Kelle, Udo; Metje, Brigitte

Mixed Methods in der Evaluationsforschung. Das Verhältnis zwischen Qualität und Quantität in der Wirkungsanalyse

formal überarbeitete Version der Originalveröffentlichung in:
formally revised edition of the original source in:
Knolle, Niels [Hrsg.]: Evaluationsforschung in der Musikpädagogik. Essen : Die Blaue Eule 2010, S. 9-39. - (Musikpädagogische Forschung; 31)

Musikpädagogische Forschung

Niels Knolle
(Hrsg.)

Evaluationsforschung in der Musikpädagogik

verlag DIE BLAUE EULE essen

Themenstellung: Evaluationsforschung ist zu einem bedeutsamen Zweig der Bildungsforschung geworden, die Vielfalt der Beiträge zur 31. AMPF-Tagung >Evaluationsforschung in der Musikpädagogik< macht deutlich, dass die musikpädagogische Forschung hierzu einen bedeutsamen Beitrag zu liefern in der Lage ist. So zielen die Beiträge dieses Bands darauf, die Voraussetzungen, Inhalte, Methoden und Resultate von musikunterrichtlichen Reformansätzen und Innovationen im Blick auf die mit ihnen verbundenen Ziele zu überprüfen und zu bewerten, um so zu einer Verbesserung des musikbezogenen Handelns bzw. entsprechender Lehr-Lern-Prozesse zu gelangen.

Der Herausgeber: *Niels Knolle*, geb. 1944. Arbeitsschwerpunkte: Multimedia als Instrument, Werkzeug und Thema des Musikunterrichts; Didaktik der Populären Musik; Bildungsreformen in der Musikpädagogik. Langjährige Arbeit in den Vorständen der BFG Musikpädagogik, des AMPF, der Bundesfachausschüsse >Musikpädagogik< und >Musik und Medien< des Deutschen Musikrats. 1999 - 2003 Mitherausgeber der Zeitschrift >Musik in der Schule<. Von 1996 bis 2010 Universitätsprofessor für Musikpädagogik an der Otto-von-Guericke-Universität Magdeburg.

Inhalt

Niels Knolle:

Vorwort 7

Beiträge zum Tagungsthema

Udo Kelle, Brigitte Metje:

Mixed Methods in der Evaluationsforschung. Das Verhältnis zwischen Qualität und Quantität in der Wirkungsanalyse 9

Susanne Naacke:

Schulentwicklung mit Chor- und Bläserklassen. Eine qualitative Fallstudie am „Evangelischen Gymnasium am Dom zu Brandenburg“ 41

Forschungspreis 2009 Hösbach

Jens Knigge, Anne Niessen, Anne-Katrin Jordan:

Erfassung der Kompetenz „Musik wahrnehmen und kontextualisieren“ mit Hilfe von Testaufgaben - Aufgabenentwicklung und -analyse im Projekt KoMus 81

Anne-Katrin Jordan, Andreas C. Lehmann, Jens Knigge:

Kompetenzmodellierung mit Methoden der Item-Response-Theorie (IRT) - Erste Ergebnisse der Validierung eines Modells für den Bereich „Musik wahrnehmen und kontextualisieren“ 109

Jürgen Oberschmidt:

Metaphorischer Sprachgebrauch im Unterricht - Überlegungen zur Evaluierung der Schülersprache 131

Kai Stefan Lothwesen:

Musikalisches Erleben und Lernen zwischen Musikschule und Grundschule. Methodenkritische Reflexionen am Beispiel der Evaluation des Programms „Monheimer Modell – Musikschule für alle“ 155

Dirk Bechtel:

„Wie Lehrer lieber lernen“ - Eine qualitative Studie über die Rolle von Fortbildungen aus der Sicht von Musiklehrerinnen und -lehrern 179

Eva Mödinger, Gabriele Hofmann:

Lampenfieber und Aufführungsängste bei Kindern und Jugendlichen - Erhebungen zur Selbstwahrnehmung im Rahmen musikalischer Vortragssituationen 201

Matthias Stubenvoll:

Qualität entsteht beim Lernen - Lerner integrierende Qualitätsbeurteilung beim E-Learning 211

Wibke Gütay:

Darf es noch ein bisschen mehr sein? Auswirkungen von Stimmtraining bei Chorklassenkindern 229

Freie Beiträge

Robert Lang:

Musiktheorie in musizierpraktischem Schulunterricht. Zur Effizienz basaler Harmonielehre für das Improvisieren mit Keyboards 255

Konsortium des JeKi-Forschungsschwerpunkts:

Der BMBF-Forschungsschwerpunkt zu „Jedem Kind ein Instrument“ in Nordrhein-Westfalen und Hamburg 275

Richard von Georgi, Kai Stefan Lothwesen:

Handlungskompetenzen und Studiumsmotivation von Musikstudierenden 305

UDO KELLE, BRIGITTE METJE

Mixed Methods in der Evaluationsforschung

Das Verhältnis zwischen Qualität und Quantität in der Wirkungsanalyse

1 Einleitung

Evaluation, verstanden als Wirkungsmessung, gewinnt eine rasch wachsende Bedeutung in sehr unterschiedlichen Politikfeldern, in den letzten Jahren auch und gerade in der Bildungsforschung: Die großen ländervergleichenden Schulstudien wie TIMMS oder PISA stellen Evaluationen ganzer Bildungssysteme im internationalen Vergleich dar (z.B. Prenzel et al. 2007; Bundesministerium für Bildung und Forschung 2001), bildungspolitische Programme wie die nordrhein-westfälische Initiative >Jedem Kind ein Instrument< werden mit umfangreichen (und teuren) Evaluationsprogrammen begleitet (Kranefeld 2009), in den Hochschulen wird im Zuge der Bologna-Reformen und der immer stärker an einer Kultur der „*accountability*" orientierten permanenten Studienreform die Bewertung von Lehrveranstaltungen und die evaluationsgestützte Verbesserung der Qualität von Lehre und Studium zu einer Selbstverständlichkeit, die auch gesetzlich abgesichert ist (Landeshochschulgesetze - eine Übersicht findet sich unter http://www.hof.uni-halle.de/steuerung/lhg_uebersicht. htm).

Eine solche Verwissenschaftlichung politischen Handelns weckt Hoffnungen auf eine stärkere Versachlichung und Rationalität solcher Maßnahmen; sie führt natürlich auch zu der Frage, ob und in welchem Umfang die Sozialwissenschaften mit ihrem Methodeninstrumentarium die an sie gerichteten Anforderungen erfüllen können. Evaluationen untersuchen die Wirkungen spezifischer Maßnahmen – das bedeutet, dass es hier immer um die Analyse von Kausalbeziehungen gehen muss. Unter erkenntnistheoretischer und methodologischer Perspektive ist Kausalität ein äußerst komplexer Gegenstand, bei der Aufdeckung und Beschreibung von Kausalbeziehungen gibt es zahlreiche Probleme, Fehler- und Irrtumsmöglichkeiten. Dabei ist es Konsens, dass eine adäquate Wirkungsmessung quantitative Methoden und eine (quasi)experi-

mentelle Logik erfordert, weil nur standardisierte Messungen eine objektive Feststellung zulassen, ob die für den Erfolg einer Maßnahme relevanten Kriterien erreicht wurden, und nur der in (quasi)experimentellen Untersuchungen mögliche kontrollierte Vergleich zwischen Situationen mit und ohne Interventionen kausale Fehlschlüsse zu vermeiden hilft.

Allerdings bringt eine Konzentration auf quantitative Methoden auch Gefahren im Hinblick auf die Validität der Evaluationsdaten mit sich, wie wir im Folgenden anhand einer ausführlichen Diskussion von Stärken und Schwächen qualitativer *und* quantitativer Forschungsmethoden zeigen wollen. So können in ausschließlich quantitativen Erhebungen kaum die Itemverständnisprobleme der Befragten oder auch deren möglicherweise sozial erwünschtes Antwortverhalten deutlich werden und Nebeneffekte von Interventionen, von denen der Forscher keine Kenntnis hat und die damit auch nicht operationalisierbar sind, können oft gar nicht in den Blick geraten. Es lohnt sich deshalb, hier aktuelle Diskussionen in den Sozialwissenschaften über die Bedeutung von „Mixed Methods“ zu berücksichtigen: Durch eine Kombination qualitativer und quantitativer Methoden in einem Forschungsdesign wird es möglich, Schwächen der beiden Methodentraditionen durch die Stärken der jeweils anderen Tradition auszugleichen.

In unserem Beitrag werden wir in einem ersten Schritt kurz auf die aktuelle Debatte um *Mixed Methods* eingehen. Danach werden wir Validitätsbedrohungen und Methodenprobleme qualitativer und quantitativer Forschung bei der Untersuchung von Kausalbeziehungen diskutieren und zeigen, wie die Integration von Methoden dabei helfen kann, solche Methodenprobleme zu entdecken und ggf. zu überwinden. Abschließend werden wir – auch anhand eines eigenen empirischen Beispiels – zeigen, dass die Integration von qualitativen Erhebungsverfahren in eine quantitativ angelegte Untersuchung das Potential bietet, Messprobleme standardisierter Instrumente zu identifizieren und partiell zu überwinden und die „*outcomes*“ von Interventionen und vor allem deren unbeabsichtigte und nicht antizipierte Nebeneffekte zu explizieren.

2 Der Methodenstreit in der empirischen Sozialforschung

Die Forschungsmethodik der Sozialwissenschaften ist seit vielen Jahren durch einen Streit zwischen zwei Traditionen geprägt, die mit jeweils unterschiedlichen Daten und Auswertungsmethoden arbeiten und in denen sich verschiedene Modelle des sozialwissenschaftlichen Forschungshandelns und differierende Qualitätskriterien für gute Forschung entwickelt haben (vgl. Kelle 2008).

Die *quantitative Tradition* orientiert sich am Modell der experimentellen Naturwissenschaften und arbeitet vorzugsweise mit standardisierten Daten, weil nur solche Daten ohne eine weitere Bearbeitung mit statistischen Verfahren analysiert werden können. In dieser Tradition wird üblicherweise ein Modell des sozialwissenschaftlichen Forschungsprozesses verfolgt, bei dem empirische Untersuchungen als Prozesse deduktiver Theoriebildung und anschließenden Theorieprüfung mit Hilfe quantifizierbarer Informationen verstanden werden. In Abgrenzung zu diesem hypothetiko-deduktiven („HD“-)Modell hat sich in den Sozialwissenschaften ein zweiter Traditionsstrang entwickelt, der seine Wurzeln in der geisteswissenschaftlichen Hermeneutik einerseits und der modernen Ethnographie andererseits hat und der den Forschungsprozess als Erkundung eines relativ fremden Terrains versteht. Die in solcher qualitativen Forschung gesammelten Daten sind unstandardisiert und oft auch kaum standardisierbar, das Datenmaterial bilden Texte (und in jüngerer Zeit zunehmend Bilder), die mit interpretativen Verfahren einer sorgsamen hermeneutischen Analyse unterzogen werden müssen.

Das Verhältnis zwischen diesen beiden methodologischen Traditionen ist schon lange angespannt und von wechselseitiger Kritik und Abgrenzung gekennzeichnet. Oftmals ist die These vertreten worden, dass qualitative und quantitative Forschungsmethoden notwendigerweise logisch inkompatibel und inkommensurabel seien, weil sie auf jeweils verschiedenen, miteinander unvereinbaren erkenntnistheoretischen Grundpositionen aufbauen würden (vgl. dazu etwa Lamnek 1995, Seite 253; Guba, Lincoln 1988, Seite 93). Allerdings haben seit den 1930er Jahren zahlreiche klassische und neuere sozialwissenschaftliche Untersuchungen, die sowohl zu empirisch neuen und überraschenden Einsichten geführt als auch oft die Theorieentwicklung des Faches außerordentlich stimuliert haben, qualitative und quantitative Forschungsmethoden parallel und gemeinsam in einem Forschungsdesign genutzt. Hierzu zählen etwa die für die Arbeitslosigkeitsforschung paradigmatische „Marienthalstudie“ (Jahoda, Lazarsfeld, Zeisel 1933/1982), die „Hawthorne Study” (Roethlisberger, Dickson 1939), die die Entwicklung der Industriesoziologie stark beeinflusst hat, die Studie von Festinger, Riecken und Schachter über Weltuntergangskulte (1956), welche die Theorie kognitiver Dissonanz (mit)begründet hat oder Zimbardos bekanntes „Gefängnisexperiment“ über die Folgen institutioneller Deindividuation (Zimbardo 1969).

