Rapid #: -27595462

CROSS REF ID: **8015463530005510**

LENDER: **RZSLUHSA (RZSLUHSA) :: Main Library**

BORROWER: **ZHAWP_RS (ZHAWP_Resource Sharing Library) :: Main Library**

TYPE:	Book Chapter
BOOK TITLE:	Sozialforschung Methoden und Anwendungen : ein Überblick für die BA-Studiengänge
USER BOOK TITLE:	Sozialforschung Methoden und Anwendungen : ein Überblick für die BA-Studiengänge
CHAPTER TITLE:	Kapitel 10 Was ist gute Forschung?- Bewertung empirischer Studien
BOOK AUTHOR:	Flick, Uwe
EDITION:	Originalausgabe
VOLUME:	
PUBLISHER:	Rowohlt Taschenbuch Verlag
YEAR:	2009
PAGES:	261-280
ISBN:	9783499557026
LCCN:	
OCLC #:	(SLSP)991148669279705501

Processed by RapidX: 10/7/2026 5:51:48 AM

This material may be protected by copyright law (Title 17 U.S. Code)

Zentral- und Hochschulbibliothek Luzern

Auftragsnummer

18895757790005505

Exemplarstrichcode

ILUM02470689

Rapido Digitization

Luzern, 07/10/2026

Ihre Benutzernummer: RZSLUHSA
Ihre Mail-Adresse:

Lieferschein Digitalisierungsauftrag

Sehr geehrte Benutzerin, sehr geehrter Benutzer

Wir danken für Ihren Digitalisierungsauftrag. Beiliegend erhalten Sie die gewünschten Seiten gemäss Ihrer Bestellung Nr. 18895757790005505 vom 07/10/2026.

Sozialforschung Methoden und Anwendungen : ein Überblick für die BA-Studiengänge

MR 2000 2009

ILUM02470689 / [ZHB] Speicherbibliothek / Büron

Autor:

Titel:

Seiten: 261-280

Ausgabe/Band:

Bemerkung:

Bestellnotiz: Titel:Sozialforschung, Kapitel-Titel:Kapitel 10 Was ist gute Forschung?- Bewertung empirischer Studien, Kapitel-Verfasser:Flick, Uwe, Seiten:261-280, Veröffentlichungsdatum:2009

Die entstandenen Gebühren (max. 50 Seiten pro Bestellung: CHF 5 ; für Firmen max. 50 Seiten pro Bestellung: CHF 25) werden innerhalb einer Woche in Ihrem Benutzerkonto einsehbar sein.

Das Gebühreninkasso wird je nach Bibliothek entweder vom Bibliothekssystembetreiber SLSP, swisscovery (per Email) oder aber über den Kanton Luzern (per Brief) betrieben. Die Rechnungen werden jeweils im übernächsten Monat mit einem Einzahlungsschein verschickt.

Die Verantwortlichkeit für die Einhaltung des Urheberrechts liegt bei der bestellenden Person.

Freundliche Grüsse

Zentral- und Hochschulbibliothek, Benutzungsdienste

Uwe Flick

Sozialforschung

Methoden und Anwendungen
Ein Überblick für die BA-Studiengänge



rowohlts enzyklopädie
im Rowohlt Taschenbuch Verlag

rowohlts enzyklopädie
Herausgegeben von Burghard König

Originalausgabe
Veröffentlicht im Rowohlt Taschenbuch Verlag,
Reinbek bei Hamburg, September 2009
Copyright © 2009 by Rowohlt Verlag GmbH,
Reinbek bei Hamburg
Umschlaggestaltung any.way, Walter Hellmann
Satz Proforma, TheSans (InDesign)
bei Pinkuin Satz und Datentechnik, Berlin
Druck und Bindung CPI – Clausen & Bosse, Leck
Printed in Germany
ISBN 978 3 499 55702 6

10 Was ist gute Forschung? – Bewertung empirischer Studien

Bewertung empirischer Forschung 261
Qualität und Bewertung quantitativer Forschung 262
Qualität und Bewertung qualitativer Forschung 270
Verallgemeinerung 274
Leitfragen für die Bewertung konkreter Projekte 278

Ziele: Nachdem Sie dieses Kapitel gelesen haben, sollten Sie

- die wichtigsten Kriterien zur Bewertung empirischer Forschung kennen,
- verstanden haben, dass diese Kriterien ursprünglich für standardisierte Forschung entwickelt wurden,
- sehen, dass für qualitative Forschung andere Kriterien und Bewertungsansätze angewendet werden,
- erkannt haben, dass ein wesentlicher Bestandteil der Bewertung von Sozialforschung die Verallgemeinerbarkeit der Ergebnisse ist, und
- wissen, welche Formen und Probleme dabei für quantitative und qualitative Forschung relevant sind.

Bewertung empirischer Forschung

Neben der Frage der Nützlichkeit empirischer Forschung und ihrer Ergebnisse ist auch die nach deren Güte zu stellen. Hier wird geprüft, ob die eingesetzten Methoden verlässlich sind und inwieweit die erzielten Ergebnisse den Anspruch auf Gültigkeit und Objektivität erheben können. Letzteres meint, inwieweit die Ergebnisse unabhängig vom einzelnen Forscher auch durch andere Forscher hätten erzielt werden können. Zur Bestimmung der Qualität empirischer Forschung werden Gütekriterien formuliert, mit denen vor allem die methodischen Vorgehensweisen überprüft werden, die zu ihnen geführt haben. Dabei ist

zu fragen, ob ein einheitlicher Satz von Gütekriterien für jede Form der empirischen Forschung geeignet ist oder ob hier zwischen den für quantitative und qualitative Forschung angemessenen Kriterien differenziert werden sollte. Entsprechend behandelt dieses Kapitel zunächst die in der quantitativen Forschung allgemein anerkannten Gütekriterien Reliabilität, Validität und Objektivität, bevor es auf spezielle Ansätze in der qualitativen Forschung eingeht.

