
Empirische Forschungsmethoden

Udo Kelle, Florian Reith und Brigitte Metje

Einleitung

Die Sozialwissenschaften sowie die pädagogische Psychologie im Allgemeinen und die Forschung zur Lehrer-Schüler-Interaktion im Besonderen verwenden eine breite Palette unterschiedlicher Verfahren der Datenerhebung und Datenauswertung. Zur Systematisierung dieser Verfahren werden wir uns an die gängige Unterscheidung zwischen „quantitativen“ und „qualitativen“ Methoden halten: quantitative Forschung erhebt standardisierte Daten mit eigens konstruierten Instrumenten (etwa Fragebögen oder Beobachtungsinventaren), welche mit statistischen Verfahren analysiert werden. In der qualitativen Forschung wird gering strukturiertes Text-, Bild- und Videomaterial gesammelt, welches dann interpretiert und kategorisiert wird. In der neueren Literatur wird oft betont, dass die strikte Trennung zwischen quantitativer und qualitativer Forschung zu kurz greift und Möglichkeiten sinnvoller Kombination außer Acht lässt (vgl. etwa Kelle & Erzberger 2013; Kelle 2008; Tashakkori & Teddlie 2010; Flick 2011a; zu einem Überblick über die Integration quantitativer und qualitativer Methoden in der empirischen Bildungsforschung siehe Gläser-Zikuda u.a. 2012). Dennoch ist es sinnvoll, quantitative und qualitative Forschung zunächst getrennt voneinander zu betrachten, denn in beiden Traditionen wurden unterschiedliche Modelle des Forschungsprozesses und verschiedene Qualitätskriterien der Forschung entwickelt und vorgeschlagen. Die Kenntnis dieser Modelle und Kriterien ermöglicht es, die Stärken und Schwächen qualitativer und quantitativer Forschungsmethoden zu verstehen und auch Möglichkeiten zur Methodenkombination einzuschätzen.

1 Quantitative Forschungsmethoden

Kennzeichnend für die quantitative Methodentradiation ist die Forderung nach einer theoriegeleiteten, objektiven und präzisen Messung sozialer und psychologischer Merkmale. Damit verbinden sich ein besonderes Konzept wissenschaftlicher Erklärung und ein Modell des Forschungshandelns, das auch als „hypothetiko-deduktives Modell“ bezeichnet wird. Dieses Modell werden wir am Anfang kurz darstellen, bevor wir auf das Problem der Messbarmachung („Operationalisierung“) und dann auf verschiedene Untersuchungsdesigns, auf Techniken der Stichprobenziehung und auf Verfahren der Datenerhebung eingehen. Methoden zur Analyse quantitativer Daten werden wir nicht behandeln, weil dies den Rahmen eines solchen Handbuchkapitels sprengen würde – hierzu kann man auf die verfügbaren Statistiklehrbücher für die Sozialwissenschaften zurückgreifen (etwa auf Bortz & Schuster 2010; Kühnel & Krebs 2014).

1.1 Methodologische Grundlagen: Wissenschaftliche Erklärung und das deduktive Modell des Forschungsprozesses

Ein grundlegendes Schema wissenschaftlicher Erklärung, das eine wesentliche methodologische Grundlage quantitativer Forschung bietet, ist das Modell „deduktiv-nomologischer Erklärung“, das nach seinen Autoren auch „Hempel-Oppenheim-Schema“ (oder „HO-Schema“) genannt wird (Hempel & Oppenheim 1948). Eine Erklärung nach dem HO-Schema besteht aus einem „Explanandum“ (dem „zu Erklärenden“) und einem „Explanans“ (dem „Erklärenden“), das wiederum einerseits aus allgemeinen Gesetzen und andererseits aus den spezifischen Randbedingungen (auch „Antezedensbedingungen“) der Situation aufgebaut ist.

Dies lässt sich anhand eines Beispiels, dem Zusammenhang zwischen der intrinsischen Lernmotivation von Grundschulkindern und der Feedbackkultur der Lehrpersonen beschreiben: Hellmich & Hoya (2014) stellen in einer Studie zur Leseleistung von Kindern aus dritten und vierten Grundschulklassen fest, dass positive Rückmeldungen seitens der Lehrerinnen und Lehrer sowohl das Selbstkonzept der Schüler stärken als auch deren Lernmotivation erhöhen. Soll also eine Erklärung für die mangelnde Lesemotivation von Schülern gefunden werden, so könnte diese im HO-Schema folgendermaßen aussehen:

Explanandum: Schüler sind im Leseunterricht manchmal unmotiviert.

Eplanans: *Allgemeines Gesetz*

Wenn positive Rückmeldungen von Lehrern im Leseunterricht ausbleiben, sind Schüler unmotiviert.

Randbedingung

Manche Lehrer geben zu wenig positive Rückmeldungen.

Die mangelnde positive Rückmeldung wird aber nur ein Aspekt von Unterricht sein, der sich auf die Schülermotivation auswirkt. Darüber hinaus sind wahrscheinlich zahlreiche weitere Variablen wie bspw. Persönlichkeitsmerkmale der Schüler, Interesse am Fach, die Sympathie, die der Lehrer genießt usw. wirksam. Die hier angebotene wissenschaftliche Erklärung kann also ohne Weiteres in Zweifel gezogen werden. Solange nicht gesichert ist, dass die allgemeine Gesetzmäßigkeit wahr ist und dass die Randbedingungen zutreffen, handelt es sich nur um eine „Erklärungshypothese“, die mit anderen Hypothesen konkurriert. Die Geltung von Hypothesen und Forschungsergebnissen kann mit weiteren empirischen Daten überprüft und abgesichert werden – dies ist die Quintessenz des „hypothetiko-deduktiven“ (HD-)Ansatzes in der Methodologie (vgl. Popper 1963/1994, S. 321f. u. S. 349f.). Idealtypisch umfasst ein hypothetiko-deduktiver Forschungsprozess folgende Schritte:

1. Die Formulierung eines Forschungsproblems und einer Fragestellung (bspw. *Warum sind manche Schüler im Unterricht unmotiviert?*),
2. die Suche nach einer Theorie, die das Problem erklärt,
3. die Ableitung von Hypothesen aus dieser Theorie,
4. die Übersetzung der Begriffe, aus denen die Hypothesen bestehen, in messbare Konzepte („Operationalisierung“),
5. die Konstruktion eines „Forschungsdesigns“,
6. die Auswahl der Untersuchungseinheiten bzw. die Ziehung der „Stichprobe“,
7. die Erhebung der Daten
8. und deren Auswertung.

Die Auswertung der Daten soll dann empirische „Evidenz“ für die Gültigkeit der erklärenden Theorien liefern. Hierzu ist es notwendig, dass die aus den Theorien abgeleiteten Hypothesen „empirisch gehaltvoll“ sind, das heißt prinzipiell an der empirischen Realität scheitern können müssen. Allerdings ist es nicht möglich, eine wissenschaftliche Gesetzhypothese durch empirische Daten *endgültig* zu beweisen, da Gesetzesaussagen sich auf eine (prinzipiell) *unendliche* Anzahl von

Fällen (*alle denkbaren unmotivierten Schüler zu allen denkbaren Zeiten*) beziehen, während empirische Untersuchungen immer nur eine *endliche* Anzahl von Beobachtungen umfassen können¹. Aus der universellen Theorieaussage „*Wenn im Leseunterricht positive Rückmeldungen ausbleiben, dann sind Schüler unmotiviert*“ lassen sich nur Einzelaussagen über konkrete Vorgänge ableiten, die überprüft werden können, bspw.: „*Wenn in einer konkreten Schulklasse mehr positive Rückmeldungen gegeben werden, steigt die Zahl der motivierten Schüler*“. Empirische Überprüfungen solcher Hypothesen stellen dann Tests für die Theorie dar, bei denen sie sich entweder (vorläufig) bewähren oder wo sie scheitern kann, d.h. dass die Theorie falsifiziert wird.

Eine zentrale Voraussetzung hierfür ist die sog. „Operationalisierung“, d.h. die „Messbarmachung“ der in der Hypothese enthaltenen Begriffe – so muss, um in unserem Beispiel zu bleiben, als erstes geklärt werden, in welcher Weise man „Motivation“ von Schülern und „positive Rückmeldungen“ von Lehrern beobachten und messen kann. Die Operationalisierung betrifft also direkt die Art des *Untersuchungsinstrumentes* (sollen z.B. Schüler zu ihrer Motivation im Unterricht selbst befragt werden, sollen Lehrer Auskunft über die Schülermotivation geben oder soll ein Mitglied des Forschungsteams im Unterricht hospitieren?) sowie dessen Konstruktion (d.h. die Fragen eines Fragebogens oder die Beobachtungskategorien).

Als Kriterien zur Beurteilung der Qualität einer Operationalisierung können die sog. klassischen „*Gütekriterien*“ dienen: *Objektivität*, *Reliabilität* und *Validität*. Diese Begriffe sind nicht einfach zu definieren, weil sich hieran Fragen nach der „Wahrheit“ von Aussagen und dem Verhältnis von Wirklichkeit und Theorie knüpfen, die in der Wissenschaftsphilosophie kontrovers diskutiert werden. Hier kann man sich jedoch mit pragmatischen Lösungen helfen.

1. Dies ist insbesondere bei dem Konzept der *Objektivität* möglich. Forschungspraktisch meint man damit weniger die Neutralität und Sachlichkeit des Wissenschaftlers (eine sicher notwendige, aber schwer belegbare Eigenschaft) sondern die *Intersubjektivität* des Forschungsprozesses: so sollen etwa unterschiedliche Forscher bei dem Vorliegen derselben Sachverhalte zu denselben Beobachtungen und Daten gelangen.
2. Unter *Reliabilität* versteht man die Zuverlässigkeit von Messinstrumenten und Datenerhebung – diese lässt sich etwa anhand der Übereinstimmung von Mes-

1 Wenn also die Daten die Hypothese nicht widerlegen („falsifizieren“), spricht man davon, dass die Hypothese „gestützt“ wurde oder sich „bewährt“ hat, nicht jedoch von einem „Beweis“ (einer „Verifikation“).

sungen beurteilen. Wird der Rückmeldungsstil eines Lehrers zum Zeitpunkt t_1 bewertet, so muss das Lehrerverhalten zum Zeitpunkt t_2 , wenn sich nichts Grundlegendes geändert hat, vom Beobachter in derselben Weise eingeordnet werden.

3. Unter Validität wird die Gültigkeit von Messinstrumenten, Daten und Ergebnissen verstanden. Misst ein Instrument tatsächlich, was es vorgibt zu messen? Geben die Daten relevante Aspekte des Sachverhaltes wieder?

Objektivität ist eine notwendige, aber nicht hinreichende Voraussetzung für Reliabilität und Reliabilität wiederum eine notwendige, aber nicht hinreichende Bedingung für Validität. Messungen, Beobachtungen und Daten können deshalb objektiv und reliabel (zuverlässig), aber trotzdem nicht valide (gültig) sein: das wäre z.B. dann der Fall, wenn man mit einem Beobachtungsinstrument vermeintlich mangelnde Schülermotivation erfasst, tatsächlich aber nur Müdigkeit misst. Trotzdem können solche invaliden Messungen objektiv und zuverlässig sein, wenn sie nämlich bei dem gleichen Grad von Müdigkeit der Schüler auch immer das gleiche Ergebnis erbringen.

1.2 Die Umsetzung der methodologischen Standards

Die Konzepte Objektivität, Reliabilität und Validität sind nur auf den ersten Blick unproblematisch. Schwierigkeit deuteten sich schon bei der Definition der Reliabilität an: wie soll man die Reliabilität bestimmen, wenn es um Merkmale geht, die nicht stabil sind? Im Folgenden sollen verschiedene theoretische Vertiefungen dieser Konzepte behandelt und technische Verfahren dargestellt werden, mit denen die sehr allgemeinen Qualitätskriterien empirischer Forschung in der Forschungspraxis umgesetzt werden können. Dabei werden wir uns auf drei Felder konzentrieren, von denen die stärksten „Validitätsbedrohungen“ quantitativer Forschung ausgehen: *messtheoretische Konzepte*, *Forschungsdesigns* und Verfahren der *Stichprobenziehung*.

1.2.1 Reliabilität und Konzepte sozialwissenschaftlicher Messung

Eine Messung stellt eine Zuordnung von Symbolen zu Objekten oder Ereignissen nach festen Regeln dar (vgl. Stevens 1951, S. 22). Für eine Messung müssen beobachtete Eigenschaften der untersuchten Objekte oder Ereignisse nach genauen Anweisungen den Ausprägungen oder Werten einer *Variablen* zugeordnet werden.

Eine Variable ist eine definierte veränderliche Größe, die einen festgelegten Wertebereich besitzt. Je nachdem, wie dieser, auch als „Skala“ bezeichnete Wertebereich beschaffen ist, kann man zwischen verschiedenen Arten von Variablen – etwa zwischen *kategorialen* und *metrischen* Variablen – unterscheiden: bei einer kategorialen Variable wird die Skala aus einfachen Kategorien oder Klassenbezeichnungen gebildet. Das theoretische Konzept Motivation lässt sich im einfachsten Fall durch eine solche kategoriale Variable mit nur zwei Ausprägungen bzw. Kategorien darstellen: die Motivation ist entweder „vorhanden“ oder „nicht vorhanden“. Die Abbildung von Schülermotivation auf einer *metrischen* Skala (etwa von 0 bis 100) würde höhere Anforderungen stellen – so müsste z.B. sichergestellt werden, dass mit der Differenz zwischen 0 und 20 Skalenpunkten ein ebenso großer Abstand gemessen wird wie mit der Differenz zwischen 80 und 100. Eine gute Operationalisierung einer kategorialen Variable² erfordert, dass die Kategorien erstens *erschöpfend*, zweitens *disjunkt* und drittens *exklusiv* sind.