Aus Bereichen heraus, die ein besonderes Interesse an der praktischen Anwendung sozialwissenschaftlicher Forschungsmethoden haben, wie insbesondere die erziehungswissenschaftliche Forschung, hat sich seit Ende der

1980er Jahre ein methodologischer Diskurs entwickelt, bei dem auf den forschungspraktischen und theoretischen Ertrag einer Kombination von Verfahren beider Traditionen hingewiesen wird – dabei hat sich (vor allem im anglo-amerikanischem Sprachraum) mittlerweile bereits eine „Mixed Methods Bewegung“ etabliert mit eigenen Konferenzen, Handbüchern und Zeitschriften (vgl. etwa Tashakkori und Teddlie 2003; Greene 2007; Kelle 2006, 2008; Creswell 2009; insbesondere auch die nun im dritten Jahr erscheinende Zeitschrift „*Journal of Mixed Methods Research*“). Seit den 1980er und insbesondere in den 1990er Jahren wurde eine wachsende Anzahl von sozialwissenschaftlichen Forschungsprojekten durchgeführt, in denen qualitative und quantitative Methoden miteinander kombiniert wurden - in den deutschen Sozialwissenschaften hat in den 1990er Jahren insbesondere der Sonderforschungsbereich 186 der DFG Pionierarbeiten geleistet (vgl. Kelle 2008, S. 231 ff.; Kluge, Kelle 2001; Heinz, Marshall 2003; internationale Überblicke über Mixed Methods Studien in verschiedenen Disziplinen finden sich etwa bei Tashakkori, Teddlie 2003, S. 491 ff. oder bei Greene 2007). Aus der Mixed Methods Bewegung liegen mittlerweile eine Fülle von Publikationen mit Vorschlägen zur Integration qualitativer und quantitativer Methoden in konkreten Forschungsdesigns vor (ein Überblick etwa bei Creswell et al. 2003; Seipel, Rieker 2003, S. 236 ff.; Greene 2007; Kelle 2008 S. 282 ff.).

Ist der alte Methodenstreit damit obsolet geworden? Ein genauerer Blick auf aktuell laufende Diskussionen zeigt, dass sich die Sache keinesfalls so einfach verhält. Die vorwiegend im anglo-amerikanischen Raum geführte Debatte über Mixed Methods hat noch eine große Anzahl von Problemen zu lösen. Der größte Teil der vorliegenden Arbeiten beschränkt sich auf methodisch-praktische Fragen der Gestaltung von Forschungsdesigns, wobei tiefer liegende Probleme der Relevanz und vor allem Gültigkeit von Forschungsergebnissen oft vernachlässigt werden. Dabei beklagen die Diskutanten oft die vorherrschende begriffliche Unklarheit und eine nach wie vor uneinheitliche Terminologie (Tashakkori, Teddlie 2003). Und Alan Bryman, einer der Gründerväter der Mixed Methods Bewegung, zeigte in einem systematischen Reviewartikel über Arbeiten, die mit einem qualitativ-quantitativen Methodenmix arbeiten, dass in solchen Studien qualitative und quantitative Methoden oft nur nebeneinander eingesetzt, aber nicht wirklich miteinander verbunden oder aufeinander bezogen würden (Bryman 2007).

Die Verdrängung der wissenschaftstheoretischen Argumente und Auseinandersetzungen (unter deren Überbetonung die alte Methodendebatte in den Sozialwissenschaften gelitten hatte) zugunsten einer reinen Forschungsprag-

matik führt also ebenso in eine Sackgasse wie der nie gelöste Methodenstreit (vgl. Kelle 2008, Seite 26ff.). Dagegen weist ein von Vertretern der Mixed Methods Bewegung häufig vorgebrachtes Argument, dass eine Methodenkombination sinnvoll ist, um die *Stärken und Schwächen* quantitativer und qualitativer Forschung auszugleichen (vgl. etwa Johnson, Turner 2003, S. 299) in die richtige Richtung. Das bedeutet aber auch: Es müssen zuerst die Schwächen qualitativer und quantitativer Forschung genau beschrieben und geklärt werden, damit dann in einem zweiten Schritt diese Schwächen durch die Stärken der jeweils anderen Tradition ausgeglichen werden können – erst dann kann eine sinnvolle Methodenkombination stattfinden. Arbeiten, die so vorgehen, sind nun aber auch in der gegenwärtigen Debatte über Mixed Methods nicht immer leicht zu finden, so dass man in der entsprechenden Literatur kaum methodologische Regeln findet, die eine Entscheidung darüber zulassen, an welchen Punkten des Forschungsprozesses für welche (Teil)fragen welche Methoden einzusetzen sind – so dass bei der Anwendung eines Mixed Methods Designs bei der Auswahl von Methoden letztlich *ad hoc* Entscheidungen getroffen werden müssten. Die Frage nach den Schwächen und Grenzen beider Traditionen, die im aktuellen Mixed Methods Diskurs oft ausgeklammert bleibt, wurde im klassischen Methodenstreit der Sozialwissenschaften aber zumindest formuliert und diskutiert. Das Problem dieser Debatten war nur, dass beide Lager sich stets auf Schwächen der Gegenseite konzentriert haben, während man die von dort aus vorgebrachte Kritik an der eigenen Tradition ignoriert oder ihr auch manchmal durch sprachliche Kunstgriffe auszuweichen trachtet (indem man etwa versucht, der Besorgnis über das Problem der mangelnden Generalisierbarkeit von Fallstudien durch sprachliche Manipulation am Begriff wissenschaftlicher Gültigkeit zu begegnen, etwa den Begriff der „Generalisierbarkeit“ durch den der „Übertragbarkeit“ von Ergebnissen zu ersetzen, vgl. Lincoln, Guba 2000 und zur Kritik an diesen Positionen Kelle 2008, S. 37 ff.).

Die Diskussion um Mixed Methods eröffnet aber die Möglichkeit, die Methodendebatte aus dieser Sackgasse zu führen, indem Kritik als Ressource genutzt wird, um die Schwächen bestimmter Ansätze zu erkennen, zu bearbeiten und (ggf. durch die Verwendung von Methoden aus einer alternativen Tradition) zu beheben. Auf diese Weise kann sich das methodische Instrumentarium der Sozialwissenschaften durch kritische Kooperation weiter entwickeln, was Streit nicht ausschließt, aber eine selbstkritische Einstellung erfordert, die bei einer ernsthaften Bestandsaufnahme der methodischen Probleme und Grenzen des eigenen Vorgehens und bei der Verbesserung methodischer Werkzeuge hilft. Auf dieser Grundlage kann das Querschnittfach „Methodologie“ seine

klassische Aufgabe erfüllen, „Fehlertheorien“ zu formulieren und zu klären, welche Schwächen konkrete Methoden bei der Untersuchung spezifischer sozialwissenschaftlicher Gegenstände jeweils haben und wo dies zu blinden Flecken bei der empirischen Forschung führt.

Im Folgenden möchten wir diese Strategie auf das Feld der Evaluationsforschung beziehen, indem wir herausarbeiten,

1. wo die spezifischen Schwächen und Probleme qualitativer und quantitativer Methoden bei der empirischen Analyse kausaler Beziehungen liegen, und

2. wie diese Schwächen durch die Einbeziehung von Methoden der jeweils anderen Tradition bearbeitet und vielleicht auch ausgeglichen werden können.

3 Kausale Analyse und die Stärken und Schwächen qualitativer und quantitativer Forschung

Evaluationsforschung, insbesondere verstanden als „strenge Wirkungsevaluation“ („*rigorous impact evaluation*“, vgl. bspw. White 2006) untersucht Kausalzusammenhänge: eine bestimmte Intervention oder Maßnahme soll „wirken“, das heißt sie soll einen beschreibbaren und messbaren „Effekt“ haben. Dabei können die Vorstellungen, die diejenigen von diesen Wirkungen haben, die eine Intervention planen und durchsetzen, mehr oder weniger genau und spezifisch sein, so dass Evaluatoren oft erst gemeinsam mit den Gestaltern von *policies* genau heraus arbeiten müssen, wie man Indikatoren für Wirksamkeit in dem konkreten Untersuchungsbereich entwickelt, das heißt, woran man genau erkennen kann, dass eine bestimmte Maßnahme wirkt. Bei einer strengen Wirkungsmessung setzen wissenschaftliche Evaluatoren und politische Entscheider oft gleichermaßen ihr Vertrauen auf standardisierte quantitative Messverfahren und das nicht ohne Grund: in der quantitativen Methodentradition gibt es seit den 1950er Jahren eine lange Tradition in der Diskussion kausaler Zusammenhänge und in der Bearbeitung der dort auftretenden Probleme (vgl. bspw. Lazarsfeld 1955; Simon 1954; Blalock 1985), während dieses Themenfeld in der qualitativen Forschung bisher eher gemieden wurde. Dort wird der Kausalitätsbegriff bisher lediglich in den Arbeiten von Anselm Strauss (Strauss/Corbin 1990, 101) und ansatzweise in Konzepten, die auf das Verhältnis von „Erklären“ und „Verstehen“ eingehen (vgl. bspw. Lamnek 1995, 74), verwendet. Weil darüber hinaus die Standards quantitativer Erhe-

bungen mit ihren Gütekriterien der Objektivität und Reliabilität sowie der Repräsentativität der Stichprobenziehung auch für viele Auftraggeber von Evaluationsprojekten starke und nachvollziehbare Argumente bieten, ist eine Dominanz quantitativer Verfahren und eine eher marginale Bedeutung qualitativer Erhebungsmethoden im Rahmen der Evaluationsforschung durchaus erklärbar. Dabei sind qualitative Methoden unabdingbar, um die grundlegenden Defizite, die quantitative Erhebungen bei der Untersuchung von Kausalbeziehungen aufweisen, aufzudecken und zumindest partiell zu überwinden.

3.1 Die Stärken quantitativer und experimenteller Ansätze

Aber in welchen Punkten sind quantitativ-standardisierte und quasi-experimentelle Verfahren den qualitativen Methoden tatsächlich überlegen? Folgt man der Argumentation eines wissenschaftsphilosophischen Fallibilismus, wie er in den Arbeiten von Charles Sanders Peirce und Karl Popper (für die Sozialwissenschaften von Hans Albert) begründet und dann in verschiedenen (auch liberaleren) Fassungen von Imre Lakatos, Nicolas Rescher und Larry Laudan weiterentwickelt wurde (z.B. Chalmers 2007), unterscheidet sich wissenschaftliche Forschung von Alltagsstrategien des Wissenserwerbs vor allem durch ihre besondere methodische Kontrolle, mit der versucht wird, typische Fehler zu vermeiden, die der *common sense* bei einer Formulierung und Überprüfung von allgemeingültigen Aussagen, bei der Untersuchung empirischer Gegenstände und bei der Formulierung von Schlussfolgerungen leicht begeht. Zur konkreten Sicherung methodischer Kontrolle in der Forschungspraxis kann Methodologie (als Querschnittsdisziplin zwischen Sozial- und Naturwissenschaften) einerseits „Fehlertheorien“ formulieren, also Theorien darüber, welche Fehler bei der Untersuchung der empirischen Realität auftreten können, und andererseits Vorschläge ausarbeiten, wie diese Fehler vermieden werden können.

