afraid of being wrong and consequently equivocate or exercise undue caution probably should seek other worthwhile work, such as bookkeeping, proofreading, window washing, watch repair, or house painting.

Projektcontrolling

Während der gesamten Durchführung des Evaluationsprojektes, v. a. aber während der Realisierungsphase, muss die Projektabwicklung im Hinblick auf die Projektziele und Projektpläne fortlaufend und im Detail kontrolliert und gesteuert werden (◘ Abb. 18.5). Auf diese Weise sollen Probleme bei der Projektabwicklung frühzeitig erkannt und gelöst werden. Zudem sichert das Projektcontrolling die Transparenz gegenüber dem Auftraggeber und anderen Stakeholdern: Die Projektleitung kann zu jedem Projektzeitpunkt auf der Basis systematischer Controllingmethoden (Soll-Ist-Vergleiche, Kennzahlen) z. B. den Arbeitsstand oder Kostenstand berichten (Kuster, et al., 2006, S. 143ff.). Der Einsatz von Projektcontrollingmethoden muss in Abhängigkeit vom jeweiligen Projekt skaliert werden und darf nicht zu einer unnötigen Bürokratisierung der Abläufe führen.



◘ **Abb. 18.5** Bei größeren Evaluationsprojekten sind Projektmanagement und Projektcontrolling nicht zu vernachlässigen. (© Stephen Bolton/Search ID: sbln136, Rights Available from CartoonStock.com)

18.3.5 Abschlussphase der Evaluation

Die in der Realisierungsphase der Evaluation gewonnenen Ergebnisse gilt es in der Abschlussphase an unterschiedliche Stakeholder der Intervention zu **vermitteln** und ihre **Nutzung** durch intendierte Nutzerinnen und Nutzer vorzubereiten. Auch eine wissenschaftliche Bewertung der Evaluation in Form von systematischer **Metaevaluation** kann sinnvoll oder notwendig sein. Schließlich ist es wünschenswert, Evaluationsstudien auch zu **publizieren**.

Abgabe der Evaluationsberichte

Bei einer **begleitenden Evaluation** erfolgt typischerweise eine fortlaufende, häufig auch **informelle Rückkopplung** von Evaluationsergebnissen an den Auftraggeber bzw. die Interventionsentwickler etwa in Meetings, Telefonaten, E-Mails, Newslettern etc. In der Abschlussphase des Evaluationsprojektes wird üblicherweise erwartet, dass der Ertrag der begleitenden Evaluation in einem **Abschlussbericht** dokumentiert wird, zum einen zur Ergebnissicherung, zum anderen als Beleg des Arbeitsauf-

wandes sowie als Basis für eine mögliche Metaevaluation. Bei einer **retrospektiven Evaluation** erfolgt die Kommunikation der Ergebnisse ohnehin stärker gebündelt und formalisiert in Form von **Präsentationen und Berichten,** wobei Zwischenberichte in einen abschließenden Bericht integriert werden.

Die Darstellung von Evaluationsergebnissen sollte ausgewählten **Evaluationsstandards** der Nützlichkeit (z. B. Auswahl und Umfang der Informationen), der Fairness (z. B. Schutz individueller Rechte wie z. B. Datenschutz, Vertraulichkeit) und Genauigkeit (z. B. Angabe von Informationsquellen, begründete Schlussfolgerungen) genügen (Gesellschaft für Evaluation e. V. [DeGEval], 2008b, S. 22ff.). Die Kommunikation von Evaluationsbefunden muss zudem – gerade mit Blick auf die Nützlichkeit – **zielgruppenorientiert und verständlich** erfolgen, d. h. z. B. gut strukturiert und ohne überflüssigen Wissenschaftsjargon. Ablaufdiagramme, Tabellen, Grafiken, Fotografien und andere Varianten der Veranschaulichung sind empfehlenswert. Für alle Fragen und Hypothesen der Evaluation sollten eindeutige und wissenschaftlich fundierte Antworten formuliert werden.

Sofern dies im Evaluationsvertrag vorgesehen ist, muss der Evaluationsbericht dem Auftraggeber auch konkrete **Entscheidungs- und Handlungsempfehlungen** an die Hand geben. Schließlich muss ein umfangreicher Evaluationsbericht die zentralen Inhalte stets in einer **aussagekräftigen Zusammenfassung** („executive summary") bündeln, die über Ausgangslage, Ziel, Methodologie und Methoden, wichtigste Ergebnisse, Schlussfolgerungen und Empfehlungen informiert. Bei Bedarf enthält der Evaluationsbericht auch ein **Glossar** mit Abkürzungen und Fachbegriffen sowie einen **Materialanhang** mit einschlägigen Dokumenten (z. B. Evaluationsauftrag, Fragebögen, Interviewleitfäden, Literaturverzeichnis etc.).

Organisation der Evaluationsnutzung

Der Evaluationsbericht ist eine wichtige Basis für die praktische Umsetzung und Nutzung der Befunde einer Evaluationsstudie, allerdings ist er allein meist nicht ausreichend. Vielmehr müssen aus dem Bericht die zentralen Botschaften für unterschiedliche intendierte Evaluationsnutzende extrahiert und in geeigneter Form (z. B. Merkblatt, Checkliste, Workshop, Seminar etc.) zum passenden Zeitpunkt an die jeweiligen Adressaten weitergegeben werden. Es kann sinnvoll sein, dass Auftraggebende und Evaluationsteam zusammenarbeiten, wenn es um die Evaluationsnutzung geht. Die folgende Checkliste fasst für die Evaluationsnutzung zentrale Fragen zusammen (Bundesamt für Gesundheit [BAG], 1997, S. 61, Checkliste 5.1):

- Haben wir die für unsere Planung und Entwicklung wichtigsten Aussagen der Studie identifiziert?
- Ist klar, was aus der Studie für die Umsetzung von Projekten folgt?
- Haben wir mögliche Folgen für die Aktivitäten verschiedener beteiligter Gruppen abgeschätzt?
- Haben wir überlegt, was nun zu tun ist?
- Ist klar, welche Rahmenbedingungen dafür geschaffen werden müssen?
- Wissen wir, welcher Aufwand mit der Durchführung solcher Folgearbeiten verbunden ist (personelle und finanzielle Ressourcen)?
- Haben wir die Personen oder Gruppen identifiziert, die am ehesten geeignet sind, die nächsten Schritte zu unternehmen?
- Haben wir überlegt, welche Personen/Gruppen Priorität haben? (Praktikabilität vs. Ideal!)
- Ist uns klar, wie wir am besten an diese Gruppe(n) herantreten?
- Wer ist am ehesten geeignet, an diese Gruppe(n) heranzutreten?

Durchführung einer Metaevaluation

Mit **Metaevaluation** ist die Evaluation der Evaluation gemeint, es geht also um eine Bewertung einer einzelnen Evaluationsstudie oder einer Menge von Evaluationsstudien. Bei der Metaevaluation werden Evaluationsstudien im Hinblick auf die Einhaltung und Umsetzung von Evaluationsstandards oder anderer einschlägiger Kriterien beurteilt (z. B. Nützlichkeit der Evaluationsbefunde für Auftraggebende und andere Anspruchsgruppen; Genauigkeit der Ergebnisse etc.).

Ebenso wie jede Grundlagenforschung eine kritische Reflexion von Methodenproblemen enthalten und die Aussagekraft ihrer Befunde kritisch diskutieren soll, ist es im Sinne der Qualitätssicherung und Transparenz wünschenswert, dass in Evaluationsprojekten reflektiert wird, welche Evaluationsstandards in welcher Weise umgesetzt werden konnten und wo und warum ggf. Probleme aufgetreten sind. Darüber hinaus sollte eine systematische Evaluation des Evaluationsprojektes erfolgen, sei es als selbst- oder fremdgesteuerte Metaevaluation, als interne oder externe Metaevaluation, als formative oder summative Metaevaluation. Eine Metaevaluation kann am Ende eines Evaluationsprojektes retrospektiv erfolgen oder auch die Evaluationsforschung von Anfang an begleiten (Stufflebeam, 2001; Stufflebeam & Shinkfield, 2007).

Eine Metaevaluation kann für die Evaluierenden wichtige Erkenntnisse liefern, Lern- und Dialog- sowie Optimierungsfunktionen erfüllen. Es kann aber auch sein, dass sie von Auftraggebenden eingefordert wird, sei es bereits mit dem Evaluationsmandat oder im Nachhinein: Nach Abschluss einer Evaluation kann es zwischen Auftraggeber, den für die Intervention Verantwortlichen und dem Evaluationsteam manchmal zu Auseinanderset-

zungen über die Qualität der Evaluationsstudie kommen. Auftraggebende könnten z. B. zu der Ansicht gelangen, die Studie sei ihr Geld nicht wert gewesen oder Interventionsentwickler könnten sich, ihre Arbeit bzw. Arbeitsergebnisse als unfair bewertet ansehen. Schließlich können Urheberrechts- oder Datenschutzfragen virulent werden. Ist informell keine Konfliktlösung zu erzielen, so kann es wünschenswert sein, die Evaluationsstudie von externen Evaluationsexperten evaluieren und begutachten zu lassen, wobei typischerweise die gängigen Evaluationsstandards heranzuziehen und anhand von empirischen Daten (z. B. Dokumentenanalyse, Befragung etc.) zu prüfen sind. Zur Anfertigung von Gutachten im Rahmen externer Metaevaluationen liegen von der Gesellschaft für Evaluation einschlägige Empfehlungen vor (Deutsche Gesellschaft für Evaluation e.V. [DeGEval], 2002). Hilfreich ist es, die Bedingungen einer möglichen Metaevaluation bereits im Evaluationsvertrag festzuhalten. Dies betrifft die für die Bewertung der Evaluationsarbeit anzulegenden Kriterien (z. B. Auswahl aus den Evaluationsstandards) sowie Vereinbarungen über die Dokumentation der für die Metaevaluation notwendigen Daten.

Schließlich kann die externe Metaevaluation auch eingesetzt werden, um die – z. B. in der Routinephase von Interventionen etablierte – **Selbstevaluation** der Beteiligten (z. B. Lehr- und Forschungsevaluation) periodisch oder inzidenziell zu prüfen.

Publikation der Evaluationsstudie

Deutlich seltener als in der Grundlagenforschung werden die Ergebnisse der Evaluationsforschung der wissenschaftlichen Fachöffentlichkeit vorgestellt. Dies liegt u. a. an der Vertraulichkeit der gewonnenen Daten, an der Projektspezifik von Erkenntnissen (speziell bei formativer Evaluation), am Imageschutz für Auftraggebende bzw. Interventionsverantwortliche (speziell im Fall ungünstiger Evaluationsbefunde) sowie nicht selten auch am Zeitmangel der Evaluierenden, für die die praktische Evaluationsarbeit in der Regel eine höhere Priorität besitzt als das akademische Publizieren. Doch Evaluationsforschung könnte sich nicht als wissenschaftliches Feld behaupten, wenn eine kritische Auseinandersetzung in der Scientific Community und auch in der breiten Öffentlichkeit fehlen würde. Dafür ist die Publikation von Evaluationen sowie von Metaevaluationen notwendig.

Die Publikation von Evaluationsforschung auf **Tagungen** der Fachgesellschaften sowie auf **fachbezogenen Onlineportalen**, in **Fachbüchern** und insbesondere in **Fachzeitschriften** dient der akademischen Erkenntnisweitergabe und im Rahmen des Peer-Review-Verfahrens auch der wissenschaftlichen Qualitätssicherung. So können aus Evaluationsforschungsprojekten theoretische, methodische und inhaltliche Erkenntnisse gewonnen werden, die dazu dienen, das Feld der Evaluationsforschung weiterzuentwickeln.

Die Publikation bzw. Zugänglichkeit von einzelnen Evaluationsstudien erlaubt **Forschungssynthese** in doppelter Hinsicht:

- Sobald zu einer bestimmten Intervention eine kritische Menge von publizierten Evaluationsstudien vorliegt, sind **Metaanalysen** (▶ Kap. 16) im Sinne einer systematischen Zusammenfassung der Befunde über das Evaluationsobjekt möglich. So legten z. B. Grawe, Donati und Bernauer (2001) eine sehr umfassende und seit der Erstveröffentlichung im Jahr 2001 kontrovers diskutierte Metaanalyse zur Wirksamkeit der verschiedenen Psychotherapiemethoden vor. Die Metaanalyse von Evaluationsstudien will in der Regel zusammenfassend Auskunft über die Wirkung von Interventionen geben („what works, for whom, under what conditions, how, and why?"; Beelmann & Lipsey, in press). Neben quantitativen Metaanalysen von Evaluationsstudien sind auch qualitative Forschungssynthesen möglich (Widmer & De Rocchi, 2012, S. 161).
- Im Unterschied zur Metaanalyse als Zusammenfassung der Ergebnisse von Evaluationsstudien (metaanalytische Aussage über das Evaluationsobjekt) zielt die **Metaevaluation** auf eine Bewertung der **Methodik** der jeweiligen Evaluationsstudie ab (z. B. Grad der Einhaltung der Evaluationsstandards). In der Praxis erfolgt die Metaevaluation meist zum Zweck der Qualitätskontrolle bzw. Qualitätssicherung eines konkreten Projektes und bleibt unveröffentlicht (▶ Abschn. 18.3.5 „Durchführung einer Metaevaluation"). Es lassen sich aber auch empirische Metaevaluationsstudien durchführen, die z. B. mit Methoden der Dokumentenanalyse oder mit Experten- und Teilnehmerbefragungen operieren und eine Stichprobe von Evaluationsprojekten untersuchen (für eine Metaevaluation von 11 Evaluationsstudien im Gender-Mainstreaming-Bereich in Spanien Bustelo, 2003; für eine Metaevaluation von 14 Evaluationsstudien im HIV-/Aids-Bereich in Afrika Patel, 2002; für eine publizierte Metaevaluation von 56 Evaluationsstudien im Bildungsbereich in der Schweiz Stamm, 2003). Derartige Metaevaluationen befassen sich z. B. mit der Frage, wie nützlich die gewonnenen Evaluationsbefunde sind und in welchem Maße sie erfolgreich in die Praxis umgesetzt werden konnten. Metaevaluationen dienen auch dazu, die Qualität von unterschiedlichen Evaluationsformen oder Evaluationsmethoden zu bewerten (für eine Evaluation der Methode der Kurzevaluation Widmer, Rothmayr, & Serdült, 1996). Empirische Metaevaluationsstudien sollen verallgemeinerbare Erkenntnisse über die Durchführung und Qualität von Evaluationsstudien liefern und

auch dabei helfen, die Brauchbarkeit von Evaluationsstandards zu bewerten (für eine vergleichende Metaevaluation von 10 schweizerischen Evaluationsstudien aus dem Politikbereich Widmer, 1996).