Mit der Qualität empirischer Forschung ist ferner die Frage verbunden, inwieweit die Ergebnisse verallgemeinert, d. h. auf andere Situationen als die konkrete Untersuchungssituation übertragen werden können. Auch hier ist wieder zu prüfen, ob Vorgehensweisen und Ansprüche an die Generalisierung bzw. Verallgemeinerung für alle Arten empirischer Forschung einheitlich oder differenziert zu formulieren sind.

Qualität und Bewertung quantitativer Forschung

Reliabilität

Das erste allgemein anerkannte Kriterium zur Geltungsprüfung von Untersuchungen stammt ursprünglich aus der Testtheorie: «Die Reliabilität (Zuverlässigkeit) gibt den Grad der Messgenauigkeit (Präzision) eines Instrumentes an. Die Reliabilität ist umso höher, je kleiner der zu einem Messwert X gehörende Fehleranteil E ist» (Bortz & Döring 2006, S. 196).

Die Reliabilität einer Messung wird auf verschiedenen Wegen bestimmt.

(1) Retest-Reliabilität

Zur Prüfung der Retest-Reliabilität wird eine Messung (ein Test, ein Fragebogen etc.) mehrfach auf dieselbe Stichprobe angewendet und die Korrelation zwischen den Ergebnissen beider Anwendungen berechnet. Im Idealfall ergeben sich identische Ergebnisse. Dies setzt allerdings voraus, dass das gemessene Merkmal selbst stabil ist und sich nicht zwischen den zwei Messzeitpunkten verändert hat. Bei der Wiederholung von Leis-

tungstests können Unterschiede zwischen den Ergebnissen auch darin begründet sein, dass sich die Leistungsfähigkeit der Untersuchten (z. B. durch zusätzlich erworbenes Wissen) zwischenzeitlich verändert hat. Außerdem können Unterschiede daraus resultieren, dass Fragen wiedererkannt werden und Lerneffekte einsetzen.

(2) Paralleltest-Reliabilität
Bei der Bestimmung der Paralleltest-Reliabilität werden zwei verschiedene Instrumente (die das gleiche Konstrukt operationalisieren) parallel angewendet. «Je ähnlicher die Ergebnisse beider Tests ausfallen, umso weniger Fehlereffekte sind offensichtlich im Spiel» (Bortz & Döring 2006, S. 197).

(3) Testhalbierungs-Reliabilität (Split-half)
Bei einem Test (oder Fragebogen) werden jeweils für die Hälfte der Items (Fragen) Ergebniswerte berechnet und miteinander verglichen. Dabei hängen die Ergebnisse auch von der Art der Teilung des Tests in zwei Hälften (die erste und die zweite Hälfte der Fragen, gerade und ungerade Fragen, zufällige Zuordnung der Fragen zur einen oder der anderen Hälfte) ab. Um Einflüsse der Fragenanordnung auszuschließen, wird die *interne Konsistenz* berechnet. Dabei wird jede Frage als eigener Teiltest behandelt und die Korrelation zwischen den Ergebnissen ermittelt.

(4) Interkoder-Reliabilität
Bei der Inhaltsanalyse wird schließlich die Interkoder-Reliabilität bestimmt, um zu überprüfen, inwieweit verschiedene Auswerter gleiche Aussagen den gleichen Kategorien zuordnen und verlässlich das Kategoriensystem und der Umgang damit in einer Untersuchung sind.

(1) Validität von Untersuchungsanordnungen

Validität wird für Untersuchungsdesigns und für Messinstrumente geprüft. Bei Untersuchungsdesigns geht es um die Überprüfung der Gültigkeit von Ergebnissen. Einerseits wird die *interne* Validität eines Untersuchungsdesigns bestimmt. Interne Validität kennzeichnet, inwieweit Ergebnisse einer Untersuchung eindeutig interpretierbar sind. Sollen beispielsweise die Wirkungen einer Intervention untersucht werden, so ist für die Messung zu prüfen, ob sich Veränderungen abhängiger Variablen ursächlich auf Veränderungen unabhängiger Variablen zurückführen lassen. Zusammenhänge lassen sich dann eindeutig interpretieren (vgl. Abbildung 10.1).

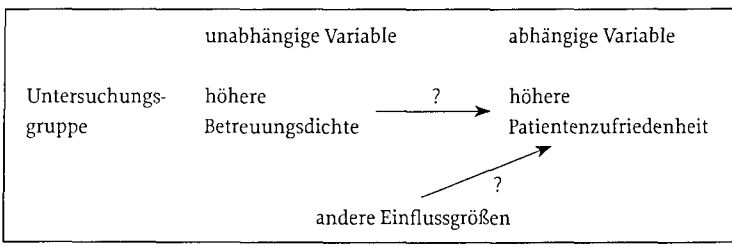


Abbildung 10.1: Interne Validität

Die Einführung einer höheren Betreuungsdichte wäre z. B. die unabhängige Variable, die Zufriedenheit der Patienten die abhängige Variable. Will man den Zusammenhang: «Höhere Betreuungsdichte führt zu höherer Zufriedenheit bei Patienten» untersuchen, so ist bei der empirischen Beantwortung der Frage zu klären, wie eindeutig dieser Zusammenhang gemessen werden kann. Zur Bestimmung dieser internen Validität wird versucht, andere Einflussgrößen auszuschließen – inwieweit sich also andere Bedingungen parallel zur Erhöhung der Betreuungsdichte verändert haben, auf die eine höhere Zufriedenheit der Patienten eigentlich zurückzuführen ist.

Zur Sicherung interner Validität werden entsprechend Bedingungen

isoliert und kontrolliert. Ein Weg zur Überprüfung der Wirkung einer Intervention ist das *Kontrollgruppendesign* (vgl. Kapitel 4): In einer zweiten so weit wie möglich vergleichbaren Gruppe wird die Intervention nicht eingeführt – die Betreuungsdichte nicht erhöht – und geprüft, ob der in der Untersuchungsgruppe festgestellte Effekt höhere Zufriedenheit dann ausbleibt.