Fehlermöglichkeiten in der Evaluationsforschung können bspw. darin bestehen, dass entweder eine Veränderung konstatiert wird, wo tatsächlich gar keine Veränderung stattgefunden hat, oder aber dort wo eine (möglicherweise nicht-intendierte) Veränderung stattgefunden hat, diese Veränderung übersehen wird.

Der erste Fehler wird leicht dort geschehen, wo keine angemessenen (oder falsche) Beobachtungsverfahren und Messinstrumente eingesetzt worden sind. Wahrnehmungs- und Kognitionspsychologen haben in ihren empirischen Forschungen gezeigt, wie stark die menschliche Wahrnehmung dazu tendiert, Einzelbeobachtungen, die mit Vorannahmen oder auch Stereotypen konsistent

sind, zu registrieren und Gegenevidenz (selbst wenn sie häufig und massiv auftritt), zu übersehen (z.B. Kahneman 1982/2008; Festinger 1957). Diese kognitive Tendenz, bestätigende Evidenz für die eigenen Annahmen zu stark und Falsifikatoren und Anomalien zu gering zu bewerten, kann zum „Rosinenpicken" führen, bei dem vermeintliche Belege für die Wirksamkeit der eigenen Maßnahme zusammengesucht und präsentiert werden.

Beliebt ist dieses Vorgehen etwa bei der Präsentation von pseudo- und alternativmedizinischen „Wunderkuren" oder im Marketing, wo gerne unterstützende und bestätigende Berichte einiger „begeisterter" Patienten oder Kunden präsentiert werden. Im Kontext der „evidenzbasierten Medizin" steht deshalb die einfache Präsentation von Fallgeschichten auf der niedrigsten Stufe wissenschaftlicher Evidenz. Aber auch im Bereich von sozialpolitischen Interventionen ist ein solches Vorgehen – Beleg der Wirksamkeit durch Einzelfallberichte – häufig. Was hier dann übersehen wird: möglicherweise stellt sich der beobachtete Effekt nur unter ganz bestimmten (sehr eingeschränkten) Bedingungen ein, die gar nicht genau bekannt sind, während die Intervention für große Teile der betreffenden Zielgruppe völlig wirkungslos ist.

Die Wahrnehmung (oder die Leugnung) von Veränderungen kann auch die Folge einfachen Wunschdenkens (vgl. dazu etwa Babad, Katz 1991 zur Bedeutung von „*wishful thinking*") sein.

Diese Art der Wahrnehmungsverzerrung hat einer der Begründer der Logikerschule von Port Royal, Antoine Arnauld bereits 1662 mit seinen Aphorismen aufgespießt: „Wir beurteilen die Dinge, nicht das zugrunde legend, was sie an und für sich sind, sondern das, was sie in der Beziehung zu uns sind." (Arnauld 1972/1662, S. 253) Und: „Wenn sie jemanden lieben, ist er von jeder Art Fehler befreit. Alles, was sie wünschen, ist gerecht und leicht, alles, was sie nicht wünschen, ist ungerecht und unmöglich, ohne dass sie irgendeinen Grund für alle diese Urteile vorzubringen vermöchten, außer der Leidenschaft, von der sie besessen sind." (ebd., S. 254)

Neben der Unter- oder Überschätzung von (erwünschten oder unerwünschten) Veränderungen von Interventionen gibt es Fehlermöglichkeiten, die den Kausalnexus zwischen Intervention und Veränderung betreffen: wenn die erwünschte Veränderung nicht wegen der Intervention, sondern aufgrund anderer Ursachen eintritt, kann eine kausale Verbindung zwischen Intervention und Veränderung unterstellt werden, die gar nicht existiert. Eine solche kausale Fehlattribution wird (nicht nur) in der methodologischen Literatur häufig beschrieben, es handelt sich um das klassische Fehlurteil des „post hoc, ergo propter hoc", das in nahezu allen Handlungsfeldern, in denen Interventionen stattfinden, eine Rolle spielen kann und das unter verschiedenen begrifflichen Etiketten diskutiert wird:

In der Medizin gehört das Wissen um „Spontanremissionen" zur Folklore: zahlreiche Krankheiten heilen auch ohne medizinische Interventionen („Ein Schnupfen dauert mit

Behandlung eine Woche, ohne Behandlung sieben Tage“). Das bedeutet aber auch: bei jeder medizinischen Intervention (auch in größeren Gruppen), der eine Heilung oder Veränderung folgt, muss die Frage gestellt werden, ob die Heilung nicht auch ohnehin geschehen wäre. Für den Bereich der Sozial- und Erziehungswissenschaften sprechen Campbell und Stanley in ihrer klassischen Arbeit über „Experimentelle und quasi-experimentelle Designs“ von „Reifungseffekten“ – eine bestimmte Lehrmethode wird vielleicht vor allem deswegen als erfolgreich angesehen, weil Schüler aufgrund normaler kognitiver Reifungsprozesse einen Kompetenzzuwachs erfahren (vgl. Campbell, Stanley 1963).

In der Tradition der quantitativ-standardisierten Forschung wurde eine Reihe von Vorsichtsmaßnahmen vorgeschlagen und ausgearbeitet, um diese einfachen Beobachtungsfehler zu beherrschen und zu begrenzen:

- So kann man Fehler, die durch „*wishful thinking*“ entstehen, das heißt Wahrnehmungsverzerrungen, die dazu führen, dass man dort eine positive Veränderung sieht, wo eine solche Veränderung gar nicht stattgefunden hat, oder die eine tatsächlich stattgefundene Veränderung übertreiben und überzeichnen, durch Verfahren einer objektiven (das heißt: beobachterunabhängigen) standardisierten Beobachtung und Messung zu begrenzen versuchen. Ein zuverlässiger empirischer Befund über eine bestimmte Veränderung liegt demnach eben nur dann vor, wenn mehrere neutrale Beobachter (auch solche, die die Vorannahmen und Wünsche des Untersuchers nicht teilen) bei dem Vorliegen derselben Daten und Befunde übereinstimmend zu demselben Urteil über diese Veränderung gelangen. In den Sozialwissenschaften ist die Methode der standardisierten Befragung ein Mittel, um Wahrnehmungsverzerrungen und Wunschdenken von Untersuchern zu vermeiden: Täuschungen darüber, ob Befragte an bestimmten Stellen ein Kreuz gemacht haben, sind selten. Demgegenüber lassen die interpretativen und hermeneutischen Verfahren, die in der qualitativen Sozialforschung angewendet werden, *prima facie* stets Interpretationsspielräume bei der Auswertung des Datenmaterials zu– die hierdurch aufgeworfenen Probleme müssen dann durch sehr komplexe und aufwändige Methoden (bspw. in Gruppenauswertungen oder durch eine EDV-gestützte synoptische Analyse des Materials) bearbeitet werden.
- Das Risiko des „Rosinenpickens“ lässt sich dadurch vermeiden, dass die Anzahl der empirischen Beobachtungen systematisch erhöht wird, indem also etwa alle Betroffenen einer Maßnahme befragt werden, oder indem aus dieser Gruppe eine (Zufalls)Stichprobe *lege artis* gezogen wird. Wenn alle diese Fälle dann auch untersucht werden, können Gegenbeispiele und *negative cases* (d.h. Fälle, bei denen keine Veränderung stattgefunden hat), nicht mehr ohne weiteres übersehen und ausgeblendet werden. Die Stichproben-

größe, die notwendig ist, um Gegenbeispiele zu erfassen, hängt nun auch von der Heterogenität des Untersuchungsfeldes (oder, technisch gesprochen, von der Varianz des untersuchten Merkmals) ab - da qualitative Erhebungs- und Analyseverfahren sehr aufwändig sind, lassen sich hiermit allerdings immer nur kleine Fallzahlen bearbeiten, so dass in dieser Tradition das Risiko einer einseitigen und verzerrenden Fallauswahl im Allgemeinen höher ist als in der quantitativen Forschung.

- Schließlich bieten experimentelle Interventionen ein sehr brauchbares Werkzeug, um klassische kausale Fehlschlüsse zu erkennen und zu vermeiden: Kausale Fehlschlüsse, bei denen Spontanremissionen oder Reifungsprozesse übersehen wurden und deshalb eine Veränderung, die ohnehin von selber eingetreten wäre, auf eine Intervention zurückgeführt wird, obwohl diese Intervention hierfür gar nicht ursächlich war, lassen sich durch die Bildung von Kontrollgruppen (also von Gruppen, bei denen keine Intervention stattfindet), vermeiden. Auch andere Hypothesen über alternative Ursachen für die interessierenden Veränderungen lassen sich durch experimentelle Kontrollen prüfen. Nun sind bei der Durchführung politischer Interventionen und sozialtechnologischer Maßnahmen nur selten strenge experimentelle Designs einsetzbar (eine zufällige Aufteilung von Betroffenen auf Versuchs- und Kontrollgruppen ist etwa oft aus politischen, praktischen oder ethischen Gründen gar nicht möglich), jedoch gibt es in der sozialwissenschaftlichen Methodenforschung seit den bahnbrechenden Arbeiten von Campbell und Stanley (1963) eine recht umfangreiche Literatur über verschiedene Möglichkeiten der „quasi-experimentellen“ Kontrolle.

Die quantitative Methodentradition stellt eine Reihe von ausformulierten Fehlertheorien (etwa inferenzstatistische Theorien des Stichprobenfehlers) und darauf aufbauende Verfahren zur Fehlerkennung und Fehlervermeidung zur Verfügung, deren routinemäßige Anwendung in der Praxis der empirischen Sozialforschung (zumindest dort, wo *lege artis* gearbeitet wird) verbreitet ist. Qualitative Verfahren werfen gerade in Bereichen, in denen die quantitative Tradition Lösungen anbietet, oft besondere Methodenprobleme auf: so ist hier häufig allein aus forschungspraktischen Gründung die Prüfung, ob verschiedene Forscher unabhängig voneinander zu denselben Ergebnissen kommen würden (Kelle 2008, 28ff.), nur schwer möglich. Fallzahlen in der qualitativen Forschung (aufgrund des besonderen Aufwands der hier eingesetzten Erhebungs- und Auswertungsmethoden) sind oft sehr klein, so dass sich sofort Fra-

gen nach der Verallgemeinerbarkeit der hier gewonnenen Befunde aufdrängen (Kelle 2008, S. 231).