1. Die Kategorien der betreffenden Variablen müssen *erschöpfend* sein, das heißt: alle empirisch registrierten Sachverhalte müssen sich in *wenigstens eine Kategorie* einordnen lassen.
2. Des Weiteren müssen die Kategorien *disjunkt* sein, das heißt, sie dürfen sich nicht überschneiden. Ein Ereignis oder Objekt darf in jedem Fall höchstens einer Kategorie zugeordnet werden: Wird etwa Schülermotivation durch Verhaltensbeobachtung erfasst, so darf jedes Schülerverhalten, das berücksichtigt wird, höchstens einer der Kategorien der Variable „Motivation“ (bspw. „motiviert“, „nicht motiviert“ oder „neutral“ bzw. „unklar“) zugeordnet werden können.
3. Schließlich müssen Kategorien *exklusiv* sein, das heißt, mit ihrer Hilfe dürfen nur solche Ereignisse oder Objekte erfasst werden, die theoretisch relevante Eigenschaften besitzen. Bei der Untersuchung über den Zusammenhang zwischen Lehrerverhalten und Schülermotivation soll bspw. nicht das Verhalten der Schüler auf dem Pausenhof erfasst werden, da die Schülermotivation bei Rangelieben dort kaum etwas mit dem Lehrerverhalten im Unterricht zu tun hat.

Ein gutes Beispiel für Schwierigkeiten der Kategorienbildung liefert Tennant (2004) in seiner Untersuchung über Ethnizität und Lehrer-Schüler-Interaktion: die Einordnung von Schülern in ethnische Kategorien kann bspw. hochgradig problematisch sein, weil solche Einordnungen höhere kulturelle Homogenitäten suggerieren, als für viele Gruppen angemessen ist (man denke nur an stark intern

2 Die folgenden Überlegungen lassen sich auch auf metrische Variablen übertragen.

differenzierte Gruppen wie „Asiaten“). Mit Hilfe dieser Überlegungen lassen sich Objektivität und Reliabilität von Messungen nun genauer definieren: Messungen sind objektiv, wenn verschiedene Forscher dieselben Ereignisse und Objekte denselben Kategorien bzw. Skalenwerten zuordnen. Reliabel sind sie, wenn bei mehrfacher Messung die gleichen Sachverhalte immer denselben Variablenwerten zugeordnet werden.

Nun lassen sich bei allen sozialwissenschaftlichen Messungen kleinere oder größere Messfehler nie vermeiden, wobei man *systematische* von *unsystematischen*, *zufälligen* Fehlern unterscheiden kann. Ein systematischer Fehler liegt vor, wenn bspw. ein bestimmtes unmotiviertes Schülerverhalten aufgrund der Messanweisungen grundsätzlich als „motiviert“ eingeordnet wird. Anders als systematische Fehler, die sich durch gründlichere Theoriearbeit und saubere Operationalisierung vermeiden lassen, sind Zufallsfehler nahezu unausrottbar: sie entstehen bspw. durch kleine Unaufmerksamkeiten von Forschern oder Befragten, die die falschen Kategorien ankreuzen. Im Gegensatz zu systematischen Fehlern haben Zufallsfehler allerdings den Vorteil, dass sie sich (weil sie nicht systematisch in eine Richtung erfolgen) bei vielen Messwiederholungen im Durchschnitt ausgleichen und man deshalb Fehlerwahrscheinlichkeiten statistisch berechnen kann. Die Grundlagen für solche Berechnungen und damit für die statistische Bestimmung der Reliabilität liefert die sog. „Klassische Testtheorie“³:

Ein beobachteter Messwert (X) setzt sich zusammen aus dem „wahren Wert“ (T) – z.B. dem tatsächlichen Motivationsgrad eines Schülers – und dem Messfehler (E). Messfehler, die zufällig entstehen, treten unabhängig voneinander auf und können deshalb addiert werden. Bei einer wachsenden Zahl von Messungen wird der Gesamtmessfehler kleiner (weil sich ja die einzelnen Messfehler ausgleichen) und der Messwert X nähert sich dem wahren Wert. Auf dieser Grundlage kann das Konzept der Reliabilität statistisch operationalisiert werden. Verschiedene Messungen an demselben Gegenstand variieren untereinander, wobei ein Teil dieser Variation auf Grund systematischer Einflüsse und ein anderer Teil auf Grund von Messfehlern zustande kommt. Diese Unterschiede werden statistisch mit Hilfe der „Varianz“ – dem Durchschnitt der (quadrierten) Abweichungen aller Messwerte von ihrem gemeinsamen Mittelwert – erfasst. Die Reliabilität kann gemessen werden als Verhältnis zwischen der Varianz der wahren Werte t_1, \dots, t_n (also der wirklichen Unterschiede zwischen den Schülern) und der Varianz der gemessenen

3 Neben der „Klassischen Testtheorie“ wird heutzutage bei der Konstruktion von sozialwissenschaftlichen Skalen häufig die „Probabilistische Testtheorie“ eingesetzt, die teilweise von realistischeren Annahmen ausgeht – deren Darstellung hier aber den Rahmen sprengen würde (vgl. hierzu etwa Bühner 2011).

Werte x_1, \dots, x_n (in denen die Messfehler enthalten sind). Eine Reliabilität von .80⁴ würde also bedeuten, dass der Anteil der „wahren“ Varianz an der Gesamtvarianz bei 80% liegt.

Die Reliabilität eines Instruments (das in der Regel aus mehreren Variablen besteht, die eine Eigenschaft messen sollen) wird dann durch unterschiedliche Vergleiche zwischen Messungen bestimmt.

Bei der Erfassung der *Test-Retest-Reliabilität* werden Messungen wiederholt durchgeführt und gelten bei einer hohen Übereinstimmung zwischen den einzelnen Messwerten als reliabel. Ein gutes Beispiel im Feld der Lehrer-Schüler-Interaktion liefern Fish und Dane (2000), die die Test-Retest-Reliabilität für eine Beobachtungsskala zu den Eigenschaften „Zusammenhalt“, „Flexibilität“ und „Kommunikation“ im System „Klassenzimmer“ berechnen. Die Test-Retest-Reliabilität lässt sich allerdings nicht sinnvoll bei Variablen einsetzen, die zeitlich nicht stabil sind. Will man etwa Lernfortschritt von Schülern erfassen, ist die Änderung des Messwertes beabsichtigt und eine einfache Wiederholungsmessung sollte dann nach einiger Zeit keine ähnlichen Messwerte mehr liefern. Weitere Formen der Reliabilitätsmessung erfassen die *Homogenität* eines Messinstruments (das man dann, wenn verschiedene Variablen zu einem Index aufaddiert werden, auch als Skala ansehen kann), bspw. die *Split-Half-Reliabilität*, bei der ein Test in zwei äquivalente Testhälften unterteilt wird oder die *Paralleltest-Reliabilität*, bei der Parallelformen eines Testes hergestellt und verglichen werden.

Validität wird stärker als die Reliabilität durch „inhaltslogische“ Überlegungen definiert und ist deswegen leider nicht so einfach statistisch zu operationalisieren und in Messwerte umzusetzen. Grundsätzlich unterscheidet man drei Arten von Validität: *Inhalts-*, *Kriteriums-* und *Konstruktvalidität*.

1. *Inhaltsvalidität* kommt der klassischen Definition⁵ (Validität erfasst das Ausmaß der Genauigkeit, mit dem ein Instrument tatsächlich das misst, was es messen soll) am nächsten: ein Instrument zur Messung von kognitiver Leistungsfähigkeit, das stattdessen das Sprachverständnis erfasst, das zum Verstehen der Aufgaben nötig ist, ist nicht inhaltsvalide. Im konkreten Fall ist Inhaltsvalidität aber oft schwer bestimmbar, weil sie sich nur durch fachspezifische, gegenstandsbezogene und logische Überlegungen erfassen lässt.

4 Reliabilität wird normalerweise mit der Pearson'schen Produkt-Momentkorrelation gemessen. Die „erklärte Varianz“ ergibt sich, wenn man den Pearson'schen Koeffizienten r quadriert. Diese Zusammenhänge werden ausführlicher in jedem Statistiklehrbuch dargestellt.

5 angelehnt an Lienert (1967)

2. Die *Kriteriumsvalidität* ist demgegenüber leichter empirisch und messtheoretisch zu operationalisieren, nämlich als Korrelation zwischen Ergebnissen, die mit einem Messinstrument erhoben wurden und einem externen Kriterium. So validieren bspw. Blankemeyer und Kollegen (2002) ihren „*Relationship with teacher*“-Fragebogen erstens anhand des Kriteriums „Grad, mit dem die Schüler ihre Schule mögen“ und zweitens anhand der Antworten der Schüler auf die Frage, wie glücklich sie jeden Tag in der Schule sind. Grundsätzlich kann man bei der Kriteriumsvalidität zwischen *prädiktiver* und *konkurrenter* Validität unterscheiden: bei der Erfassung der *prädiktiven Validität* erfolgt die Validierung mit Hilfe eines in der Zukunft beobachtbaren Kriteriums (wenn man etwa untersucht, ob die Abiturnote ein valider Indikator für Studienerfolg ist, indem man Abiturnoten mit Abschlussnoten der Universität vergleicht, vgl. u.a. Rindermann & Oubaid 1999; Gold & Souvignier 2005). *Konkurrente Validität* untersucht man mit Hilfe von Variablen, die zum selben Zeitpunkt erhoben werden wie das zu validierende Merkmal. Neben diesen beiden Hauptformen der Kriteriumsvalidität werden weitere in der Literatur diskutiert (vgl. z.B. Bühner 2011; Diekmann 2008; Schnell u.a. 2013), von denen hier nur die *inkrementelle Validität* erwähnt werden soll, weil sie einen Sonderstatus einnimmt: sie gibt an, welchen zusätzlichen Erkenntnisgewinn ein neues Messinstrument gegenüber bereits etablierten Verfahren erbringt.
3. Die empirische Bestimmung der *Konstruktvalidität* ist deutlich komplizierter. Hierzu muss *a priori* festgelegt werden, welche Eigenschaftsdimensionen das gemessene Konstrukt aufweist und mit welchen Eigenschaftsdimensionen anderer Konstrukte sie in Beziehung stehen. Sowohl Abweichungen als auch Übereinstimmungen zwischen Messinstrumenten, mit denen verschiedene Konstrukte gemessen werden, können dabei erwünscht sein. Eine hohe Konvergenz deutet nämlich auf eine hohe Übereinstimmung zweier Konstrukte hin und zeigt damit deren „*konvergente Validität*“. Abweichungen zu Messinstrumenten, mit denen andere Konstrukte erfasst werden, führen zur Bestimmung der „*diskriminanten Validität*“. Ein etabliertes Instrument zur Überprüfung der Konstruktvalidität ist das Konzept der *Multitrait-Multimethod-Matrix* (MTMM) von Campbell und Fiske (1959) (ausführliche Darstellungen bei Schermelleh-Engel & Schweitzer 2012, S. 346ff.); hierbei werden unterschiedliche Konstrukte gleichzeitig mit verschiedenen Messinstrumenten erfasst und es wird eine Matrix der Korrelationen aller Messungen erstellt.

Unter messtheoretischer Perspektive bleibt Validität ein schwieriges Konzept: Inhaltsvalidität lässt sich empirisch und statistisch nicht erfassen und die Messung der Konstrukt- oder Kriteriumsvalidität führt nur zur Prüfung abgeleiteter Aus-

sagen, aber nicht zur Testung der Validität des Messinstruments selber (vgl. dazu auch Bühner 2011, S.60ff.).

1.2.2 Konzepte zur Bestimmung interner und externer Validität

Einen anderen Zugang zum Validitätsbegriff eröffnet die Unterscheidung zwischen „externer Validität“ und „interner Validität“, die Campbell und Stanley in einer Grundlagenarbeit zu experimenteller und quasi-experimenteller Forschung in den Sozial- und Erziehungswissenschaften entwickeln (Campbell & Stanley 1963):

1. *Externe Validität* bezeichnet das Ausmaß, in dem ein bestimmtes Messergebnis auf andere Populationen und auf andere Settings verallgemeinert werden kann.
2. *Interne Validität* bezieht sich auf die Frage, ob das Ergebnis eines Experiments tatsächlich auf den experimentellen Einfluss (auf das „treatment“) zurückgeführt werden kann.

Da Campbell und Stanley ihre Vorstellungen von Validität ausgehend von der Logik des Experiments entwickeln, werden wir zuerst zentrale Aspekte dieser Logik und das hier zugrunde liegende Konzept von „Kausalität“ diskutieren, um den Begriff der „internen Validität“ zu erläutern. Anschließend werden wir auf „externe Validität“ und auf der Problem der Verallgemeinerung eingehen.