Nicht zuletzt ist es notwendig, die **Wirksamkeit von Evaluationsstudien** empirisch auf den Prüfstand zu stellen. Denn selbst wenn laut Metaevaluation die Evaluationsstudien in einem bestimmten Teilbereich die Evaluationsstandards erfüllen und die Evaluationsergebnisse sogar von den intendierten Evaluationsnutzern in intendierter Weise genutzt werden (z. B. Diskussion der Ergebnisse unter Entscheidungsträgern, Ableitung und Umsetzung von Verbesserungsmaßnahmen), ist damit noch nichts darüber gesagt, ob diese Aktivitäten der Evaluationsnutzenden nachweislich zu praktischen Verbesserungen führen. Für diese Bewertung müssen die aus den Evaluationsergebnissen gezogenen praktischen Konsequenzen ihrerseits evaluiert werden.

Lernquiz

1. Worin besteht der Unterschied zwischen formativer und summativer Evaluation?
2. Was kennzeichnet interne versus externe Evaluation und Selbst- versus Fremdevaluation?
3. Grenzen Sie Interventions- und Evaluationsforschung voneinander ab!
4. Was versteht man unter dem CIPP-Modell?
5. Was unterscheidet die nutzungsorientierte von der partizipativen Evaluationsforschung?
6. Welche Ziele im Kontext der Evaluationsforschung verfolgen Metaanalyse und Metaevaluation?
7. Gute grundlagenwissenschaftliche Studien müssen Kriterien der methodischen und ethischen Strenge genügen. Wo sind diese in den Evaluationsstandards abgebildet?
8. Mit welchen Methoden wird die Effizienz eines Evaluationsgegenstandes im Kontext der ökonomischen Evaluation ermittelt?
9. Angenommen, die zahnmedizinische Versorgung soll dadurch verbessert werden, dass bei schmerzhaften Behandlungen auf die Injektion von Schmerzmitteln verzichtet und stattdessen mit modernen Hypnosetechniken gearbeitet wird. Hypnosetherapie hätte den Vorteil, dass Medikamente eingespart werden (Kostenersparnis, geringere Belastung des Organismus) und sich die Patienten während und nach der Behandlung evtl. besser fühlen. In einer Großstadt werden zufällig 5 Zahnarztpraxen, die mit Hypnose arbeiten, und 3 Praxen, die herkömmliche Methoden der Schmerzbehandlung einsetzen, ausgewählt. Während einer 4-wöchigen Untersuchungsphase wird direkt nach jeder Behandlung auf gesonderten Erhebungsbögen notiert, wie unangenehm die Behandlung für den Patienten war (gar nicht, wenig, teils teils, ziemlich, völlig) und ob der Patient eine bessere Schmerzversorgung wünscht (ja, nein). Zudem werden Alter (in Jahren), Geschlecht (weiblich/männlich; zur Operationalisierung von Geschlecht ► Abschn. 8.5.1) und Art der Behandlung (z. B. Wurzelbehandlung, Krone, Inlay) aufgezeichnet. Am folgenden Tag wird telefonisch nachgefragt, ob nach der Behandlung noch unangenehme Nachwirkungen auftraten (gar nicht, wenig, teils teils, ziemlich, völlig). Wie würden Sie diese Evaluationsstudie anhand der folgenden Merkmale charakterisieren?
   1. Evaluationsfrage?
   2. Unabhängige Variable (und Skalenniveau)?
   3. Moderatorvariablen (und Skalenniveau)?
   4. Abhängige Variablen (und Skalenniveau)?
   5. Datenerhebungsmethode?
   6. Untersuchungsdesign?
   7. Verhältnis von Interventions- und Evaluationsstichprobe?
   8. Erfolgskriterien?
10. Diskutieren Sie für das obige Beispiel die Problematik der internen und externen Validität. Angenommen, es stellt sich heraus, dass die Hypnosegruppe tatsächlich weniger Beschwerden berichtet als die Kontrollgruppe; welche Störeinflüsse könnten die interne Validität beeinträchtigt haben? Inwiefern sind Ergebnisse dieser Studie generalisierbar?

Übungsaufgaben

Dem Freundes- und Förderkreis einer Gesamtschule ist es gelungen, einen beträchtlichen Spendenbetrag einzuwerben, um damit ein pädagogisch wertvolles außerschulisches Programm mit regelmäßigen Arbeitsgruppen (z. B. Schach-AG, Foto-AG, Poetry-Slam-AG, Hockey-AG) zu gestalten. Die SpenderInnen stellen in Aussicht, ein erfolgreiches Programm auch langfristig fördern zu wollen, knüpfen ihre weitere Spendenbereitschaft jedoch an die Bedingung, dass die Maßnahme wissenschaftlich evaluiert wird. Die Schulleitung ist begeistert von dieser Möglichkeit. Über Kontakte zur benachbarten Hochschule gelingt es, ein externes Evaluationsteam anzuwerben. Unter Leitung einer Evaluationsexpertin sollen im Rahmen eines zweisemestrigen Praxisseminars 15 Studierende an dem Evaluationsprojekt mitarbeiten. Der Evaluationsauftrag wird in 3 Teilevaluationen gegliedert.

Planen Sie die einzelnen Teilevaluationen möglichst präzise und verschriftlichen Sie Ihre Ideen (arbeiten Sie dazu möglichst im Team)!

- **Konzeptevaluation** (Context- und Input-Evaluation gemäß CIPP-Modell): Fragestellung summativ: Ist das Programmkonzept durchführbar und erfolgversprechend? Fragestellung formativ: Wie können Konzeptideen hinsichtlich Durchführbarkeit und Erfolgsorientierung verbessert werden? Um diese Evaluationsfragen zu beantworten, sind u. a. folgende Punkte zu klären: Wer sollte und könnte an der Konzeptentwicklung des Programms mitwirken? Welche Stakeholder sind für das Projekt besonders relevant? Wie lassen sich die Bedürfnisse und Erwartungen der Zielgruppen sowie die für das Programm an der Schule nutzbaren Ressourcen erfassen und bewerten? Angenommen, die Schulleitung plädiert für ein Programm mit Arbeitsgruppen, die einen engen thematischen Bezug zu Schulfächern aufweisen (z. B. Englisch-Konversations-AG), während die Schülervertretung sich v. a. AG-Angebote mit hohem Freizeitwert wünscht (z. B. Tanz-AG). Wie kann das Evaluationsteam entsprechend divergierende Konzeptideen bewerten und integrieren? Ist eine Orientierung an Best-Practice-Beispielen anderer Schulen möglich? Diskutieren und ergänzen Sie diese Fragen und planen Sie das empirische Vorgehen zu deren Beantwortung! Behalten Sie den Zeitrahmen im Auge: Bis wann muss die Konzeptevaluation abgeschlossen sein und ein verbessertes Konzept vorliegen, um rechtzeitig in die Umsetzungsphase überzugehen (zu beachten: Taktung Schuljahr, Taktung Vorlesungszeiten, Evaluationsdauer zwei Semester)? Wer sind die intendierten Nutzer/innen der Resultate der Konzeptevaluation und was sollten sie mit den Resultaten tun?
- **Implementationsevaluation** (Process-Evaluation gemäß CIPP-Modell): Fragestellung summativ: Wie erfolgreich wird das Programm im laufenden Betrieb umgesetzt? Fragestellung formativ: Wie lassen sich Probleme und Hindernisse bei der Programmumsetzung überwinden? Um diese Evaluationsfragen zu beantworten, sind u. a. folgende Punkte zu klären: Welche Merkmale (z. B. Alter, Geschlecht, sozioökonomischer Hintergrund, Schulleistung) haben SchülerInnen, die an den Arbeitsgemeinschaften regelmäßig teilnehmen im Vergleich zu denen, die nicht oder nicht mehr oder nur sporadisch teilnehmen? Wie hoch ist die Zufriedenheit der Schülerschaft mit dem laufenden Programm? Was genau wird wo und wie in den AG gemacht (Raum-, Materialbedarf etc.)? Wie bewerten die AG-LeiterInnen die Arbeit in und mit den Gruppen? Wie unterscheiden sich die einzelnen AG-Angebote hinsichtlich ihrer Akzeptanz und wie sind die Unterschiede erklärbar? Diskutieren und ergänzen Sie diese Fragen und planen Sie das empirische Vorgehen zu deren Beantwortung! Wer sind die intendierten Nutzer/innen der Resultate der Implementationsevaluation und was sollten sie mit den Resultaten tun?
- **Ergebnisevaluation** (Product-Evaluation gemäß CIPP-Modell). Fragestellung summativ: Werden die Ziele des Programms erreicht (Effektivität) und wenn ja, wie hoch ist der damit verbundene Aufwand (Effizienz)? Fragestellung formativ: Wie lassen sich Effektivität und Effizienz des Programms zukünftig steigern? Um diese Evaluationsfragen zu beantworten, sind u. a. folgende Punkte zu klären: Wie sind die Ziele und Erfolgskriterien des Programms festzulegen (Vorgehensweise und exemplarische Ziele und Kriterien). Welche Output- und/oder Outcome- und/oder Impact-Variablen sollten berücksichtigt werden und wie sind sie zuverlässig zu erfassen (Vorgehensweise und exemplarische Variablen)? Welche Gesamtkosten fallen für das Programm an, und in welche Kostenarten sind sie gegliedert? Ist ergänzend zur zielorientierten Evaluation auch eine zielfreie Evaluation sinnvoll (wenn ja, wie und wann und von wem sollte sie durchgeführt werden)? Diskutieren und ergänzen Sie diese Fragen und planen Sie das empirische Vorgehen zu deren Beantwortung! Wer sind die intendierten Nutzer/innen der Resultate der Ergebnisevaluation und was sollten sie mit den Resultaten tun?

18

Weiterführende Literatur

- Fitzpatrick, J., Sanders, J., & Worthen, B. (2011). *Program evaluation. Alternative approaches and practical guidelines* (4th ed.). Boston: Pearson.
- Flick, U. (Hrsg.) (2006). *Qualitative Evaluationsforschung: Konzepte, Methoden, Umsetzungen*. Hamburg: Rowohlt.
- Gollwitzer, M. & Jäger, R. S. (2009). *Evaluation kompakt*. Weinheim: Beltz.
- Holling, H. (Hrsg.) (2009). *Enzyklopädie der Psychologie: Evaluation – Band 1*. Göttingen: Hogrefe.

- Patton, M. Q. (2008). *Utilization-focused evaluation: The new century text* (4th ed.). Thousand Oaks: Sage.
- Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (2003). *Evaluation. A systematic approach* (7th ed). Thousand Oaks: Sage.
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park: Sage.
- Stockmann, R. & Meyer, W. (2010). *Evaluation. Eine Einführung*. Opladen: Budrich.
- Stufflebeam, D. L. & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco: Wiley.
- Weiss, C. (1998). *Evaluation: Methods for studying policies and programs* (2nd ed.). Upper Saddle River: Prentice Hall.
- Widmer, T. & De Rocchi, T. (2012). *Evaluation: Grundlagen, Ansätze und Anwendungen*. Zürich: Ruegger.
- Widmer, T., Beywl, W. & Fabian, C. (Hrsg.) (2009). *Evaluation. Ein systematisches Handbuch*. Wiesbaden: VS.
- Wottawa, H. & Thierau, H. (2003). *Lehrbuch Evaluation* (3. Aufl.). Göttingen: Huber.

Literatur

Alkin, M. C. (1990). *Debates on evaluation*. Thousand Oaks: Sage.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (2002). *Standards for Educational and Psychological Testing*. Washington: American Education Research Association.

American Evaluation Association (AEA). (2004). *The American evaluation association guiding principles for evaluators.* Retrieved November 14, 2012, from http://www.eval.org/publications/guidingprinciples.asp

Backhaus, K., Erickson, B., Plinke, W., & Weiber, R. (1994). *Multivariate Analysemethoden*. Heidelberg: Springer.

Balzer, L. (2005). *Wie werden Evaluationsprojekte erfolgreich? Ein integrierender theoretischer Ansatz und eine empirische Studie zum Evaluationsprozess*. Landau: Empirische Pädagogik.

Bateman, A. W. (2000). Integration in psychotherapy: An evolving reality in personality disorder. *British Journal of Psychotherapy, 17*(2), 147–156.

Beelmann, A. & Lipsey, M. W. (in press). Meta-analysis of effect estimates from multiple studies. In M. W. Lipsey & D. S. Conrady (Eds.), *Field experimentation: Methods for evaluating what works, for whom, under what circumstances, how, and why*. Thousand Oaks: Sage.

Beywl, W., Bestvater, H., & Friedrich, V. (2011). *Selbstevaluation in der Lehre. Ein Wegweiser für sichtbares Lernen und besseres Lehren*. Münster: Waxmann.

Birckmayer, J. D. & Weiss, C. (2000). Theory-based evaluation in practice. *Evaluation Review, 24*(4), 407–431.

Bohnsack, R. & Nentwig-Gesemann, I. (Hrsg.). (2010). *Dokumentarische Evaluationsforschung. Theoretische Grundlagen und Beispiele aus der Praxis*. Opladen: Budrich.

Bolton, S. (Artist). (2012). *"I'm in constant contact with everyone involved, every step of the way. Well . . . I left a message."*.

Bortz, J. & Lienert, G. A. (2003). *Kurzgefaßte Statistik für die klinische Forschung. Ein praktischer Leitfaden für die Analyse kleiner Stichproben* (2. Aufl.). Heidelberg: Springer.

Brunswik, E. (1955). Representative design and probability theory in a functional psychology. *Psychological Revue, 62*(3), 193–217.

Bundesamt für Gesundheit (BAG). (1997). *Leitfaden für die Planung von Projekt- und Programmevaluation*. Bern: Bundesamt für Gesundheit.

Bussmann, W. (1995). *Evaluationen staatlicher Maßnahmen erfolgreich begleiten und nutzen: Ein Leitfaden*. Zürich: Ruegger.

Bustelo, M. (2003). Evaluation of gender mainstreaming: Ideas from a meta-evaluation study. *Evaluation, 9*(4), 383–403.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31*(2), 199–218.

Chambers, R. (2009). Making the poor count: Using participatory options for impact evaluation. In R. Chambers, D. Karlan, M. Ravallion, & P. Rogers (Eds.), *Designing Impact evaluations: Different perspectives.* New Delhi, India: *international initiative for impact evaluation.* Retrieved November 11, 2013, from http://www.3ieimpact.org/admin/pdfspapers/50.pdf.