3.2 Methodenprobleme und Validitätsbedrohungen quantitativer Untersuchungen

Angesichts dieser Situation stellt sich natürlich die Frage, warum es dann überhaupt sinnvoll sein kann, qualitativ zu forschen. Nun haben auch quantitative Methoden ganz spezifische Beschränkungen und Methodenprobleme, unter anderem dort, wo Kausalität als sozialwissenschaftliches Konzept verwendet wird, das heißt dort, wo die untersuchten Ursache- und Wirkungsphänomene *soziale Handlungen* sind. Interessanterweise stammt der Begriff *causa* historisch auch gar nicht aus einem naturwissenschaftlichen Kontext, sondern wurde zuerst in sozialen Zusammenhängen in enger Verbindung mit dem Konzept des Handelns verwendet, wie R.G. Collingwood in einer sprachgeschichtlichen Untersuchung gezeigt hat (vgl. Collingwood 1937/1938). Begriffe, die auf Ursachen verweisen, wie „bewirken", „verursachen", „beeinflussen" oder „führen zu" werden seit der Antike (und bis heute nicht nur im alltäglichen Sprachgebrauch, sondern auch in vielen wissenschaftlichen Kontexten, die sich auf menschliches Handeln beziehen, etwa in den Rechtswissenschaften) nicht primär auf Naturprozesse, sondern auf freie und intendierte Handlungen von verantwortlichen Akteuren bezogen. Soziales Handeln kann somit in der ursprünglichen Bedeutung des Begriffs *causa* durchaus als (wenn auch nicht deterministisch) beeinflusst oder verursacht verstanden werden. Folgt man einer bereits bei Aristoteles getroffenen Unterscheidung, besteht die *causa* einer Handlung aus der *causa ut,* einer bestimmten Absicht eines Akteurs, und der *causa quod,* einer bestimmten Situation oder einem bestimmten Sachverhalt. Diese Differenzierung lässt sich auch in modernen handlungstheoretischen Konzeptionen finden: Alfred Schütz zufolge erfordert das Verstehen und die Erklärung sozialen Handelns, dass einerseits „um-zu Motive", d.s. Ziele, Motive oder Absichten des Handelnden (die *causa ut*) und andererseits „weil-Motive", d.s. Situationsmerkmale (die *causa quod*) einbezogen werden (vgl. Schütz 1971). Natürlich ist aber ein solches Verstehen sozialen Handelns nicht möglich ohne eine gewisse Kenntnis allgemeiner Regeln, mit denen sich Handlungsziele (die „um-zu-Motive") und Handlungsbedingungen (die „weil-Motive") miteinander verknüpfen lassen. Aus der Perspektive nahezu aller aktuellen Handlungstheorien muss dabei davon ausgegangen werden, dass die *Handlungsziele* von Akteuren plastisch und flexibel sein können und sich an veränderte *Handlungsbedingungen* anpassen, dass Handlungsbedingungen

wiederum oft variabel sind und von kreativen, entscheidungsfähigen Akteuren veränderbar und dass schließlich jene *Handlungsregeln*, mit denen Handlungsbedingungen und Handlungsziele miteinander verbunden werden, nicht universell sind, sondern kulturspezifisch und lokal gelten und soziokulturellem Wandel unterliegen.

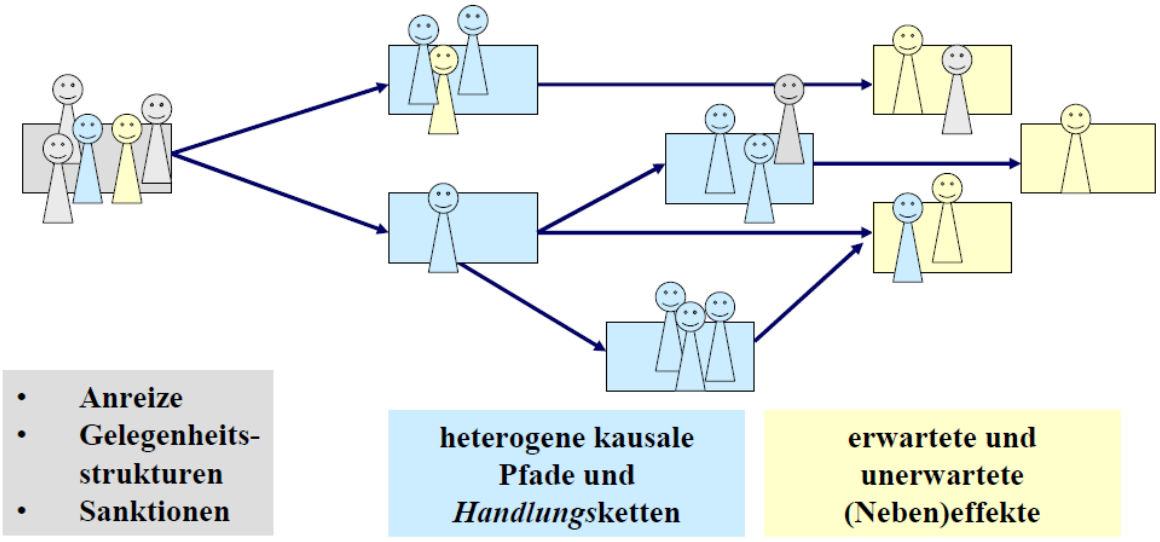
Das klassische Kausalmodell der quantitativ orientierten Evaluationsforschung kann mit solch einer Flexibilität und Kontingenz von Handlungszielen, Handlungsbedingungen und Handlungsregeln oft nur schwer umgehen. Das Standardmodell nimmt hier die folgende Form an: Zur Überprüfung der Hypothese, dass eine Intervention einen bestimmten *outcome* erzeugt, müssen die Intervention als unabhängige Variable und der *outcome* als abhängige Variable operationalisiert werden. Eine kausale Wirkung wird dann angenommen, wenn die Veränderung der unabhängigen Variablen (die idealerweise durch ein experimentelles Design erfolgt, in dem eine randomisierte Zuordnung der Untersuchungseinheiten zu den Ausprägungen der unabhängigen Variablen erfolgt) auch eine messbare Änderung der abhängigen Variablen nach sich zieht (Grafik 1).

Grafik 1: Intervention und Outcome

Bildungs- und sozialpolitische Interventionen[1] involvieren nun komplexes soziales Handeln und erfordern lange instrumentelle Ketten, die zahlreiche Einzelhandlungen einer großen Anzahl individueller Akteure miteinander vernetzen, welche jeweils verschiedene Handlungsziele und einen unterschiedlichen kulturellen Hintergrund haben. Die verschiedenen Einzelhandlungen generieren dabei große Mengen verschiedener Handlungsresultate, die erst zusammengenommen, im Aggregat, darstellbar etwa durch statistische Mittelwertsbildung, als Gesamt-*outcome* betrachtet werden können (siehe Grafik 2). Das grundlegende Problem besteht nun darin, dass der größte Teil des Netzwerks von Handlungsketten, die sich durch die Interventionsmaßnahme her-

1 Etwa: die Einführung des Programms „Jedem Kind ein Instrument“ an nordrhein-westfälischen Grundschulen oder auch: die Einführung eines einheitlichen Systems gestufter Studiengänge und Abschlüsse in der Universitätsausbildung in verschiedenen Staaten, die sich darauf verständigen.

ausbilden, niemals völlig durch jene Akteure kontrolliert werden kann, die die Intervention planen und anordnen. Durch eine Interventionsmaßnahme lassen sich nur bestimmte *Handlungsbedingungen* der Akteure vor Ort *direkt* beeinflussen, nicht deren Handlungen selber. Eine Intervention kann Anreize schaffen (etwa: *ökonomische Anreize für Musikschullehrer*), sie kann Gelegenheitsstrukturen institutionalisieren (etwa: *Angebote an Musikunterricht in Grundschulen* usw.) oder sie kann administrative Sanktionsdrohungen (etwa: die Kürzung staatlicher *Mittel*) verhängen.



Grafik 2: Komplexe sozialpolitische Intervention mit zahlreichen Akteuren

Hierdurch werden Akteure aber nicht mechanisch gezwungen, sondern nur motiviert, im Sinne des gewünschten *outcome* zu handeln, weil dies bestimmten Handlungszielen dient (bspw.: *in den Genuss ökonomischer Vorteile zu gelangen*, *Qualifikationszertifikate zu erwerben*, oder *die Kürzung finanzieller Mittel vermeiden*), die ihnen von denjenigen, die die Intervention planen, mehr oder weniger sinnvoll unterstellt werden können. Es ist allerdings durchaus möglich, dass Akteure, die die Interventionen umsetzen sollen, andere und konkurrierende Handlungsziele verfolgen. Akteure sind in der Regel in lokale kulturelle und institutionelle *settings* eingebunden, in denen bestimmte Handlungsziele und die zu ihrer Erreichung notwendigen Mittel und Wege bekannt sind. In vielen Fällen kennen aber weder diejenigen, die die Intervention planen, noch diejenigen, die sie evaluieren, solche lokalen Organisations- oder Professionskulturen. Zudem können findige Akteure Strategien entwickeln, um die Handlungsbedingungen, die durch die Intervention vorgegeben wurden, in einem Sinn zu beeinflussen, der den Intentionen der sozialpolitisch

Intervenierenden zuwiderläuft. Administrative Sanktionen können umgangen werden, Anreize können konsumiert werden, ohne dass entsprechende Gegenleistungen erbracht werden, Angebote können ignoriert werden, dadurch können Mitnahmeeffekte entstehen u.v.a.m.. Auf diese Weise können Interventionen zahlreiche unbeabsichtigte Effekte nach sich ziehen, einerseits dadurch, dass die an der Intervention Beteiligten und von ihr Betroffenen in lokale kulturelle und institutionelle Kontexte eingebunden sind, die sie mit Handlungszielen und Handlungswissen ausstatten, die Planern und Evaluatoren nicht bekannt sind, und andererseits dadurch, dass die beteiligten Personen entscheidungs- und handlungsfähige Subjekte sind.

3.2.1 Das Problem der übersehenen (Neben)wirkungen

Hiermit gelangen wir zum ersten grundlegenden Problem quantitativer „Monomethodendesigns“ in der Evaluationsforschung: das *Problem der übersehenen (Neben)wirkungen* von Evaluationen. Quantitative Forschung folgt dem Grundprinzip des „hypothetiko-deduktiven Modells“ des Forschungsprozesses. Das bedeutet, dass am Anfang empirischer Forschung stets die Formulierung von präzisen Hypothesen (in der Evaluationsforschung sind dies spezifische, genau ausformulierte Annahmen über konkrete Wirkungen der Interventionen) stehen muss, mit deren Hilfe dann die gemessenen Variablen definiert und die Messinstrumente konstruiert werden. Ohne ein fertiges Messinstrument können standardisierte Daten nicht erhoben werden, was bedeutet, dass die Forscher schon vor der empirischen Untersuchung wissen müssen, welche konkreten Folgen der Intervention sie untersuchen wollen.

Wenn wir allerdings davon ausgehen können, dass die Intervention durch das Handeln entscheidungsfähiger und kreativer Akteure auf der Grundlage von (den Forschern oft gar nicht bekannten) lokalen Wissensbeständen umgesetzt werden (und auch bei dieser Umsetzung in einer Weise verändert werden, die den Intentionen derjenigen, die die Intervention geplant haben, gar nicht entspricht), müssen wir stets mit einer Reihe von (vor einer Intervention nicht antizipierten und vielleicht auch gar nicht antizipierbaren) Nebenfolgen rechnen.

Bemühungen zu einer verstärkten Integration und Inklusion behinderter Schüler in die Regelschule können etwa neue psychosoziale Probleme bei besonders vulnerablen Individuen oder Gruppen aufwerfen: so ist bspw. für stotternde Kinder in Regelschulen das Risiko besonders hoch, während der Schulzeit Opfer von Mobbing, verbalem und physischen Missbrauch durch nichtbehinderte Mitschüler zu werden (vgl. Benecken, Spindler 2004).

Solche unerwarteten Effekte und Folgen können bei einem rein hypothetiko-deduktiven Design jedoch überhaupt nicht in den Blick geraten, zu ihrer Erforschung braucht es explorative, qualitative Methoden.

3.2.2 Das Problem der unbekannten Störvariablen

Eng mit dem ersten Problem verbunden ist das *Problem der unbekannten Störvariablen*, die die Umsetzung einer Intervention und die Erreichung der geplanten Wirkungen be- oder verhindern können. Da soziale Interventionen immer durch das Handeln ganz unterschiedlicher Akteure mit verschiedenen Interessen, Zielen und Wissensbeständen umgesetzt werden müssen, erklären sich die Wirkungen von Störvariablen auch oft durch die (mehr oder weniger gezielten) Handlungen der bei der Umsetzung der Intervention beteiligten Akteure und Akteursgruppen.