1.2.2.1 Das Kausalitätsparadigma

Um Erkenntnismöglichkeiten und Probleme experimenteller Forschung zu verstehen, sind Grundkenntnisse über das Konzept der Kausalität hilfreich, wie es in der modernen Wissenschaftstheorie im Anschluss an Humes *Regularitätstheorie der Kausalität* (Hume 1748/1957) entwickelt wurde. Naturvorgänge werden dann sinnvollerweise als kausal interpretiert, so Hume, wenn Ereignisse immer wieder in derselben Weise in zeitlicher Aufeinanderfolge beobachtet werden können: auf eine Ursache X muss also immer und ohne Ausnahme eine bestimmte Wirkung Y folgen. Dieses strikt deterministische Kausalitätsverständnis wirft jedoch Probleme auf, die unter dem Begriff der „Hintergrundbedingungen“ (vgl. Mackie 1980) oder der „*ceteris paribus* Klausel“⁶ diskutiert werden. Selbst bei einem so trivialen Beispiel wie der Betätigung eines Lichtschalters lassen sich prinzipiell unendlich viele Hintergrundbedingungen finden, die gegeben sein müssen, damit tatsächlich Licht aufleuchtet, wenn der Lichtschalter gedrückt wird: so darf der Stromkreis-

6 „*ceteris paribus*“ bedeutet: bei ansonsten gleichen (Hintergrund)bedingungen

lauf nicht unterbrochen sein, die Glühbirne muss intakt sein, das Elektrizitätswerk muss Strom liefern u.v.a.m.. Die Betätigung des Schalters kann also nur dann als eine Hume'sche Ursache für das Aufleuchten des Lichts angesehen werden, wenn alle diese Bedingungen immer konstant sind und deshalb in einer „*ceteris paribus* Klausel“ vernachlässigt werden dürfen. Variieren die Hintergrundbedingungen aber im Untersuchungsfeld, dann müssten alle Bedingungen bekannt sein, damit man eine Kausalaussage unter Beachtung der Hume'schen Regularitätstheorie formulieren darf (denn wenn die Glühbirne nicht *immer* bei Betätigung des Schalters aufleuchtet, ist die Anforderung eines „konstanten Zusammenhangs“ zwischen Ursache und Wirkung ja verletzt). Dies ist in den Sozialwissenschaften und der Psychologie eine unangemessen strenge Anforderung, deren genaue Beachtung die wissenschaftliche Arbeit sehr behindern würde. Die Lernmotivation eines Schülers ist natürlich nicht nur vom Lehrerverhalten, sondern von zahlreichen anderen Einflüssen (seinem häuslichen Umfeld, seinem momentanen Gesundheitszustand, Ablenkung durch Mitschüler usw.) abhängig, die stark variieren können. Trotzdem ist es sinnvoll, von einer kausalen Wirkung des Lehrerverhaltens auf die Schülermotivation auszugehen, ohne bei der Untersuchung dieses Zusammenhangs alle diese Einflüsse immer einzubeziehen.

Dieses Problem lässt sich theoretisch durch eine *statistische* und durch eine *interventionistische Theorie* der Kausalität lösen. Beide Ansätze führen zu verschiedenen (aber miteinander verknüpfbaren) methodischen Strategien:

1. Dem Prinzip der probabilistischen oder statistischen Kausalität zufolge wirkt ein Ereignis X auf ein Ereignis Y kausal, wenn X die Auftretenswahrscheinlichkeit (bzw. die Häufigkeit des Auftretens) von Y beeinflusst⁷ (Suppes 1970, S. 10). Die Variable, die das Ereignis X erfasst, bezeichnet man dabei als unabhängige Variable, die Variable Y als abhängige Variable (weil sich Y in Abhängigkeit von X verändern soll). Ein bestimmtes Lehrerverhalten kann demnach als kausaler Faktor für die Schülermotivation gelten, wenn Schüler häufiger (oder seltener) motiviert sind, wenn Lehrer dieses Verhalten zeigen. Auf diese Weise wird die deterministische Regularitätstheorie abgeschwächt und es wird bspw. möglich, Hintergrundbedingungen, die für eine sozialwissenschaftliche Fragestellung nicht relevant sind (situative Bedingungen, die Laune bestimmter Schüler usw.) zu vernachlässigen.
2. Schon der Philosoph Mill hatte darauf hingewiesen, dass Beobachtung ohne experimentelle Einflussnahme zwar eine konstante Aufeinanderfolge von Er-

7 Außerdem darf keine zusätzliche Bedingung Z existieren, die im gemeinsamen Auftreten mit X die Auftretenswahrscheinlichkeit von Y verändert.

eignissen zeigen kann, aber nicht eine Ursache-Wirkungsbeziehung (Mill 1874, S. 277). So lässt sich ein im Unterricht beobachteter Zusammenhang zwischen positiven Rückmeldungen und Schülermotivation dadurch erklären, dass das Lehrerverhalten das Schülerverhalten, aber auch dadurch, dass das Schülerverhalten das Lehrerverhalten beeinflusst: möglicherweise geben Lehrer deswegen viele positive Rückmeldungen, weil die Schüler bereits motiviert mitarbeiten. Eine experimentelle Intervention könnte diese Unklarheit beseitigen. Wenn Lehrer von einem bestimmten Zeitpunkt an gezielt positive Rückmeldungen geben und die Schüler daraufhin motivierter arbeiten, lassen sich Hypothesen über die Richtung der Kausalbeziehung besser absichern. Die *Interventionstheorie der Kausalität* (vgl. Woodward 2001) geht (in ihrer starken Fassung) sogar davon aus, dass sich Kausalbeziehungen nur durch Interventionen nachweisen lassen (Holland 1985, S. 51).

1.2.2.2 Experimentelle und quasi-experimentelle Designs

Die Interventionstheorie der Kausalität führt konsequenterweise zum Experiment, dessen Anwendung in der sozialwissenschaftlichen und psychologischen Forschung aber oftmals schwierig ist, weil der Experimentator dort nicht immer alle kausalen Bedingungen beeinflussen kann. Dies betrifft insbesondere die Lehrer-Schüler-Interaktion – Theorien über den kausalen Einfluss eines bestimmten Lehrerverhaltens können nicht im Labor, sondern nur im natürlichen Umfeld der Schule geprüft werden.

Campbell und Stanley wenden sich genau diesem Problem zu und kritisieren bestimmte Arten der empirischen Forschung, die sie als „vor-experimentelle Designs“ bezeichnen und deren prinzipiellen Aufbau sie mit „X-O Diagrammen“ veranschaulichen: „X“ steht dabei für die Intervention (das „treatment“) und „O“ für die jeweilige Messung („observation“). Bei dem vorexperimentellen Design der „one shot case study“ wird auf eine Erhebung des Wertes der abhängigen Variablen (zum Zeitpunkt t_1) ganz verzichtet, das *treatment* durchgeführt und anschließend der Wert der abhängigen Variablen ermittelt (vgl. Tabelle 1).

Tabelle 1 one shot case study

	t_1	t_2	t_3
		X	O

Die zentrale Schwäche dieses Designs besteht darin, dass kein Vergleich möglich ist. Die Feststellung, dass die durchschnittliche Schülermotivation bspw. in einer bestimmten Schulklasse nach einer Schulung der Lehrer einen bestimmten

Wert hat, hat für sich genommen gar keine Aussagekraft. Die einfachste Möglichkeit, einen Vergleichsmaßstab herzustellen, bestünde darin, dass in anderen Schulklassen (in denen keine Intervention stattfand) geprüft wird, wie hoch die durchschnittliche Schülermotivation dort ist. Aber auch die Schlussfolgerungen, die man aus solch einer *static group comparison* (Tabelle 2) ziehen kann, ist eingeschränkt: es bleibt die Möglichkeit, dass die Schulklasse, in der die Intervention stattfand (die „Versuchsgruppe“) von vornherein einen höheren Durchschnittswert als die Vergleichsgruppe (die „Kontrollgruppe“) aufwies.

Tabelle 2 static group comparison

	t ₁	t ₂	t ₃
		X	O ₁
			O ₂

Diese Validitätsbedrohung wird bei der dritten Form des vor-experimentellen Designs, beim sog. „*Eingruppen-Pretest-Posttest-Design*“ (vgl. Tabelle 3) zwar vermieden. Weil hier aber wiederum eine Kontrollgruppe fehlt, kann man nicht feststellen, ob eine gemessene Veränderung nicht ohnehin (etwa durch einen „Reifungseffekt“, d.h. aufgrund der Tatsache, dass die Schüler bei der zweiten Messung wegen ihres höheren Alters psychosozial gereift sind) stattgefunden hätte.

Tabelle 3 Eingruppen-Pretest-Posttest-Design

	t ₁	t ₂	t ₃
	O ₁	X	O ₂

Ein echtes experimentelles Design (Fisher 1935) erfordert demgegenüber, dass die Werte der abhängigen Variablen sowohl in der Versuchs- als auch in der Kontrollgruppe zweimal (nämlich vor *und* nach dem *treatment*) gemessen werden. Idealerweise erfolgt die Aufteilung der Versuchspersonen auf die Versuchs- und Kontrollgruppe dabei „randomisiert“, also zufällig (in der Tabelle 4 wird dies dargestellt durch das „R“ in der ersten Spalte): Dabei muss jeder Beteiligte am Experiment die Chance haben, der Versuchsgruppe oder der Kontrollgruppe zugeordnet zu werden, weil nur so „Selektionseffekte“ ausgeschlossen werden können: Lässt man bspw. Lehrern die Wahl, ob sie eine neue Methode der Schülermotivation im Unterricht einführen oder nicht, ist es wahrscheinlich, dass nur die von der Methode überzeugten Lehrer die Versuchsgruppe bilden. Nun ist es aber leicht möglich, dass gerade die von der Methode überzeugten Lehrer die Schüler auf-

grund ihres eigenen Engagements mitreißen. Der entscheidende kausale Faktor wäre dann nicht das *treatment*, sondern eine andere Variable (hier: das Engagement der Lehrer).

Tabelle 4 Randomisiertes Pretest-Posttest-Kontrollgruppendesign

	t ₁	t ₂	t ₃
R	O ₁	X	O ₂
R	O ₃		O ₄

Eine weitere wichtige Strategie ist die „Verblindung“: den Versuchspersonen und/oder dem Versuchsleiter wird verschwiegen, wer das *treatment* erhält. Hierdurch hofft man „Versuchsleitereffekten“ entgegenzuwirken: Wissen oder Überzeugungen der Versuchsleiter über die Versuchspersonen kann nämlich die Versuchsergebnisse mehr oder weniger stark beeinflussen: In lang laufenden experimentellen Studien an Schulen zeigten Schüler, von deren besonderer Leistungsstärke die Versuchsleiter überzeugt waren, tatsächlich nach einer gewissen Zeit durchschnittlich bessere Leistungen (Rosenthal & Jacobson 1974). Ein Beispiel für die (in sozialwissenschaftlichen Studien oft sehr schwer umzusetzende) Verblindung liefern ter Laak und Kollegen (2001), die in ihrer Untersuchung über Lehrerurteile Schülergruppen gebildet haben, deren Zusammensetzung aus „normalen“ und „schwierigen“ Schülern den Beobachtern nicht mitgeteilt wurde (vgl. ter Laak u.a. 2001, S. 261).

Letztendlich ist die randomisierte Aufteilung der Versuchspersonen auf Versuchs- und Kontrollgruppe für ein Experiment sogar noch bedeutsamer als die Durchführung von Pretest-Messungen⁸, so dass auch ein randomisiertes *Nur-Posttest-Kontrollgruppendesign* (vgl. Tabelle 5), der Gruppe der „echten“ experimentellen Designs zugeordnet wird.

Tabelle 5 Randomisiertes Nur-Posttest-Kontrollgruppendesign

	t ₁	t ₂	t ₃
R		X	O ₂
R			O ₄

8 Unterschiede zwischen Versuchs- und Kontrollgruppe, die eine Vorhermessung zeigen könnte, lassen sich durch echte Randomisierung mit einer statistisch bestimmaren Wahrscheinlichkeit ausschließen.

Nun ist aber eine zufällige Aufteilung der Versuchspersonen in den Sozialwissenschaften oft entweder aus ethischen Gründen nicht vertretbar oder nicht praktikabel – schließlich lassen sich bspw. Kinder nicht nach dem Zufallsprinzip in „normale“ und „schwierige“ Schüler umwandeln. Muss man auf eine Randomisierung verzichten, so können die besonderen Schwächen vorexperimenteller Designs vermieden werden, indem zumindest eine Vorher- und Nachhermessung und eine Aufteilung in Versuchs- und Kontrollgruppe stattfindet. Campbell und Stanley (1963) sprechen hier von einem „quasi-experimentellen Design“ (vgl. Tabelle 6).

Tabelle 6 Quasi-experimentelles Pretest-Posttestdesign mit nicht-äquivalenter Kontrollgruppe

	t ₁	t ₂	t ₃
	O	X	O
	O		O

Durch den Vergleich der nicht-randomisierten Gruppen können zumindest viele Fehlschlüsse vermieden werden, die durch Selektion oder Reifung entstehen können – wichtig ist, dass Versuchs- und Kontrollgruppe sich in ihrer Zusammensetzung, so weit es geht, ähneln. Untersucht man bspw. Einflüsse der Klassengröße auf das Verhalten von Lehrern (Blatchford 2003), so muss man versuchen, Unterschiede zwischen Versuchs- und Kontrollklassen zu vermeiden, die durch andere Faktoren entstehen (etwa weil die kleinen Schulklassen aus einer Grundschule in einem bildungsbürgerlich geprägten Stadtteil stammen und die großen Schulklassen in einem sozialen Brennpunkt liegen).⁹

Die schlimmsten Mängel vorexperimenteller Designs können aber auch in Einzelgruppenexperimenten beherrschbar werden, etwa in *Zeitreihenexperimenten ohne Kontrollgruppe* (vgl. Tabelle 7). Durch zahlreiche Messungen vor und nach dem *treatment* lassen sich Reifungseffekte gut erkennen. Wenn die Veränderung der abhängigen Variable (etwa: Lernerfolg) zwischen O₃ und O₄ (also vor und nach dem *treatment*, etwa der Einführung einer neuen Lernmethode) genauso groß ist wie zwischen O₁ und O₂, ist das ein wichtiger Hinweis darauf, dass das *treatment* hier keinen eigenständigen Effekt hat.