«Interne Validität liegt dann vor, wenn Veränderungen in den abhängigen Variablen eindeutig auf den Einfluss der unabhängigen Variablen zurückzuführen sind bzw. wenn es neben der Untersuchungshypothese keine besseren Alternativerklärungen gibt» (Bortz & Döring 2006, S. 53).

Interne Validität ist im Labor und in der experimentellen Forschung am ehesten sicherzustellen und zu prüfen. Dies geht jedoch auf Kosten der zweiten Form von Validität von Untersuchungsanordnungen, der *externen Validität.* Hier ist die generelle Frage, inwieweit Ergebnisse über Situationen und Personen, an denen sie erhoben wurden, auf Situationen und Personen außerhalb der Forschung übertragen werden können. Lässt sich ein gefundener Zusammenhang zwischen der Erhöhung der Betreuungsdichte und der Patientenzufriedenheit auf andere Stationen, Kliniken oder auf Versorgungssituationen generell übertragen, oder gilt er nur unter den konkreten Bedingungen, unter denen er festgestellt wurde (vgl. Abbildung 10.2)?

Während im Labor und unter weitgehend kontrollierten Bedingungen die interne Validität hoch ist, ist die externe Validität eher gering. Bei Untersuchungen im Feld und unter natürlichen Bedingungen ist dagegen die externe Validität höher und die interne Validität geringer, da dort die Kontrolle der Bedingungen nur eingeschränkt möglich ist.

«Externe Validität liegt vor, wenn das in einer Stichprobenuntersuchung gefundene Ergebnis auf andere Personen, Situationen oder Zeitpunkte generalisiert werden kann» (Bortz & Döring 2006, S. 53).

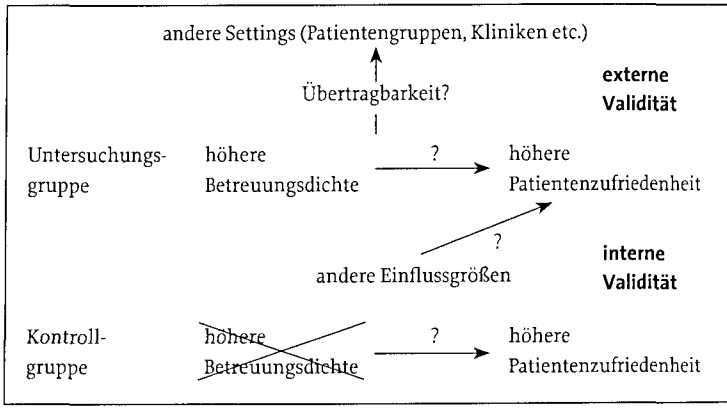


Abbildung 10.2: Externe Validität

Beide Kriterien in einem Untersuchungsdesign gleichzeitig und gleichermaßen einzulösen gilt als schwierig. Hier liegt ein Dilemma empirischer Forschung vor, das in einem Untersuchungsdesign nur schwer aufgelöst werden kann. Externe und interne Validität werden für Untersuchungsdesigns geprüft, aber auch für Messinstrumente.

(2) Validität von Messinstrumenten

Die Validität einer Forschungsmethode lässt sich in der Frage bündeln: *Misst die Methode, was sie messen soll?* Zur Beantwortung dieser generellen Frage werden verschiedene Formen von Validitätsprüfungen vorgenommen: Inhaltsvalidität, Kriteriumsvalidität und Konstruktvalidität.

- *Inhaltsvalidität* ist dann gegeben, wenn das Messinstrument oder der Test den zu untersuchenden Gegenstand in seinen wesentlichsten Aspekten erschöpfend erfasst. Dies kann man entweder subjektiv selbst oder besser durch Einschätzungen des Messinstrumentes durch Experten oder Laien prüfen. Fehler sollten bei solchen Prüfungen ins Auge fallen, weshalb man dies auch *Augenscheinvalidität* nennt. Wird z. B. die Betreuungsdichte in den relevanten Situationen des Tages-

ablaufs in der Klinik erhoben oder nur zu einem bestimmten Zeitpunkt – etwa bei der Neuaufnahme eines Patienten?

- *Kriteriumsvalidität* ist dann gegeben, wenn das Ergebnis einer Messung mit einem Außenkriterium übereinstimmt: Dies ist erfüllt, wenn z. B. die Ergebnisse eines Berufseignungstests mit dem beruflichen Erfolg des Getesteten übereinstimmt (vgl. Abbildung 10.3). Solche Außenkriterien können entweder parallel bestimmt werden, wobei dann die *Übereinstimmungsvalidität* geprüft wird – etwa durch Verhaltensbeobachtung. Oder sie werden zeitlich später bestimmt, wobei dann *prognostische Validität* geprüft wird – erlauben beispielsweise die Ergebnisse eines Berufseignungstests die Vorhersage des beruflichen Erfolgs?

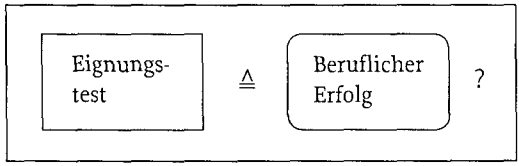


Abbildung 10.3: Kriteriumsvalidität

Ein Problem ist dabei, dass das Außenkriterium – zur Messung des tatsächlichen beruflichen Erfolgs – selbst valide sein muss, damit sich daraus ableiten lässt, dass die eigentliche Messung valide ist. Ferner ist noch die *differenzielle Validität* zu berücksichtigen – Übereinstimmungen zwischen dem Testwert und dem Außenkriterium können in unterschiedlichen Populationen unterschiedlich ausfallen. Messinstrumente sollen in der Regel geeignet sein, Unterschiede bei verschiedenen Gruppen zu erfassen.