Ein berühmtes Beispiel aus der Geschichte der Bildungsforschung, das bis heute Gegenstand wissenschaftlicher, politischer und juristischer Auseinandersetzungen ist, zeigt, wie die Intentionen von Forschern und politischen Akteuren durch Betroffene konterkariert werden können (vgl. Coleman 1976, 1977; Hage, Meeker 1988, S27 ff.). Angeregt durch empirische Evidenz, wonach die schlechte Performanz schwarzer Schüler eine Folge ihrer Beschulung in sozial und ethnisch homogenen Schulen war, versuchten US-amerikanische Schulreformer in den 1960er Jahren eine ethnische Desegregation durchzusetzen. Die Erfolge der drastischen Maßnahmen (etwa „busing“, bei dem Schulkinder manchmal länger als eine Stunde von der elterlichen Wohnung zu ihrer neuen Schule transportiert wurden) blieben äußerst beschränkt. Weil viele weiße Mittelschichteltern mit Schulflucht reagierten, erreichten die sozialpolitischen Maßnahmen in manchen Städten sogar das Gegenteil des angestrebten *outcomes*, indem ethnische und soziale Ungleichheiten verstärkt und nicht reduziert wurden.

In der erkenntnistheoretischen Diskussion um das Kausalitätskonzept sind solche Phänomene unter der Bezeichnung „Simpsons Paradoxon“ (Simpson 1951) bekannt. Wir haben es hier – technisch gesprochen – mit einer intervenierenden Variable (der Gegenwehr bestimmter Akteure gegen die Intervention) zu tun, deren Wirkung umso stärker wird, je stärker auch die unabhängige Variable (also die Intervention) wirkt. Die Intervention kann also sehr effektiv sein und gerade deshalb keine messbaren Wirkungen auf der Ebene empirischer Beobachtungen erzielen, weil sie soziale Gegenkräfte auf den Plan ruft. Auch solche Phänomene lassen sich nur mit Hilfe intensiver und aufwändiger Explorationen im Feld, nicht jedoch mit vorab entwickelten Standardmessinstrumenten, in den Blick nehmen.

3.2.3 Das Problem der unbeobachteten Heterogenität

Interventionen sind stets in bestimmten Situationen und mit einem bestimmten Klientel besonders effektiv, und erzielen unter anderen Bedingungen und mit anderen Zielgruppen kaum nennenswerte Erfolge. Oft sind die ansprechbaren und die nicht ansprechbaren Zielgruppen gemischt und können (bereits aus praktischen Gründen) nicht sinnvoll voneinander getrennt werden. Manche dieser intervenierenden Variablen, die die „Ansprechbarkeit" des Klientels oder die konkreten Rahmenbedingungen der Intervention vor Ort betreffen, sind den Planern der Intervention und den Evaluatoren bekannt, andere nicht.

Ein Programm wie „Jedem Kind ein Musikinstrument" etwa wird in bestimmten Schulen, bei bestimmten Lehrern und bei bestimmten Schülern besonders starke, in anderen Fällen gar keine oder nur sehr schwache Effekte erzielen. Manche der intervenierenden Variablen lassen sich dabei leichter antizipieren und erfassen als andere: so liegt die Hypothese nahe, dass sich besonders dort starke Effekte erzielen lassen, wo die Eltern das Programm unterstützen und ihre Kinder zu Hause zum Üben bewegen. Diese Unterstützung der Eltern lässt sich mit vertretbarem Aufwand durch standardisierte Items in einem Fragebogen erfassen. Andere Variablen, wie das Engagement von Lehrern, die Unterstützung durch die Schulleitung u.ä. lassen sich weitaus schwerer, teilweise gar nicht, mit quantitativen Methoden untersuchen.

Quantitative Forscher verfügen zwar mit ausgearbeiteten Stichprobentheorien und den Methoden der Stichprobenziehung über einige Werkzeuge, um mit unbeobachtbarer Heterogenität umzugehen: solange es sinnvoll ist, von der klassischen statistischen Normalverteilungstheorie der Fehler auszugehen (das heißt von der Annahme, dass sich verstärkende und hemmende Effekte intervenierender Variablen *grosso modo* ausgleichen), und anzunehmen, dass quantitative Untersuchungen bei genügendem Stichprobenumfang Ergebnisse liefern, die brauchbare Schätzer für Durchschnittseffekte der Intervention abgeben. Nur ist leider ein solches Denken in Mittelwerten nicht immer ohne Probleme: denn die Schwankungsbreite der Effekte (technisch: die Fehlervarianz) kann so hoch werden, dass die Effekte bei nicht sehr umfangreichen Stichprobengrößen nicht mehr signifikant sind (und damit die Intervention als nicht oder kaum erfolgreich gilt, obwohl sie bei bestimmten Gruppen sehr starke Effekte zu erzielen vermag). Und Planer von Interventionen und Evaluatoren sind oft nicht daran interessiert, Minimaleffekte im Durchschnitt zu konstatieren, sondern eher daran, herauszufinden, *wo* (d.h. unter welchen Bedingungen und bei welchem Klientel) eine Maßnahme besonders gut wirkt.

3.2.4 Das Problem systematischer Messfehler und von Methodenartefakten

Das Messparadigma quantitativ-standardisierter Sozialforschung zielt auf eine objektive (d.h. beobachterunabhängige), wiederholbare und möglichst fehlerfreie Erfassung individueller Merkmale (also z.Bsp. der Leistungszuwächse von Schulkindern, oder der Zufriedenheit von Leistungsempfängern), die als Indikatoren für den Erfolg bestimmter Interventionen gelten können. Doch erfassen die hierzu verwendeten Instrumente, etwa Fragebögen, tatsächlich das, was sie messen sollen (Leistung, Zufriedenheit…)? Diese notwendige Frage nach der Validität von Messinstrumenten kann mit den verfügbaren teststatistischen Kontrollprozeduren, welche helfen, die Qualität von Fragebogeninstrumenten zu bewerten, in der Regel kaum beantwortet werden. Diese Prozeduren richten sich nämlich in der Regel stets nur auf die Gütekriterien der *Objektivität* und der *Reliabilität*, die mithilfe statistischer Koeffizienten (etwa anhand von Messwiederholungen) statistisch abgesichert werden können. Demgegenüber ist die „Inhaltsvalidität" eines Fragebogens (also das Ausmaß, in dem bspw. ein Instrument zur Lehrevaluation wirklich das Kriterium „Lehrqualität" misst), nicht wirklich rechnerisch erfassbar. Amerikanische Autoren versuchen in jüngerer Zeit diesen Aspekt mit dem Begriff der „*substantive validity*" (Onwuegbuzie et al. 2007, 117) zu erfassen, welcher primär die Qualität der verwendeten Items und Verständnisdivergenzen betrifft. Um mit Hilfe von Fragebögen valide Ergebnisse zu erzielen, müssen Forscher und Befragte nämlich bspw. kulturelle Bedeutungen miteinander teilen (Cicourel 1974, 29), indem sie Begriffe (wie sie bspw. in einem Fragebogen verwendet werden) übereinstimmend verstehen und interpretieren. Weiterhin müssen beide Gruppen auf dieselben Sprachsysteme und Wissensbestände zurückgreifen können. Lassen unpräzise Formulierungen die Möglichkeit für ein unterschiedliches Begriffsverständnis zu und eröffnen sie Interpretationsspielräume, sind die Ergebnisse nicht mehr eindeutig interpretierbar (Prüfer/ Rexroth 2005, 6). Nur dann, wenn alle Befragten alle Items auf dieselbe Weise und in der vom Forscher intendierten Bedeutung verstehen und alle Begriffe eindeutig interpretierbar sind, können valide Ergebnisse erwartet werden. Weiterhin müssen die Befragten über das grundlegende Wissen verfügen, um die Items adäquat beantworten zu können und hinreichend motiviert sein, den Fragebogen (auch ehrlich) zu beantworten (Fowler 1995, 4).

Außerdem können sich Befragte bei der Beantwortung der Items aus verschiedenen Gründen dafür entscheiden, sozial erwünscht zu antworten oder ihre tatsächliche Einstellung nicht ehrlich preiszugeben (Prüfer/ Rexroth 2005,

4), so dass auch in diesen Fällen die Validität der Ergebnisse beschränkt bleibt.

Aus unserer eigenen Forschung können wir hierzu einige empirische Beispiele berichten. In Untersuchungen über Statuspassagen zwischen Hochschulausbildung und Erwerbstätigkeit in der ehemaligen DDR, bei denen ostdeutsche Hochschulabsolventen kurz nach der Wende befragt wurden, wie sie zu ihrer ersten Arbeitsstelle gelangt sind, neigten die Befragten in standardisierten Fragebögen sehr stark dazu, die „offizielle" Darstellung, wonach Hochschulabsolventen in der DDR durch zentrale Vermittlungsbüros bei den Universitäten ihren Arbeitsstellen zugeteilt wurden, zu bestätigen (vgl. Kelle 2008, S. 256f.). In intensiven qualitativen Interviews zeigte sich dann aber, dass der Prozess der Arbeitsvermittlung in der Praxis oftmals ganz anders verlief und die Absolventenvermittlungen vielfach nur eine legitimierende Funktion übernahmen für einen „grauen Arbeitsmarkt", der im Sozialismus in dieser Form gar nicht existieren sollte. Ein anderes Beispiel kommt aus dem Bereich stationärer Altenpflege. Hier zeigen empirische Untersuchungen, dass die zur Messung von „Kundenzufriedenheit" in Altenheimen oft eingesetzten Fragebogen-Instrumente zu grob verzerrten Ergebnissen führen (das heißt konkret: tatsächliche Zufriedenheit überschätzen), weil die Befragten sich oft nicht trauen, ihre Unzufriedenheit mit der Pflegesituation zu benennen (Kelle 2007; Kelle, Niggemann, Metje 2008).

Im Kontext von Evaluation und evidenzbasierter Politik lassen sich solche Fehler in Anknüpfung an eine bekannte, zum Sprichwort geronnene historische Anekdote als „Potemkin-Effekt" bezeichnen: durch elaborierte methodische Werkzeuge wird dann (oft gar nicht beabsichtigt) ein vordergründiger Schein von Veränderung und positiver Entwicklung erzeugt. Solche Probleme werden in einer ausschließlich quantitativen Erhebung häufig nicht sichtbar.

4 Methodenintegration in der Evaluationsforschung

Wie bereits erläutert, wird in der Diskussion um Mixed Methods davon ausgegangen, dass die Kombination quantitativer und qualitativer Erhebungsmethoden das Potential bietet, die Schwächen beider Methodentraditionen durch die Stärken der jeweils anderen zu kompensieren. Die Zielrichtung kann dabei zum einen eine gegenseitige Methodenkritik sein, wenn Methodenprobleme und Validitätsbedrohungen, die bei der qualitativen oder quantitativen Forschung auftauchen, mit Verfahren der jeweils anderen Forschungstradition identifiziert und in vielen Fällen auch behoben werden. Hierbei kann zum anderen eine Ergänzung von Perspektiven erfolgen, weil qualitative und quantitative Methoden den Blick oft auf jeweils unterschiedliche Phänomene oder zumindest unterschiedliche Aspekte desselben Phänomens zu richten helfen (Kelle 2008, S. 54). Diese beiden Funktionen der Methodenkombination, wechselseitige Methodenkritik und Validierung von Ergebnissen einerseits,

Ergänzung von Perspektiven andererseits, widersprechen sich keineswegs, sondern lassen sich oft gemeinsam nutzen – dies möchten wir im letzten Abschnitt unseres Beitrags zuerst methodologisch erläutern und dann anhand eines empirischen Beispiels zeigen.