9 Manche Selektionseffekte, die durch das Forschungsdesign nicht ausgeschlossen werden können, lassen sich allerdings nachträglich statistisch kontrollieren, zumindest dann, wenn man die Störvariablen kennt.

Tabelle 7 Zeitreihenexperiment ohne Kontrollgruppe

O ₁	O ₂	O ₃	X	O ₄	O ₅	O ₆
----------------	----------------	----------------	---	----------------	----------------	----------------

1.2.2.3 Stichprobenziehung und das Problem der Verallgemeinerbarkeit

Die Auswahl der Untersuchungseinheiten hat eine zentrale Bedeutung sowohl für die Aussagekraft von Experimenten als auch in der (nicht-experimentellen) Umfrageforschung – eine brauchbare Aussage über die *Grundgesamtheit* oder *Population* (etwa: alle Schüler der 9. Klassen von Gymnasien) ist nur möglich, wenn die in der Untersuchung verwendete Stichprobe *hinreichend* „repräsentativ“ ist. Der Begriff der Repräsentativität (die statistische Verteilung der Merkmale in der Stichprobe entspricht der Verteilung der Merkmale in der Population) lässt sich nur merkmalsbezogen sinnvoll verwenden – eine Stichprobe kann repräsentativ bezogen auf eine bestimmte Variable (bspw. Geschlecht), aber nicht-repräsentativ bezogen auf viele andere (etwa: soziale Herkunft, Motivation) sein. Leider wird der Begriff oft missbraucht, indem die Repräsentativität einer Stichprobe bezogen auf wenige (z.B. sozialstatistische) Merkmale bestimmt und dann der Eindruck erweckt wird, die Stichprobe sei deshalb umfassend (d.h. auch bezogen auf andere Variablen, deren Verteilungen in der Grundgesamtheit unbekannt sind) repräsentativ. Entscheidend für die Einschätzung der Repräsentativität ist also letztendlich nicht ein Vergleich mit Werten aus einer Grundgesamtheit, sondern das *Ziehungsverfahren*. Dabei sind zwei Verfahren gebräuchlich: die *bewusste Auswahl* und die *Zufallsauswahl*.

1. Bei einer *bewussten Auswahl* werden bestimmte Merkmale festgelegt, die für die Fragestellung relevant sind und auf die bezogen Repräsentativität hergestellt werden soll. Man kann die Kenntnis der Verteilung dieser Merkmale in der Population bspw. dazu nutzen, Interviewern „Quotenpläne“ auszuhändigen (sie etwa instruieren, eine bestimmte Anzahl von Frauen und von Männern zu befragen). Ein Nachteil solcher Quotenstichproben besteht darin, dass man schon im Vorfeld ein sehr umfassendes Wissen über den Gegenstandsbereich haben muss, um die relevanten Merkmale für die Ziehung auswählen zu können, wobei nie ausgeschlossen werden kann, dass nicht doch unbekannte Faktoren die Stichprobe verzerren.
2. Die *Zufallsziehung* bietet den Vorteil, dass sich hierbei die Wahrscheinlichkeit, eine bezüglich aller möglichen Merkmale repräsentative Stichprobe zu ziehen, berechnen lässt, wobei diese Wahrscheinlichkeit mit steigendem Stichproben-

umfang wächst. Definiert wird eine Zufallsstichprobe dadurch, dass jedes Element der Grundgesamtheit eine von 0 verschiedene, angebbare Chance hat, in die Stichprobe zu gelangen. Realisiert werden kann eine Zufallsstichprobe im einfachsten Fall durch eine sog. „Listenauswahl“, bei der aus einer vollständigen Liste der Grundgesamtheit die gewünschte Anzahl von Elementen zufällig ausgewählt wird.

In der Praxis der empirischen Sozialforschung werden beide Verfahren oftmals kombiniert, wobei eine reine Zufallsauswahl oft schwer zu realisieren ist¹⁰. Man behilft sich hier mit Verfahren einer „mehrstufigen“ Zufallsauswahl, wie sie etwa in der „Allgemeinen Bevölkerungsumfrage in den Sozialwissenschaften“ (ALLBUS), einer regelmäßig wiederholten Befragung einer Bevölkerungsstichprobe in Deutschland, stattfindet (vgl. Wasmer u.a. 2014): in einer ersten Stufe der Zufallsauswahl werden sog. „*sampling points*“ festgelegt – dies waren für die Befragung 2012 162 Ortsgemeinden in ganz Deutschland. Anschließend werden aus den Einwohnermelderegistern dieser Gemeinden zufällig Personenadressen gezogen.

Ein solches Vorgehen gestaltet sich komplexer, wenn Teile der Stichprobe „geschichtet“, d.h. getrennt nach Subpopulationen, gezogen werden. So wurde in der deutschen Erweiterung der PISA-Studie zuerst nach Bundesländern geschichtet, indem für jedes Bundesland eine eigene Stichprobe gezogen und innerhalb der Länder die Stichproben differenziert nach Schulformen geschichtet wurden. Dabei wurden die Teilstichproben „disproportional“ geschichtet – das heißt, dass der relative Umfang der Länderstichproben nicht den realen Bevölkerungszahlen entspricht und dass die Stichproben der einzelnen Schulformen nicht die tatsächliche Verteilung der Schüler widerspiegeln: die Stichprobe in Hamburg ist etwa genauso groß wie die Stichprobe in Nordrhein-Westfalen, die Stichprobe der Gymnasiasten genauso groß wie die Stichprobe der Hauptschüler (obwohl wesentlich mehr Schüler ein Gymnasium besuchen als eine Hauptschule). Hierdurch sollen differenzierte statistische Analysen auch für kleine Subpopulationen ermöglicht werden. Die realen Verhältnisse können dann bei der statistischen Auswertung durch eine proportionale „Gewichtung“¹¹ wieder hergestellt werden.

10 Für die Ziehung einer Bevölkerungsstichprobe durch ein Listenverfahren fehlt bspw. in Deutschland ein einheitliches bundesweites Melderegister, hinzukommen oft Probleme des *nonresponse* (s.u.).

11 Bei den Schätzungen von Kennwerten für Bundesländer und Schulformen wird ein Korrekturfaktor verwendet, der der jeweiligen Größe des Bundeslandes und der jeweiligen Anzahl von Schülern in den verschiedenen Schulformen gerecht wird (vgl. Deutsches PISA-Konsortium online, o.J.).

Die Repräsentativität einer Stichprobe kann entweder durch *Zufallsfehler* oder durch *systematische Fehler* eingeschränkt werden:

1. *Zufallsfehler* lassen sich nie vollständig vermeiden – die einzige Möglichkeit, das Risiko für solche Fehler möglichst gering zu halten, besteht darin, dass der Stichprobenumfang hinreichend groß gehalten wird. Mit Hilfe wahrscheinlichkeitstheoretischer Modelle lässt sich die Größe des Stichprobenfehlers bei gegebener Stichprobe jedoch schätzen.
2. Bei *systematischen Fehlern* („*biases*“) wird das Prinzip, dass jedes Element der Grundgesamtheit eine Chance haben muss, in die Stichprobe zu gelangen, systematisch verletzt. Dies ist etwa bei Passantenbefragungen der Fall (Personen, die sich tagsüber seltener am Befragungsort aufhalten, etwa Vollzeitberufstätige, sind hier stark unterrepräsentiert). Grundsätzlich können aber auch solche Stichproben nützlich sein, solange der Verallgemeinerungsanspruch nicht über die faktische Grundgesamtheit (z.B. „*Personen, die sich tagsüber in Einkaufspassagen bewegen und sich von Interviewern ansprechen lassen*“) hinausgeht. Im Gegensatz zu Zufallsfehlern lassen sich systematische Fehler nicht statistisch beherrschen, sondern nur methodisch (bspw. durch praktische Vorkehrungen bei der Stichprobenziehung) begrenzen.

1.3 Die Erhebung standardisierter Daten

Standardisierte Daten werden in den Sozialwissenschaften in der Regel entweder durch *Befragung* oder durch *Beobachtung* gewonnen. Diese beiden zentralen Erhebungsverfahren und ihre Probleme sollen im Folgenden besprochen werden.

1.3.1 Befragung

Die Erhebung standardisierter Befragungsdaten erfolgt in der Regel durch Fragebögen. Die einzelnen Fragen (auch „*Items*“ genannt) repräsentieren dabei jene Variablen, die durch die Operationalisierung der theoretischen Konzepte entstanden sind. Im Idealfall kennt man bereits deren mögliche Ausprägungen, die dann dem Befragten in Form fester Antwortalternativen vorgelegt werden. In standardisierten Fragebögen vermeidet man, viele „*offene Fragen*“ zu stellen, deren Auswertung (durch eine Entwicklung von Kategorien nach der Erhebung) aufwändig ist. Die Konstruktion valider Fragebögen ist anspruchsvoll und zeitraubend, deshalb werden (wenn die Untersuchungsfragestellung dies zulässt) gerne bereits erprobte Instrumente genutzt. Klassische Fragebogenformen für die Untersuchung der

Lehrer-Schüler-Interaktion lassen sich im ICEQ (Individualised Classroom Environment Questionnaire) von Fraser (1990) oder dem QTI (Questionnaire on Teacher Interaction) von Wubbels & Levy (1993) finden. Der ICEQ widmet sich dem Lernumfeld¹², der QTI, den auch Mellor und Moore (2003) verwenden, dem Verhalten des Lehrers gegenüber seinen Schülern¹³. Bei komplexeren Instrumenten werden, manchmal mit Hilfe spezifischer statistischer Verfahren wie der Faktorenanalyse, Items zu „Itembatterien“ zusammengefasst, die dann eine neue Variable oder Skala repräsentieren. Für die Messung der Qualität solcher Skalen existieren verschiedene statistische Verfahren und Koeffizienten, die die Reliabilität (genauer: die interne Konsistenz, s.o.) einer Skala erfassen sollen (ein Überblick hierzu findet sich bei Bühner 2011).

Hinsichtlich der Interaktion zwischen Interviewer und Befragtem lassen sich verschiedene Befragungsformen unterscheiden: die *schriftliche Befragung* (z.B. als Befragung im Klassenzimmer oder als postalische Befragung), die *Telefonbefragung*, die direkte mündliche *face-to-face-Befragung* und in neuerer Zeit die *Online-Befragung*.

Eine eigene Methodenforschung (Porst 2000; ein Überblick findet sich bei Engel & Schmidt 2014) befasst sich mit den besonderen Problemen der Befragung, etwa mit dem sog. *nonresponse*: bei zahlreichen Untersuchungen nehmen viele der ursprünglich vorgesehenen Personen nicht teil – Gründe hierfür sind *Nichtbefragbarkeit* (etwa aufgrund von Krankheit), *Nichterreichbarkeit* oder *Verweigerung*. *Nonresponse* ist für Bevölkerungsumfragen bedeutsamer als bei Befragungen in institutionellen Kontexten, bspw. in Schulen – aber auch hier kann offene und verdeckte Verweigerung bspw. den Rücklauf von Fragebögen, die an Lehrer oder Schulleitungen ausgegeben werden, erheblich senken und eine ganze Studie gefährden. Problematisch wird *nonresponse*, wenn die *Nonresponder* sich hinsichtlich relevanter Merkmale deutlich von der Gruppe der Befragten unterscheiden und auf diese Weise systematische Stichprobenfehler erzeugen. Hinweise auf Merkmalshäufungen in der Gruppe der Befragten oder bei *Nonrespondern* müssen deshalb sehr aufmerksam registriert werden. Neben dem „*unit-nonresponse*“ (komplette Fragebögen gelangen nicht in die „Nettostichprobe“), tritt ebenfalls häufig das

12 Beispielitems: Gefragt wird danach wie oft der Schüler möchte, dass die im Item angesprochenen Dinge in seiner Klasse passieren: „Students would be punished if they behaved badly in class“ (Item 28); „Students would work on their own speed“ (Item 5). Die Items wurden der „Long Form“ des „Preferred Classroom“ Fragebogens entnommen (Fraser 1990, S. 30)

13 Beispielitems: Die Aussagen der Items sollen anhand einer 5-stufigen Likert-Skala (Likert 1932, vgl. Diekmann 2008, S. 240ff.) von A=Never bis E=Always eingeordnet werden: „He (der Lehrer) thinks we cheat“ (Item 19); „We are afraid of him“ (Item 61).

Problem des „*item-nonresponse*“ auf: Befragte verweigern z.B. die Beantwortung bestimmter, zumeist sensibler Fragen (etwa Fragen nach dem Einkommen) oder brechen die Beantwortung des Fragebogens aus mangelnder Motivation ab.

