Martin Wellenreuther

Quantitative Methoden[1]

1. Was ist quantitative Forschung

Quantitative Forschung umfasst (1) deskriptive Forschung (z. B. internationale Vergleichsstudien wie TIMSS und PISA), (2) hypothesenprüfende Forschung (vor allem Experimente und Längsschnittuntersuchungen) sowie (3) Evaluations- und Entwicklungsforschung (Entwicklung und Prüfung pädagogischer Programme). Sie wird von qualitativer Forschung abgegrenzt, zu der z. B. explorative Studien, naturalistische Feldstudien, ethnographische Studien und Fallstudien zählen. Quantitative und qualitative Forschungen ergänzen sich gegenseitig. Häufig bereiten qualitative Forschungen quantitative Forschungen vor. »Kontroverse Diskussionen bewegen sich nicht mehr an der Trennlinie ›quantitativ / qualitativ‹, sondern richtigerweise entlang der Unterscheidung von guter Forschung / schlechter Forschung« (Terhart 1997, S. 33).

Quantitative pädagogische Forschung zielt auf die Weiterentwicklung unserer Erkenntnisse über die pädagogische Wirklichkeit. Verschiedenen Fragestellungen sind unterschiedliche Forschungstypen zugeordnet:

(1) Was ereignet sich? Wie sieht die Wirklichkeit aus? (Deskriptive Forschung)

Beispiel: Welches Kenntnis- und Kompetenzniveau haben deutsche und japanische Schüler in einem bestimmten Fach am Ende der 9. Klasse?

(2) Was sind die systematischen (kausalen) Prozesse bzw. Zusammenhänge? Und welche Mechanismen stehen dahinter? (Hypothesenprüfende Forschung)

Beispiel: Welche Faktoren steuern in den Phasen des Leselernens die Lesefähigkeit?

(3) Wie sind diese Prozessfaktoren miteinander zu kombinieren, um mit möglichst geringen Kosten die gewünschten Effekte zu erzielen? (Entwicklungs- bzw. Evaluationsforschung)

Beispiel: Wie können Faktoren, die in einer bestimmten Phase des Leselernprozesses eine positive Wirkung haben, am besten miteinander in einem Lesetraining kombiniert werden?

Diese Fragen sind miteinander verbunden. Nach der Identifikation eines Problems im Rahmen deskriptiver Forschung (z. B. Kinder aus bildungsfernen Schichten werden in Deutschland vergleichsweise schlecht gefördert) kann durch experimentelle Forschung geklärt werden, welche Bedingungen und Ursachen zu Grunde liegen, um dann aufgrund dieser Erkenntnisse wirksame und kostengünstige Methoden entwickeln zu können.

Weil sie der Begrenzung unseres Wahrnehmungsapparates bei der Beurteilung der Wirksamkeit einer Methode Rechnung tragen, bieten experimentelle Methoden eher als die Praxiserfahrungen von Pädagogen eine Gewähr, dass die Spreu (falsche Theorien; unwirksame Methoden) vom Weizen getrennt wird. Schon Herbart hat auf diese Begrenzung hingewiesen: »Wollten wir nur sämtlich bedenken, dass jeder nur erfährt, was er versucht, ein neunzigjähriger Dorfschulmeister hat die Erfahrung seines neunzigjährigen Schlendrians, er hat das Gefühl seiner langen Mühe. Aber hat er auch die Kritik seiner Leistungen und seiner Methode?« (Herbart 1982, Bd. 2, S. 19).[2] Lehrer mögen es als Erfolg werten, wenn ihre Schüler nach 10 Stunden Stationenarbeit im Durchschnitt einen Testpunktwert von 20 Punkten erreichen. Dieses Ergebnis wird erst dann kritischer bewertet, wenn gezeigt wird, dass unter vergleichbaren Voraussetzungen durch direkte Instruktion ein um 10 Punkte höheres Ergebnis erzielt wird (vgl. Wellenreuther 2004).