Chen, H. (1990). *Theory-driven evaluations*. Thousand Oaks: Sage.

Chen, H. (2005). *Practical program evaluation: Assess and improve program planning, implementation, and effectiveness.* Thousand Oaks: Sage.

Cho, H. & Witte, K. (2005). Managing fear in public health campaigns: A theory-based formative evaluation process. *Health Promotion Practice, 6*(4), 482–490.

Chouinard, J. A. (2013). The case for participatory evaluation in an era of accountability. *American Journal of Evaluation, 34*, 237–253.

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. *New Directions for Evaluation, 2003*(97), 7–36.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions: An empirical examination. *American Journal of Evaluation, 28*(1), 8–25.

Clemens, M. A. & Demombynes, G. (2011). When does rigorous impact evaluation make a difference? The case of the millennium villages. *Journal of Development Effectiveness, 3*(3), 305–339.

Cook, T. D., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). *„Sesame Street" revisited.* New York: Sage.

Coryn, C. L. S., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32*(2), 199–226.

Cousins, J. B. & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis, 14*(4), 397–418.

Creswell, J. W. & Plano Clark, V. L. (Eds.). (2010). *Designing and conducting mixed methods research*. Tousand Oaks: Sage.

Daigneault, P.-M., Jacob, S., & Tremblay, J. (2012). Measuring stakeholder participation in evaluation: An empirical validation of the participatory evaluation measurement instrument (PEMI). *Evaluation Review, 36*(4), 243–271.

Datta, L. (2013). Paradox lost and paradox regained. Comments on Chouinard's „The case for participatory evaluation . . .". *American Journal of Evaluation, 34*(2), 254–260.

Deutsche Gesellschaft für Evaluation e.V. (DeGEval). (2002). Angefragte Gutachten/Meta-Evaluationen auf Basis der DeGEval-Standards. Abgerufen 31. Juli, 2012, unter http://www.degeval.de/standards/Richtlinien_Meta_evaluation.pdf

Deutsche Gesellschaft für Evaluation e.V. (DeGEval). (2004). Empfehlungen zur Anwendung der Standards für Evaluation im Handlungsfeld der Selbstevaluation [Broschüre. Abgerufen 14. November, 2012, unter http://www.degeval.de/publikationen/selbstevaluation

Dienel, P. C. (1978). *Die Planungszelle*. Opladen: Westdeutscher Verlag.

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review, 70*(11), 35–36.

Döring, N. (2005). Für Evaluation und gegen Evaluitis. Warum und wie Lehrevaluation an deutschen Hochschulen verbessert werden sollte. In B. Berendt, H.-P. Voss & J. Wildt (Hrsg.), *Neues Handbuch Hochschullehre*. (S. 1-22). Berlin: Raabe.

Döring, N. (2009). Phasen der Evaluationsforschung. In H. Holling & R. Schwarzer (Hrsg.), *Evaluation. Band 1. Grundlagen und statistische Methoden der Evaluationsforschung* (S. 99–134). Göttingen: Hogrefe.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., & O'Brien, B. J. (2005). *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237–256.

Earl, L. M. & Cousins, J. B. (Eds.). (1995). *Participatory evaluation in education: Studies in evaluation use and organizational learning*. London: Falmer.

Eisenführ, F. & Weber, M. (1993). *Rationales Entscheiden*. Heidelberg: Springer.

Engelhardt, H. D., Trojan, A., & Nickel, S. (2009). Leistungen von Selbsthilfegruppen und deren ökonomische Bewertung. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz, 52*(1), 64–70.

Fetterman, D. M. (1994). Empowerment evaluation. *Evaluation Practice, 15*(1), 1–15.

Fetterman, D. M. (2001). the transformation of evaluation into a collaboration: A vision of evaluation in the 21st century. *The American Journal of Evaluation, 22*(3), 381–385.

Fleischer, D. N. & Christie, C. A. (2009). Evaluation use. Results fraim a survey of U.S. American Evaluation Association Members. *American Journal of Evaluation, 30*(2), 158–175.

Flick, W. (Hrsg.). (2006). *Qualitative Evaluationsforschung. Konzepte – Methoden – Umsetzung*. Reinbek: Rowohlt.

Gesellschaft für Evaluation e. V. (DeGEval). (2008a). Recommendations on education and training in evaluation [Broschüre]. Abgerufen 5. November, 2013, unter http://www.degeval.de/publikationen/requirementprofilesforevaluators

Gesellschaft für Evaluation e. V. (DeGEval). (2008b). *Standards für Evaluation.* Abgerufen 5. November, 2013, unter http://www.alt.degeval.de/calimero/tools/proxy.php?id$=$24065

Gesellschaft für Evaluation e. V. (DeGEval). (o. J.). *Empfehlungen für Auftraggebende von Evaluationen. Eine Einstiegsbroschüre für den Bereich der Öffentlichen Verwaltung* [Broschüre]. Abgerufen 5. November, 2013, unter http://www.degeval.de/publikationen/auftraggebende-von-evaluationen

Giel, S. (2013). *Theoriebasierte Evaluation: Konzepte und methodische Umsetzungen.* Münster: Waxmann.

Goldfried, M. R. (1991). Research issues in psychotherapy integration. *Journal of Psychotherapy Integration, 1*(1), 5–25.

Gollwitzer, M. & Jäger, R. S. (2009). *Evaluation kompakt*. Weinheim: Beltz.

Grawe, K., Donati, R., & Bernauer, F. (2001). *Psychotherapie im Wandel – Von der Konfession zur Profession* (5. Aufl.). Göttingen: Hogrefe.

Greene, J. C. (1997). Evaluation as advocacy. *American Journal of Evaluation, 18*(1), 25–35.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255–274.

Gregory, A. (2000). Problematizing participation: A critical review of approaches to participation in evaluation theory. *Evaluation, 6*(2), 179–199.

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth Generation Evaluation*. Thousand Oaks: Sage.

Gutknecht-Gmeiner, M., Hense, J. U., Kihm, S., Lüth, K., & Steckhan, H. (2011). *Didaktik der Aus- und Weiterbildung in der Evaluation. Ein Empfehlungspapier des Arbeitskreises Aus- und Weiterbildung in der Evaluation in der Gesellschaft für Evaluation – DeGEval e. V.* [Broschüre]. Mainz: Gesellschaft für Evaluation – DeGEval e. V.

Hager, W., Patry, J.-L. & Brezing, H. (Hrsg.). (2000). *Evaluation psychologischer Interventionsmaßnahmen. Standards und Kriterien. Ein Handbuch*. Bern: Huber.

Hansen, M. B. & Vedung, E. (2010). Theory-based stakeholder evaluation. *American Journal of Evaluation, 31*(3), 295–313.

Heiner, M. (1988). Von der forschungsorientierten zur praxisorientierten Selbstevaluation. Entwurf eines Konzeptes. In M. Heiner (Hrsg.), *Selbstevaluation in der sozialen Arbeit. Fallbeispiele zur Dokumentation und Reflexion beruflichen Handelns* (S. 7–40). Freiburg: Lambertus.

Heiner, M. (Hrsg.). (1996). *Selbstevaluation als Qualifizierung in der Sozialen Arbeit. Fallstudien aus der Praxis*. Freiburg: Lambertus.

Hense, J. U. (2006). *Selbstevaluation – Erfolgsfaktoren und Wirkungen eines Ansatzes zur selbstbestimmten Qualitätsentwicklung im schulischen Bereich*. Frankfurt/Main.: Lang.

Hense, J. U. (2010). Nutzen und Nutzung von Evaluation zur Unterrichtsverbesserung. In W. Schönig, A. Baltruschat & G. Klenk (Hrsg.), *Dimensionen pädagogisch akzentuierter Schulevaluation* (S. 125–144). Baltmannsweiler: Schneider Verlag Hohengehren.

Hense, J. U. & Mandl, H. (2003). *Selbstevaluation. Ein Ansatz zur Qualitätsverbesserung pädagogischer Praxis und seine Umsetzung am Beispiel des Modellversuchprogramms SEMIK* (Forschungsbericht Nr. 162). München: Ludwig-Maximilians-Universität, Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie.

Hense, J. U. & Mandl, H. (2011). Wissensmanagement und Evaluation. *Zeitschrift für Evaluation, 10*(2), 267–301.

Herrmann, T. (1976). *Die Psychologie und ihre Forschungsprogramme*. Göttingen: Hogrefe.

Herrmann, T. (1979). *Psychologie als Problem*. Stuttgart: Klett.

Hiatt, K., Stelle, C., Muslow, M., & Scott, J. P. (2007). The importance of perspective: Evaluation of hospice care from multiple stakeholders. *American Journal of Hospice and Palliative Medicine, 24*(5), 376–382.

House, E. R. & Howe, K. R. (1998). *Deliberative democratic evaluation in practice*. Boulder: University of Colorado.

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods, 7*, 178–193.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377–410.

Joint Committee on Standards for Educational Evaluation (Hrsg.). (2006). *Handbuch der Evaluationsstandards* (3. Aufl.). Opladen: Leske & Budrich.

Jungermann, H., Pfister, H. R., & Fischer, K. (2005). *Die Psychologie der Entscheidung* (2. Aufl.). Heidelberg: Spektrum.

Kandler, J. (1995). Rückflusseffekte der städtischen Zuschüsse für Selbsthilfe-Initiativen für die öffentliche Hand. In H. D. Engelhardt (Ed.), *Was Selbsthilfe leistet. . . : ökonomische Wirkungen und sozialpolitische Bewertung* (S. 76–85). Freiburg im Breisgau: Lambertus.

Karlan, D. (2009). Thoughts on randomised trials for evaluation of development: presentation to the Cairo evaluation clinic. *Journal of Development Effectiveness, 1*(3), 237–242.

Keeney, R. L. (1992). *Value-focused thinking*. Cambridge: Harvard University Press.

Keßler, H. & Winkelhofer, G. (2004). *Projektmanagement. Leitfaden zur Steuerung und Führung von Projekten* (4. Aufl.). Berlin: Springer.

Keuth, H. (1989). *Wissenschaft und Werturteil. Zu Werturteilsdiskussion und Positivismusstreit*. Tübingen: Mohr.

Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2009). *Handbook on impact evaluation: Quantitative methods and practices*. Washington: World Bank Pubn.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation, 2000*(88), 5–23.

Kromrey, H. (1987). Zur Verallgemeinerbarkeit empirischer Befunde in nichtrepräsentativen Stichproben. *Rundfunk & Fernsehen, 4*, 478–499.

Kromrey, H. (2003). Evaluation in Wissenschaft und Gesellschaft. *Zeitschrift für Evaluation, 2*(1), 93–116.

Kuckartz, U. & Busch, J. (2012). Mixed Methods Evaluationen. In U. Kuckartz & S. Rädiker (Hrsg.), *Erziehungswissenschaftliche Evaluationspraxis. Beispiele – Konzepte – Methoden* (S. 145–158). Weinheim: Beltz Juventa.

Kuster, J., Huber, E., Lippmann, R., Schmid, A., Schneider, E., Witschi, U., et al. (2006). *Handbuch Projektmanagement* (3. Aufl.). Berlin: Springer.

Lewin, K. (1946). Action research & minority problems. *Journal of Social Issues, 2*(4), 34–46.

Lewin, K. (1951). *Field theory in social science*. New York: Harper & Row.

Lüders, C. (2006). Qualitative Evaluationsforschung – was heißt hier Forschung? In U. Flick (Hrsg.), *Qualitative Evaluationsforschung: Konzepte - Methoden - Umsetzung* (S. 33–62). Reinbek: Rowohlt.

MacDonald, B. (1976). Evaluation and the control of education. In D. Tawney (Eds.), *Curriculum evaluation today: trends and implications* (pp. 125–134). London: MacMillan Education.

Mackenzie, M. & Blamey, A. (2005). The practice and the theory: Lessons from the application of a theories of change approach. *Evaluation: The International Journal of Theory, 11*(2), 151–168.

McCabe, A. & Horsley, K. (2008). *The evaluator's cookbook: Exercises for participatory evaluation with children and young people*. New York: Routledge.

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation, 31*(3), 390–399.

Miller, R. L. & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation, 27*(3), 296–319.

Mittag, W. & Hager, W. (2000). Ein Rahmenkonzept zur Evaluation psychologischer Interventionsmaßnahmen. In W. Hager & H. Brenzing (Hrsg.), *Evaluation psychologischer Interventionsmaßnahmen* (S. 102–128). Bern: Huber.

Moosbrugger, H. & Schweizer, K. (2002). Evaluationsforschung in der Psychologie. *Zeitschrift für Evaluation, 1*(1), 19–37.

Morrongiello, B. A., Schwebel, D. C., Bell, M., Stewart, J., & Davis, A. L. (2012). An evaluation of the great escape: Can an interactive computer game improve young children's fire safety knowledge and behaviors? *Health Psychology, 31*(4), 496–502.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315–340.

Müller-Kohlenberg, H. & Beywl, W. (2003). Standards der Selbstevaluation. Abgerufen 5. November, 2013, unter http://www.seval.ch/documents/unterlagenstandards/weiterereferenzen/w10_muellerkohlenberg_beywl_od_selbstevaluation.pdf

Neber, H. & Heller, K. A. (2002). Evaluation of a summer-school program for highly gifted secondary-school students: The German Pupils Academy. *European Journal of Psychological Assessment, 18*(3), 214–228.

Nelson, T. D. & Steele, R. G. (2006). Beyond efficacy and effectiveness: A multifaceted approach to treatment evaluation. *Professional Psychology: Research and Practice, 37*(4), 389–397.

Patel, M. (2002). A meta-evaluation or quality assessment of the evaluations in this issue based on the African evaluation guidelines: 2002. *Evaluation and Program Planning, 25*(4), 329–332.

Patton, M. Q. (2008). *Utilization-focused evaluation: The new century text* (4th ed.). Thousand Oaks: Sage.

Pawson, R. & Tilley, N. (1997). *Realistic Evaluation*. Thousand Oaks: Sage.

Pawson, R. & Tilley, N. (2004). Realist evaluation. Retrieved November 5, 2013, from http://www.communitymatters.com.au/RE_chapter.pdf

Poskitt, K. & Appleby, S. (1993). *Die 99 Lassedasse. Tätigkeiten, vor denen wir nur warnen können. Ein Führer* (1. Aufl.). Kiel: Achterbahn.