Weitere Fehlerquellen können sich entweder aus dem Fragebogen („*Frageeffekte*“) oder aus der Interaktion zwischen Befragtem und Interviewer ergeben, wobei hier sowohl *Merkmale der Interviewer* (die zu „*Interviewereffekten*“ führen) als auch solche der *Befragten* (die „*Befragteneffekte*“ erzeugen) eine Rolle spielen können.

1. Unter den *Befragtenmerkmalen* besonders erwähnenswert ist eine oft vorhandene Tendenz zu „sozial erwünschtem“ Antwortverhalten: Viele Interviewpartner wählen bestimmte Antwortalternativen eines Items oft nicht, weil sie sie als gesellschaftlich nicht akzeptabel einschätzen, obwohl sie ihre eigentliche Einstellung wiedergeben würden. Auch ist die generelle Befragungsbereitschaft bei potentiellen Interviewpartnern sehr unterschiedlich ausgeprägt. Der „Umfragemüdigkeit“ (Porst 1996) von Befragten kann man mit besonderen Maßnahmen begegnen (indem man etwa kleine finanzielle Anreize bietet). Das kann allerdings Zweifel an der Validität von Untersuchungen wecken, wie sie in Medienberichten zur PISA-Studie von 2006 ihren Ausdruck fanden. Hier wurde die Vergleichbarkeit der Ergebnisse angezweifelt, weil in einigen Staaten (wegen einer angeblichen „Testmüdigkeit“ der Schüler) die Teilnahme mit Beträgen von bis zu 50 € prämiert wurde (vgl. ZEIT online 2007)¹⁴.
2. Auch *Interviewermerkmale* (etwa Merkmale, die auf die soziale Herkunft des Interviewers oder auf bestimmte Einstellungen schließen lassen) können das Antwortverhalten beeinflussen und Effekte sozialer Erwünschtheit verstärken. Hierzu gehört auch der sog. „Sponsorship-Effekt“. Die Kenntnis über den Auftraggeber kann das Antwortverhalten beeinflussen – bei der Befragung von Schülern über Eigenschaften ihrer Lehrer wird es bspw. eine Rolle spielen, ob der Interviewer von der Schülerzeitung kommt oder vom Kultusministerium beauftragt wurde.
3. In der Literatur wurden bislang sehr zahlreiche und unterschiedliche *Frageeffekte* beschrieben (vgl. etwa Groves 2004). Ergebnisse einer Befragung können oft durch minimale Umformulierungen stark verändert werden, wie Krämer (2015, S. 129ff.) deutlich macht. Eine Übersicht über solche Frageeffekte findet sich bspw. bei Porst (2000).

14 Bei der ersten PISA Studie wurde vom deutschen Konsortium der verfälschende Einfluss motivationaler Anreize in einer gesonderten Untersuchung ausgeschlossen (vgl. Max-Planck-Institut für Bildungsforschung online, o.J.)

Fehler in Befragungen können auch durch das Zusammenwirken mehrerer dieser Fehlerquellen zustande kommen. Ein gutes Beispiel hierfür liefert das Problem der *Validität von Selbstbeschreibungen*, etwa zum Arbeitsverhalten von Schülern: in der ersten PISA-Studie von 2000 wurden bspw. Skalen zum „selbstregulierten Lernen“ (Boekaerts 1999; Zimmerman 1999) eingesetzt, die die Schüler direkt nach ihrem Arbeitsverhalten fragten. Solche Selbstbewertungen können valide Daten liefern, wie Schneider (1996) zeigt. Andere Forscher haben aber auch Probleme dabei entdeckt, etwa Blankemeyer und Kollegen (2002) in ihrer Studie zu Aggression und sozialer Kompetenz, die mit Hilfe von Schüler-Selbsteinschätzungen gemessen wurden.

1.3.2 Beobachtung

Insbesondere die Forschung zur Lehrer-Schüler-Interaktion erfordert oft die Beobachtung konkreten Verhaltens in konkreten Situationen. Das Spektrum solcher Verhaltensbeobachtungen reicht von echten experimentellen Designs, bei denen das *treatment* im psychologischen Labor erfolgt (bspw. die „Bobo-Doll Studie“ zum Modelllernen von Bandura und Kollegen 1961 (Bandura u.a. 1993)), bis hin zu Beobachtungen in alltäglichen Handlungskontexten wie einem Klassenzimmer (etwa die Studien von ter Laak u.a. (2001) über das Beurteilungsverhalten von Lehrern oder von Blatchford (2003) über die Abhängigkeit des Lehrerverhaltens von der Klassengröße). Bei einer Verhaltensbeobachtung nach dem hypothetiko-deduktiven Modell werden theoretische Konzepte mit Hilfe von Beobachtungsinventaren operationalisiert, die nach ähnlichen Regeln konstruiert werden müssen wie standardisierte Fragebögen: hierbei müssen unter Beachtung der Gütekriterien *Objektivität*, *Reliabilität* und *Validität* trennscharfe und erschöpfende Beobachtungskategorien konstruiert werden. Dabei erfordert insbesondere die präzise Ausformulierung der Kategorien sowie der Bedingungen, die festlegen, wann ein beobachtetes Verhalten einer Kategorie zugeordnet werden soll, die Festlegung detaillierter Regeln, die sicherstellen, dass die Beobachter dasselbe Verhalten in genau derselben (vom Forscher intendierten) Weise beurteilen (vgl. Kern 1997, S. 34ff.). Zwar werden auch in Fragebögen Probanden manchmal um die Beurteilung von Beobachtungen gebeten – hier ist es aber unproblematisch (manchmal sogar erwünscht), dass sich Urteile unterscheiden. So ist bspw. die mit Fragebögen arbeitende klassische Einstellungsforschung ja gerade an (subjektiven) Unterschieden zwischen menschlichen Urteilen und Bewertungen interessiert, während man die Beobachter bei der Verwendung eines Beobachtungsinventars trainieren muss, Dinge einheitlich (objektiv) zu beurteilen. Zudem müssen Beobachtungsschemata einfach konstruiert sein und dürfen v.a. nicht zu viele Kategorien enthalten, damit

die Beobachter jederzeit für ein spontan auftretendes Verhalten die entsprechenden Codes auf dem Kodierbogen finden. Die besonderen Probleme bei der Konstruktion von Beobachtungsinventaren für die Analyse der Lehrer-Schüler-Interaktion werden in dem Erfahrungsbericht von Fish und Dane (2000) dargestellt.

2 Qualitative Forschungsmethoden

Qualitative Forschung unterscheidet sich in drei wesentlichen Aspekten von quantitativen Methoden:

1. Das *Ziel des Forschungsprozesses* ist nicht die Testung von präzise formulierten Theorien und Hypothesen – vielmehr werden in der qualitativen Forschung auf der Grundlage allgemeiner theoretischer Vorannahmen konkrete Kategorien und theoretische Annahmen erst unter Zuhilfenahme von empirischen Daten entwickelt.
2. Die *Daten* werden nicht mit besonderen Messinstrumenten standardisiert erhoben, sondern durch „offene Verfahren“, deren Ergebnis wenig strukturierte „Textdaten“, Bilder oder Videoaufzeichnungen sind.
3. Diese Daten werden nicht mit Hilfe statistischer Methoden, sondern durch *interpretative* und *kategorienbildende* Verfahren ausgewertet.

Im Folgenden wollen wir zuerst methodologische Grundlagen qualitativer Forschung skizzieren, um dann zentrale Verfahren qualitativer Datenerhebung und Datenanalyse kurz darzustellen.

2.1 Methodologische Grundlagen

Die qualitative Forschungstradition verfügt nicht über ein ähnlich einheitliches Modell des Forschungshandelns wie das HD-Modell – unter dem Etikett „qualitative Methoden“ werden vielmehr viele unterschiedliche Methoden der Datenerhebung und -auswertung mit verschiedenen theoretischen Wurzeln zusammengefasst. Dennoch gibt es bestimmte Gemeinsamkeiten, insbesondere, was die methodologische Begründung des Forschungshandelns betrifft: demnach ist der Gegenstand der Sozialwissenschaften und der Psychologie, das menschliche Erleben, Denken und (soziale) Handeln, durch Eigenschaften gekennzeichnet, die die durchgehende Anwendung hypothetiko-deduktiver Forschungsstrategien nicht sinnvoll erscheinen lassen. Qualitative Ansätze betonen, dass Menschen sich in

der Welt orientieren und handeln aufgrund der *subjektiven Bedeutungen*, die die Dinge in ihrer Umgebung und das Verhalten ihrer Mitmenschen für sie haben. Solche Prozesse der Bedeutungszuschreibung werden auch beeinflusst durch gesellschaftliche Regeln und Verhaltensvorschriften. Solche Regeln gelten jedoch im Gegensatz zu Naturgesetzen nicht universell, wie qualitativ ausgerichtete, „interpretative“ sozialwissenschaftliche Theorieansätze (bspw. der symbolische Interaktionismus (Mead 1934) oder die soziologische Phänomenologie (Schütz 1974)) deutlich machen – sie sind vielmehr oft mehrdeutig und müssen von den Handelnden *situationsgebunden interpretiert* werden. Hierdurch können soziale Regeln auch Neuinterpretationen erfahren, die sie dauerhaft verändern. Dass *Subjektivität*, *Situativität* und *Flexibilität* menschlichen Handelns und Erlebens nicht durch starre Gesetze determiniert, sondern durch (prinzipiell veränderbare) Regeln und Strukturen beeinflusst werden, hat bedeutsame methodologische Konsequenzen: Forscher besitzen oft nicht genügend Kenntnisse über den untersuchten Gegenstandsbereich, um zu Beginn des Forschungsprozesses fundierte Hypothesen aufzustellen, zu operationalisieren und somit im Sinne des HD-Modells zu überprüfen (vgl. Gerdes 1979, S. 5). Qualitative Forschungsmethoden eröffnen Wege, mit denen man Zugang finden kann zu den subjektiven Sichtweisen, den Handlungsorientierungen und dem Alltagswissen der Akteure im Feld sowie zu gruppen- und kulturbezogenen Normen, Werten und Handlungspraktiken.

Während man also in einem rein quantitativen Forschungsprojekt zur Lehrer-Schüler-Interaktion eine kausale Hypothese („Lehrerrückmeldung erhöht die Schülermotivation“) aufstellen, operationalisieren und prüfen würde, versucht man in der qualitativ orientierten „interpretativen Unterrichtsforschung“ (vgl. Krummheuer & Naujok 1999) Fragen zu beantworten wie: „*Wie erleben und interpretieren Schüler bestimmte Verhaltensweisen ihrer Lehrer?*“ oder „*Welche Gründe haben Schüler überhaupt, sich am Unterricht zu beteiligen?*“. Qualitative Forschung dient dabei einer systematischen Exploration wenig bekannter Gegenstandsbereiche und einer methodisch kontrollierten, empirisch begründeten Entwicklung von theoretischen Aussagen (vgl. Kelle 1997, S. 21).

2.2 Fallauswahl und Fallkontrastierung

Ebenso wie in quantitativen Studien sind auch in der qualitativen Forschung Fragen nach der Auswahl von Untersuchungseinheiten von zentraler Bedeutung. Mit Hilfe qualitativer Methoden kann allerdings in der Regel immer nur eine kleine Anzahl von Fällen untersucht werden (ein qualitatives Interview dauert bspw. in der Regel wesentlich länger als die Beantwortung eines Fragebogens und seine

Auswertung verlangt deutlich mehr Zeit). *Statistische Repräsentativität* wird hierbei normalerweise nicht erreicht und auch nicht angestrebt. Nun ist eine solche Repräsentativität auch in der quantitativen Forschung nicht Selbstzweck, wie bereits der Abschnitt über Stichprobenziehung deutlich gemacht hat – die Berücksichtigung der regulativen Idee der Repräsentativität soll vielmehr verhindern helfen, dass eine Stichprobe bezogen auf theoretisch relevante Merkmale verzerrt oder fehlerhaft ist.

Bei der qualitativen Fallauswahl versucht man ebenfalls, ein verzerrtes Bild des Gegenstandsbereichs zu vermeiden, indem man sich darum bemüht, die für die Forschungsfragestellung relevante Heterogenität und Varianz der Fälle im Untersuchungsfeld möglichst gut zu erfassen. Kriterium ist dabei nicht mehr die Abbildung einer bestimmten Verteilung – vielmehr kann es sogar sinnvoll sein, auch in einem sehr kleinen qualitativen Sample Extremfälle besonders zu berücksichtigen. Ein einzelnes Mädchen in einer Berufsschulklasse von männlichen Kfz-Mechanikern wäre bspw. so ein interessanter Extremfall, der Informationen über die Geschlechterproblematik in dieser Berufsausbildung liefern und deshalb auch in einer kleinen qualitativen Studie mit fünf Interviews Berücksichtigung finden kann.