- Schließlich wird noch die *Konstruktvalidität* bestimmt. Dabei geht es darum zu prüfen, inwieweit das von einer Methode erfasste Konstrukt mit möglichst vielen anderen Variablen in theoretisch begründbaren Zusammenhängen steht und hieraus Hypothesen ableitbar sind, die einer empirischen Prüfung standhalten. Dabei kommen verschiedene Messungen zum Einsatz. Dazu wird in der Regel eine *Multitrait-Multi-*

Method-Matrix (vgl. Campbell & Fiske 1959) eingesetzt: Konstrukte werden mit mehreren Methoden gemessen. *Konvergente Validität* ist dann gegeben, wenn mehrere Methoden dasselbe Konstrukt übereinstimmend messen – wenn beispielsweise Patientenzufriedenheit durch einen Fragebogen und ein Interview untersucht werden und beide Methoden übereinstimmende Ergebnisse liefern. *Diskriminante Validität* bezieht sich darauf, ob die Messungen geeignet sind, das Konstrukt von anderen Konstrukten zu unterscheiden – inwieweit tatsächlich Zufriedenheit des Patienten mit der Betreuung und nicht ein allgemeines Wohlbefinden bestimmt wird.

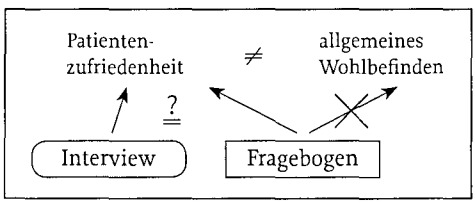


Abbildung 10.4: Konstruktvalidität

(3) Validität bei Indices

Ein Index wird gebildet, wenn eine Größe nicht direkt beobachtbar ist, vor allem wenn darin verschiedene Aspekte eines theoretischen Konstrukts zusammenfließen – z. B. Patientenzufriedenheit (vgl. Kapitel 5). Die Messung dieses Konstrukts erfordert dann die Auswahl eines oder mehrerer geeigneter Indikatoren. Um die auftretenden Fehler bei der Messung komplexer Konstrukte bzw. Größen gering zu halten, wird man mehrere Indikatoren verwenden, um die Messqualität zu erhöhen. Die Gesamtbenotung von Schulaufsätzen etwa setzt sich aus den Bewertungen der Rechtschreibung, des Stils, des Inhalts und der Form zusammen – entweder zu gleichen Teilen oder mit unterschiedlichem Gewicht, wenn z. B. der Stil stärker berücksichtigt wird als die Fehlerzahl.

So wird man auch die Patientenzufriedenheit aus verschiedenen Indikatoren zu bestimmen suchen und etwa ein Instrument zur Lebensqualitätsbestimmung verwenden sowie einen Fragebogen zur Messung

der Zufriedenheit mit den Leistungen des Klinikpersonals oder mit der Infrastruktur des Krankenhauses einsetzen. Dabei stellt sich dann die Frage, wie diese einzelnen Variablen bei der Bildung des Index *gewichtet* werden – welchen Anteil die Fragebogenresultate zur Zufriedenheit mit den Leistungen des Personals im Verhältnis zur Lebensqualitätsbestimmung bei der Bildung des Patientenzufriedenheitsindex haben. Weiterhin ergibt sich das Problem bei der Bestimmung der Validität des Index, dass die eingeflossenen Variablen – Lebensqualität, Zufriedenheit mit dem Personal – selbst valide bestimmt werden müssen, damit der Index insgesamt valide sein kann. Damit stellen sich die genannten Validitätsbestimmungen auf zwei Ebenen – auf der Ebene des einzelnen Indikators und auf der Ebene des aus den Indikatoren gebildeten Index. «Die Qualität eines Indexes hängt wesentlich davon ab, ob alle relevanten Dimensionen ausgewählt und angemessen gewichtet wurden» (Bortz & Döring 2006, S. 144). Bei Indices setzt sich die Validität entsprechend aus der Validität die einzelnen Items bzw. Fragen, der Validität der aus ihnen gebildeten Skalen und aus der angemessenen Gewichtung der Bestandteile zusammen.

Zur Validität bei Indices kann Folgendes festgehalten werden: Da die Indexbildung auf der Verwendung mehrerer Indikatoren zur Messung einer nicht direkt beobachtbaren Größe beruht, ergeben sich hier spezifische Probleme der Validität. So sollte geprüft werden, ob

- die relevanten Dimensionen ausgewählt und angemessen gewichtet wurden,
- die Instrumente zur Messung der gewählten Indikatoren valide sind und
- die Items, die in die Indikatoren einfließen, valide sind.

Validität bezeichnet also unterschiedliche Anhaltspunkte zur Prüfung der Geltung von Untersuchungsergebnissen, wobei auch Fragen ihrer Übertragbarkeit und Verallgemeinerbarkeit angesprochen sind.

Objektivität

Objektivität bei Tests oder Fragebogen gibt an, inwieweit die Anwendung des Instruments unabhängig ist von der Person, die ihn einsetzt.

Wenn verschiedene Forscher denselben Test auf dieselben Personen anwenden, müssen die Ergebnisse identisch sein. Dabei werden *drei Formen* unterschieden:

- *Durchführungsobjektivität* bezieht sich darauf, dass Antworten oder Testergebnisse des Untersuchungsteilnehmers vom Anwender des Tests (Untersuchungsleiter) unabhängig sein sollten. Vor allem wird dies durch die Standardisierung der Durchführung (standardisierte Instruktionen zur Bearbeitung des Instruments, standardisierte Bedingungen in der Anwendungssituation) erreicht.
- *Auswertungsobjektivität* bezeichnet, dass Antworten in Tests oder Fragebogen unabhängig von der Person des Auswerters klassifiziert (etwa einem bestimmten Punktwert zugeordnet) werden müssen.
- *Interpretationsobjektivität* meint, dass die Interpretation von Antworten oder Testwerten unabhängig von der Person des Auswerters bzw. Interpreten und seinen subjektiven Deutungen oder Bewertungen vorgenommen werden soll. Deshalb werden für viele Tests Normwerte (Alters-, Geschlechts- oder Bildungsnormen) anhand repräsentativer Stichproben ermittelt, die dann zur Bewertung der Leistungen bzw. Werte des Untersuchungsteilnehmers herangezogen werden.