Es gelingt dem alten Dorfschulmeister nicht, die relevanten Bedingungen unter fairen Voraussetzungen systematisch miteinander zu vergleichen. Dazu müsste er zunächst vergleichbare Klassen bilden, die dann den konkurrierenden Bedingungen bzw. Programmen ausgesetzt werden würden. Der Dorfschulmeister wendet in der Regel immer eine Methode an, wenn er am Ende eines Schuljahres auf die Erfolge seiner Methode verweist. Ob und in welchem Maße die festgestellten Wirkungen durch seinen Unterricht oder durch andere Bedingungen verursacht wurden (z.B. Anregungen durch Eltern, Freunde, Nachhilfe oder »Reifung«), kann er nicht sagen. Vor allem aber kann er nicht sagen, wie sich alternative Methoden ausgewirkt hätten.

2. Geschichte quantitativer pädagogischer Forschung

Quantitative empirische Forschung hat ihre Wurzeln in den Arbeiten von D. Hume über das Verständnis von Kausalität. Nach Valentine/Cooper (2005) kann Hume's Verständnis von Kausalität grob in folgender Weise charakterisiert werden: Ursache und Wirkung müssen (a) miteinander kovariieren bzw. korrelieren, (b) die »Ursache« muss der Wirkung vorhergehen, und (c) für den Zusammenhang darf es keine relevanten alternativen Erklärungen geben. Für die Methode des Experiments waren die Vorarbeiten von Stuart Mill maßgebend. Die Verbindung seiner Methode der Übereinstimmung mit der Methode der Differenz lässt sich in jedem Experiment wiederfinden, in dem geprüft wird, ob sich Gruppen entsprechend der gemachten Vorhersagen voneinander unterscheiden. Die Methode der strengen Prüfung gehaltvoller, kühner Theorien steht im Mittelpunkt des Erkenntnisprogramms, das von K.R. Popper in seiner »Logik der Forschung« entwickelt wurde. Er grenzt hier quantitative Forschung nicht nur von essentialistischen und holistischen Ansätzen ab, die in manchen qualitativen Ansätzen zum Tragen kommen, sondern auch von empiristischer Fliegenbeinzählerei. Danach können wir durch strenge experimentelle Prüfungen niemals die Wahrheit von Theorien endgültig – z.B. durch induktives Schließen – beweisen, sondern lediglich zeigen, welche Theorien wiederholt bei strengen Prüfungen bestätigt wurden oder gescheitert sind. Für die statistische Auswertung quantitativer Forschungen waren ferner die Arbeiten von R. Fisher über das Experiment mit Zufallszuteilung von Versuchseinheiten grundlegend. Die konkreten Konsequenzen für die Planung und Durchführung von Experimenten in der Pädagogik wurden von D.T. Campbell und J.C. Stanley (1963) ausgearbeitet. In jüngster Zeit wurde diese experimentelle Methodologie durch die Methode der Meta-Analyse ergänzt, in der die Ergebnisse von Experimenten in einem bestimmten Bereich systematisch ausgewertet werden (vgl. Glass 2000).

Das Feld der quantitativen empirisch-pädagogischen Forschung hat sich seit den sechziger Jahren des vorigen Jahrhunderts erheblich differenziert.

- Im Bereich der Wissenschaftstheorie gab es den sog. Positivismusstreit, der zwischen Vertretern der sog. »kritischen Theorie« (z.B. J. Habermas und T. Adorno) und des kritischen Rationalismus (z.B. K.R. Popper und H. Albert) ausgetragen wurde (vgl. Adorno/Dahrendorf/Pilot/Albert/Habermas/Popper 1969/1972). Diesem folgte die Auseinandersetzung zwischen verschiedenen Formen des Konstruktivismus (vgl. Anderson/Reder/Simon 1999).
- Im Bereich der Messung zeigte sich diese Ausdifferenzierung in einer kritischen Bewertung von »paper and pencil Tests« (klassische Fragebogentests; vgl. Bühner 2006) und einer stärkeren Gewichtung von nicht-reaktiven Messverfahren (z.B. Inhaltsanalysen, Beobachtungsverfahren; vgl. Webb/Campbell/Schwarz/Seechrest 1966) sowie in der Entwicklung probabilistischer Testmodelle als Alternative zur klassischen Testtheorie (vgl. Bühner 2006).