Insgesamt lassen sich drei Arten der qualitativen Fallauswahl unterscheiden: die *Suche nach Gegenbeispielen*, das „*theoretical sampling*“ nach Glaser und Strauss (1967) sowie die *Konstruktion qualitativer Stichprobenpläne*:

1. Die *Suche nach Gegenbeispielen* ist ein schon lange bewährtes Verfahren qualitativer Fallauswahl (Znaniecki 1934; Lindesmith 1947; Cressey 1971). Die Analyse eines ersten Falls führt dabei zu einer (kausalen) Hypothese, die dann die Suche nach Gegenbeispielen („*crucial cases*“) anregt, also nach Fällen, die möglicherweise empirische Gegenevidenz enthalten. Wird ein solcher Fall gefunden, muss die Hypothese modifiziert oder die Fragestellung umformuliert werden. Dieser Prozess der sukzessiven Modifikation und Prüfung der Hypothese wird solange fortgeführt, bis keine Gegenbeispiele mehr gefunden werden können. Dieses Verfahren, das in vieler Hinsicht dem klassischen HD-Modell ähnelt (auch wenn es nicht nur der Bestätigung oder Widerlegung von Theorien, sondern vor allem deren systematischer Weiterentwicklung dient) hat allerdings den Nachteil, dass bereits am Anfang des Forschungsprozesses eine sehr präzise (da widerlegbare) Hypothese formuliert werden muss.
2. Bei dem Verfahren des *theoretical sampling* (Glaser & Strauss 1967) ist es leichter möglich, Hypothesen erst während des Forschungsprozesses zu entwickeln. Hierbei werden bei der Analyse der ersten Fälle allgemeine Kategorien oder Merkmale (nicht bereits spezifische Hypothesen) gefunden, die die

- Auswahl weiterer Fälle anleiten, die hinsichtlich eines oder mehrerer Merkmale große Ähnlichkeiten oder große Unterschiede zum Vorgänger aufweisen. Gewinnt ein Unterrichtsforscher also bspw. durch die ersten Interviews den Eindruck, dass die Geschlechterproblematik in seinem Untersuchungsfeld eine große Rolle spielt, wird er in späteren Interviews systematisch das Geschlecht seiner Interviewpartner berücksichtigen. Die Auswahl möglichst ähnlicher Fälle („*minimization*“) kann die theoretische Relevanz einer bestimmten Kategorie erhärten, die Auswahl anders gelagerter Fälle („*maximization*“) die Heterogenität im Untersuchungsfeld abbilden.
3. Das dritte Verfahren qualitativer Fallauswahl, die Konstruktion qualitativer Stichprobenpläne vor der Erhebung, erfordert eine Kenntnis theoretisch relevanter Merkmale bereits zu Beginn des Forschungsprozesses. Bei der Aufstellung qualitativer Stichprobenpläne werden dann Fälle ausgewählt, die eine bestimmte Kombination forschungsrelevanter Merkmale aufweisen: so kontrastieren etwa ter Laak und Kollegen (2001) Schülergruppen danach, ob es sich um „normale“ oder „schwierige“ Schüler handelt; oder Woods und Jeffrey (2002) kontrastieren Schulen, die sie in ihr Untersuchungsdesign aufnehmen, nach deren Größe, Lage und der Menge der zu Beginn eines Schuljahres aufgenommenen Schüler (vgl. Abb 1).

		Trägerschaft	
		Staatlich	privat
Schülerzahl (gesamt)	≥ 1000	Stadt/Land	Stadt/Land
	< 1000	Stadt/Land	Stadt/Land

Abbildung 1 Differenzierung nach drei Merkmalen (Schulträger, Schulgröße und Stadt-Land-Unterschied), angelehnt an Woods & Jeffrey (2002)

2.3 Methoden qualitativer Datenerhebung

Qualitative Daten sind unstandardisierte Daten, zumeist freie Texte, die der Forscher selber oder seine Informanten und Interviewpartner schriftlich oder mündlich produzieren. Ähnlich den standardisierten Daten der quantitativen Forschung werden qualitative Daten zumeist auf zwei Wegen, durch *Befragung* und *Beobachtung* gewonnen (zu seltener eingesetzten Methoden wie der qualitativen Dokumentenanalyse vgl. Flick 2011, S. 321ff.), ggf. auf Tonträger und/oder visuell aufgezeichnet und verschriftlicht.

2.3.1 Offene Interviews

Zentrales Merkmal eines qualitativen Interviews ist seine „Offenheit“, die dem Befragten die Möglichkeit lässt, eigene Wahrnehmungen, Sichtweisen und Orientierungen zu entfalten, ohne an feste Frageschemata und Antwortvorgaben gebunden zu sein. Es existiert eine große Vielzahl von (oft für besondere Forschungskontexte entwickelten) Interviewformen, die manchmal schwer zu unterscheiden sind. Eine Möglichkeit zur Systematisierung qualitativer Interviews bietet deren *Grad an Strukturiertheit*, der von unvorbereiteten informellen Gesprächen im Lauf eines Feldaufenthalts bis hin zu strukturierten Interviewtechniken reicht wie dem „*Teacher Relationship Interview*“, bei dem Fragen und Nachfragen sowie deren Reihenfolge fest vorgeschrieben sind, der Befragte aber frei antworten kann (Stuhlman & Pianta 2002).

Hopf (2013) nennt drei Kriterien für die Strukturiertheit qualitativer Interviews:

1. Durch die *Art der gestellten Fragen* können den Interviewpartnern unterschiedliche starke Vorgaben gemacht werden: so kann ein kurzer „Erzählstimulus“ gegeben werden, wie beim „narrativen Interview“ (s.u.), auf den hin der Befragte das Gespräch möglichst selbst in die Hand nehmen soll, oder es wird ein „Leitfaden“ formuliert, der zentrale Themengebiete enthält (wobei das Interviewer flexibel bleibt in der Reihenfolge, in der er die Themen anspricht). In manchen qualitativen Interviews werden auch in einer bestimmten Reihenfolge vorformulierte Fragen ohne Antwortvorgaben gestellt, die dann offen beantwortet werden sollen.
2. Ein weiteres Unterscheidungskriterium ist die *thematische Eingrenzung des Interviews*, Thematiken qualitativer Interviews können von ganzen Lebensläufen (wie in „biographischen Interviews“) bis zu eng umgrenzten Themenfeldern (etwa in „Experteninterviews“) reichen.
3. Die *Rolle des Interviewers* kann vom aktiven Zuhören (mit gelegentlich zustimmende Lauten und Gesten) bis hin zu einem aktiven Interviewerverhalten reichen, bei dem auch konfrontative Fragen gestellt werden (Scheele & Groeben 1988).

Eine Sonderstellung nimmt das *narrative Interview* ein, bei dem die Rolle des Interviewers sehr stark eingeschränkt wird: außer einem einzelnen Erzählanreiz ganz zu Anfang beschränkt sich der Interviewer weitgehend auf die Rolle des Zuhörers und kommt erst in der Schlusssequenz auf bestimmte Bereiche der Erzählung zurück, für die er Nachfragen formuliert. Das Anwendungsgebiet des klassischen narrativen Interviews liegt im Wesentlichen in der Biographie- und

Lebenslaufforschung, wo Interviewpartner ihr ganzes Leben (oder wesentliche Abschnitte davon) erzählen und reflektieren sollen. Diese Interviewform wurde inspiriert durch eine sozialwissenschaftliche Erzähltheorie, der zufolge selbst berichtete Ereignisse und Geschehnisse die Orientierungsstrukturen des faktischen Handelns weit genauer reflektieren als Meinungen, Argumente oder zusammenfassende Berichte (vgl. Schütze 1977). Dies ist auch der Grund, warum Elemente narrativer Interviews (nämlich Anreize für kurze Erzählungen konkreter Begebenheiten) häufig auch in andere qualitative Interviewformen¹⁵ integriert werden. So arbeiten etwa Bulterman-Bos und Kollegen (2003) in den Interviews, die sie mit Lehrern über deren Einstellungen gegenüber nationalen Vergleichstests und Leistungsstandards führen, intensiv mit Erzählanreizen, um die Lehrerperspektive besser in den Blick zu bekommen, ohne klassische narrative Interviews im eigentlichen Sinne durchzuführen.

In der psychologischen Forschung sind strukturierte und thematisch fokussierte qualitative Interviewformen beliebt, etwa *Struktur- oder Dilemma-Interviews* (vgl. Colby & Kohlberg 1987), bei denen die Befragten spezifische Aufgaben lösen, indem sie etwa ein bestimmtes moralisches Dilemma bewerten. Der Begriff des „fokussierten Interviews“ wiederum geht zurück auf medienanalytische Forschungen von Merton und Kendall (1979): hierbei sollen die Interviewpartner bestimmte Reize bewerten (in der klassischen Form waren dies Filme oder Fernsehbeiträge). In ähnlicher Weise werden bei focus groups ganze Gruppen zu bestimmten Themen interviewt (vgl. Flick 2011, Kapitel 15; oder Ernst 2006), Neben diesen klassischen Interviewformen existieren heute zahlreiche Mischformen, von denen sich die meisten als „Leitfadeninterviews“ bezeichnen lassen. Einer der häufigsten Interviewerfehler bei diesen Interviews ist mangelnde Geduld: so werden Leitfäden manchmal wie standardisierte Fragebögen genutzt – dort, wo eigentlich intensive Nachfragen oder Erzählanreize zur Vertiefung („*Können Sie hierzu ein Beispiel erzählen?*“) angebracht wären, vollziehen Interviewer Themenwechsel, um ihren Leitfaden „abzuarbeiten“ (vgl. Hopf 1978).

2.3.2 Teilnehmende Beobachtung

Die Teilnehmende Beobachtung, eine Methode der ethnologischen Feldforschung, bei der sich Wissenschaftler zum Teil viele Monate im Feld aufhalten und intensiven Kontakt zu Mitgliedern einer fremden Kultur aufbauen (Malinowski 1979),

15 Das führt allerdings auch zu dem weit verbreiteten Missverständnis, dass Mischformen wie Leitfadeninterviews mit geringen narrativen Anteilen fälschlicherweise als narratives Interview bezeichnet werden.

gelangte in die Sozialwissenschaften durch die Untersuchungen der sog. „Chicago School“, die in der ersten Hälfte des 20. Jahrhunderts die Lebenswelten städtischer Subkulturen erforschte. Heute ist auch der Begriff der „Ethnographie“ gebräuchlich, der manchmal unzulässigerweise mit qualitativer Sozialforschung überhaupt gleichgesetzt wird (Lüders 2013). Methodologisch zentrale (und teilweise äußerst kontrovers diskutierte) Aspekte ethnographischer Forschung in den Sozialwissenschaften betreffen die *Rolle des teilnehmenden Beobachters im Feld* sowie die *Vertrauenswürdigkeit der ethnographischen Daten*.

Die möglichen Rollen teilnehmender Beobachter lassen sich durch verschiedene Gegensatzpaare systematisieren (vgl. Friedrichs 1999, S. 272ff.), insbesondere

- zwischen *offener* und *verdeckter Beobachtung*, also betreffend das Ausmaß, in dem die Akteure im Feld Kenntnis von der Rolle des Beobachters haben,
- und zwischen *teilnehmender* und *nicht-teilnehmender Beobachtung*, also bezüglich des Grades, mit dem der Beobachter an den Alltagsinteraktionen des Feldes selber teilnimmt¹⁶.

Als ein Vorteil des Verfahrens wird oft dessen Offenheit angesehen, von der man annimmt, sie fördere die Unvoreingenommenheit des Feldforschers. Das Konzept des wissenschaftlich neutralen, objektiven, der fremden Kultur gegenüber aufgeschlossenen bis wohlwollend gegenüberstehenden Ethnographen ist allerdings umstritten, denn die Frage, wie stark Voreingenommenheiten der Forscher die Untersuchungsergebnisse beeinflussen und auch verfälschen, muss immer gestellt werden. Mögliche Fehldeutungen, die durch eine unangemessene Übertragung der Normen und Werte des Beobachters auf eine fremde Kultur entstehen, haben schon etliche Male zu heftigen Kontroversen geführt: so wurde der Kulturanthropologin Margaret Mead vorgeworfen, in ihrer berühmten Studie über die psychosexuelle Entwicklung von Jugendlichen auf Samoa seien ihr krasse Fehldeutungen unterlaufen (Shankman 2000). Fehlwahrnehmungen können manchmal auch entstehen durch eine starke Identifikation der Forscher mit ihrem Feld (in der klassischen Literatur als „going native“ bezeichnet), die durch den ethnographischen Forschungsstil gefördert wird, dessen Ziel ja darin besteht, sich mit den Perspektiven der Akteure im Feld vertraut zu machen. Arbeiten aus den 1980er Jahren haben zudem gezeigt, wie die Vorstellung eines neutralen ethnographischen Beobachters durch rhetorische Mittel erzeugt werden kann (van Maanen 1988).

16 Gold (1958) unterscheidet den vollständigen Teilnehmer, den Teilnehmer als Beobachter, den Beobachter als Teilnehmer und den vollständigen Beobachter.

Angesichts dieser Probleme ist die Qualität der aus dem Feld mitgebrachten Daten – also der Feldprotokolle und Aufzeichnungen der Forscher – von herausragender Bedeutung. In der entsprechenden Literatur haben deshalb Fragen der Protokollierung und Aufzeichnung von Ereignissen eine wichtige Bedeutung (Hammersley & Atkinson 1983, S. 144ff.). Das zentrale Ziel der Protokollierung muss in jedem Fall sein, dass die Ergebnisse des Feldaufenthaltes für den Rezipienten nachvollziehbar werden.

2.4 Qualitative Datenanalyse

Die Auswertung qualitativen Datenmaterials, das manchmal viele hundert Seiten transkribierte Interviewtexte oder Feldprotokolle umfasst, stellt besondere Herausforderungen, weil hier – anders als bei der statistischen Analyse quantitativer Daten – keine standardisierten Verfahren existieren, die den Forschern einen schnellen Überblick über die Daten (etwa in Form von statistischen Kennziffern) ermöglichen. Der folgende Überblick über Strategien der Textinterpretation und der Kategorienbildung soll eine allererste Orientierung bieten über Möglichkeiten zur Strukturierung und Auswertung qualitativen Materials, ersetzt aber keinesfalls das vertiefte Studium entsprechender Literatur vor der Durchführung eines Forschungsvorhabens.