Objektivität wird vor allem durch die Standardisierung der Durchführung einer Untersuchung und der Auswertung und Interpretation ihrer Ergebnisse angestrebt, um subjektive oder individuelle Einflüsse des Untersuchungsleiters oder Beeinflussungen in der konkreten Situation auszuschließen.

Qualität und Bewertung qualitativer Forschung

Anwendung der klassischen Kriterien

Die bislang behandelten Kriterien sind vor allem in der quantitativen Forschung anerkannt und finden dort ihre Anwendung. Sie basieren alle mehr oder minder auf der Standardisierung der Untersuchungssituation.

Teilweise wird ihre Anwendung und Einhaltung auch für qualitative Untersuchungen diskutiert (vgl. Kirk & Miller 1986). Hier wird jedoch die

Frage aufgeworfen, inwieweit sie gerade aufgrund der starken Betonung der Standardisierung des Vorgehens und der Ausschaltung der kommunikativen Einflüsse des Forschers dem besonders auf Kommunikation, Interaktion und subjektiver Deutung und Interpretation durch den Forscher basierenden Vorgehen qualitativer Forschung gerecht werden können. Wobei hier Kommunikation, Interpretation und Deutung nicht als Störvariablen, sondern als Stärken und Voraussetzung der Forschung gesehen werden. Entsprechend fragen schon Glaser und Strauss (1993, S. 92), «ob der Kanon quantitativer Sozialforschung als Kriterium (...) auf qualitative Forschung (...) anwendbar ist. Die Beurteilungskriterien sollten vielmehr auf einer Einschätzung der allgemeinen Merkmale qualitativer Sozialforschung beruhen – der Art der Datensammlung (...) der Analyse und Darstellung und der (...) Weise, in der qualitative Analysen gelesen werden.»

Aus dieser Skepsis resultieren im Laufe der Zeit eine Reihe von Versuchen, eine *eigenständige* Diskussion über Kriterien qualitativer Forschung zu initiieren (vgl. Flick 2007a, S. 487–510), und Versuche, methodenangemessene Kriterien (Flick 2008c) zu entwickeln und diese an die Stelle von Kriterien wie Validität und Reliabilität zu setzen.

Reformulierung klassischer Kriterien

Vorschläge zur Reformulierung von Reliabilität im Sinne einer stärker prozeduralen Konzeption zielen darauf ab, das Zustandekommen der Daten dahin gehend zu explizieren, dass überprüfbar wird, was noch Aussage des jeweiligen Subjekts ist und wo die Interpretation des Forschers bereits begonnen hat. Schließlich soll sich die Reliabilität im gesamten Prozess durch dessen reflexive Dokumentation erhöhen.

Auch Validität wird reformuliert. So wird eine spezifische Analyse der Interviewsituation, ausgehend von den verschiedenen Geltungsansprüchen, die in Habermas' «Theorie des kommunikativen Handelns» (1981) unterschieden werden, vorgeschlagen (vgl. dazu Legewie 1987). Eine weitere Möglichkeit der Bestimmung von Validität zielt auf die Einbeziehung der untersuchten Subjekte oder Gruppen in den Forschungsprozess. Einen Weg hierzu bietet die Einführung kommunikativer Vali-

dierung in einem zweiten Termin nach Abschluss des Interviews und der Transkription (für konkrete Vorschläge vgl. Scheele & Groeben 1988). Eine Zeit lang wurde kommunikative Validierung auch in Bezug auf Interpretationen von Texten diskutiert. Nicht zuletzt aufgrund der bei der Konfrontation mit Interpretationen auftretenden ethischen Probleme hat dieses Verständnis an Bedeutung verloren. Für eine allgemeine Anwendung solcher Strategien sind zwei Fragen noch nicht befriedigend beantwortet:

- Wie ist das methodische Vorgehen bei der kommunikativen Validierung zu gestalten, damit es den untersuchten Sachverhalten und der Sicht der Subjekte tatsächlich gerecht wird?
- Wie lässt sich jenseits der Zustimmung der Subjekte die Frage der Geltungsbegründung weiter gehend beantworten?

Mishler (1990) geht einen Schritt weiter in der Reformulierung des Konzepts der Validität. Er setzt am Prozess der Validierung an (statt am Zustand der Validität) und definiert «Validierung als soziale Konstruktion von Wissen» (S. 417), durch die wir «Behauptungen über die ‹Vertrauenswürdigkeit› berichteter Beobachtungen, Interpretationen und Verallgemeinerungen aufstellen und diese bewerten» (S. 419). Schließlich umgeht «Validierung, verstanden als der soziale Diskurs, durch den Vertrauenswürdigkeit hergestellt wird, solche vertrauten Konventionen wie Reliabilität, Falsifikation und Objektivität». Als empirische Basis für diesen Diskurs und die Konstruktion von Vertrauenswürdigkeit erörtert Mishler die Verwendung von Beispielen aus narrativen Studien.