- Im Bereich der Forschung gab es – teilweise als Folge des Sputnik-Schocks – ein Aufblühen der Curriculumforschung, deren Ergebnisse allerdings die in sie gesetzten hohen Erwartungen nicht erfüllte (vgl. Cronbach/Ambron/ Dornbusch/ Hess/ Hornik/ Phillips/ Walker/ Weiner 1980; Hager 1995). Als Alternative zur klassischen Entwicklungsforschung (vgl. 3.3.) wurden Modelle der Handlungsforschung[3] vorgeschlagen sowie das »design experiment« (vgl. Brown 1992; Levin/O'Donnell 1999) entwickelt. Beide Ansätze forderten, dass der Forscher handelnd in den Forschungsprozess eingreifen darf.
- Im Bereich international-vergleichender pädagogischer Forschung haben Ländervergleiche in zentralen Leistungsbereichen wie »reading literacy«, Mathematik und Naturwissenschaften (vgl. TIMSS und PISA) sowie die dazu begleitend durchgeführten Untersuchungen zu Lernkulturen in verschiedenen Ländern eine wachsende Bedeutung (vgl. Stevenson/Stigler 1992; Stigler/Hiebert 1998).

Diese Phase der Differenzierung, die durch konkurrierende Forschungsparadigmen und heftige Auseinandersetzungen um die einzig richtige Methode gekennzeichnet war, scheint nun in eine Phase der Konsolidierung getreten zu sein. Dies zeigt sich in einer Annäherung wissenschaftstheoretischer Positionen (vgl. Anderson/Greeno/Reder/Simon 2000). Ferner werden qualitative Verfahren nicht mehr als Gegenpol zu quantitativen Verfahren verstanden, sondern in systematischer Weise in quantitative Forschungsprogramme integriert. Zur abschließenden Beurteilung der Wirksamkeit eines Programms sind dann echte Experimente unerlässlich (vgl. dazu Feuer/Towne/Shavelson 2002; Phye/Robinson/Levin 2005; Raudenbush 2004).

Der Streit um das richtige Forschungsparadigma hat in verschiedenen Ländern unterschiedliche Spuren hinterlassen. So gibt es einige Hinweise, dass eine partielle Abkehr vom quantitativen Paradigma in den USA erst vergleichsweise spät in den 80er-Jahren des letzten Jahrhunderts eintrat (vgl. Gersten/Baker/Smith-Johnson/Flojo/Hagan-Burke 2004; Hsieh/ Acee/ Chung/ Hsieh/ Kim/ Thomas/ You/ Levin/ Robinson 2005). So ist der Anteil echter pädagogischer Experimente in der wichtigsten Zeitschrift für empirisch-pädagogische Forschung – im »Journal of Educational Psychology« – seit 1983 stetig zurück gegangen (vgl. Hsieh et al. 2005, S. 527).

In Deutschland gab es schon früh Vertreter einer empirischen Pädagogik wie z.B. Wilhelm August Lay, Ernst Meumann und Peter Petersen, doch haben deren Arbeiten kaum Einfluss auf eine geisteswissenschaftlich geprägte Pädagogik genommen. Quantitative Forschung hat im Zuge der Bildungsexpansion zwischen 1965 und 1975 eine kurze Blüte erfahren (Sinnbild dafür ist das Schlagwort von H. Roth (1962) » Die realistische Wende in der Erziehungswissenschaft«). Danach war quantitative Forschung aufgrund der Studentenbewegung und der sich durchsetzenden Frankfurter Schule (»Kritische Theorie«) für die deutsche Pädagogik weitgehend bedeutungslos. Hypothesenprüfende, experimentelle Forschung geriet unter Ideologieverdacht: Durch sie würde die Lebenswelt unzulässig zergliedert, die Engführung durch Hypothesen würde die wahren Zusammenhänge verschleiern. Dabei wurde unterstellt, nur durch die Brille der kritischen Theorie wäre ein adäquates Erfassen der Wirklichkeit möglich (vgl. Terhart 1997, S. 27).[4] In dem Bemühen, sich durch Verwendung qualitativer Methoden den anderen Wissenschaften wie der Soziologie oder der Ethnologie anzunähern, »musste der hohe Preis einer Aufgabe der bisherigen Identität gezahlt werden, in der der leitende Begriff des ›Pädagogischen‹ verloren ging«(Oelkers 1998, S. 224).