Rädiker, S. & Niklesz, A. (2012). Kompetenzen von Evaluierenden. In U. Kuckartz & S. Rädiker (Hrsg.), *Erziehungswissenschaftliche Evaluationspraxis. Beispiele – Konzepte – Methoden.* (S. 200–220). Weinheim: Beltz Juventa.

Reade, N. (2008). Konzept für alltagstaugliche Wirkungsevaluierungen in Anlehnung an Rigorous Impact Evaluations – Erprobung der Durchführung im Rahmen von GTZ Unabhängingen Evaluierungen, *Ceval–Arbeitspapiere*. Saarbrücken: Centrum für Evaluation.

Reininger, B. M., Barroso, C. S., Mitchell-Bennett, L., Cantu, E., Fernandez, M. E., Gonzalez, D. A., et al. (2010). Process evaluation and participatory methods in an obesity-prevention media campaign for Mexican Americans. *Health Promotion Practice, 11*(3), 347–357.

Robinson, M. J. (1976). Public affairs television and the growth of political malaise. The case of „Selling the Pentagon". *American Political Science Review, 70*(2), 409–432.

Rossi, P. H. & Freeman, H. E. (1993). *Evaluation*. Thousand Oaks: Sage.

Schaumburg, H. (2001). Neues Lernen mit Laptops? Ein Überblick über Forschungsergebnisse zur Nutzung mobiler Computer in der Schule. *Zeitschrift für Medienpsychologie, 13*(1), 11–21.

Schöffski, O. & Schulenburg, J. M. (2000). *Gesundheitsökonomische Evaluationen*. Berlin: Springer.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *AERA monograph series on curriculum evaluation* (Vol. 1, pp. 39–83). Chicago: Rand McNally.

Scriven, M. (1974). Evaluation perspectives and procedures. In J. W. Popham (Ed.), *Evaluation in education* (pp. 91). Berkeley: McCutchan.

Scriven, M. (2007). The Logic and Methodology of Checklists. Retrieved November 5, 2013, from http://www.wmich.edu/evalctr/archive_checklists/papers/logic&methodology_dec07.pdf

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation. Theories of practice.* Newbury Park: Sage.

Shaw, I. (1999). *Qualitative evaluation. Introducing qualitative methods*. London: Sage.

Slater, J. K. (2006). Review of „Evaluation social programs and problems: Visions for the new millennium". *American Journal of Evaluation, 27*(1), 128–129.

Soellner, R. (2010). Evaluationsmodelle. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (S. 233–243). Göttingen: Hogrefe.

Soellner, R., Hapkemeyer, J., & Scheibner, N. (2010). Evaluationsausbildung in der Psychologie – Eine Bestandsaufnahme. *Zeitschrift für Evaluation, 9,* 297–308.

Spiel, C. (2003). Wissenschaftliche Evaluation. Eine Maßnahme der Qualitätssicherung. *Erziehung und Unterricht, 153*(7–8), 883–888.

Spiel, C., Gradinger, P., & Lüftenegger, M. (2010). Grundlagen der Evaluationsforschung. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (S. 223–232). Göttingen: Hogrefe.

Stake, R. E. (1975). *Program evaluation, particularly responsive evaluation*. Kalamazoo: Evaluation Center, Western Michigan University.

Stake, R. E. (2004). *Standards-based and responsive evaluation.* Thousand Oaks: Sage.

Stamm, M. (2003). *Evaluation und ihre Folgen für die Bildung. Eine unterschätzte pädagogische Herausforderung*. Münster: Waxmann.

Stockmann, R. (2002). *Was ist eine gute Evaluation? Ceval-Arbeitspapiere* (Vol. 9). Saarbrücken: Centrum für Evaluation.

Stufflebeam, D. L. (2001). The metaevaluation imperative. *The American Journal of Evaluation, 22*(2), 183–209.

Stufflebeam, D. L. (2003). The CIPP Model for Evaluation. Annual Conference of the Oregon Program Evaluators Network (OPEN). Abgerufen 6. Februar, 2013, unter https://goeroendeso.files.wordpress.com/2009/01/cipp-modeloregon10-031.pdf.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., et al. (1971). *Educational evaluation and decision making*. Itasca: Peacock.

Stufflebeam, D. L. & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco: Wiley.

The New Yorker. (1993). *Die schönsten Katzen Cartoons.* München: Knaur.

Tscheulin, D. K. (1991). Ein empirischer Vergleich der Eignung von Conjoint-Analysen und „Analytic Hierarchy Process" (AHP) zur Neuproduktplanung. *Zeitschrift für Betriebswirtschaft, 61*(11), 1267–1280.

Uhl, A. (1997). Probleme bei der Evaluation von Präventionsmaßnahmen im Suchtbereich. *Wiener Zeitschrift für Suchtforschung, 20*(3/4), 93–109.

Walter, E. & Zehetmayr, S. (2006). *Guidelines zur gesundheitsökonomischen Evaluation*. Wien: IPF Institut für Pharmaökonomische Forschung.

Wandersman, A., Snell-Johns, J., Lentz, B. E., Fetterman, D. M., Keener, D. C., Livet, M., et al. (2005). The principles of empowerment evaluation. In D. M. Fetterman & A. Wandersman (Eds.), *Empowerment evaluation principles in practice* (pp. 27–41). New York: Guilford.

Weiß, M. & Preuschoff, C. (2004). *Kosten und Effizienzanalysen im Bildungsbereich. Referenzpapier für das Projekt „ Bildungssteuerung"*. Frankfurt/Main: Deutsches Institut für Internationale Pädagogische Forschung.

White, H. (2013). An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness, 5*(1), 30–49.

Wholey, J. S. (1979). *Evaluation. Promise and performance*. Washington, DC: The Urban Institute.

Whyte, W. (Ed.). (1991). *Participatory action research*. Thousand Oaks: Sage.

Widmer, T. (1996). *Metaevaluation. Kriterien zur Bewertung von Evaluationen.* Bern: Haupt.

Widmer, T. (2000). Qualität der Evaluation – Wenn Wissenschaft zur praktischen Kunst wird. In R. Stockmann (Hrsg.), *Evaluationsforschung* (S. 77–102). Opladen: Leske & Budrich.

Widmer, T. (2005). Leitfaden für Wirksamkeitsüberprüfungen beim Bund – Instrument zur Qualitätssicherung gestützt auf die Evaluationsstandards der Schweizerischen Evaluationsgesellschaft (SEVAL–Standards). Abgerufen 5. November, 2013, unter http://www.ejpd.admin.ch/content/dam/data/staat_buerger/evaluation/umsetzung/leitfaden-wirksamkeitsueberpruefung-d.pdf

Widmer, T. & De Rocchi, T. (2012). *Evaluation. Grundlagen, Ansätze und Anwendungen.* Zürich: Rüegger.

Widmer, T., Rothmayr, C., & Serdült, U. (1996). *Kurz und gut? Qualität und Effizienz von Kurzevaluationen*. Zürich: Rüegger.

Wilkens, I. (2002). Kosten und Nutzen der öffentlichen Selbsthilfeförderung: eine Fallstudie am Beispiel der Selbsthilfeförderung der Stadt Göttingen. *Zeitschrift für Sozialreform, 48*, 601–619.

Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505–560). New York: Plenum.

Wittmann, W. W. (1990). Brunswik-Symmetrie und die Konzeption der Fünf-Datenboxen. Ein Rahmenkonzept für umfassende Evaluationsforschung. *Zeitschrift für Pädagogische Psychologie, 4*, 241–251.

Wittmann, W. W., Nübling, R., & Schmidt, J. (2002). Evaluationsforschung und Programmevaluation im Gesundheitswesen. *Zeitschrift für Evaluation, 1*, 39–60.

Wottawa, H. & Thierau, H. (2003). *Lehrbuch Evaluation* (3. Aufl.). Bern: Huber.

Wright, G. (Ed.). (1993). *Behavioral decision making*. New York: Plenum.

Ziegele, F. & Hener, Y. (2004). Benchmarking in der Hochschulpraxis. Effizienzsteigerung und Prozessoptimierung durch kooperatives Handeln. In B. Berendt, H.-P. Voss & J. Wildt (Hrsg.), *Handbuch Qualität in Studium und Lehre* (S. E7.2, 1–20). Berlin: Raabe.

Zukoski, A. & Luluquisen, M. (2002). Participatory evaluation: What is it? Why do it? What are the challenges? *Policy & Practice, 5*, 3–9.

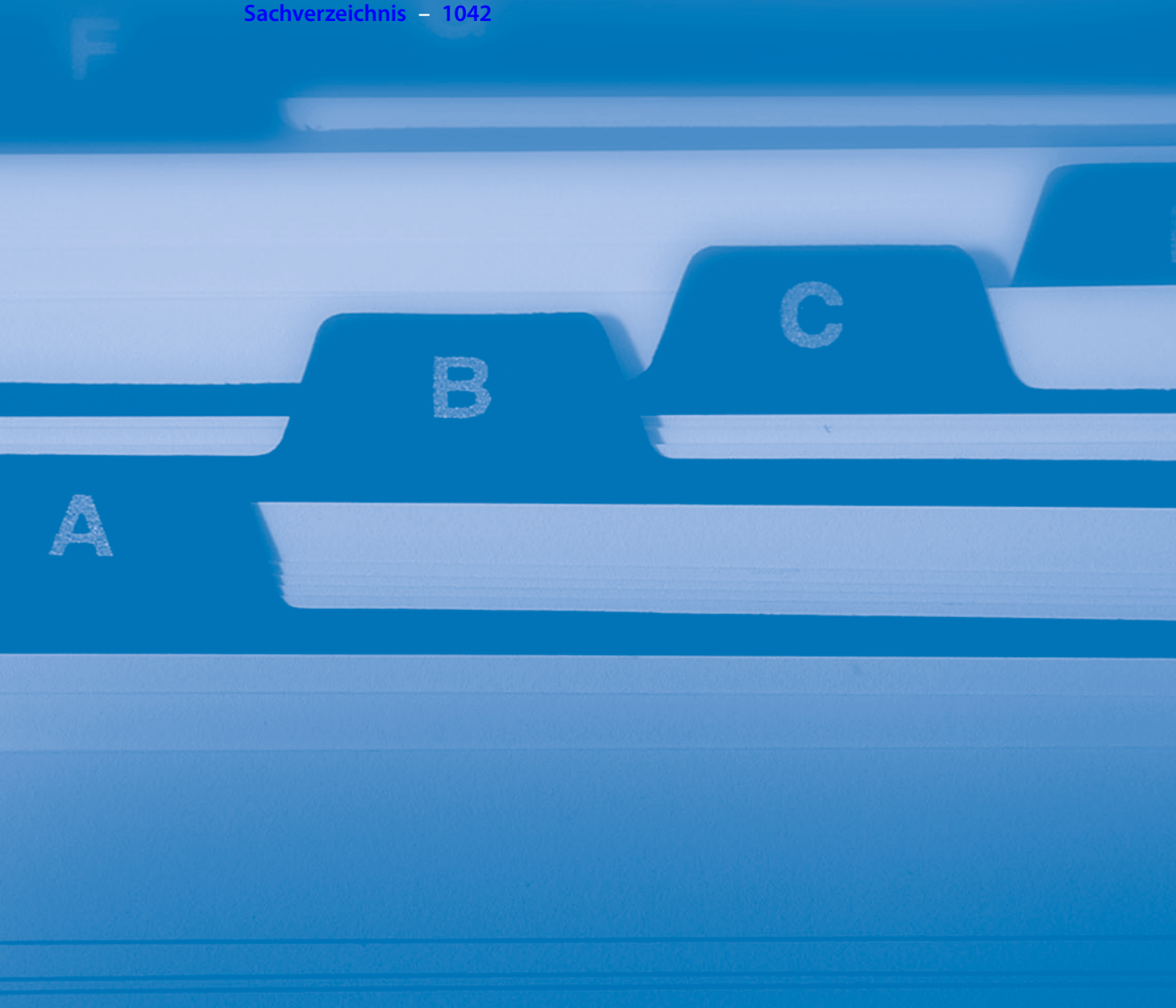
Serviceteil

Anhang – 1038

A.1 SAS-Syntax für die Berechnung einiger Konfidenzintervalle – 1038

A.2 Griechisches Alphabet – 1041

Sachverzeichnis – 1042



Anhang

A.1 SAS-Syntax für die Berechnung einiger Konfidenzintervalle[1]

A.1.1 Konfidenzintervall für δ (► Kap. 14)

```
/*Berechnung der Nichtzentralitätsparameter*/
data nc_ci_u;
/*t-Test für unabhängige Stichproben*/
t=3.44;
df=162;
n1=82;
n2=82;
/*Berechnung der Nichtzentralitätsparameter*/
ncp_low=tnonct(t, df, 0.975);
ncp_up=tnonct(t, df, 0.025);
/*Berechnung des Konfidenzintervalls für Delta*/
delta_lo=ncp_low*sqrt((n1+n2)/(n1*n2));
delta_up=ncp_up*sqrt((n1+n2)/(n1*n2));
run;
proc print;
run;
```

OBS	T	DF	N1	N2	NCP_LOW	NCP_UP	DELTA_LO	DELTA_UP
1	3.44	162	82	82	1.43949	1.43025	0.22481	0.84806

A.1.2 Konfidenzintervall für δ′ (► Kap. 14)

```
data nc_ci_a;
/*t-Test für abhängige Stichproben*/
t=4.38;
df=29;
n=30;
/*Berechnung der Nichtzentralitätsparameter*/
ncp_low=tnonct(t, df, 0.975);
ncp_up=tnonct(t, df, 0.025);
/*Berechnung des Konfidenzintervalls für Delta*/
delta_lo=ncp_low*(1/sqrt(n));
delta_up=ncp_up*(1/sqrt(n));
run;
proc print;
run;
```

OBS	T	DF	N	NCP_LOW	NCP_UP	DELTA_LO	DELTA_UP
1	4.38	29	30	2.09292	6.60989	0.38211	1.20679

[1] Nach Kline (2004). Die fett gedruckten Werte sind frei wählbar (Voreinstellung: 95 %ige Konfidenzintervalle).

A.1.3 Konfidenzintervall für η^2 (▶ Kap. 14)

```
data nc_ci_es;
/*Varianzaufklärung Eta-Quadrat*/
F=4.90;
df1=2;
df2=153;
/*Berechnung der Nichtzentralitätsparameter*/
ncp_low=fnonct(F, df1, df2, 0.975);
ncp_up=fnonct(F, df1, df2, 0.025);
/*Berechnung des Konfidenzintervalls für Eta-Quadrat*/
r_sq_lo=(ncp_lo)/(ncp_lo+df1+df2+1);
r_sq_up=(ncp_up)/(ncp_up+df1+df2+1);
output
run;
proc print;
run;
```

OBS	F	DF1	DF2	NCP_LOW	NCP_UP	R_SQ_LO	R_SQ_UP
1	4.9	2	153	0.67067	24.9075	0.0042808	0.13768

Bei kleinen F-Werten kann es sein, dass kein entsprechender Nichtzentralitätsparameter existiert und ein Konfidenzintervall somit nicht berechnet werden kann.