Methodenangemessene Kriterien

Schließlich wird der Gedanke der Gegenstandsangemessenheit dahin gehend ausgeweitet, dass methodenangemessene Gütekriterien für qualitative Methoden gefordert werden. Neben der bereits vorgestellten kommunikativen Validierung und der in Kapitel 8 behandelten *Triangulation* lässt sich gerade in der amerikanischen Diskussion eine Vielzahl von neuen Kriterien finden (vgl. Flick 2007a, S. 511–530; Flick 2008c). So propagieren Lincoln und Guba (1985) Vertrauenswürdigkeit, Glaubwürdigkeit, Übertragbarkeit, Zuverlässigkeit und Bestätigbarkeit als

Kriterien qualitativer Forschung, wobei «Vertrauenswürdigkeit» im Vordergrund steht. Um die Glaubwürdigkeit qualitativer Forschung, Daten und Ergebnisse zu erhöhen, skizzieren sie fünf Strategien:

- Aktivitäten zur Erhöhung der Wahrscheinlichkeit, dass glaubwürdige Erkenntnisse produziert werden durch ein «verlängertes Engagement» und «ausdauernde Beobachtungen» im Feld und die Triangulation verschiedener Methoden, Forscher und Datensorten.
- «Peer debriefing»: regelmäßige Besprechungen mit nicht an der Forschung beteiligten Personen, um die eigenen blinden Flecke aufzudecken sowie Arbeitshypothesen und Ergebnisse zu überprüfen.
- Die Analyse abweichender Fälle.
- Angemessenheit in den Bezugspunkten von Interpretationen und ihrer Überprüfung.
- «Member-Checks» – kommunikative Validierung von Daten und Interpretationen mit den Mitgliedern der untersuchten Felder.

Für die Überprüfung der Verlässlichkeit qualitativer Daten wird in diesem Zusammenhang ein Prozess des «Auditing» vorgeschlagen, der am Vorgang der Buchprüfung im Finanzwesen orientiert ist. Dafür wird ein *Überprüfungspfad* (Auditing Trail) skizziert:

Ein Auditing Trail erfasst

- die Rohdaten, ihre Erhebung und Aufzeichnung;
- Datenreduktion und Ergebnisse von Synthesen durch Zusammenfassung, theoretische Notizen, Memos, Zusammenfassungen, Kurzdarstellungen von Fällen etc.;
- Datenrekonstruktionen und Ergebnisse von Synthesen anhand der Struktur entwickelter und verwendeter Kategorien (Themen, Definitionen, Beziehungen), Erkenntnisse (Interpretationen und Schlüsse) sowie die erstellten Berichte mit ihren Integrationen von Konzepten und den Bezügen zu existierender Literatur;
- Prozessnotizen, d. h. methodologische Notizen und Entscheidungen auch hinsichtlich der Herstellung von Vertrauens- und Glaubwürdigkeit der Erkenntnisse;
- Materialien in Bezug auf Absichten und Anordnungen wie die Forschungskonzeption, persönliche Aufzeichnungen und Erwartungen der Beteiligten;

- Informationen über die Entwicklung der Instrumente einschließlich der Pilotversionen und vorläufigen Pläne.

Die Frage der Übertragbarkeit von Ergebnissen wurde bereits unter der Überschrift «externe Validität» diskutiert. Sie wird aber auch generell zu einem Aspekt der Bewertung empirischer Forschung, weshalb sie erneut für quantitative und qualitative Forschung behandelt werden soll.

Verallgemeinerung

Verallgemeinerung in der quantitativen Forschung

In der quantitativen Forschung wird die Verallgemeinerbarkeit der Ergebnisse einer Untersuchung im Wesentlichen in zwei Richtungen überprüft. Durch die bereits dargestellte Prüfung der externen Validität soll sichergestellt werden, dass die in der Stichprobe gefundenen Ergebnisse auch für die Grundgesamtheit gelten und geprüft wird, inwieweit sie auf ähnliche Grundgesamtheiten übertragbar sind. Bortz und Döring (2006, S. 335) halten fest, dass «Generalisierbarkeit in der quantitativen Forschung durch den wahrscheinlichkeitstheoretisch abgesicherten Schluss von Zufallsstichproben (bzw. Stichprobenkennwerten) auf Populationen (bzw. Populationsparameter) erreicht wird».

Die verschiedenen Stichprobenverfahren (vgl. Kapitel 4) sollen dies gewährleisten. Am konsequentesten wird dies mit der *Zufallsstichprobe* angestrebt, bei der jedes Mitglied der Population die gleiche Chance hat, in die Stichprobe aufgenommen zu werden. Durch dieses Zufallsverfahren können Verzerrungen, die sich durch die im Vergleich zur Grundgesamtheit unverhältnismäßig gewichtete Verteilung von Merkmalen in der Stichprobe ergeben, ausgeschlossen werden. Damit ist die Stichprobe für die Grundgesamtheit repräsentativ, und der Schluss von Stichprobe auf die Grundgesamtheit ist hinsichtlich der Gültigkeit der Ergebnisse zulässig. Andere Verfahren versuchen, die Verteilung in der Grundgesamtheit in der Stichprobe gezielter abzubilden, indem etwa eine *geschichtete Stichprobe* gezogen wird:

«Man zieht eine geschichtete Stichprobe, indem man die Zielpopulation auf der Basis einer oder mehrerer Merkmale in Teilpopulationen (Schichten) einteilt – pro Merkmalsausprägung bzw. Merkmalskombination entsteht eine Teilpopulation – und aus jeder dieser Schichten eine Zufallsstichprobe nimmt» (Bortz & Döring 2006, S. 425).

Verallgemeinerung über die Prüfung der externen Validität einer Untersuchung basiert auf der Frage nach der Ähnlichkeit zwischen den Untersuchungsteilnehmern und den Populationen, für die eine Untersuchung mit ihren Ergebnissen Gültigkeit aufweisen soll, das heißt, auf die sie verallgemeinert werden soll. Entsprechend spricht Campbell (1986) von «proximal similarity» statt von externer Validität: In den (für die Untersuchung und ihre Ergebnisse) relevanten Dimensionen muss größtmögliche Ähnlichkeit zwischen Stichprobe und der Gesamtheit (auf die Ergebnisse übertragbar sein sollen) herrschen – dann ist ihre Verallgemeinerung möglich.