Die Konsequenzen für die Schulpädagogik sind noch bis heute gravierend. Wer sich mit den Prozessen des Lehrens und Lernens im Unterricht befasste, galt als Technokrat. Statt auf strenge empirische Belege stützte man sich auf eine Ratgeberliteratur (wichtigste Repräsentanten: Hilbert Meyer und Herbert Gudjons). Offene Lernformen wurden propagiert, andere Unterrichtsmethoden (Frontalunterricht bzw. direkte Instruktion) wurden verdammt. Eine Unterrichtsmethode war theoretisch gerechtfertigt, wenn sie Selbstständigkeit und Mündigkeit zu

fördern schien. Der Wechsel vom quantitativen zum qualitativen Forschungsparadigma war in Wahrheit die Durchsetzung einer Position[5], in deren Rahmen Forschungstypen wie »Querschnittsuntersuchungen [dominierten], während experimentell angelegte Studien und Interventionsstudien nur eine schmale Bedeutung hatten« (Leschinsky 2004, S. 79). Die Ergebnisse dieser »kritischen Pädagogik« sind – z.B. im Hinblick auf die Realisierung von Chancengleichheit – im internationalen Maßstab ernüchternd.

3. Forschungstypen und Gütekriterien

Im Folgenden gehe ich auf die drei Grundtypen quantitativer Forschung in der Pädagogik ein (vgl. Wellenreuther 2000), weil je nach Forschungstyp unterschiedliche Gütemaßstäbe gelten.

3.1. Deskriptive Forschung

Beispiele sind internationalen Vergleichsuntersuchungen zur Wirksamkeit des deutschen Bildungswesens wie TIMSS[6] und PISA[7]. Solche Untersuchungen haben das Ziel, möglichst präzise den Leistungsstand deutscher Schüler im Vergleich zu Schülern anderer Länder zu beschreiben. Für solche präzisen Beschreibungen sollten mindestens zwei Kriterien erfüllt sein:

- *Repräsentativität:* Die Stichproben der Schüler eines Landes sollten innerhalb festgelegter Fehlergrenzen Aussagen über alle Schüler diesen Landes erlauben, d.h. sie sollten repräsentativ für die Schüler eines Landes sein. Um dieses Kriterium der Repräsentativität zu erreichen, kann man eine Zufallsstichprobe aus allen Schülern einer bestimmten Klassenstufe ziehen.
- *Messvalidität:* Die Messinstrumente zur Erfassung der Leistungen und Kompetenzen der Schüler sollten in fairer Weise ein internationales Kerncurriculum präzise und gültig erfassen. Die Gültigkeit der Messinstrumente sollte durch zusätzliche empirische Forschungen belegt werden (vgl. das Beispiel in Box 1).

Box 1: Deskriptive Studie bzw. Beobachtungsstudie

Die TIMSS-Videostudie[8] zum Mathematikunterricht in den USA, Japan und Deutschland

Ziel: Hinweise für Erklärungen der Leistungsunterschiede zwischen Japan, den USA und der BRD finden.

Methode der Untersuchung: In die Stichprobe gelangten 50 japanische Schulklassen der achten Klassenstufe, 81 Klassen aus den USA und 100 Klassen aus Deutschland. Man bemühte sich um Repräsentativität, wobei die Klassen per Zufall aus den Klassen der TIMSS-Hauptstudie gezogen werden sollten.

Methodische Probleme: Aufgrund von Teilnahmeabsagen mussten in Deutschland die Hälfte der Klassen durch vorher festgelegte Ersatzklassen ersetzt werden. In den USA haben statt der geplanten hundert Klassen nur 81 teilgenommen. In Japan wurde überwiegend Geometrieunterricht aufgezeichnet (in Japan 78% der aufgezeichneten Stunden, in den USA 12%). In Deutschland und in den USA wurde in den Klassen, in denen die Videoaufzeichnungen durchgeführt wurden, auch später die Leistung durch Tests ermittelt. In Japan wurden die Klassen entweder vom Schulleiter oder vom nationalen japanischen Institut für Bildungsforschung ausgewählt. Es kann nicht geklärt werden, in welchem Umfang durch diese Besonderheiten der Stichprobenbildung die Repräsentativität der Ergebnisse gefährdet wurde.