2.4.1 Grundprinzipien der Textinterpretation

Verfahren zur Erschließung der Bedeutung von Texten werden unter dem Oberbegriff „Hermeneutik“ zusammengefasst, ein Begriff, der im 19. Jahrhundert von dem Theologen Schleiermacher geprägt und dann durch den Philosophen Dilthey in die Kultur- und Sozialwissenschaften eingeführt wurde (Dilthey 1900/1924). Auf Schleiermacher geht auch das Modell des „hermeneutischen Zirkels“ zurück, das sich gut von der Analyse historischer Texte auf die Interpretation sozialen Handelns übertragen lässt.

Die Bedeutung einer Textstelle oder Sinn einer Handlung einer Person in einer sozialen Situation (z.B. die Mitarbeit eines Schülers im Unterricht) erschließt sich demnach immer nur aus ihrem Kontext (bei einer Bibelstelle etwa die gesamte Bibel, bei einem Schülerhandeln das gesamte Unterrichtsgeschehen). Weil der Kontext aber selber wiederum aus einzelnen Bausteinen (nämlich vielen Textstellen oder Handlungen) besteht, steckt der Interpret, der beides nicht versteht, in einem Zirkel. Dieser Zirkel kann erst durchbrochen werden, wenn man mit einem (ggf. nur rudimentären) Vorverständnis an das Material herangeht. Beim sozialwissen-

schaftlichen Verstehen kann man zum Beispiel im einfachsten Fall auf allgemeine Wissensbestände zurückgreifen, zu denen man als Gesellschaftsmitglied Zugang hat: Forscher wissen in der Regel zumindest, was Lehrer, Schüler und Schulunterricht sind. Dieses alltagsweltliche Vorverständnis kann und sollte natürlich ergänzt werden durch das Wissen über didaktische Theorien und über Unterrichtskonzepte von Lehrern. Wissenschaftliche Beobachtung und Interpretation ist immer „theoriebeladen“ (Hanson 1958): so wird jemand mit einem bestimmten Vorwissen in der Unterrichtsforschung dort ein innovatives Lehrkonzept sehen, wo ein Beobachter ohne entsprechendes Vorwissen nur eine „Horde schreiender Kinder“ wahrnimmt.

Das Entscheidende ist nun, das Vorwissen so offen zu halten, dass es durch neue Textstellen und Beobachtungen erweitert und verändert werden kann. Dann wird der hermeneutische Zirkel zur „hermeneutischen Spirale“, die auf einen (imaginären) Punkt zuläuft – der „wirklichen“ Bedeutung der Textstelle oder Handlung. Im Idealfall tastet sich der Interpret, ausgehend von seinem Vorwissen immer weiter in Richtung dieses Punktes vor. Das praktische Vorgehen hierzu wurde vor allem in Arbeiten zu „sequenzanalytischen Methoden“ beschrieben (vgl. etwa Oevermann u.a. 1980), die für die Interpretation von Interaktionsprotokollen (wie sie bei der Videoaufzeichnung von Unterrichtssituationen anfallen können) entwickelt wurden. Hierbei werden in einer Forschergruppe zu jeder Textstelle (bspw. jeder Äußerung einer Person in einer Interaktionssituation) mögliche Deutungshypothesen oder „Lesarten“ formuliert, d.h. es werden gedankenexperimentell mögliche Kontexte konstruiert, in der genau diese Textstelle bzw. Äußerung Sinn macht. Diese Hypothesen werden bei dem weiteren Durchgang durch das Material (das beim sequenzanalytischen Vorgehen in der exakten Reihenfolge der Interaktionssequenzen durchgegangen wird) weiter erhärtet oder sie erweisen sich als nicht haltbar.

2.4.2 Die empirisch begründete Konstruktion von Kategorien und Typen

Abhängig von der Art des Datenmaterials können hermeneutische Analysen in der oben beschriebenen Art sehr aufwändig sein. Aber das qualitative Methodenreservoir enthält nicht nur Verfahren, um kurze Texte extensiv auszudeuten, sondern auch Methoden, mit denen umfangreiche Datenmengen durch theoretische Kategorien auf den „Punkt gebracht“ werden können.

Die grundlegende Operation der *Kategorienbildung*, die den Ausgangspunkt für eine empirisch begründete Typenbildung und Theoriekonstruktion bildet, wird als „Kodierung“ bezeichnet – eine Zuordnung von Kategorien zu Daten (meist

Textpassagen), wobei sich eine *offene Kodierung* von der *Kodierung anhand eines vorbereiteten Kategorienschemas* unterscheiden lässt:

- Bei der *offenen Kodierung*, vorgeschlagen von Glaser und Strauss (1967) als Verfahren zur Entwicklung empirisch begründeter Theorie („*grounded theory*“), wird das Datenmaterial sequentiell (Zeile für Zeile bzw. Satz für Satz) durchgearbeitet, wobei passende Begriffe gesucht werden, die wichtige Aspekte bezeichnen. Oft sind dies sog. „*in-vivo Kodes*“, Begriffe, die die Akteure im Feld verwenden. Aber auch hier gilt, dass der Untersucher sein alltagsweltliches und theoretisches Vorwissen einsetzen muss, um passende Benennungen zu finden. Glaser und Strauss nennen dies „theoretische Sensibilität“: die Fähigkeit, empirische Daten in theoretische Konzepte zu fassen. Allerdings darf hierbei die wesentliche Funktion qualitativer Forschung, die Exploration bislang unbekannter Sachverhalte, nicht außer Kraft gesetzt werden, indem den Daten unpassende Kategorien aufgezwungen werden. Eine theoretisch sensibilisierte Kodierung erfordert dabei, dass der Untersucher über einen sehr großen Wissensfundus an theoretischen Konzepten – Glaser (1978) nennt dies „theoretische Kodes“ – verfügen muss, aus denen er die dem Material angemessenen auswählen kann – insbesondere für Anfänger stellt dies eine große Herausforderung dar.
- Bei der *Kodierung anhand eines vorbereiteten Kategorienschemas* wird die Suche nach Begriffen durch die Konstruktion eines heuristischen Rahmens von Kategorien erleichtert. Will man dabei aber das Potential der qualitativen Methode zur Entdeckung ausnutzen (und nicht zu einem hypothetiko-deduktiven Vorgehen übergehen), dürfen diese Kategorien nicht zu spezifisch, sondern müssen hinreichend offen sein (vgl. hierzu Kelle & Kluge 2010, S. 56ff.). Diese Bedingung erfüllen bspw. viele alltagsnahe Begriffe, allgemeine „thematische Kategorien“ (bspw. die Themen eines Leitfadens) oder manche sehr allgemeinen Theoriekonzepte, denen sich sehr verschiedene empirische Sachverhalte zuordnen lassen – ein Beispiel hierfür bilden handlungstheoretische Begriffe wie „situative Bedingungen“, „Intentionen“ oder „Handlungskonsequenzen“, die sich auf ganz verschiedene Handlungen beziehen lassen.

Um Subkategorien für die bereits gefundenen Kategorien zu definieren und um Gemeinsamkeiten zwischen Kategorien zu entdecken, die die Bildung von Oberkategorien anregen können, werden dann Textstellen miteinander verglichen (wobei der Einsatz spezieller Software sinnvoll ist, vgl. Kelle 2008; Kelle 2013). Das sich bildende Kategoriensystem soll hierbei zunehmend Struktur erhalten, einerseits durch eine Festlegung von „Kernkategorien“ und andererseits, indem sozialwis-

senschaftliche und psychologische Theorien herangezogen werden, um die Kodes und ihre Beziehungen untereinander zu ordnen. Detailliertere Beschreibungen der Kategorien- und Typenbildung finden sich bei Strauss und Corbin (2010), Kelle und Kluge (2010) oder bei Kuckartz (2010).

3 Die Verbindung qualitativer und quantitativer Methoden

Quantitativen und qualitativen Methoden liegen verschiedene Modelle des Forschungsprozesses zugrunde, wobei beide Methodentraditionen differierende Ziele verfolgen: quantitative Forschung soll der Prüfung von vorab formulierten Hypothesen dienen, dabei wird statistische Generalisierbarkeit angestrebt und ein besonderes Augenmerk auf die Objektivität und Zuverlässigkeit der Daten und Prozeduren gelegt. Qualitative Forschung dient der Erkundung subjektiver Sichtweisen von Akteuren und der Entdeckung bislang unbekannter kultureller Regelbestände, was ein offenes und exploratives Vorgehen erfordert. Aufgrund dieser unterschiedlichen Ziele lassen sich verschiedene Kriterien für gute Forschung definieren, die miteinander in Konflikt geraten können: die Kontrolle von Störvariablen im Experiment kann dazu führen, dass eine hochgradig lebensferne Situation geschaffen wird und quantitative Forscher können mit ihren präzisen Hypothesen an den Relevanzsetzungen und Handlungsorientierungen der Befragten im Feld völlig vorbeigehen. Qualitative Feldforscher wiederum stehen in der Gefahr, dass sie aufgrund zu kleiner Fallzahlen und der mangelnden Objektivität ihrer Feldprotokolle empirische Phänomene in ihrer Bedeutung falsch einschätzen.

Diese Probleme haben zu einer langen Kontroverse zwischen Vertretern beider Methodentraditionen über den „richtigen Weg“ in der Forschung geführt. Inzwischen mehren sich aber die Stimmen, die betonen, dass diese Unterschiede zwar bedeutsam sind, jedoch eine pragmatische Forschungspraxis nicht behindern sollten (vgl. etwa Hammersley 1995; Seale 2000). Eine wachsende Zahl von Sozialwissenschaftlern betont heute sogar den Nutzen von „*Mixed-Methods-Designs*“ (vgl. Tashakkori & Teddlie 2010; Kelle 2008): durch eine Kombination qualitativer und quantitativer Forschung können nämlich die Schwächen der einen Methodentradition durch Verfahren der anderen Tradition ausgeglichen werden. So können etwa theoretische Aussagen, die anhand qualitativer Daten entwickelt wurden, anhand umfangreicher Stichproben quantitativ erhärtet werden, quantitative Daten können genutzt werden, um die qualitative Fallauswahl zu unterstützen oder qualitative Interviews können eingesetzt werden, um Erklärungen für verwirrende und schwer verständliche statistische Befunde zu finden. Diese und andere Möglich-

keiten der Methodenkombination werden heute zunehmend in der Schulforschung eingesetzt und erprobt.

Abschließend möchten wir daran erinnern, dass die hier vorgestellten Verfahren, Strategien und Techniken nur einen beschränkten Ausschnitt aus der Mannigfaltigkeit sozialwissenschaftlicher Methoden repräsentieren. Über die Brauchbarkeit einer Methode sollte dabei immer konkret, das heißt vor dem Hintergrund eines bestimmten Forschungsproblems und einer spezifischen Fragestellung entschieden werden.

Literatur

- Bandura, A., Ross, D. & Ross, S. (1993): Imitation of film-mediated aggressive models. *Journal of Abnormal Social Psychology*, 66, 3-11.
- Blankemeyer, M., Flannery, D. & Vazsonyi A.T. (2002): The role of aggression and social competence in childrens' perceptions of the child-teacher relationship. *Psychology in the Schools*, 39(3), 293-304.
- Blatchford, P. (2003): A systematic observational study of teachers' and pupils' behaviour in large and small classes. *Learning and Instruction*, 13, 569-595.
- Boekaerts, M. (1999): Self-regulated learning: Where we are today. *International Journal of Educational Research*, 31, 445-457.
- Bortz, J. & Schuster, C. (2010): *Statistik für Human- und Sozialwissenschaftler*. (7. Aufl.). Berlin, Heidelberg: Springer.
- Bühner, M. (2011): *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.
- Bulterman-Bos, J., Verloop, N., Terwel, J. & Wardekker, W. (2003): Reconciling the pedagogical goal and the measurement goal of evaluation: The perspectives of teachers in the context of national standards. *Teachers College Record*, 105(3), 334-374.
- Campbell, D.T. & Fiske, D. (1959): Convergent and discriminant validation by the multi-trait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Campbell, D.T. & Stanley, J.C. (1963): *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
- Colby, A. & Kohlberg, L. (1987): *The measurement of moral judgment*. Cambridge, N.Y.: Cambridge University Press.
- Cressey, D.R. (1971): *Other people's money: A study in the social psychology of embezzlement*. Belmont, Calif.: Wadsworth Pub. Co.
- Deutsches PISA – Konsortium. *Zur Stichprobenziehung innerhalb der PISA-Erweiterung*. Verfügbar unter: <http://www.mpib-berlin.mpg.de/pisa/Stichprobe.pdf> [5.11.2015].
- Diekmann, A. (2008): *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (19. Aufl.). Reinbek: Rowohlt.
- Dilthey, W. (1900/1924): *Die Entstehung der Hermeneutik. Gesammelte Schriften. Die geistige Welt. Einleitung in die Philosophie des Lebens. Bd. 5*. Leipzig.
- Engel, U. & Schmidt, B.O. (2014): Unit- und Item-Nonresponse. In: Baur, N. & Blasius, J. (Hrsg.): *Handbuch Methoden der empirischen Sozialforschung* (S. 331-348). Wiesbaden: Springer.