Verallgemeinerung in der qualitativen Forschung

Während bei quantitativer Forschung die Verallgemeinerung vor allem ein *numerisches* Problem darstellt, das mit *statistischen* Mitteln gelöst werden soll, ist diese Vorgehensweise bei der qualitativen Forschung diffiziler. Zunächst stellt sich die Frage der Verallgemeinerbarkeit in ähnlicher Weise: Eine begrenzte, nach bestimmten Kriterien ausgewählte Zahl von Fällen (oder ein einzelner Fall) wurde untersucht, und die Ergebnisse beanspruchen über das Untersuchungsmaterial hinaus Gültigkeit. Der Fall oder die Fälle stehen für allgemeine Zusammenhänge. Die Frage der Verallgemeinerbarkeit stellt sich in der qualitativen Forschung aber auch in grundsätzlich anderer Hinsicht, als sie teilweise der Theorieentwicklung aus empirischem Material dient (im Sinne von Glaser & Strauss 1967/2005). Dann ist zu prüfen, auf welche weiteren Kontexte die entwickelte Theorie übertragen werden kann bzw. für welche weiteren Kontexte sie gültig ist.

Entsprechend ist ein Ansatzpunkt zur Beurteilung qualitativer Forschung, zu fragen, welche Überlegungen und Schritte unternom-

men wurden, um den Geltungsbereich empirischer Ergebnisse oder daraus entwickelter Theorien zu bestimmen bzw. zu erweitern. Dabei sind die Ausgangspunkte die Analyse von Fällen und die Wege, die von ihnen ausgehend zu allgemeinen Aussagen beschritten werden. Das Problem der Generalisierung liegt bei qualitativer Forschung unter anderem darin, dass ihr Ansatzpunkt häufig gerade die auf einen Kontext, auf einen konkreten Fall bezogene Analyse von Bedingungen, Zusammenhängen, Verläufen etc. ist. Durch diesen Kontextbezug gewinnt qualitative Forschung (oft) eine spezifische Aussagekraft. Im Schritt der Generalisierung wird dieser Kontextbezug aufgegeben, um zu untersuchen, inwieweit die gefundenen Zusammenhänge auch unabhängig und außerhalb von spezifischen Kontexten gelten. In der Zuspitzung dieses Dilemmas behandeln Lincoln und Guba (1985) das Problem unter der Überschrift «Die einzige Generalisierung ist: Es gibt keine Generalisierung». Jedoch skizzieren sie mit der *Übertragbarkeit* (transferability) von Erkenntnissen aus einem Kontext in einen anderen und der *Passung* (fittingness) als Grad der Vergleichbarkeit verschiedener Kontexte Kriterien und Wege für die Verallgemeinerung von Erkenntnissen über einen Kontext hinaus.

So werden verschiedene Möglichkeiten diskutiert, wie der Weg vom Fall zur Theorie so abgesteckt werden kann, dass sich zumindest eine gewisse Verallgemeinerung erreichen lässt. Ein erster Schritt ist die Klärung der Frage, welcher *Grad an Verallgemeinerung* mit der jeweiligen Studie überhaupt angestrebt wird und erreichbar ist, um angemessene Ansprüche an die Generalisierung abzuleiten. Ein zweiter Schritt ist die sorgfältige *Einbeziehung von unterschiedlichen Fällen und Kontexten*, in denen die untersuchten Zusammenhänge empirisch analysiert werden. Die Generalisierbarkeit der Ergebnisse hängt häufig eng mit der Realisierung der Auswahl zusammen, wobei das theoretische Sampling eine Strategie anbietet, die Variation der Bedingungen, unter denen ein Phänomen empirisch untersucht wird, möglichst breit zu gestalten. Der dritte Ansatzpunkt ist der *systematische Vergleich erhobenen Materials.*

(1) Methode des konstanten Vergleichs

Neben der Methode des «theoretischen Sampling» schlägt Glaser (1969) für den Prozess der Theorieentwicklung die «Methode des konstanten Vergleichs» als Verfahren bei der Auswertung von Texten vor. Es besteht im Wesentlichen aus *vier Stufen*: «(1) Vergleich von Ereignissen, die in die jeweilige Kategorie passen, (2) Integration von Kategorie und ihren Inhalten, (3) Eingrenzung der Theorie und (4) Formulierung der Theorie» (S. 220). Zentral ist dabei für Glaser die systematische Zirkularität dieses Prozesses: «Obwohl es sich bei dieser Methode um einen kontinuierlichen Wachstumsprozess handelt – jede Stufe transformiert sich nach einer Weile in die nächste –, bleiben vorherige Stufen während der gesamten Analyse wirksam und sorgen für eine kontinuierliche Entwicklung zur nächsten Stufe, bis die Analyse abgeschlossen ist» (S. 220).

Zur Methode des konstanten Vergleichs wird dieses Vorgehen, wenn Interpreten darauf achten, dass sie Kodierungen immer wieder mit bereits vollzogenen Kodierungen und Zuordnungen vergleichen, dass bereits kodiertes Material mit seiner Zuordnung nicht ‹erledigt› ist, sondern weiter im Prozess des Vergleichs einbezogen bleibt.