4.1 Mixed Methods - Validierungsstrategie oder Perspektivenergänzung?

In der methodologischen Debatte wird oft der Begriff der „Triangulation“ für eine Kombination von Methoden verwendet (vgl. Kelle 2008, S. 49; Kelle, Erzberger 2001; Erzberger, Kelle 2002). Diese aus der Trigonometrie und Landvermessung entlehnte Metapher wird allerdings von unterschiedlichen Autoren in verschiedener Weise interpretiert. Eine mit den Arbeiten Donald Campbells begonnene Tradition versteht unter Triangulation die Überprüfung und Validierung von Ergebnissen, die mit bestimmten Verfahren der Datenerhebung und –auswertung gewonnen wurden, mit Hilfe anderer Methoden (Campbell, Fiske 1959). Gegen diese Sichtweise ist eingewandt worden, dass qualitative und quantitative Verfahren Prämissen jeweils unterschiedlicher *Theorie*traditionen in den Forschungsprozess einbringen, und deshalb weniger zur gegenseitigen Validierung als zur gegenseitigen Ergänzung geeignet seien (z.B. Fielding & Fielding 1986, S. 33; Flick 1992). Hier zeigen sich die Grenzen des Triangulationsbegriffs ebenso wie seine systematische Vieldeutigkeit: der Begriff 'Position eines Ortes', klar verständlich im Kontext der Landvermessung, ist in der empirischen Sozialforschung nicht genau definiert. Dabei existieren zwei Lesarten der Triangulationsmetapher: Triangulation als kumulative Validierung von Forschungsergebnissen und Triangulation als Ergänzung von Perspektiven, die eine komplexere Erfassung, Beschreibung und Erklärung eines Gegenstandsbereichs ermöglichen. Erfahrungen aus der Forschungspraxis (Kelle 2008, S. 227ff.) machen jedoch deutlich, dass sich das Verhältnis zwischen qualitativen und quantitativen Forschungsergebnissen nicht auf Grund eines einzelnen Modells bestimmten lässt - etwa in dem Sinne, dass Ergebnisse qualitativer und quantitativer Methoden grundsätzlich konvergent sind und deswegen zur gegenseitigen Validierung verwendet werden können oder in dem Sinn, dass sich qualitative und quantitative Ergebnisse unter jeweils verschiedenen Bedingungen zu einem stimmigen Gesamtbild verbinden lassen. Bei einem parallelen Einsatz qualitativer und quantitativer Verfahren in einem gemeinsamen Untersuchungsdesign sind vielmehr drei Ausgänge möglich (ebd., S. 232). In manchen Fällen *konvergieren* qualitative und quantitative Forschungsergebnisse, in anderen Fällen verhalten sie sich *komplementär* zueinander (ergänzen sich also), manchmal sind sie auch *diver-*

gent, d.h. sie widersprechen sich gegenseitig. Insbesondere das Vorliegen von Divergenzen ist für unsere Überlegungen von besonderem Interesse – in der Regel machen solche Widersprüche nämlich Probleme und Grenzen von Methoden, die einer der beiden Traditionen entstammen, deutlich. Hiermit wird also jeweils ein methodisches Problem aufgeworfen, wobei eine Erklärung der Divergenz (die ggf. eine Hinzuziehung weiterer externer Gesichtspunkte, etwa in Form neuer theoretischer Erklärungen, erfordern kann) dann wiederum zu einer Komplementarität der qualitativen und quantitativen Forschungsergebnisse führt. Das bedeutet: Die zusätzlichen Untersuchungen mit Methode 2, deren Resultate den Ergebnissen von Methode 1 widersprechen, können helfen, die Daten von Methode 1 in einem neuen und anderen Licht zu beurteilen und zu interpretieren, so dass beide Befunde aufeinander bezogen werden können und auf diese Weise ein neues und vollständigeres Bild des untersuchten Gegenstandsbereichs ergeben.

Dies lässt sich gut zeigen anhand empirischer Beispiele aus Projekten, in denen Methodenkombination genutzt wurde, um die Begrenzungen, welche die „Monomethodenforschung“ bei der Erfassung und Beschreibung konkreter empirischer Phänomene hat, zu erkennen und auszugleichen. Wir möchten uns dabei im Folgenden auf die Nutzung qualitativer Forschung für die Identifikation von Validitätsbedrohungen und Methodenproblemen quantitativer Methoden konzentrieren, weil eine solche Form der Methodenkombination gerade in quantitativen Evaluationsprojekten bislang nur selten systematisch genutzt wird.

4.2 Qualitative Forschung zur Identifikation von Methodenproblemen und Validitätsbedrohungen quantitativer Forschung

Quantitative Methoden reichen oftmals nicht aus, um alle potentiellen Wirkungen von Interventionsmaßnahmen zu erfassen, weil zur Konstruktion der in der quantitativen Forschung verwendeten Instrumente immer die interessierenden Effekte *vor* einer empirischen Datenerhebung expliziert werden müssen. Somit sind quantitative Methoden nicht geeignet, um unbekannte Phänomene und Handlungsweisen von Akteuren im Feld zu entdecken. Wenn diese Handlungen dann in der oben beschriebenen Weise zu Störvariablen werden, die dem Effekt der unabhängigen Variablen (also der Intervention) entgegen wirken, steigt das Risiko, dass die Nullhypothese (also die Annahme, dass die Maßnahme keinen Erfolg erzielt hat) beibehalten wird und die Hypothese, dass die Maßnahme einen positiven Effekt erzeugt hat, dann zurückgewiesen

werden muss (obwohl der Effekt tatsächlich nur durch den Einfluss von Störvariablen, die in vielen Fällen eliminiert werden könnten, vermindert wird).

Eine wesentliche Stärke qualitativer Verfahren in der empirischen Sozialforschung besteht darin, dass hiermit bislang *unbekannte und nicht in den Vorannahmen des Forschers explizit benannte Phänomene* systematisch in den Blick genommen werden können. Auf diese Weise können etwa die (unbeabsichtigten) erwünschten und unerwünschten Nebeneffekte von Interventionen untersucht werden.

Werden bei der Evaluation eines Modellprojektes wie „Jedem Kind ein Instrument" nicht nur standardisierte Fragebögen und andere Messinstrumente eingesetzt zur Erhebung vorab definierter Variablen, deren Beeinflussung man sich durch die Intervention erhofft (wie etwa die bessere Befähigung zur Stressbewältigung, die verbesserte auditive Wahrnehmung die musikalischen Präferenzen u.a.m. betreffend), sondern auch Verfahren der qualitativen Datenerhebung (wie offene, teilstandardisierte Interviews), kann hiermit ein breites Spektrum an Effekten exploriert werden, die die Maßnahme aus Sicht der Betroffenen (also der unterrichteten Kinder, ihrer Eltern, den Lehrern und Schulleitungen) hat. Manche unerwünschten Nebeneffekte von sozial- und bildungspolitischen Maßnahmen (etwa ein erhöhtes Mobbingrisiko für hörgeschädigte Kinder in Integrationsklassen) lässt sich nur mit Hilfe von qualitativen Methoden (also etwa durch die Erhebung von Tiefeninterviews oder durch intensive teilnehmende Beobachtung vor Ort) überhaupt beobachten und einer wissenschaftlichen Untersuchung zugänglich machen.

Insbesondere die *Entdeckung unbekannter Störvariablen* durch die Handlungen von Akteuren, die der Erreichung der Programmziele beabsichtigt oder unbeabsichtigt entgegen wirken, ist ohne einen intensiven Kontakt mit dem Forschungsfeld oft gar nicht möglich. Die beteiligten Forscher müssen hier oft vor Ort Beobachtungen machen und Gespräche führen und zum Teil Beziehungen zu den Akteuren im Feld aufbauen, um Informationen über Prozesse zu erhalten, die die Umsetzung von Maßnahmen blockieren, behindern oder verzögern können.

Dies ist vor allem dann der Fall, wenn die untersuchten Interventionen noch nicht in der Sphäre der öffentlichen Meinung mit Rede und Gegenrede debattiert werden (wie dies etwa bei den Maßnahmen des „busing" der Fall war, wo eine empirische Untersuchung der Gegenstrategien von betroffenen Akteuren nicht ohne weiteres durch eine Inhaltsanalyse von Medienberichten erfolgen kann). Wenn das Untersuchungsfeld etwa eine bestimmte Organisation bildet, in der eine Maßnahme durchgeführt wird, dürfen etwa die „mikropolitischen" Strategien der beteiligten Akteure nicht vernachlässigt werden – dies ist oft ohne professionelle ethnographische Strategien des Feldzugangs gar nicht möglich.

Schließlich lassen sich qualitative Methoden auch gut dafür einsetzen, um *Quellen unbeobachteter Heterogenität* zu finden. Werden, angeleitet durch eine geschickte Fallkontrastierung nach der Methode des von Glaser und Strauss

vorgeschlagenen „*theoretical sampling*“ (Glaser, Strauss 1967, S. 45 ff.) systematisch unterschiedliche Akteure befragt und verschiedene Situationen und *settings* im Feld vergleichend analysiert, lassen sich auf diese Weise Informationen sammeln über unterschiedliche Bedingungskonstellationen, unter denen bestimmte Maßnahmen besonders gut oder besonders schlecht wirken.

Im Fall des Modellprojekts „Jedem Kind ein Instrument“ würde das bedeuten, dass mit Angehörigen der verschiedenen Akteursgruppen Interviews durchgeführt und in sehr unterschiedlichen Schulen und Unterrichtssituationen Beobachtungen gemacht werden. Auch hier hat ein Mixed Methods Design wiederum besondere Stärken gegenüber einem Monomethodendesign: wenn mit Hilfe quantitativer Untersuchungen und Mittelwertsvergleichen systematisch Schulen oder Schülergruppen, in denen das Programm besonders großen Erfolg und solche Einrichtungen und Schülergruppen, in denen es geringen oder keinen Erfolg bezogen auf die untersuchten outcome-Variablen aufweist, verglichen werden, kann in einem nächsten Untersuchungsschritt mit qualitativen Methoden untersucht werden, *aus welchen Gründen* (das heißt: unter welchen Bedingungen) die Intervention erfolgreich ist oder nicht.

Darüber hinaus können im Rahmen der Evaluationsforschung systematische Messfehler und Methodenartefakte standardisierter Messinstrumente durch ergänzend eingesetzte qualitative Verfahren aufgespürt und auch beseitigt werden. Dabei kann die gesamte Palette der in der Literatur seit langem bekannten Methodenprobleme standardisierter Befragung mit unterschiedlichen qualitativen Verfahren aufgedeckt und bearbeitet werden: Verständnisprobleme der Befragten, Interviewereffekte, Effekte sozialer Erwünschtheit u.a.m. lassen sich auf diese Weise in den Blick nehmen und es können Maßnahmen getroffen werden, um diese Probleme in den Griff zu bekommen.

Validitätsprobleme, die quantitative Befragungen bei der Erhebung der Zufriedenheit von Bewohnern stationärer Einrichtungen der Altenpflege aufweisen (siehe oben) konnten wir bspw. in einer Mixed Methods Studie detailliert beschreiben und darstellen. In dieser Untersuchung wurden qualitative und quantitative Methoden kombiniert, um die Qualität von stationären Pflegedienstleistungen zu untersuchen. Hierzu wurden qualitative Leitfadeninterviews sowie eine Befragung mit einem standardisierten Fragebogen in einer Reihe von Pflegeeinrichtungen mit unterschiedlicher Bettenzahl und Trägerschaft in ganz Deutschland durchgeführt. Ein Teil der standardisierten *face-to-face* Interviews wurde dabei auf Tonträger aufgezeichnet und transkribiert. Erst durch diese Kombination qualitativer und quantitativer Methoden konnte aufgedeckt werden, dass eine standardisierte Befragung bezogen auf Variablen, mit denen die Zufriedenheit der Heimbewohner mit bestimmten Dienstleistungen der Einrichtung gemessen werden sollten, ein aus der Perspektive der untersuchten Heimbewohner hochgradig verzerrtes Bild wiedergab: Zahlreiche Befragte neigen in standardisierten Interviews systematisch dazu, aus Furcht vor Sanktionen Kritik und Unzufriedenheit an der Einrichtung und dem Pflegepersonal zu verschweigen – in qualitativen Intensivinterviews werden solche Probleme aber dann sichtbar und

einer Untersuchung zugänglich (vgl. Kelle, Niggemann 2002; Kelle, Niggemann, Metje 2008).