- Ernst, S. (2006): Die Evaluation von Qualität – Möglichkeiten und Grenzen von Gruppendiskussionsverfahren. In: Flick, U. (Hrsg.): *Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen* (S.183-213). Reinbek: Rowohlt.
- Fish, M. & Dane, E. (2000): The classroom systems observation scale: Development of an instrument to assess classroom using a systems perspective. *Learning Environment Research*, 3, 67-92.
- Fisher, R.A. (1935): *The design of experiments*. Edinburgh, London: Oliver and Boyde.
- Flick, U. (2011): *Qualitative Sozialforschung: Eine Einführung* (4. Aufl.). Reinbek: Rowohlt.
- Flick, U. (2011a): *Triangulation: Eine Einführung*. Wiesbaden: VS Verlag.
- Fraser, B.J. (1990): *Individualised classroom environment questionnaire*. Hawthorn Victoria: Australian Council for Educational Research ACER.
- Friedrichs, J. (1999): *Methoden empirischer Sozialforschung* (15. Aufl.). Opladen: Westdt. Verl.
- Gerdes, K. (1979): *Explorative Sozialforschung: Einführende Beiträge aus "Natural sociology" und Feldforschung in den USA*. Stuttgart: Enke.
- Gläser-Zikuda, M., Seidel, T., Rohlf, C., Gröschner, A. & Ziegelbauer, S. (2012): *Mixed Methods in der empirischen Bildungsforschung*. Münster: Waxmann.
- Glaser, B.G. (1978): *Theoretical sensitivity: Advances in the Methodology of Grounded Theory*. Mill Valley Calif.: Sociology Press.
- Glaser, B.G. & Strauss, A. L. (1967): *The Discovery of Grounded Theory: Strategies for qualitative Research*. New York: Aldine.
- Gold, A. & Souvignier, E. (2005): Prognose der Studierfähigkeit: Ergebnisse aus Längsschnittanalysen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 37(4), 214-222.
- Gold, R.L. (1958): Roles in sociological field observations. *Social Forces*, 36, 217-223.
- Groves, R.M. (2004): *Survey errors and survey costs*. Hoboken: Wiley.
- Hammersley, M. (1995): *The politics of social research*. London: Sage Publications.
- Hammersley, M. & Atkinson, P. (1983): *Ethnography: Principles in practice*. London: Tavistock.
- Hanson, N.R. (1958): *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge, NY: Cambridge Univ. Press.
- Hellmich, F. & Hoya, F. (2014): Die Rolle von Klassenklima und Rückmeldungen für Selbstkonzept, Motivation und Leseleistung bei Kindern im Grundschulunterricht – Ergebnisse aus einer empirischen Studie. In: Tillack, C., Fetzner, J. & Fischer, N. (Hrsg.): *Beziehungen in Schule und Unterricht. Teil 2: Soziokulturelle und schulische Einflüsse auf pädagogische Beziehungen* (S. 163-185). Immenhausen: Prolog-Verlag.
- Hempel, C.G. & Oppenheim, P. (1948): Studies in the Logic of Explanation. *Philosophy of science*, 15(2), 135-175.
- Holland, P. (1985): *Statistics and Causal Inference*. Princeton, New Jersey: Educational Testing Service.
- Hopf, C. (1978): Die Pseudo-Exploration – Überlegungen zur Technik qualitativer Interviews in der Sozialforschung. *Zeitschrift für Soziologie*, 7, 97-115.
- Hopf, C. (2013): Qualitative Interviews – Ein Überblick. In: Flick, U., Kardorff, E.v. & Steinke, I. (Hrsg.): *Qualitative Forschung. Ein Handbuch* (S. 349-360). Reinbek: Rowohlt.
- Hume, D. (1748/1957): *An inquiry concerning human understanding*. New York: The Liberal Arts Press.

- Kelle, U. (1997): *Empirisch begründete Theoriebildung: Zur Logik und Methodologie interpretativer Sozialforschung* (2. Aufl.). Weinheim: Dt. Studien-Verl.
- Kelle, U. (2008): *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung: Theoretische Grundlagen und methodologische Konzepte*. Wiesbaden: VS Verlag.
- Kelle, U. (2013): Computergestützte Analyse qualitativer Daten. In: Flick, U., Kardorff, E.v. & Steinke, I. (Hrsg.): *Qualitative Forschung. Ein Handbuch* (S. 485-502). Reinbek: Rowohlt.
- Kelle, U. & Erzberger, C. (2013): Qualitative und quantitative Methoden: kein Gegensatz. In: Flick, U., Kardorff, E.v. & Steinke, I. (Hrsg.): *Qualitative Forschung. Ein Handbuch* (S. 299-309). Reinbek: Rowohlt.
- Kelle, U. & Kluge, S. (2010): *Vom Einzelfall zum Typus: Fallvergleich und Fallkontrastierung in der qualitativen Sozialforschung*. Wiesbaden: VS Verlag.
- Kern, H.J. (1997): *Einzelfallforschung: Eine Einführung für Studierende und Praktiker*. Weinheim: Beltz.
- Krämer, W. (2015): *So lügt man mit Statistik*. Frankfurt, New York: Campus.
- Krummheuer, G. & Naujok, N. (1999): *Grundlagen und Beispiele interpretativer Unterrichtsforschung*. Leverkusen: Leske + Budrich.
- Kuckartz, U. (2010): *Einführung in die computergestützte Analyse qualitativer Daten*. Wiesbaden: VS Verlag.
- Kühnel, S.-M. & Krebs, D. (2014): *Statistik für die Sozialwissenschaften: Grundlagen, Methoden, Anwendungen*. Reinbek: Rowohlt.
- Lienert, G.A. (1967): *Testaufbau und Testanalyse*. Weinheim, Berlin, Basel: Beltz.
- Likert, R. (1932): A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140, 1-55.
- Lindesmith, A.R. (1947): *Opiate addiction*. Bloomington, Ind.: Principia Press.
- Lüders, C. (2013): Beobachten im Feld und Ethnographie. In: Flick, U., Kardorff, E.v. & Steinke, I. (Hrsg.): *Qualitative Forschung. Ein Handbuch* (S. 384-401). Reinbek: Rowohlt.
- Mackie, J.L. (1980): *The cement of the universe: A study of causation*. Oxford: Clarendon Press.
- Malinowski, B. (1979): *Argonauten des westlichen Pazifik: ein Bericht über Unternehmungen und Abenteuer der Eingeborenen in den Inselwelten von Melanesisch-Neuguinea*. Schriften in vier Bänden, Bd. 1. Frankfurt a. M.: Syndikat.
- Max-Planck-Institut für Bildungsforschung, B. (o.J.): *Kurzbericht. Die Effekte von Anreizen auf Motivation und Leistung in Mathematiktests*. Verfügbar unter: <https://www.mpib-berlin.mpg.de/Pisa/KurzberichtMotivation.pdf> [12.11.2015].
- Mead, G.H. (1934): *Mind, self & society*. Chicago, Ill.: Univ. of Chicago Press.
- Mellor, D. & Moore, K. (2003): The Questionnaire on Teacher Interaction: Assessing Information Transfer in Single and Multi-Teacher Environments. *Journal of classroom Interactions*, 38(2), 29-35.
- Merton, R.K. & Kendall, P.L. (1979): Das fokussierte Interview. In: Hopf, C. & Weingarten, E. (Hrsg.): *Qualitative Sozialforschung* (S. 169-204). Stuttgart: Klett-Cotta.
- Mill, J.S. (1874): *A system of logic*. New York: Harper & Brothers, Publishers.
- Oevermann, U., Allert, T. & Konau, E. (1980): Zur Logik der Interpretation von Interviewtexten. In: Heinze, T., Klusemann, H.-W. & Soeffner, H.-G. (Hrsg.): *Interpretationen*

- einer Bildungsgeschichte. Überlegungen zur sozialwissenschaftlichen Hermeneutik (S. 15-69). Bensheim: Päd.-extra-Buchverl.
- Popper, K. R. (1963/1994): *Vermutungen*. Tübingen: Mohr.
- Porst, R. (1996): *Ausschöpfungen bei sozialwissenschaftlichen Umfragen. Die Sicht der Institute*, ZUMA. Zuma-Arbeitsbericht: 96/07. Verfügbar unter: http://www.gesis.org/Publikationen/Berichte/ZUMA_Arbeitsberichte/96/96_07.pdf [12.11.2015].
- Porst, R. (2000): *Praxis der Umfrageforschung* (2. Aufl.). Stuttgart u.a.: Teubner.
- Rindermann, H. & Oubaid, V. (1999): Auswahl von Studienanfängern durch Universitäten.: Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. *Zeitschrift für Differenzielle und Diagnostische Psychologie*, 20(3), 172-191.
- Rosenthal, R. & Jacobson, L. (1974): *Pygmalion im Unterricht: Lehrererwartungen und Intelligenzentwicklung der Schüler*. Weinheim: Beltz.
- Scheele, B. & Groeben, N. (1988): *Dialog-Konsens-Methoden zur Rekonstruktion subjektiver Theorien: Die Heidelberger Struktur-Lege-Technik (SLT), konsensuale Ziel-Mittel-Argumentation und kommunikative Flussdiagramm-Beschreibung von Handlungen*. Tübingen: Francke.
- Schermelleh-Engel, K. & Schweizer, K. (2012): Multitrait-Multimethod-Analysen. In: Moosbrugger, H. & Kelava, A. (Hrsg.): *Testtheorie und Fragebogenkonstruktion* (S. 346-362). Berlin, Heidelberg: Springer.
- Schneider, W. (1996): Zum Zusammenhang zwischen Metakognition und Motivation bei Lern- und Gedächtnisvorgängen. In: Spiel, C. (Hrsg.): *Motivation und Lernen aus der Perspektive lebenslanger Entwicklung* (S. 121-133). Münster, New York, München, Berlin: Waxmann.
- Schnell, R., Hill, P.B. & Esser, E. (2013): *Methoden der empirischen Sozialforschung* (10. Aufl.). München: Oldenbourg.
- Schütz, A. (1974): *Der sinnhafte Aufbau der sozialen Welt: Eine Einleitung in die verstehende Soziologie*. Frankfurt am Main: Suhrkamp.
- Schütze, F. (1977): *Das narrative Interview in Interaktionsfeldstudien*. Universität Bielefeld: Mimeo.
- Seale, C. (2000): *The quality of qualitative research*. London: Sage.
- Shankman, P. (2000): Culture, Biology and Evolution: The Mead Freeman Controversy Revisited. *Journal of Youth and Adolescence*, 29, 539-556.
- Stevens, S.S. (1951): Mathematics, Measurement and Psychophysics. In: Stevens, S.S. (Hrsg.): *Handbook of experimental psychology* (S. 1-49). New York: Wiley.
- Strauss, A. & Corbin, J. (2010): *Grounded Theory: Grundlagen Qualitativer Sozialforschung*. Weinheim: Beltz.
- Stuhlman, M.W. & Pianta, R.C. (2002): Teachers' narratives about their relationships with children. Associations with behaviour in classrooms. *School Psychology Review*, 31(2), 148-163.
- Suppes, P. (1970): *A probabilistic theory of causality*. Amsterdam: North-Holland Pub. Co.
- Tashakkori, A. & Teddlie, C. (2010): *Sage handbook of mixed methods in social & behavioral research*. Los Angeles, Calif.: Sage.
- Tennant, G. (2004): Differential classroom interactions by ethnicity: A quantitative approach. *Emotional and Behavioural Difficulties*, 9(3), 191-204.
- ter Laak, J., Goede, M. de & Brugman, G. (2001): Teachers' judgement of pupils: Agreement and accuracy. *Social Behaviour and Personality*, 29(3), 257-270.

- van Maanen, J. (1988): *Tales of the field: On writing ethnography*. Chicago [u.a.]: Univ. of Chicago Press.
- Wasmer, M., Blohm, M., Walter, J., Scholz, E. & Jutz, R. (2014): *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (ALLBUS) 2012*: Köln: GESIS.
- Woods, P. & Jeffrey, B. (2002): The reconstruction of primary teachers' identities. *British Journal of Sociology of Education*, 23(1), 89-106.
- Woodward, J. (2001): *Causation and Manipulability*. Verfügbar unter: <http://plato.stanford.edu/archives/fall2001/entries/causation-mani/> [12.11.2015].
- Wubbels, T. & Levy, J. (1993): The questionnaire on Teacher Interaction (American version, male teacher). In: Wubbels, T. & Levy, J. (Hrsg.): *Do you know what you look like? Interpersonal relations in education* (S. 163-166). London: The Falmer Press.
- ZEIT online. (2007, 04. Dezember): *Bildung: Pisa Forscher: Geldprämien verzerren Ergebnisse nicht*. Verfügbar unter: <http://images.zeit.de/text/news/artikel/2007/12/04/2432307.xml> [6.12.2007].
- Zimmerman, B.J. (1999): Commentary: toward a cyclically interactive view of self-regulated learning. *International Journal of Educational Research*, 31, 545-551.
- Znaniecki, F. (1934): *The method of sociology*. New York: Farrar & Rinehart.

Lehrer-Schüler-Interaktion

Inhaltsfelder, Forschungsperspektiven und
methodische Zugänge

Schweer, M.K.W. (Hrsg.)

2017, XIII, 633 S. 23 Abb., Softcover

ISBN: 978-3-658-15082-2