A.1.4 Konfidenzintervall für einen Einzelvergleich ($\delta\psi$) für $p = 3$ (▶ Kap. 14)

```
data nc_ci_dp;
/*Delta-Psi einfaktorielle Varianzanalyse*/
/*3 unabhängige Stichproben*/
t_cont=1.168; df=12;
n1=5; n2=5; n3=5;
c1=0.5; c2=–1; c3=0.5;
/*Berechnung der Nichtzentralitätsparameter*/
ncp_lo=tnonct(t_cont, df, 0.975);
ncp_up=tnonct(t_cont, df, 0.025);
/*Berechnung des Konfidenzintervalls für Delta-Psi*/
d_psi_lo=ncp_low*sqrt(c1**2/n1 + c2**2/n2 + c3**2/
n3);
d_psi_up=ncp_up*sqrt(c1**2/n1 + c2**2/n2 + c3**2/
n3);
run;
proc print;
run;
```

OBS	T_CONT	DF	N1	N2	N3	C1	C2	C3	NCP_LO	NCP_UP	D_PSI_LO	D_PSI_UP
1	1.168	12	5	5	5	0.5	–1	0.5	–0.86852	3.15894	–0.47571	1.73022

A.1.5 Konfidenzintervall für einen Einzelvergleich ($\delta\psi$) für $p = 4$ (▶ Kap. 14)

```
/*Delta-Psi einfaktorielle Varianzanalyse*/
data nc_ci_d4;
/*4 unabhängige Stichproben*/
t_cont=3.77; df=16;
n1=5; n2=5; n3=5; n4=5;
c1=1; c2=–1/3; c3=–1/3; c4=–1/3;
/*Berechnung der Nichtzentralitätsparameter*/
ncp_lo=tnonct(t_cont, df, 0.975);
ncp_up=tnonct(t_cont, df, 0.025);
/*Berechnung des Konfidenzintervalls für Delta-Psi*/
d_psi_lo=ncp_lo*sqrt(c1**2/n1 + c2**2/n2 + c3**2/n3
+ c4**2/n4);
d_psi_up=ncp_up*sqrt(c1**2/n1 + c2**2/n2 + c3**2/n3
+ c4**2/n4);
run;
proc print;
run;
```

OBS	T_CONT	DF	N1	N2	N3	N4	C1	C2	C3	C4	NCP_LO	NCP_UP	D_PSI_LO	D_PSI_UP
1	3.77	16	5	5	5	5	1	–0.33333	–0.33333	–0.33333	1.37976	6.07849	0.71250	3.13892

A.2 Griechisches Alphabet

Griechisches Alphabet

Griechischer Buchstabe		Name des Buchstabens
A	α	Alpha
B	β	Beta
Γ	γ	Gamma
Δ	δ	Delta
E	ε	Epsilon
Z	ζ	Zeta
H	η	Eta
Θ	ϑ	Theta
I	ι	Jota
K	$\varkappa$	Kappa
Λ	λ	Lambda
M	μ	My
N	ν	Ny
Ξ	ξ	Xi
O	o	Omikron
Π	π	Pi
P	ρ	Rho
Σ	$\sigma\ \varsigma$	Sigma
T	τ	Tau
Υ	υ	Ypsilon
Φ	φ	Phi
X	χ	Chi
Ψ	ψ	Psi
Ω	ω	Omega

Sachverzeichnis

A

A-B-A-B-Plan 768
A-B-A-Plan 767
A-BC-B-BC-Plan 768
Abdeckungsfehler 383
Abduktion 35
A-B-Plan 767
Abstract 794
Adaptives Testen 491
Adhoc-Stichprobe 305
Adjusted-Goodness-of-Fit-Index 968
ADM-Design 392
ADM-Mastersample 316, 388
Adress-Random-Verfahren 388
AGFI 968
Akteursliste 563
Akteursorientierter Evaluationsansatz 1012
Aktive Rekrutierung 411
ALLBUS 624
Alles-auf-einen-Blick-Tabelle 871
Allgemeiner Leistungstest 455
Alltagsbeobachtung 325
Alltagsgespräch 357
Alltagswissen 87
Alpha-Fehler-Kumulation 105
Alpha-Fehler-Niveau 664
Alpha-Fehler-Wahrscheinlichkeit 664
Alternativhypothese 147, 660
Alterseffekte 745
Ambulantes Assessment 417
AMOS 969
Analyseeinheit 555, 603
Anarchismus, methodologischer 61
Anekdotische Evidenz 326
Annahmen der Klassischen Testtheorie 463
Anne-Experiment 402
Anomalistik 89
Anonymisierung 128
– qualitativer Datensätze 584
– quantitativer Datensätze 589
Ansatz, wissenschaftstheoretischer 184
Anspruchsgruppe 981, 983
– wissenschaftsexterne 84
– wissenschaftsinterne 84
Antwortvorgabe, ausbalancierte 439
Anwaltschaftliche Evaluation 1013
Anwendungsforschung 18, 91, 185
Äpfel-und-Birnen-Problem 903
A-posteriori-Einzelvergleiche 709
A-priori-Einzelvergleiche 709
A-priori-Teststärkeanalyse 814
Arbeitsbündnis 69
Arbeitsschritte
– der qualitativen Datenanalyse 603
– der quantitativen Datenanalyse 616
Archivforschung 533
ARIMA(p, d, q)-Modelle 751
Assessment, ambulantes 417
Asymptotischer Permutationstest 773
Auftragsforschung 18, 186
Augenschein-Validität 445
Ausbalancierte Antwortvorgabe 439
Ausgangswertproblematik 505
Ausreißerwert 682
Ausschlussverfahren 591
Auswahlgrundlage 294
Auswahlpopulation 294
Autoethnografie 340
Automatische Telefonbefragung 417
Availability-Bias 902
Axiales Kodieren 547
Axiologie 35
Axiom 56
– der Klassischen Testtheorie 463
– Existenz- 463
– Unabhängigkeits- 464
– Verknüpfungs- 463

B

B-A-B-Plan 767
Baseline 743
Basissatzproblem 41
Baum-Test 452
Bayes-Ansatz 615, 637
Bayes'sche Metaanalyse 905
BBC Prison Study 208
Bedeutungsanalyse 226
Befragung
– für größere soziale Gebilde 421
– halbstrukturierte schriftliche 403
– intervieweradministrierte 386
– Mixed-Mode- 384
– mobile 416
– Modi 296
– Online- 414
– per Austeilen und Einsammeln 413
– Person 362
– postalische 413
– schriftliche 399
– selbstadministrierte 386
– unstrukturierte schriftliche 401
– zu Kleingruppen 422
– zu Organisationen 424
– zu Paaren 421
– zu sozialen Netzwerken 423
Befund
– statistisch falsch-negativer 62
– theoretisch falsch-negativer 62
– theoretisch und statistisch falsch-positiver 62
Begleitende Evaluation 991
Begründungszusammenhang 45
– von Theorien 45
Beobachterübereinstimmung 346
Beobachtung
– Konstruiertheit der 328
– nonreaktive 349
– Perspektivität der 328
– qualitative 332, 335
– quantitative 342, 343
– Selektivität der 328
– -sfehler 330
– -smethode 328
– -sprotokoll 341
– -sprozess 326
– -sregel 327
– wissenschaftliche 323
Bestätigung von Nullhypothesen 885
Beta-Fehler-Wahrscheinlichkeit 671
Bewährungsgrad einer Theorie 42
Bewertungsangst 101
Bewertungsstandards 983
Bildung von neuen Variablen 592
Biopsychologie 500
Bivariate Korrelation 680
Bivariates Assoziationsmaß 681
Blickbewegungskamera 523
Blockplan 729
Blutdruckmessung 516
Bogus-Pipeline-Methode 438
Box-Jenkins-Verfahren 750
Box-Plots 622
Branched Testing 492
Brutto-Stichprobe 384

C

Call for Papers 793
Campbell Collaboration 900
Campbell-Tradition 469
CAPI-Interview 390
CAQDAS 608
CATI-Anlage 392
Ceilingeffekt 738
Ceteris-Paribus
– -Klausel 54, 196
– -Prinzip 99
CFI 968
χ^2-Test 957, 958, 966, 967
CIPP-Evaluationsmodell 1008
Citation-Bias 902
Clusteranalyse 624
Cochrane Collaboration 900
Code
– -buch 553
– -liste 550
– -plan 586
Cohens Kappa-Koeffizient 567
Cohort-Sequential-Design 747
Comparative-Fit-Index 968
Computer-Assisted Telephone Interview 392
Computermodell
– qualitatives 169
– quantitatives 169
Computerprogramm
– für die qualitative Datenanalyse 607
– für die quantitative Datenanalyse 618

– für Metaanalysen 937
Computersimulation 169
Confirmation-Bias 901
Confirmatory Data Analysis 621
Conjoint Measurements 1005
Cost-Bias 902
Cronbachs Alpha-Koeffizient 468
Cross-Lagged-Panel-Design 698
Cross-Sequential-Design 748

D

Data-Mining 628
Daten
– -aufbereitung 580
– -bereinigung 580
– -eingabe 588
– empirische 5
– -erhebung 322
– -qualität 581
– -sätze 580
– -transformation 580
– unvollständige 957
Datenanalyse 598
– deskriptive 631
– explanative 657
– explorative 621
– integrierte 505
– qualitative 599
– quantitative 612
Datenarchiv 589
– -Standard 589
Datenaufbereitung 580
Datenbereinigung 580
Datenerhebung 322
– Aufwand 373
Datenqualität 585
Deckeneffekt 738
Deduktion 35
Deduktionsschluss 37
Delphi-Methode 420
Δ-Koeffizient von Glass 823
Δ-Maß 915
δ-Maß von Cohen 822
Demokratische Evaluation 1014
Demoskopie 193
Design
– Cohort-Sequential- 747
– Cross-Sequential- 748
– Time-Sequential- 747
df 964
Diagnostik 430
– Eignungs- 456
Dialektik 60, 74
Dichotome Indikatorvariable 688
Dichotomes Raschmodell 489
Dichtefunktion 634
Differenzmaß, einfaches 732
Differenzwerte 732
Dimensionale Analyse 226
Dimensionalität eines Tests 479
Direkte Rangordnungen 240
Diskriminanzanalyse 724
Diskursanalyse, kritische 602
Disordinale Interaktion 713
Dokument
– forschungsgeneriertes 533
– Lieferdienst 163
– Qualitätsanalyse 555
– visuelles 571
– vorgefundenes 534
– vorhandenes 538
Dokumentarische Methode 602
Dokumentenanalyse 533
– genuine 534
– qualitative 540
– quantitative 552
– wissenschaftliche 536, 538
Dominanzpaarvergleich 242
Doppelblind-Versuch 198
Dreifachblind-Versuch 198
Dreifachverblindung 587
Dual-Frame-Ansatz 391

E

Ebbinghaus'sche Behaltenskurve 216
Effekt
– direkter 952, 953
– indirekter 952, 953
– totaler 952
Effektgröße 53, 668, 814, 868
– der wichtigsten Signifikanztests 821
– erwartbare 841
– Klassifikation 819
– unstandardisierte bzw. absolute 814
Effektgrößenmaß 816, 914
– Gruppendifferenz- 817
– relatives 814
– Risiko- 817
– standardisiertes 814
– Umrechnung 819
– Varianzaufklärungs- 817
– Zusammenhangs- 817
Effektmodelle der Metaanalyse 897
Effizienz 636
Effizienzorientierte Evaluationsmodelle 1003
Eichung 449
Eignungsdiagnostik 456
Eignungstest, spezieller 457
Einfaches Differenzmaß 732
Einfaktorielle Varianzanalyse 709
Einfallstudie 215
Eingabemaske 587
Ein-Gruppen-Prätest-Posttest
– -Design 202
– -Pläne 738
Einseitiger und zweiseitiger Signifikanztest 666
Einzelfalldiagnostik, Signifikanztest 778
Einzelfallhypothese 765
Einzelfallstudie 215
– quasi-experimentelle 215, 216
Einzelindikator 264
Einzelitem 407
Elektroenzephalografie (EEG) 508, 509
Elektrokardiografie (EKG) 515
Elektromyografie 520
Elektrookulografie 523
Empirischer Wert 665
Empirismus 37
Empowerment-Evaluation 1015
Entdeckungszusammenhang 45, 66
– von Theorien 46, 172
Entwicklungshypothesen 744
Entwicklungstest 456
Epidemiologie 193
Epistemologie 35
Epochale Effekte 746
Epsilon-Korrektur 884
EQS 969
Ereigniskorrelierte Potenziale 510
Ereignisstichprobe 293, 327
Ergebnis
– nicht-signifikantes 669
– signifikantes 668
Erkenntnisgewinn
– vorwissenschaftlicher 87
– wissenschaftlicher 5
Erkenntnisobjekt 59
Erkenntnissubjekt 59
Erkenntnistheoretische Frage 184
Erkenntnistheorie, feministische 71
Erklären 58
Erklärung
– intentionale 34
– kausale 34
– monokausale 55
– multikausale 55
Erklärungsmodell, deduktiv-nomologisches 48, 49
Erwartbare Effektgröße 841
Erwartungseffekt 197
Erwartungstreue 635
Erwünschtheit, soziale 437
Essenziell τ (tau-)äquivalentes Item 468
η^2-Koeffizient 871
Ethikkommission 131
Ethnografisch
– Feldbeobachtung 334
– Feldforschung 334
– Feldinterview 371
Evaluandum 979
Evaluation
– anwaltschaftliche 1013
– begleitende 991
– demokratische 1014
– Empowerment- 1015
– formative 990
– managementorientierte 1012
– Mixed-Methods- 991
– nutzenfokussierte 1013
– ökonomische 1003, 1007
– partizipative 1014
– prospektive 991
– Prozess- 990
– qualitative 991
– quantitative 991
– realistische 1012
– responsive 1014
– retrospektive 991
– -sbericht 1028
– -scheckliste 995
– -sdaten 1026
– -sergebnis 990
– -sforschung 976
– -sfrage 1022
– -sfunktion 987
– -sgegenstand 979
– -shypothese 1022
– -skonzept 1019
– -skriterium 983
– -smodell 995
– -snutzung 988, 1029
– -sobjekt 979
– -sstandard 984, 991
– -sstichprobe 1024
– summative 990
– -svertrag 1021
– Team 1019
– theoriegeleitete 1011
– Ziel 1018
– zielfreie 986
Evaluationsansatz
– akteursorientierter 1012
– ergebnisorientierter 997