Im Folgenden möchten wir noch einmal anhand eines eigenen empirischen Beispiels aus der Evaluationsforschung an Hochschulen darstellen, wie die Beschränkungen standardisierter Instrumente, die in Monomethodendesigns nahezu zwangsläufig unentdeckt bleiben, durch eine Kombination qualitativer mit quantitativen Methoden bearbeitbar werden.

4.3 Ein empirisches Beispiel: Die Identifizierung von Itemverständnisproblemen durch qualitative Methoden

Wie wir schon oben ausgeführt haben, stellt in quantitativen Untersuchungen die Verständlichkeit der Items ein grundlegendes Qualitätskriterium dar. Valide Ergebnisse sind nur dann zu erheben, wenn die Items so präzise formuliert sind, dass sie von allen Befragten auf die gleiche Weise verstanden werden und wenn ausreichendes *Wissen* zur angemessenen Beantwortung vorhanden ist (Fowler 1995, 4). Validitätsbedrohungen durch divergierende kognitive Prozesse der Befragten können allerdings nur dann identifiziert werden, wenn qualitative Methoden ergänzend zu den quantitativen Erhebungen eingesetzt werden. Das Mittel der Wahl stellt hier die Technik des kognitiven Interviews dar (vgl. Willis 2005, 42ff.), bei dem die Befragten während des Ausfüllens eines Fragebogens entweder gebeten werden, alle Gedanken laut auszusprechen (Think-Aloud-Technik), die Frage in eigenen Worten wiederzugeben, die Verlässlichkeit ihrer Antworten zu bewerten oder komplizierte Begriffe zu erklären (in der Methodenliteratur als Probing bezeichnet). Offenbaren sich dabei Verständnisdivergenzen zwischen den Befragten, müssen die entsprechenden Items überarbeitet werden.

Am Beispiel eines Fragebogenitems zur Lehrveranstaltungsevaluation soll gezeigt werden, wie die Itemqualität durch qualitative Methoden überprüft und verbessert werden kann. Eine Aussage aus den FEVOR- und FESEM-Fragebögen (Staufenbiel 2000), die sich in ähnlicher Weise auch in anderen gebräuchlichen Erhebungsinstrumenten zur Lehrevaluation findet, lautet: „Die Vorlesung/ Das Seminar gibt einen guten Überblick über das Themengebiet.“

In den Ausführungen der befragten Studierenden wird deutlich, dass sie das Item nur angemessen beantworten können, wenn sie selbst schon einen Überblick über die Breite und Komplexität des Themengebietes gewonnen haben. Somit ist das Item besonders für Studierende in den Anfangssemestern kaum verwendbar und auch in höheren Semestern wird der Horizont der Studierenden häufig nicht so weit über die Thematik der Veranstaltung hinausge-

hen, dass eine valide Einschätzung vorgenommen werden kann. In den Interviewpassagen finden sich dazu u.a. folgende Aussagen[2]:

> *„je, nachdem, welchen Dozenten sie haben, ist es auch sehr einseitig ausgerichtet. Zum Beispiel [...] der Herr B. [...], der ist halt ein großer Anhänger der Verhaltenstherapie und die Tiefenpsychologie wird so ganz ausgeblendet. Aber am Anfang merken sie das gar nicht. [...] durch Praktika oder dadurch, dass man etwas mehr Zugang zur Materie kriegt, merkt man, dass es vielleicht doch ein bisschen einseitig ist.“ (Interview Nr. 7, 4.10.2007, Abs. 30)*

> *„Ja, es ist schwierig. Es ist halt das erste Mal, dass ich mich mit dem Thema befasse. Jetzt würde ich natürlich sagen, es gibt einen sehr guten Überblick, aber, unter Umständen gibt´s ja auch was, was wir nicht behandeln und was ich dann ja auch gar nicht mitkriege.“ (Interview Nr. 18, 7.1.2008)*

> *„Man kann hoffen, dass die Dozenten das auswählen, was es da so alles gibt. Weil man ja selber da noch nicht so den Überblick hat. Und darauf hoffen, dass die eben das Relevante auswählen.“ (Interview Nr. 6, 24.9.2007)*

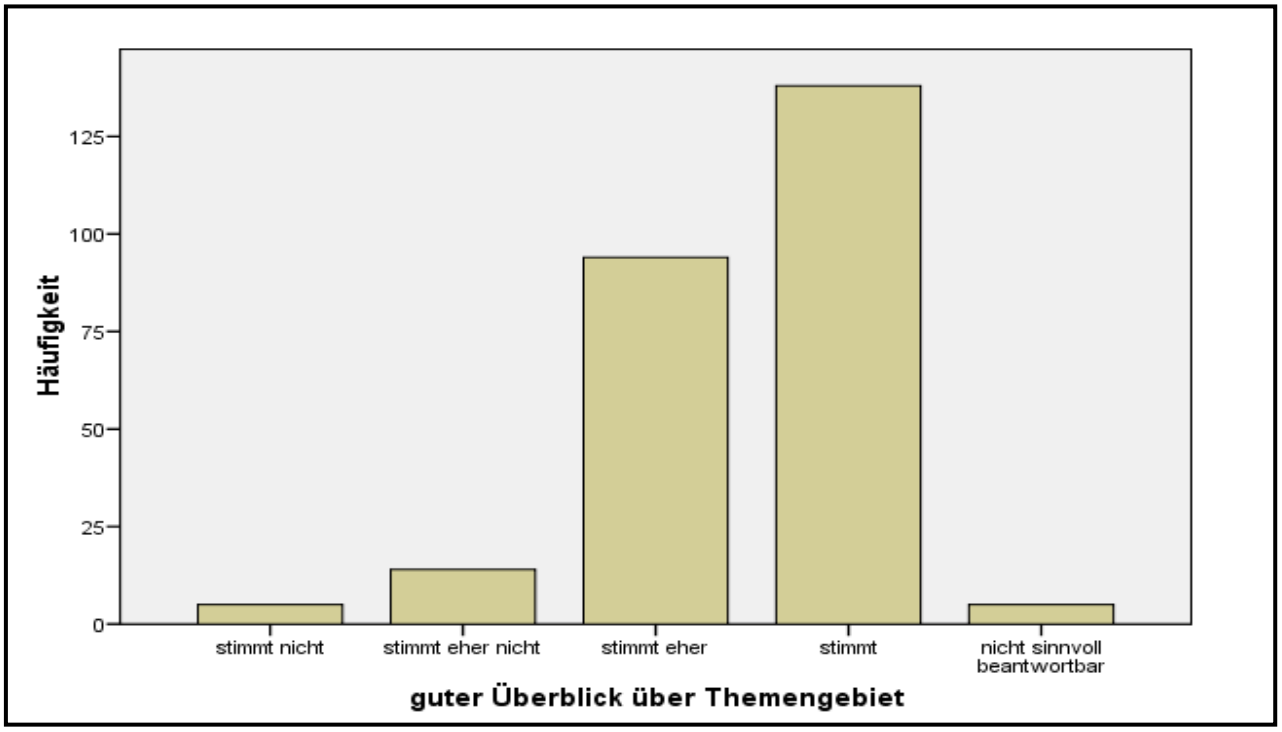
> *„Also, das ist schwer zu sagen, ob es jetzt ein guter Überblick war. Bis jetzt kennen wir ja noch nicht so viel Anderes. Also von daher „stimmt eher“.“ (Interview Nr. 5, 23.8.2007)*

> *„Kann ich bei den meisten Seminaren, glaube ich, nicht abschätzen. Weil mir da das große Ganze fehlt. Für gewöhnlich lernt man den Großteil ja erst im Seminar. Gut, da habe ich jetzt einfach mal Vertrauen. Insgesamt meine ich, es passt. [...] Deshalb "stimmt eher".“ (Interview Nr. 15, 17.12.2007, Abs. 53-58)*

Diese Unsicherheiten der Studierenden werden in den quantitativen Analysen kaum sichtbar. Selbst wenn die Studierenden sich nicht kompetent fühlen, das Item zu beantworten, wählen sie vorwiegend eine zustimmende Antwortalternative und geben dem Fragebogenentwickler nur selten den Hinweis, das Item sei „nicht sinnvoll beantwortbar“. Schauen wir uns dazu die Verteilung an, die sich in einer Evaluation von neun Veranstaltungen an der Universität Marburg ergab (vgl. Grafik 3).

---

2 Metje, 2009, 232ff.



Grafik 3

Lediglich 8% der befragten 357 Studierenden wählte eine ablehnende Antwortalternative und 4% hielt die Aussage für „nicht sinnvoll beantwortbar“. Verlässt sich der Evaluator nun nur auf die quantitativen Ergebnisse, werden die in den qualitativen Interviews offenbar gewordenen Validitätsbedrohungen nicht evident. Nur dann, wenn quantitative und qualitative Daten systematisch aufeinander bezogen werden, können Probleme dieser Art identifiziert werden. Das obengenannte Item sollte letztendlich gänzlich aus dem Fragebogen entfernt werden, weil die Studierenden offenbar nicht in der Lage sind, diesen Qualitätsaspekt von Lehre adäquat einzuschätzen.

5 Zusammenfassung und Ausblick

Mit Evaluationen werden Kausalbeziehungen untersucht, indem in der Regel ausschließlich auf der Basis quantitativer und standardisierter Erhebungsinstrumente die vorab definierten Wirkungen spezifischer Maßnahmen erhoben werden. In unserem Beitrag haben wir gezeigt, dass eine ausschließliche Konzentration des methodischen Instrumentariums auf quantitative Erhebungsinstrumente mit verschiedenen Validitätsbedrohungen einhergehen kann. So können entweder Veränderungen fälschlicherweise konstatiert oder auch übersehen werden oder es werden bestimmte nicht antizipierte Kausalzusammen-

hänge übersehen, weil die Handlungsziele von Akteuren nicht ohne weitere Exploration vorhersehbar und die Handlungsbedingungen variabel sind und darüber hinaus Handlungsregeln, denen die Akteure unterliegen, nur für eng begrenzte, lokale Handlungsfelder gelten oder unbekannten soziokulturellen Veränderungsprozessen unterliegen. Ein theoretisch angenommenes lineares Zusammenhangsmodell zwischen Intervention und vorher definiertem Outcome scheitert damit letztlich in der Realität an der Flexibilität und „Kreativität des Handelns“ (Joas 2002) der Akteure im Feld, die je eigene Interessen und Ziele verfolgen können. Entwickler quantitativer Erhebungsinstrumente können oftmals diese individuellen Wirkungspfade und unbeabsichtigten Nebeneffekte von Interventionen nicht antizipieren und damit auch nicht in Form von Items operationalisieren. An diesem Punkt werden die Vorteile einer Methodenintegration in der Evaluationsforschung offensichtlich: mit qualitativen Methoden können Informationen über die Handlungsziele und Handlungsweisen von Akteuren gesammelt und unvorhersehbare Outcomevariablen definiert, operationalisiert und präzisiert werden. Sollen die Ausmaße solcher Effekte quantifiziert werden, bietet sich dem Forscher jetzt die Möglichkeit, entsprechende Items in den Fragebogen aufzunehmen.