(2) Fallkontrastierung und Idealtypenbildung

Der konstante Vergleich wird in Strategien der Fallkontrastierung (vgl. auch Kelle & Kluge 1999) weiterentwickelt und systematisiert – am stärksten bei Gerhardt (1986) in der Idealtypenbildung. Sie umfasst folgende Schritte: Nach der Fallrekonstruktion und -kontrastierung lassen sich Typen bilden; anschließend werden «reine» Fälle ermittelt; im Vergleich zu diesen idealtypischen Verläufen lässt sich das Einzelfallverstehen systematisieren; durch weitere Typenbildungen (vgl. Kapitel 6) findet dieser Prozess über das Strukturverstehen, d. h. das Verstehen von Zusammenhängen, die über den einzelnen Fall hinausweisen, seinen Abschluss. Zentrale Instrumente sind der *minimale Vergleich* von möglichst ähnlichen Fällen und der *maximale Vergleich* von möglichst unterschiedlichen Fällen auf Unterschiede und Gemeinsamkeiten. Der Vergleich konkretisiert sich dabei zunehmend in Hinblick auf das im empirischen Material enthaltene Spektrum, dessen Endpunkte durch den maximalen

Vergleich und dessen Zentrum durch den minimalen Vergleich besondere Aufmerksamkeit erfahren. Ähnlich schlägt Schütze (1983) die *minimale und maximale Konstrastierung von Einzelfällen* für die vergleichende Auswertung narrativer Interviews vor.

Verallgemeinerung bei qualitativer Forschung liegt in der schrittweisen Übertragung von Erkenntnissen aus Fallstudien und ihrem Kontext auf allgemeine und abstraktere Zusammenhänge, z. B. eine Typologie. Die Aussagekraft solcher Muster lässt sich dann wieder danach bestimmen, inwieweit unterschiedliche theoretische und methodische Perspektiven auf den Gegenstand – nach Möglichkeit auch von verschiedenen Forschern – *trianguliert* wurden und wie mit abweichenden Fällen umgegangen wurde. Zu berücksichtigen ist dabei, welcher Grad an Verallgemeinerung überhaupt mit der konkreten Studie angestrebt wird. Dann wird die Frage, ob die angestrebte Ebene der Verallgemeinerung erreicht worden ist, zu einem weiteren Kriterium der Bewertung von Ergebnissen qualitativer Forschung und des Prozesses, der zu ihnen geführt hat.

Leitfragen für die Bewertung konkreter Projekte

Bei der Bewertung eines empirischen Projekts in der Sozialforschung können folgende Fragen eine Orientierung bieten:

Kasten 10.1: **Leitfragen für die Bewertung konkreter Projekte**

1. Wurden bei einer quantitativen Studie die Kriterien Reliabilität, Validität und Objektivität geprüft?
2. Welche Kriterien bzw. Ansätze zur Bestimmung der Qualität wurden bei einer qualitativen Studie angewendet?
3. Inwieweit sind die Studie, aber auch die Prüfung ihrer Qualität transparent und explizit gemacht – in der Darstellung des Vorgehens und der Ergebnisse?
4. Wie wurde mit abweichenden Ergebnissen umgegangen?
5. Wie wurde die Verallgemeinerbarkeit der Ergebnisse überprüft, welche Ziele wurden dabei gesetzt, und wie wurden sie eingelöst?

Diese Aspekte sind für die Einschätzung der eigenen Forschung ebenso relevant wie bei der Bewertung fremder Studien.

Kernpunkte

- Für die Bewertung von quantitativer Sozialforschung haben sich die Kriterien Reliabilität, Validität und Objektivität etabliert.
- In der qualitativen Sozialforschung lassen sich diese Kriterien nur in seltenen Fällen anwenden.
- Es liegt jedoch eine Reihe von Vorschlägen vor, wie diese Kriterien reformuliert werden können.
- Darüber hinaus wurden methodenangemessene Kriterien für die qualitative Forschung entwickelt.
- Verallgemeinerung basiert in der quantitativen Forschung in der Regel auf dem (statistischen) Schluss von der Stichprobe auf die Grundgesamtheit.
- In der qualitativen Forschung wird eher eine theoretische Verallgemeinerung angestrebt.
- Hierbei spielen Typenbildung und Fallkontrastierung eine zentrale Rolle.

Weiterführende Literatur

Bortz, J. & Döring, N. (2006). Forschungsmethoden und Evaluation für Sozialwissenschaftler (3. überarbeitete Auflage). Berlin u. a.: Springer.

Diekmann, A. (2007). Empirische Sozialforschung (17. Auflage). Reinbek bei Hamburg: Rowohlt.

Flick, U. (2008c). Managing Quality in Qualitative Research. London, Thousand Oaks, New Delhi: Sage.

Steinke, I. (1999). Kriterien qualitativer Forschung. Ansätze zur Bewertung qualitativ-empirischer Sozialforschung. Weinheim, München: Juventa.

11 Ethische Fragen der Sozialforschung

Prinzipien ethisch vertretbarer Forschung 281
Informierte Einwilligung 282
Anonymität und Datenschutz 284
Vermeidung von Schädigungen für die Beteiligten 287
Ethikrichtlinien bzw. Ethikkodizes 289
Ethikkommissionen 290
Regeln guter wissenschaftlicher Praxis 292
Forschungsethik zwischen Fall und Masse 294
Leitfaden für die Berücksichtigung ethischer Fragen in der Sozialforschung 295

Ziele: Nach der Lektüre dieses Kapitels sollten Sie
- die ethischen Fragen, die mit Sozialforschung verknüpft sind, sehen,
- eine Sensibilität für ethische Fragen in der Sozialforschung entwickeln,
- erkennen, dass es keine einfache Antwort auf diese Fragen gibt, und
- Forschung in einem ethisch legitimen Rahmen planen und durchführen können.

In Kapitel 8 wurden Grenzen der Sozialforschung behandelt. Dabei ging es vor allem um methodische bzw. technische Beschränkungen – Was kann man mit einer Methode bzw. Richtung erfassen, was bleibt eher außen vor, wie lässt sich dies durch die Verwendung mehrerer Methoden verändern? – und um eher grundsätzliche Grenzen – Wann sollte Forschung eher nicht durchgeführt werden? In diesem Kapitel steht eine andere Begrenzung der Sozialforschung im Vordergrund: Welche ethischen Probleme sollten beachtet werden, an welche ethischen Grenzen stößt sie, und wie sind ethische Belange in der Forschung zu berücksichtigen?