– systemischer 1007
– theorieorientierter 1008
Evaluationsforschung
– Phasen 1017
Evaluationsmodell
– effizienzorientiertes 1003
– klientenorientiertes 1012
– stakeholderorientiertes 1014
– wirksamkeitsorientiertes 998
Evaluierende 992
Evidenz, anekdotische 7, 326
Evidenzbasiert 88
Evidenzquelle 448
Exakter Iterationshäufigkeitstest 775
Exakter Permutationstest 772
Exhaustion 44
Exhaustivitätskriterium 238
Existenz einer Kausalbeziehung 204
Exklusivitätskriterium 238
Experiment 193
– natürliches 201
Experimentelle „Manipulation“ 194
Experimentum Crucis 63
Experten-Interview 375
Explanandum 49
Explanans 49
Exploration
– empirisch-qualitative 173
– empirisch-quantitative 173
– methodenbasierte 173
– theoriebasierte 173
Explorative Datenanalyse 621
Explorative Faktorenanalyse 624
Explorative multivariate Auswertungstechniken 623
Exploratory Data Analysis 621
Exposé 154
Ex-post-Erklärung 628
Ex-post-facto-Plan 202
Externe Fremdevaluation 989
Externe Validität 102
Extremgruppe 210
– -nvergleiche 708
Eye-Tracking 523

F

Fachgesellschaft, wissenschaftliche 793
Fachkonferenz 793
Fähigkeitstest 453
Fail-Safe-*N*-Methode 908
Faktor 951
– fester 709
– zufälliger 709
Faktorenanalyse 695, 949
– explorative 951
– konfirmatorische 950, 951
Faktorielles Design 194
Faktorladung 951, 954
Faktorladungsmatrix 625
Fallbezogene Auswertung 603
Fallibilismus 39, 40
Fall-Kontroll-Studie 204, 213
Fallstudie 215
Fallstudie, qualitative 765
Fallübergreifende Auswertung 605
Falsch-negative Entscheidung 105
Falsch-positive Entscheidung 105
Falsifikation 38
Falsifikationismus 37, 38
– methodologischer 37, 42, 43
– raffinierter methodologischer 44, 45
Falsifikationskriterium 49
Falsifikationsversuche 42
Falsifikator 43
Falsifizierbarkeit 39
F-Äquivalente 878
Fehler
– Abdeckungs- 383
– Nonresponse- 384
– Überabdeckungs- 384
– Unterabdeckungs- 383
Fehlinterpretation von Operationalisierungen 232
Fehlspezifikation 957
Feldbeobachtung, ethnografische 334
Feldexperiment 198
Feldforschung, ethnografische 334
Feldinterview, ethnografisches 371
Feldstudie 199, 206
– qualitative 207
– quantitative 207
Feldzugang 337
Fester Faktor 709
Fit-Funktion 956
Fixed-Effect-Modell 897, 913
Flexibilität, methodische 67
Floor- oder Bodeneffekt 738
Flussdiagramm 167
Fokusgruppen-Diskussion 380
Fokussiertes Interview 377
Follow-Up-Messung 209
Forest-Plot 922, 933
Formative Evaluation 990
Formatives Messmodell 230, 463
Formdeuteverfahren 450
Forscherrolle 337
Forschung
– angewandte 185
– interdisziplinäre 15
– -sdesiderat 145
– -sethik 86, 197, 581
– -skompetenz 11
– -slücke 145, 163
– -sökonomie 152, 182
– -spragmatik 152
– -sprogramm 45, 61
– -ssynthese 894
– -süberblick 187
– wissenschaftliche 7
Forschungsansatz
– qualitativer 184
– quantitativer 184
Forschungsethik 182
Forschungsfrage 145, 146, 174
Forschungsgeneriertes Dokument 533
Forschungshypothese 145, 146, 175, 658
Forschungsproblem 144
– wissenschaftliches 85
Forschungsprozess
– qualitativer (zirkulär) 27
– quantitativer 24, 27
– wissenschaftlicher 85
Forschungsstand 164
Forschungsthema 144
Frage
– halboffene 588
– offene 588
Fragebogen 405
– Feinkonzept 407
– -Grobkonzept 405
– -Pretest 410
– -rücklauf 411
– standardisierter 407
– teil-/halbstandardisierter 403
Fragebogenmethode
– qualitative 401
– quantitative 405
– wissenschaftliche 398
Frankfurter Schule 60
Freiheitsgrade 964
Freiwilligkeit 124
Fremdevaluation 989
– externe 989
– interne 989
Frequenzanalyse 559
Funktionelle Magnetresonanztomografie (fMRT) 512
Funktionsprüfungstest, spezieller 457
Funnel-Plot 907

G

G*Power 850, 851
Gabler-Häder-Design 391
Garbage-in-Garbage-out-Problem 903
Gatekeeper 337
Geburtstagsmethode 389
Gelegenheitsstichprobe 305
Genauigkeitskriterium 238
Generalisierbarkeitstheorie 464
Generalisierter $\hat{\eta}^2$-Wert 838
Genuine Dokumentenanalyse 534
Gerichtete Hypothese 668
Gesamt-$\hat{\eta}^2$-Wert 838
Geschichte der Metaanalyse 898
Geschichtete Zufallsstichprobe 312
Gesetze
– deterministische 49
– probabilistische (wahrscheinlichkeitstheoretische) 49
Gesetzmäßigkeit 58
Gesprächsnotiz 367
Gestaltungsverfahren 452
Gewichtung von Stichproben 592
Gezielte Fallauswahl 304
GFI 968
Glaubwürdigkeitsintervall 640
Gleichung, lineare 951, 954
Global-repräsentative Stichprobe 299
Globalvariable 55
Going Native 337
Good-Enough-Prinzip 869
Goodness-of-Fit-Index 968
Griechisch-Lateinisches Quadrat 720
Großtheorien 166
Grounded-Theory-Methodologie 171, 545, 603

Gründe, forschungsökonomische 208
Grundlagenforschung 17, 18, 91, 185, 977
Grundlagenwissenschaftliche Theorie 977
Gruppendifferenz-Effektgrößenmaß 817
Gruppenstudie 215
Gütekriterien 958, 965, 967
– für Punktschätzer 634
Gütekriterium 440, 448
– für Mixed-Methods-Forschung 114
– für qualitative Forschung 106, 109
– für quantitative Forschung 93
– wissenschaftliches 83
Guttman-Skala 272

H

Halbstrukturiertes Interview 371
Halo-Effekt 252
Handlungstheorie 166
HARKing 628
Haupteffekt 711
Hauptgütekriterien psychologischer Tests 440
Hautleitfähigkeit 518
Hawthorne-Effekt 101
Hempel-Oppenheim-Schema 48
Hermeneutik, objektive 602
Heywood Cases 965
Hierarchischer Plan 717
high-stakes testing 430
Hilfs- und Instrumententheorien 42, 51, 54
Hintergrund, theoretischer 165
Holsti-Formel 566
Homogenität der Items 464
Homogenitätsbasierte Metaanalyse 904
Homogenitätstest 919
Humanwissenschaften 5, 13
Hybride Interaktion 713
Hypothese
– gerichtete 147, 668
– gerichtete bzw. einseitige 147
– monokausale 54
– multikausale 54
– probabilistische 50
– spezifische 149
– ungerichtete 147, 668
– ungerichtete bzw. zweiseitige 147
– unspezifische 149
Hypothesenpaar, statistisches 661
Hypothesenprüfende Einzelfalluntersuchung 765
Hypothesenprüfung 614
– statistische 52, 53

I

Idealismus 34
Impact-Factor 787
Implementierungstreue 1001
Incentivierung 155
Index 278
– -bildung 277
– -werte 592
Indikatorvariable 950, 953
– dichotome 688
Indirekte Rangordnungen 241
Induktion 35
Induktionsschluss 37
Inferenzpopulation 294
Informationsgehalt von Theorien 43
Informierte Einwilligung 124
Inhaltsanalyse
– qualitative 541, 602
– quantitative 553, 571
Inhaltsanalytisches Kategoriensystem 553
Inhaltsvalidität 446
Inkommensurabilitäts-Problem 903
Input-Output-Analyse 555
Insiderperspektive 150
Instrumentelle Reaktivität 1001
Intelligenztest 430, 456
Intensitätsanalyse 559
Interaktion
– disordinale 713
– hybride 713
– ordinale 713
– -sdiagramm 712
– -seffekt 104, 195, 712
– vollstrukturiertes Telefon 390
– zweiter Ordnung 740
Interdisziplinarität 13
Interessen-Test 458
Inter-Kodierer-Übereinstimmung 566
Interkorrelationsmatrix 625
Interne Fremdevaluation 989
Interne Konsistenz 467
Interne Validität 99
Interpretative Phänomenologische Analyse 544, 601
Intersubjektive Nachvollziehbarkeit 786
Intervallschätzung 640
Intervallskala 244, 464
– Niveau des Items 464
Interventionsforschung 977
Interventionsmodell 757
Interview
– CAPI- 390
– Experten- 375
– fokussiertes 377
– halbstrukturiertes 371
– Leitfaden- 372, 373
– Modi 386
– narratives 369
– nicht-strukturiertes 369
– Online-Leitfaden- 375
– problemzentriertes 376
– qualitatives 365
– Schulung 362
– Telefon- 390
– telefonisches Leitfaden- 373
– unstrukturiertes 369
– unstrukturiertes/halbstrukturiertes Gruppen- 379
– unstrukturiertes/halbstrukturiertes Paar- 379
– vollstrukturiertes 361, 387
– vollstrukturiertes Online- 393
– vollstrukturiertes persönliches 387
– vollstrukturiertes Telefon- 390
– wissenschaftliches 356, 358
Interviewadministrierte Befragung 386
Interviewfehler 360
Interview-Leitfaden 372
Interview-Modus 386
Intra-Class-Correlation (ICC) 566
Intra-Klassen-Korrelation 568
Intra-Kodierer-Übereinstimmung 566
Irrtumswahrscheinlichkeit 664
Item
– τ (tau-)äquivalentes 468
– -analyse 475
– essenziell τ (tau-)äquivalente 468
– kongenerisches 468
– paralleles 468
– -Response-Theorie (IRT) 461, 482
– -schwierigkeit 476
– -Trennschärfe 478
Iterationshäufigkeitstest 775
– exakter 775

K

Kanonische Korrelation 691
Kategoriensystem 557
– inhaltsanalytisches 553
– Pretest 557
– Revision 557
Kausal
– -aussage 968
– -erklärung 99
– -hypothese 193
– -modell 696
Kausale Zusammenhangshypothesen 695
Kausalität 34
– Nachweis 99
Kerntheorie 42, 54
KiGGS-Studie 650
Klassifikation von Effektgrößen 819
Klassifikationskriterium 399, 538
– für wissenschaftliche Beobachtungsmethoden 328
– für wissenschaftliche Dokumentenanalysen 538
– für wissenschaftliche Interviews 358
– für wissenschaftliche schriftliche Befragungen 399
– für wissenschaftliche Tests 434
Klassische Testtheorie 461, 463
Klassischer Nullhypothesen-Signifikanztest 808
Klassischer Signifikanztest 868
Kleinstquadrat-Methode 637
Klientenorientierte Evaluationsmodelle 1012
Klinischer Test 460
Klumpenstichprobe 314
Kodiereinheit 555
Kodieren
– axiales 547
– offenes 546
– selektives 547

Kodiererschulung 558
Kodierparadigma 547, 551
Kodierung gemäß Grounded-Theory-Methodologie 545
Kohorteneffekte 746
Kohortenstudie, prospektive 204, 213
Kombinierter Signifikanztest 924
Kompromiss-Teststärkeanalyse 815
Konfidenzintervall 644
– als Alternative zum Signifikanztest 673
– des arithmetischen Mittels bei bekannter Varianz 640
– des arithmetischen Mittels bei unbekannter Varianz 645
– eines Populationsanteils 647
– für Populationseffektgrößen 868
Konfigurationsfrequenzanalyse 691
Konfirmatorisch 946
Konfundierung 203
Konfundierungseffekte 100
Kongenerisches Item 468
Königsweg der Erkenntnis 194
Konkordanz 242
Konsistenz 636
– interne 467
Konstanthalten 200
Konstruiertheit der Beobachtung 328
Konstrukt 946, 949
– theoretisches 41
Konstruktivität 328
Konstruktvalidität 97, 98, 100, 469
– in der Campbell-Tradition 469
– in der Messick-Tradition 469
Kontingenzanalyse 559
Kontrolle von Sequenzeffekten 730
Kontrollgruppenplan mit Prä- und Posttest 202
Kontrollskala 439
Kontrollvariable 722
Konvergenz 965
Konversationsanalyse 601
Konzept, sensibilisierendes 66
Konzeptevaluation 990
Konzeptspezifikation 99, 100, 223, 225
Korrelation 683
– bivariate 680
– kanonische 691
Korrespondenzproblem 41, 42, 98
Korrespondenztheorie der Wahrheit 38
Kosten
– -Analyse 1003
– -Effektivitäts-Analyse 1004
– -Nutzen-Analyse 1006
– -Nutzwert-Analyse 1004
Kovarianzanalyse 722
Kovarianzmatrix
– empirische 955–957, 967
– modelltheoretische 955–957, 967
Kovarianzstrukturanalyse 955
Kriterien wissenschaftlicher Qualität 90
Kriterium 146
– der Nützlichkeit 75
– der Sparsamkeit 167
Kriteriumsorientiertes Testen 433
Kriteriums-Teststärkeanalyse 815
Kriteriumsvalidität 447
Kriteriumsvariable 146, 952, 953
Kritische Theorie 60
Kritischer Rationalismus 36
Kritischer Wert 665
Kritizismus 37, 38
Kuder-Richardson-Formel 468
Kumulative Wissenschaft 868