Eine weitere Validitätsbedrohung ausschließlich quantitativer Erhebungen stellen systematische Messfehler und Methodenartefakte dar. Sollen Fragebögen zur Evaluation spezifischer Interventionen entwickelt werden, ist eine Einbeziehung qualitativer Verfahren unverzichtbar, um divergierende Iteminterpretationen der Befragten zu entdecken und uneindeutige Itemformulierungen zu präzisieren. Auch Items, die von den Befragten nicht adäquat beantwortet werden können, weil ihnen das entsprechende Wissen fehlt, können mit kognitiven Interviewtechniken identifiziert und dann entsprechend überarbeitet werden. Allerdings haben diese Validitätsproblematiken unter deutschen Evaluationforscher(inne)n bisher nur wenig Beachtung gefunden. Bisher ist es hierzulande, anders als in Amerika, wo eine Debatte über kognitive Prozesse bei der Beantwortung von Fragebögen schon seit den frühen 1980er Jahren geführt wird, nicht üblich, die Qualität quantitativer Instrumente durch qualitative Daten abzusichern. Unsere Ausführungen sollen deshalb ein Beitrag dazu sein, das Bewusstsein für die Potentiale und die Notwendigkeit einer Methodenintegration zu schärfen. Wie deutlich geworden sein sollte, bietet die Evaluationsforschung zahlreiche Anknüpfungspunkte für die Entwicklung von Konzepten und Methoden der Mixed Methods Forschung – eine stärkere Berücksichtigung der aktuellen Debatten um Methodenintegration würde die Methodenentwicklung in diesem Feld also erheblich voranbringen.

Literatur:

ARNOLD, A. (1972/1662): Die Logik oder die Kunst des Denkens. Darmstadt: Wissenschaftliche Buchgesellschaft.

BABAD, E.; KATZ, Y. (1991): Wishful Thinking – Against All Odds. In: *Journal of Applied Social Psychology*, 21, pp. 1921-1938.

BENECKEN, J.; SPINDLER, S. (2004): Zur psychosozialen Situation stotternder Schulkinder in Allgemeinschulen. In: *Die Sprachheilarbeit*, 49 (2), S. 61 – 70.

BLALOCK, H.M. (1985): Causal Models in the Social Sciences. New York: Aldine.

BRYMAN, A. (2007): Barriers to integrating quantitative and qualitative research. In: *Journal of Mixed Methods Research*, 1, pp. 1-18.

BUNDESMINISTERIUM FÜR BILDUNG UND FORSCHUNG (2001): TIMSS- Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente. Bonn: BMBF publik.

CAMPBELL, D.T.; STANLEY, J.C. (1963): Experimental and Quasi-Experimental Designs for Research. Dallas: Houghton Mifflin.

CAMPBELL, D.T.; FISKE, D.W. (1959): Convergent and Discriminant Validation by the Multitrait-Multimix Matrix. In: *Psychological Bulletin*, 56, pp. 81-105.

CHALMERS, A.F. (2007): Wege der Wissenschaft: Einführung in die Wissenschaftstheorie. Berlin: Springer.

CICOUREL, A. V. (1974): Methode und Messung in der Soziologie. Frankfurt am Main: Suhrkamp.

COLEMAN, J.S. (1976): Liberty and equality in school desegregation. In: *Social Policy,* 6, pp. 9-13.

COLLINGWOOD, R.G. (1937/1938): On the So-called Idea of Causation. In: *Proceedings of the Aristotelian Society*, 38, pp. 85-112.

CRESWELL, J.W.; PLANO CLARK, V.L.; GUTMANN, M.L.; HANSON, W.E. (2003): Advanced mixed methods research designs. In: Tashakkori, A.; Teddlie, C. (Eds.): Handbook of Mixed Methods in Social and Behavioral Sciences. Thousand Oaks, Calif.: Sage, pp. 209-240).

CRESWELL, J.W. (2009): Research design: qualitative, quantitative and mixed methods approaches. Thousand Oaks, Calif.: Sage.

ERZBERGER, C.; KELLE, U. (2002): Making Inferences in Mixed Methods: The Rules of Integration. In: Tashakkori, A. & Teddlie, C. (Eds.). Handbook of mixed methods for the social and behavioural sciences. Thousand Oaks, Calif.: Sage, pp. 457 – 490.

FESTINGER, L. (1957): A theory of cognitive dissonance. Stanford, Cal.: Stanford University Press.

FESTINGER, L.; RIECKEN, H.W.; SCHACHTER, S. (1956): When Prophecy Fails. Minneapolis, Mi.: University of Minnesota Press.

FIELDING, N.G.; FIELDING, J.L. (1986): Linking Data. (*Qualitative Research Methods Vol. 4*). London: Sage.

FLICK, U. (1992): Triangulation Revisited: Strategy of Validation or Alternative? In: *Journal for the Theory of Social Behaviour*, 22, pp. 175-197.

FOWLER, F.J. (1995): Improving survey questions: design and evaluation. Thousand Oaks: Sage.

GLASER, B.; STRAUSS, A. (1967): The Discovery of Grounded Theory. Strategies for qualitative Research. New York: Aldine.

GREENE, J.C. (2007): Mixed Methods in Social inquiry. San Francisco, Calif.: Jossey Bass.

GUBA, E.G.; LINCOLN, Y.S. (1988): Do inquiry paradigms imply inquiry methodologies? In: Fetterman, D.M. (Ed.) Qualitative approaches to evaluation in education: The silent scientific revolution. New York: Praeger, pp. 88-115.

HAGE, J.; MEEKER, B.F. (1988): Social Causality. Boston: Unwin Hyman.

HEINZ, W.R.; MARSHALL, V.W. (2003): Social Dynamics of the Life Course: Transitions, Institutions, and Interrelations. New York: Aldine de Gruyter.

JAHODA, M.; LAZARSFELD, P.F.; ZEISEL, H. (1933/1982): Die Arbeitslosen von Marienthal. Frankfurt: Suhrkamp.

JOAS, H. (2002): Die Kreativität des Handelns. Frankfurt/M.: Suhrkamp.

JOHNSON, B.; TURNER; L.A. (2003): Data Collection Strategies in Mixed Methods Research. In: Tashakkori, A.; Teddlie, C. (Eds.): Handbook of Mixed Methods in Social and Behavioral Sciences. Thousand Oaks, Calif.: Sage, pp. 297-319.

KAHNEMAN, D. (1982/2008): Judgment under uncertainty: heuristics and biases. Cambridge: Cambridge University Press.

KELLE, U. (2006): Combining Qualitative and Quantitative Methods in Research Practice – Purposes and Advantages. In: Gürtler, L.; Huber, G.L. (Eds.). Special Guest Issue on Mixed Methods. Qualitative Research in Psychology. Vol. 3 (4), pp. 293-311.

KELLE, U. (2007): "Kundenorientierung" in der Altenpflege? Potemkinsche Dörfer sozialpolitischen Qualitätsmanagements. In: *PROKLA 146, Zeitschrift für kritische Sozialwissenschaft*, 37 (1), S. 113-128.

KELLE, U. (2008): Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. Wiesbaden: VS Verlag für Sozialwissenschaften.

KELLE, U.; ERZBERGER, C. (2001): Die Integration qualitativer und quantitativer Forschungsergebnisse. In: KLUGE, S.; KELLE, U. (Hg.): Methodeninnovation in der Lebenslaufforschung. Integration qualitativer und quantitativer Verfahren in der Lebenslauf- und Biographieforschung. Weinheim; München: Juventa, S. 89- 133.

KELLE, U.; NIGGEMANN, C. (2002): „Wo ich doch schon einmal vor zwei Jahren verhört worden bin...": Methodische Probleme bei der Befragung von Heimbewohnern. In: Motel-Klingebiel, A.; Kelle, U. (Hg.): Perspektiven der empirischen Alternssoziologie. Opladen: Leske und Budrich, S. 99 – 132.

KELLE, U.; NIGGEMANN, C.; METJE, B. (2008): Datenerhebung in totalen Institutionen als Forschungsgegenstand einer kritischen gerontologischen Sozialforschung. In: Amann, A.; Kolland, F. (Hg.): *Das erzwungene Paradies des Alters?*, S. 163-193.

KLUGE, S.; KELLE, U. (2001) (Hg.): Methodeninnovation in der Lebenslaufforschung. Integration qualitativer und quantitativer Verfahren in der Lebenslauf- und Biographieforschung. (Statuspassagen und Lebenslauf, Band 4). Weinheim; München: Juventa.

KRANEFELD, U. (2009): Perspektivwechsel: Den musikalischen Denkwegen der Schülerinnen und Schülern folgen. In: Greuel, T., Kranefeld, U.; Szczepaniak, E. (Hg.): Jedem Kind (s)ein Instrument. Die Musikschule in der Grundschule. (Musik im Diskurs; 23). Aachen: Shaker.

LAMNEK, S. (1995): Qualitative Sozialforschung. Bd. 1: Methodologie. Weinheim: Beltz, Psychologie-Verlags-Union.

LANDESHOCHSCHULGESETZE:

http://www.hof.uni-halle.de/steuerung/lhg_uebersicht.htm, letzter Zugriff am 6.4.2010)

LAZARSFELD, P.F. (1955): Interpretation of Statistical Relations as Research Operation. In: Lazarsfeld, P.D.; Rosenberg, M. (Eds.): The Language of Social Research. New York: John Wiley and Sons, pp. 115-125.

LINCOLN, Y.S.; GUBA, E.G. (2000): Paradigmatic controversies, contradictions, and emerging confluences. In: Denzin, N.K.; Lincoln, Y.S. (Eds.): Handbook of Qualitative Research. Thousand Oaks: Sage, pp. 163-188.

METJE, B. (2009): Evaluation universitärer Lehrveranstaltungen. Standpunkte von Studierenden und Systemdefizite im Fokus. Marburg: Tectum.

ONWUEGBUZIE, A.J.; WITCHER, A.E.; COLLINS, K.M.T.; FILER, J.D.; WIEDMAIER, C.D.; MOORE, C.W. (2007): Students´ Perceptions of Characteristics of Effective College Teachers: A Validity Study of a Teaching Evaluation Form Using a Mixed-Methods Analysis. In: *American Educational Research Journal,* 44 (1), pp. 113-160.

PRENZEL, M.; ARTELT, C.; BAUMERT, J.; BLUM, W.; HAMMANN, M.; KLIEME, E.; PEKRUN, R. (2007) (Hg.): Pisa 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie. Münster; New York; München; Berlin: Waxmann.

PRÜFER, P.; REXROTH, M. (2005): Kognitive Interviews. ZUMA How-to-Reihe, Nr. 15. [http://www.gesis.org/Publikationen/Berichte/ZUMA_How_to/Dokumente/pdf/How_to15PP_MR.pdf, 14.3.2007]

ROETHLISBERGER, F.J.; DICKSON, W.J. (1939): Management and the worker. Cambridge, Ma.: Harvard University Press.

SCHÜTZ, A. (1971): Gesammelte Aufsätze. Bd. 1: Das Problem der sozialen Wirklichkeit. Den Haag: Nijhoff.

SEIPEL, C.; RIEKER, P. (2003): Integrative Sozialforschung. Konzepte und Methoden der qualitativen und quantitativen empirischen Forschung. Weinheim; München: Juventa.

SIMON, H.A. (1954): Spurious Correlation: A Causal Interpretation. In: *Journal of the American Statistical Association*, 49, pp. 467-479.

SIMPSON, C.H. (1951): The Interpretation of Interaction in Contingency Tables. In: *Journal of the Royal Statistical Society*, 13, pp. 238-241.

STAUFENBIEL, T. (2000): Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. In: *Diagnostica*, 46 (4), S. 169-181.

STRAUSS, A.; CORBIN, J. (1990): Basics of Qualitative Research. Grounded Theory Procedures and Techniques. Newbury Park, Calif.: Sage.

TASHAKKORI, A.; TEDDLIE, C. (2003) (Eds.): Handbook of Mixed Methods in Social and Behavioral Sciences. Thousand Oaks, Calif.: Sage.

WHITE, H. (2006): Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank. Washington D.C.: World Bank.

WILLIS, G. B. (2005): Cognitive interviewing: a tool for improving questionnaire design. Thousand Oaks; London; New Delhi: Sage.

ZIMBARDO, P.G. (1969): The human choice: Individuation, reason and order versus deindividuation, impulse and chaos. In: Arnold, W.T.; Levine, D. (Eds.): Nebraska Symposium on Motivation, 17, pp. 237-307. Lincoln, Ne.: University of Nebraska Press.