L

Labor 196
Laborstudie 206
– qualitative 207
– quantitative 206
Längsschnittstudie 210, 211, 697
– qualitative 214
– retrospektive (rückschauende) 213
Language-Bias 902
Lateinisches Quadrat 720
Latent-Class-Modell 485
Latente Merkmale 224
Latente Variable 483
Latent-Trait-Modell 485
lavaan 959, 970
Law of Categorial Jugdement 260
Law of Comparative Jugdement 262
Lebenswelt, alltägliche 64
Lehrererwartungs-Effekt 927
Leistungstest 440, 453–455
– allgemeiner 455
Leitfaden-Interview 372, 373
Likelihood 638
– -Ratio-Test 958, 966
Likert-Skala 269
LISREL 969
Literatur
– -beschaffung 162
– -datenbank, wissenschaftliche 159, 160
– -recherche 158, 906
– -verwaltung 789
– -verwaltungsprogramm 161
Local Molar Causal Validity 100
Logik 35
– des statistischen Signifikanztests 659
Log-lineares Modell 691
Lokale stochastische Unabhängigkeit 464, 490
Longitudinalstudie 745
Lost-Letter-Technik 351
Lügendetektor 502, 526

M

Managementorientierte Evaluation 1012
MANCOVA 723
Mangelnde Teststärke 810
Manifeste Merkmale 224
Manifeste Variable 483
Manipulationszwecke 581
Mann-Zeichen-Test (MZT) 452
MANOVA 723
Marienthal-Studie 336
Masterdatensatz 590
Matched Samples 200
Matching von Stichproben 200
Maximalstandard 83
Maximum Likelihood (ML) 957, 966
Maximum-Likelihood-Methode 637, 638
Mcdonalds Omega 468
Mediatoreffekt 697
Mediatorvariable 952, 953
Meehl'sches Paradoxon 692
Mehrfachkodierung 561
(Mehr-)Faktorielle Pläne 710
Mehrfaktorielle Pretest-Posttest-Pläne 740
Mehrgruppenpläne 708
Mehrstufige Zufallsstichprobe 315
Memos 547
Menschenbild 56, 65
Merkmalsspezifisch-repräsentative Stichprobe 298
Messfehler 946, 949
– Unabhängigkeit 464
– -variable 950, 953
– -varianz 951
Messick-Tradition 469
Messmethode, objektive 502
Messmodell 950, 953
– formatives 463
– reflektives 463
Messtheorie 236
Messung 232, 235
– der Hautleitfähigkeit 518
– physiologische 207, 499
– psychophysiologische 499
– sexueller Erregung 524
– soziodemografischer Merkmale 266
Messwiederholung 728
Metaanalyse 166, 187, 191, 894
– Bayes'sche 905
– Gesamteffekt 918
– homogenitätsbasierte 904
– Homogenitätstests für 919
– multivariate 913
– psychometrische 905
– qualitative 894
– quantitative 894
– zur Psychotherapie-Wirkung 898
Metaevaluation 1029
Meta-Metaanalysen 897
Methode
– des lauten Denkens 370
– des permanenten Vergleichs 546
– Kleinstquadrat- 637
– Kompetenz 11
– Maximum-Likelihood- 637, 638
– Momenten- 637
– -nmix 72
– -nstudie 187
– -nvergleichsstudie 187
– Stouffer- 926
– wissenschaftliche 9
Methoden
– -effekt 386
– -forschung 582
Methodologie 35, 85
– Grounded-Theory- 545, 603

– kritische 71
– wissenschaftliche 9
Mikrozensus 390
Milde-Härte-Fehler (Leniency-Severity-Fehler) 253
Milgram-Experiment 126
Mindeststandard 83
Minimum-Effektgrößen-Test 614
Minimum-Effekt-Nullhypothese 870
Mischverteilungs-Modell 485
Missing Data 590
Missing-Data-Problem 904
Mixed-Effects-Modell 897, 914
Mixed-Methods
– -Ansatz 184
– -Evaluation 991
– -Forschung, Gütekriterien 114
– -Paradigma 72
– -Sozialforschung 17
Mixed-Mode-Befragung 384
ML *siehe* Maximum Likelihood
Mobile Befragung 416
Modell
– -anpassungstest 890
– -bildung, grafische 167
– deduktiv-nomologisches 57
– -fehler 957
– -Fit 967
– Fixed-Effect- 897, 913
– genau identifiziertes 964
– -gleichungsfehlervariable 953
– -güte 967
– hierarchisch geschachteltes 966
– -identifikation 964
– Interventions- 757
– konzeptuelles 167
– log-lineares 691
– Metaanalyse-Effekt- 897
– Mixed-Effects- 897, 914
– Random-Effects- 897, 913
– Transferfunktions- 758
– überidentifiziertes 964
– unteridentifiziertes 964
– Vorhersage- 756
Moderator
– -effekt 697
– -variable 146, 895
– -variablen-Analyse 921
Modifikationsindizes 966
Modus Tollens 48
Modus-Effekt 386
Momenten-Methode 637
Mono-Method-Design 184
Mono-Methoden-Bias 100
Mono-Operationalisierungs-Bias 100
Mortalität, experimentelle 103
Motivations-Test 458
Mozart-Effekt 190
Mplus 970
MTMM-Matrix 473
Multidimensionale Skalierung 626
Multidisziplinarität 13
Multi-Facetten-Modell der Evaluation 1007
Multi-Mode-Befragung 296
Multiple Korrelations- und Regressionsanalyse 626
Multiple Regression 690
Multiple Zusammenhänge 687
Multiple-Baseline-Plan 768
Multitrait-Multimethod-Analyse 472
Multivariate Metaanalyse 913
Multivariate Pläne 723
Multizentrische Studie 718
Münchhausen-Trilemma 39

N

Nachvollziehbarkeit, intersubjektive 786
Namengenerator 425
Namensinterpretator 426
Narrative Analyse 601
Narratives Interview 369
Narratives Review 898
Nebengütekriterien psychologischer Tests 448
Nested Models 966
Netto-Stichprobe 384
Netzwerkanalyse 423
Neuropsychologie 500
Neurotizismus 458, 480
Nichtnormalität 966, 967
Nicht-signifikantes Ergebnis 669
Nicht-strukturiertes Interview 369
Nicht-Verfälschbarkeit 449
Nicht-Wissenschaft 86, 89
Nil-Nullhypothese 868
– Signifikanztest 808
Nominaldefinition 224
Nominalskala 237
Nonreaktive Beobachtung von Verhaltensspuren 349
Non-Response-Fehler 296, 384
Non-Statement-View 62
Normalverteilung 966
Normierung 449
Normorientiertes Testen 433
Novitätseffekt 101
Nullhypothese 147, 660
– Nil- 868
H_0-Modell 662
Nullhypothesen-Signifikanztest
– Klassischer 808
– Nil 808
Nullmodell 967
Nutzenfokussierte Evaluation 1013
Nützlichkeit 449

O

Objektiv
– Hermeneutik 602
– Messmethode 502
– Persönlichkeitstest 438
Objektivität 46, 47, 184, 442, 443
– Auswertungs- 443
– Durchführungs- 443
– Interpretations- 443
Ockhams Rasiermesser 57
Odds Ratio 818
Offenes Kodieren 546
Offenheit, theoretische 66
Ökonomische Evaluation 1003, 1007
Omega-Quadrat-Maß 836
Omnibustest 921
One Stop F Table 871
One-Shot Case Study 102, 202
Online-Befragung 414
Online-Experiment 199
Online-Interview
– Leitfaden- 375
– vollstrukturiertes 393
Ontologie 34
Open-Access-Initiative 159
OpenMx 970
Operationale Definition 228
Operationalisierung 228, 946
– von abhängigen Variablen 230
– von unabhängigen Variablen 231
Optimaler Stichprobenumfang 671, 815, 841
– für Konfidenzintervalle 648
– für Signifikanztests 843
Ordinale Interaktion 713
Ordinalskala 239
Originalstudie 188
Originäre Rangreihe 240
Othering 337
Outsider-Perspektive 337
Overall-Signifikanz 709

P

Paarvergleiche 241
Panel 212
– sozioökonomisches 213
Panelstudie 210
Paradigma 9
– mixed-methods- 72
– qualitatives 63
– quantitatives 36
Paralleles Item 468
Parallelisierung von Stichproben 200
Paralleltest-Reliabilität 466
Parameterschätzung 613
– standardisierte 962
Parametrisierung 506
Parapsychologie 89
Parawissenschaft 86, 89
Partialkorrelation 684, 685
Partizipative Evaluation 1014
Passive Rekrutierung 411
Peer-Review 794
– Prozess 791
– Verfahren 84
Persönliche Umfrage 387
Persönlichkeits-Struktur-Tests 458
Persönlichkeitstest 436, 437, 457
– objektiver 438
Perspektivität der Beobachtung 328
Perzentil 259
Pfad
– -analyse 699, 949, 952
– -diagramm 949, 950, 953
– -koeffizient 954
– -modell 167
Pfeildiagramm 167
Phasen der Evaluationsforschung 1017
Physiologische Messung 499
Physiologische Psychologie 500
Pläne mit Kontrollvariablen 721
Plausibilitätschecks 590
Polygraf 502
Population 292
Populationsbeschreibende Studien 593
Populationsparameter 613

Positivismus 37
Positivismusstreit 60
Post-/Neopositivismus 38
Postalische Befragung 413
Poster, wissenschaftliches 799
Post-hoc-Teststärkeanalyse 809
Postskriptum 367
Power-Test 453
Prä- und Post-Messung 209
Prädiktoren 146
Prädiktorvariable 146, 952, 953
Pragmatismus 74
Präsentationsfolie 795
Präsentationsqualität 92
Praxeologie 602
Pretest 557
Pretest und Revision des Kategoriensystems 557
Pretest-Effekt 103
Pretest-Posttest
– Design 102
– Kontrollgruppendesign 102
Primacy-Recency-Effekt 254
Primäranalyse 191
Primärquellen 159
Prinzip der lokalen stochastischen Unabhängigkeit 490
Probabilistische Stichproben 310
Probabilistische Testtheorie (PTT) 461, 482, 487
Probabilistisches Testmodell 485, 486
Problem
– der Mehrfachkodierung 561
– Garbage-in-Garbage-out 903
– Inkommensurabilitäts- 903
– Missing-Data 904
Problemzentriertes Interview 376
Programmtheorie 977, 980
Programmtreue 1001
Projektiver Test 450
Projektmanagement 153, 1019
Propensity-Score-Analyse 200
Proposition 56
Prospektive Evaluation 991
Prozentuale Übereinstimmung 566
Prozessevaluation 990
Prozessualität 67
Prüfung von
– H_{00} 872
– H_{01} 874
– H_{05} 876
Prüfverteilung 662
Pseudo-Hypothesentest 628
Pseudowissenschaft 86, 87, 89
Psychologie
– Bio- 500
– Neuro- 500
– physiologische 500
Psychologischer Test 430, 432, 436, 440, 448
Psychometrisch
– Metaanalyse 905
– Skala 267, 407
– Test 453
Psychophysiologie 500
Psychophysiologische Messung 499
Psychotest 434, 435
Publication Manual 788
Publication-Bias 901, 907
Publikationen, wissenschaftliche 787
Pulsfrequenzmessung 516
Punktschätzung 631
– Gütekriterien für 634
– Methoden 637
Pygmalion-Effekt 928

Q

QALY 1005
QDA-Software 608
Quadratische Pläne 719
Qualifikationsarbeit 786
Qualität
– qualitativer Studien 112
– wissenschaftliche 82, 84
Qualitativ
– Datenanalyse 599
– Dokumentenanalyse 540
– Evaluation 991
– Fallstudie 765
– Fragebogenmethode 401
– Inhaltsanalyse 541, 602
– Interview 365
– Metaanalyse 894
Qualitative Beobachtung 332
– mit geringem Komplexitätsgrad 335
Qualitative Forschung
– Gütekriterien 106, 109
Qualitativer Stichprobenplan 303
Qualitätsanalyse von Dokumenten 555
Qualitätsindikatoren 83
Qualitätskriterien
– wissenschaftliche 83
Qualitätsstandards 985
Quantitativ
– Beobachtung 342
– Beobachtung mit geringem Komplexitätsgrad 343
– Datenanalyse 612
– Dokumentenanalyse 552
– Evaluation 991
– Fragebogenmethode 405
– Metaanalyse 894
Quantitative Forschung
– Gütekriterien 93
Quantitative Inhaltsanalyse 553, 571
– von visuellen Dokumenten 571
Quasi-Experiment 102, 199
(Quasi-)experimentelle Studie ohne Messwiederholungen 209
Querschnittstudie 210
Querschnittuntersuchung 745
Quotenstichprobe 307

R

R (Statistikprogrammumgebung) 959, 970
Random-Digit-Dialing-Verfahren 391
Random-Effects-Modell 897, 913
Randomisierung 196
Randomize(d)-Last-Digit-Verfahren 391
Randomized-Response-Technik 188
Random-Response-Technik 438
Random-Route-Verfahren 388
Random-Walk-Verfahren 388
Rangsummentest 776
Ratekorrektur bei Testaufgaben 441
Rater-Ratee-Interaktion 254
Ratingskalen 245, 246
Rationalismus 37
Reaktivität
– der experimentellen Situation 101
– instrumentelle 1001
Realismus 34
– kritischer 40
– naiver 40
Realistische Evaluation 1012
Referenzbereich 83
Referenzwert 83
Reflektives Messmodell 230, 463
Regelstandard 83
Register-Stichprobe 388
Regression
– Analyse 947
– Diskontinuitäts-Analyse 741
– for explanation 963
– for prediction 963
– zur Mitte 210, 735
Regressionseffekt
– statistischer 103
Reifungsprozess 103
Rekrutierung 296
– aktive 411
– passive 411
Relative Risk 818
Relatives Effektgrößenmaß 814
Relevanz
– inhaltliche 90
– praktische 91
– theoretische 90
– wissenschaftliche 90, 150
Reliabilität 116, 442, 444, 465–467, 469, 946
– Paralleltest- 466
– Retest- 465
– Split-Half- 467
– Testhalbierungs- 467
– von Untertests 469
Replikation 46, 104, 190
Replikationsstudie 42, 150, 169, 188
– exakte 189, 190
– systematische 189, 190
Replizierbarkeit 184
Repositorien 159, 162
Repräsentativität 297, 298
Resampling-Verfahren 637
Residualvariable 953
Residuen 957
– standardisierte 957
Respondent Driven Sampling 308
Responsive Evaluation 1014
Retest-Reliabilität 465
Retrospektive Evaluation 991
Retrospektive Teststärkeanalyse 813
Review, narratives 898
Revision 557
Revolution, wissenschaftliche 59
Rigorous Impact Evaluation 999
Risiko-Effektgrößenmaß 817
RMSEA 967
Robustheit 636

Rohdaten 580
– -material 580
Rohwerteverteilung 476
Root Mean Square Error of Approximation 967
Rorschach-Test 451
Rosenthal-Effekt 101
Rücklauf
– -kurve 412
– -quote 296, 412
– -statistik 412

S

Saisonale Modelle 755
SARIMA$(p, d, q) - (P, D, Q)$-Modell 755
Satorra-Bentler-scaled-χ^2-Test 966, 967
Satzergänzungsmethode 402
Scatter-Plot 623
– -Matrix 623
Scenotest 453
Schätzmethode 956
Schätzprobleme 965
Schätzung, iterative 965
Scheinwerfertheorie der Erkenntnis 38
Schneeballverfahren 308
Schreiben von Memos 547
Schreibtischforschung 187
Schriftliche Befragung 399
Schultest 457
Schulung von Interviewenden 362
Schwedenschlüssel 389
Scientific Community 798
Sekundäranalyse 191, 213
Selbstadministrierte Befragung 386
Selbstevaluation 989
Selektionseffekt 103
Selektives Kodieren 547
Selektivität
– der Beobachtung 328
– des Beobachtungsprozesses 326
SEM *siehe* Strukturgleichungsmodelle
sem (R-Paket) 970
Semantisches Differenzial 276
Sensitivitäts-Teststärkeanalyse 815
Sequenzanalytisches Vorgehen 602
Sequenzeffekte 731
Signifikantes Ergebnis 668
Signifikanz
– statistische 868
Signifikanzen-Fischen 105, 628
Signifikanzentscheidung 664
Signifikanzniveau 665
Signifikanztest
– auf Probe 627, 708
– Effektgröße 821
– einseitiger und zweiseitiger 666
– für den metaanalytischen Gesamteffekt 919
– für die Einzelfalldiagnostik 778
– für Unterschiedshypothesen 706
– für Veränderungshypothesen 727
– für Zusammenhangshypothesen 677
– klassischer 53, 868
– kombinierter 924
– optimaler Stichprobenumfang 843
– statistischer 50
Simpson-Paradox 693
Single-Item-Measures 267
Skala
– Intervall- 464
– Nominal- 237
– Ordinal- 239
– psychometrische 407
– Stanine- 459
– Verhältnis- 256
Skalenbildung 267
Skalenmarken 245
Skalenniveau 232
Skalenstufen 249
Skalentransformationen 257
Skalenverankerung 248
Skalenwert 592
Skalierbarkeit 449
Skalierer 954
Skalierung 449, 464, 954
Social Desirability Scale 439
Software 969
– für die qualitative Datenanalyse 607
– für die quantitative Datenanalyse 618
– für Metaanalysen 937
Solomon-Viergruppen-Plan 210, 731
Sozial- und Humanwissenschaften 12
Soziale Erwünschtheit 437
Sozialforschung
– empirische 14
– qualitative 16
– quantitative 15
Sozialkonstruktivismus 71
Sozialwissenschaften 5
Soziogramm 422
Soziomatrix 422
Soziometrie 422
Speed-Test 453
Spezieller Eignungstests 457
Spezieller Funktionsprüfungstest 457
Spezifitätsproblematik („specificity problem“) 505
Split-Half-Reliabilität 467
Spontan-EGG 510
Sprechblasentest 402
SPSS 587
SQ3R-Methode 21
SRMR 967
Stakeholder 981, 983
Stakeholderorientierte Evaluationsmodelle 1014
Standardisierter Fragebogen 407
Standardisiertes Effektgrößenmaß 814
Standardized Root Mean Square Residual 967
Standardnormalverteilung 642
Standard-Random-Verfahren 388
Standards der Wissenschaftlichkeit 85
Standpunkt-Theorie, epistemologische 71
Stanford-Prison-Experiment 208
Stanine-Skala 459
state 460
Statistisch
– Hypothese 661
– Hypothesenpaar 661
– Signifikanz 868
– Signifikanztest 659
Statistische Validität 106
Stem-and-Leaf-Plots 622
Stichprobe
– Brutto- 384
– Erhebung 215, 294
– Fehler 957
– für qualitative Studien 302
– für quantitative Studien 305
– Gewichtung 592
– Kennwert 613
– Konstruktion 296
– -nart 294
– Netto- 384
– optimaler Umfang 294, 671, 815, 841
– Register- 388
– Umfang 294
Stichprobenverfahren 387, 390
– für persönliche Umfragen 387
– für spezielle Populationen 308
– für Telefonumfragen 390
Störungseffekt 101
Störvariable
– personenbezogene 196
– Umwelt- bzw. untersuchungsbedingte 196
– untersuchungsbedingte 197
Stouffer-Methode 926
Strenge
– ethische 91
– methodische 91
Strukturalismus, wissenschaftstheoretischer 62
Strukturgleichungsmodell 176, 591, 890, 946, 947, 949
Strukturierte Verhaltensbeobachtung 345
Strukturierung/Standardisierung des Forschungsprozesses 33
Strukturmodell 953
Studie
– anwendungswissenschaftliche 185
– deskriptive 149, 192
– Einzelfall- 215
– empirische 187, 188
– experimentelle 192, 193
– explanative 149, 192
– explorative 149, 192
– Feld- 206
– grundlagenwissenschaftliche 185
– Gruppen- 214
– Labor- 206
– Längsschnitt- 211
– Methoden- 187
– mit Messwiederholung 208
– Mixed Methods- 184
– nicht-experimentelle 194, 201
– nicht-randomisierte kontrollierte 193
– ohne Messwiederholung 208
– Primär- 191
– prospektive (vorausschauende) 213
– qualitative 184
– quantitative 184
– quasi-experimentelle 193
– Querschnitt- 210

– randomisierte kontrollierte 193
– Sekundär- 191
– Theorie- 186
– Trend- 211
Subgruppen-Analyse 921
Subjektivität
– der Forschenden 70
– reflektierte 111
Subskala 458, 480, 481
Suchbegriffe 158
Suffizienz 636
Summative Evaluation 990
Systematisches Review 900
Systemischer Evaluationsansatz 1007
Systemtheorie 166

T

Tagebuchmethode 417
– halbstrukturierte 405
Tailored Testing 492
Targeted-Sampling-Ansatz 310
τ (tau-)äquivalentes Item 468
Täuschung 197
Tea Room Trade 338
Technik des vertraulichen Kuverts 188
Technologische Theorie 977
Teilhierarchische Pläne 719
Teiltranskription 583
Telefonbefragung, automatische 417
Telefonisches Leitfaden-Interview 373
Telefonumfragen 390
Tendenz zur Mitte 253
Test
– -anwender 431
– -aufgabe 441
– Baum- 452
– -bibliothek 433
– Eignungs- 457
– Entwicklungs- 456
– Fähigkeits- 453
– -fairness 430, 449
– Funktionsprüfungs- 457
– -halbierungs-Reliabilität 467
– Intelligenz- 430, 456
– Interessen- 458
– -Item 454
– -Items in Leistungstests 454
– klinischer 460
– -kompendium 433
– Leistungs- 440, 453–455
– Mann-Zeichen- (MZT) 452
– -modell, probabilistisches 485, 486
– Motivations- 458
– -müdigkeit 210
– -ökonomie 449
– -person 431
– Persönlichkeits- 436, 437, 457
– Persönlichkeits-Struktur- 458
– Power- 453
– projektiver 450
– psychologischer 430, 432, 436, 440, 448
– psychometrischer 453
– Rorschach- 451
– Sceno- 453
– Schul- 457
– -schutz 432
– Speed- 453
– -stärke 105
– thematischer Apperzeptions- 451
– -übung 103, 210
– Wartegg-Zeichen- (WZT) 452
– -wert, Gültigkeit 448
– wissenschaftlicher 434
Testen, adaptives 491
Teststärke 669, 809, 868
– Ausreichende Sicherung 841
– Mangelnde 810
Teststärkeanalyse 809
– A-Priori 814
– Kompromiss- 815
– Kriteriums- 815
– Post-hoc 809
– Retrospektive 813
– Sensitivitäts- 815
Testtheorie 461
– klassische 461, 463
– probabilistische 461, 482, 487
Testverfälschung
– bei Leistungstests 440
– bei Persönlichkeitstests 436
Thematische Analyse 605
Thematischer Apperzeptionstest 451
Theorem 56
Theoretisch
– Begriff 222
– Konstrukt 222
– Konzept 222
– Stichprobe 302
Theorie
– Generalisierbarkeits- 464
– geringer Reichweite 166
– grundlagenwissenschaftliche 977
– Item-Response- 482
– mittlerer Reichweite 166
– -modell, integratives 170
– Pluralismus 45
– Programm- 977, 980
– -rahmen der Evaluation 996
– -studie 186
– technologische 977
– Umfrage- 384
– wissenschaftliche 56
Theoriearbeit 63
– reine 166
Theoriebildung 66, 605
– gegenstandsverankerte 170
Theoriegeleitete Evaluation 1011
Theorieorientierter Evaluationsansatz 1008
Thurstone-Skala 274
Tiefenhermeneutik 602
Time-Sequential-Design 747
Total Survey Error 383
Total-Survey-Error-Modell 416
Transdisziplinarität 14
Transferfunktionsmodell 758
Transkription 367, 583
– -ssoftware 367
– -ssystem 367
– Teil- 583
– Voll- 583
Treatment 199, 202
Treatmentdiffusion in die Kontrollgruppe 101
Trendstudie 210
Trendtest 771
Trennschärfe eines Items 478
Triangulation 72
Trim-and-Fill-Prozedur 909
Typenbildung 605

U

Überabdeckungs-Fehler 384
Übersichtsartikel 898
ULS *siehe* Unweighted Least Squares
Umfrage, persönliche 387
Umfrage-Theorie 384
Umpolung von Variablen 592
Umrechnung standardisierter Effektgrößenmaße 819
Unabhängigkeit
– lokale stochastische 464, 490, 950
– wissenschaftliche 47
Unabhängigkeitsmodell 967
Ungerichtete Hypothese 668
Unstandardisierte bzw. absolute Effektgröße 814
Unstrukturiertes Interview 369
Unterabdeckungs-Fehler 383
Unterschiedshypothese 203
Untersuchungsdesign 182, 183
Untertest 469
Unweighted Least Squares (ULS) 957, 968
Urteilsfehler 252
Usability-Test 207

V

Validität 184, 199, 445, 946
– Augenschein- 438, 445
– diskriminante 446
– externe 53, 97, 102, 104, 198, 208
– faktorielle 446
– Inhalts- 446
– inkrementelle 447
– interne 53, 97, 99, 103, 196, 206
– konkurrente 447
– Konstrukt- 97, 98
– konvergente 446
– Kriteriums- 447
– ökologische 106, 198, 206
– prognostische 447
– retrospektive 447
– -skoeffizient 469
– statistische 97, 105, 106
– Typen 93
Variable
– abhängige 194
– Bildung 592
– endogene 953
– exogene 953
– Geschlecht 204
– konfundierende 196
– latente 483, 949–951
– manifeste 483, 949, 951
– -nlabel 587
– Umpolung 592
– unabhängige 194
Varianz, gemeinsame 949
Varianzanalyse
– einfaktorielle 709
– mit Messwiederholungen 745
Varianzaufklärungs-Effektgrößenmaß 817
Veränderungshypothese 147
Veränderungsstudie 210
Verbal-thematisches Verfahren 451
Verblindung 196

Verfahren
– imputationsbasiert 591
– modellbasiert 591
– verbal-thematisches 451
– zur Erfassung aktueller Zustände 460
Verhaltensbeobachtung, strukturierte 345
Verhaltensspuren 349
Verhältnisskala 256
Verifikation 37, 38
Verlaufsanalyse 506
Verstehen 58
Versuchsleiter 197
– -effekt 101, 197
– -Erwartungseffekt 101, 927
Versuchsperson, „blinde“ 197
Versuchsplan
– Formalisierung 102
– multivariater 195
– univariater 195
Verteilung
– Funktion 634
– Rohwerte- 476
Vertiefungsmodell 185
Vertraulichkeit 128
Verwendungszusammenhang 46
Visualisierungstechnik 621
Visuelles Dokument 551, 571
Volkszählung 293
Vollerhebung 215, 292
Vollstrukturiertes Online-Interview 393
Volltranskription 583
Voraussetzung für statistische Signifikanztests 300
Vorgefundenes Dokument 534
Vorhandenes Dokument 538
Vorhersagemodell 756
Vorstudienmodell 184
Vote Counting 925

W

Wahrheitssuche 40
Wahrheitsverständnis, korrespondenztheoretisches 58
Wahrscheinlichkeitsfunktion 634
Wartegg-Zeichen-Test (WZT) 452
Web Content Mining 629
Web Structure Mining 629
Weltmodell von Forrester 169
Wert
– empirischer 665
– fehlender 590, 957
– -freiheit 47, 60
– kritischer 665
– Label 587
– -neutralität 48
– -rationalität 60
Werturteilsstreit 60
Wikipedia 159
Wilson-Online-Rechner 824
Wirkprinzip, kausales 204
Wirksamkeitsorientierte Evaluationsmodelle 998
Wirkungsrichtung, kausale 204
Wissenschaft 86, 89
– kumulative 868
– -sbildung 11
– -sdisziplinen 12
– -sethik 86
– -sfeindlichkeit 10
– -sforschung 35
– -sgeschichte 32, 35
– -sgesellschaft 5
– -sgläubigkeit 10
– -skommunikation 801
– -smarketing 801
– -spraxis 35
– -ssoziologie 32, 35
– -ssystem 12
Wissenschaftlichkeit 84, 86
– Standards 85
Wissenschaftstheorie 32
– allgemeine 34
– explizite 36
– fachspezifische 34
– implizite 36
Wissensgesellschaft 5
WLSMV 968
Wording-/Framing-Technik 188
Wunsch-Nullhypothese 885

Z

Zeichnerische Verfahren 452
Zeitreihe 749
Zeitreihenanalyse 214
Zeitschriftenartikel 787
Zeitstichprobe 327
Zentrale Tendenz 253
Zentrales Grenzwerttheorem 640
Zielfreie Evaluation 986
Zielpopulation 294
Zirkel, hermeneutischer 67
Zirkularität 67
z-Transformation 642
Zufälliger Faktor 709
Zufallsexperiment 633
Zufallsstichprobe 196
Zufallsvariable 633
Zufallszahlen 310
Zumutbarkeit 449
Zusammenhang
– faktorieller 694
– positiv gerichteter 146
Zusammenhangs-Effektgrößenmaß 817
Zusammenhangshypothese 146, 203
– bivariate 146
– hierarchische 695
– multiple 146
– negativ gerichtete 146
Zweckrationalität 60
Zwei-Gruppen-Plan 194, 707
Zwei-Gruppen-Pretest-Posttest-Plan 739