Kodierung der Ergebnisse: Zur Analyse der einzelnen Stunden wurden allgemeine Beobachtungskategorien gebildet, die mit ausreichender Übereinstimmung (mindestens 80%) von unabhängigen Beobachtern erfasst werden konnten.

Voruntersuchungen: Um eine hohe Aufnahmequalität einer Stunde in jedem Land sicherzustellen, wurden vor Beginn der Hauptuntersuchung insgesamt 28 Aufnahmen in den drei Ländern durchgeführt. 30 Unterrichtsprotokolle aus jedem Land wurden so tabellarisch zusammengefasst, dass alle Hinweise auf das Land entfielen (Blindversuch). Diese Protokolle wurden dann von amerikanischen Mathematikern und Mathematikdidaktikern hinsichtlich ihrer Qualität beurteilt. Außerdem wurden die Lehrer zusätzlich noch befragt.

Ergebnisse: Die Experten bescheinigten 30% des japanischen Unterrichts eine hohe Qualität (57% eine mittlere Qualität), verglichen mit 23% des deutschen und 0 % des amerikanischen Unterrichts. Während im deutschen Unterricht 89 % der Unterrichtszeit (USA 95%) der Stunden auf die Einübung von Routineprozeduren verwandt wurde, betrug der entsprechende Prozentsatz in Japan 42%. Hier wurde viel mehr Zeit auf die Anwendung mathematischer Konzepte (Japan 14%, BRD 6%, USA 5%) sowie auf Problemlöse- und Denkaufgaben verwandt (Japan 44%, Deutschland 5% und USA 0%).

Grenzen: Auch wenn die Stichproben völlig repräsentativ wären, bliebe eine Interpretation der einzelnen Ergebnisse schwierig, weil alternative Erklärungen nicht ausgeschlossen werden können.

Es gibt außerdem deskriptive Forschungen, die der Dokumentation herausragender Fälle dienen (z.B. »best practice« Leseunterricht in Neuseeland: Wilkinson/Townsend 2000; Beschreibung des Verhaltens von Expertenlehrern im Vergleich zu Novizen: Leinhardt 1989 oder die Beschreibung herausragender Schulen: Pressley/Gaskins/Solic/Collins 2006).

Wenn wir die Leistungen deutscher Schüler im Vergleich zu Schülern anderer Länder beschreiben, dann wissen wir noch nichts über mögliche Gründe oder Ursachen. Um Anhaltspunkte für die gefundenen Leistungsunterschiede zu erhalten, wurden flankierend zu TIMSS und PISA einige Untersuchungen durchgeführt, in denen der Zusammenhang zwischen bestimmten sozialen Merkmalen wie Sozialschicht, Lesekonsum und Geschlecht mit dem erreichten Kompetenzniveau untersucht wurden (vgl. Schümer 2004). Schümer beschrieb z.B., wie eng in Deutschland noch immer der Zusammenhang zwischen sozialer Herkunft und erreichtem Bildungsniveau ist.

Solche Forschungen können Hinweise auf mögliche verursachende Bedingungen geben, zur strengen Prüfung der Ursachen müssen zusätzlich hypothesenprüfende Untersuchungen durchgeführt werden, also vor allem Experimente.

3.2. Hypothesenprüfende Forschung

Diese sind von zentraler Bedeutung für unsere Erkenntnisse über die Wirksamkeit pädagogischer Maßnahmen (vgl. Levin 2005). Joanna Williams, ehemalige Herausgeberin des »Journal of Educational Psychology«, bemerkt zur Glaubwürdigkeit empirisch-pädagogischer Forschung: »The paradigm for instructional research is the intervention, and, without carefully conducted genuine interventions (experiments), we can't really answer the question of whether a particular practice or technique is effective. We need such studies to draw firm conclusions, and we should try hard not to compromise our methodological standards when

we do them… We should remember that in … the context of intervention research, [observational studies, case studies, ethnographies, and correlational studies, MW] remain preliminary studies« (zit. nach Levin/ O'Donnell 1999, S. 281).

Zur Erläuterung der experimentellen Methode vgl. Box 2.

Box 2: Randomisiertes Experiment

Strukturiertes oder offenes Lernarrangement? (Tuovinen/Sweller 1999)

Ziel: Überprüfung der Hypothese, dass bei der Aneignung neuen Wissens eine klar strukturierte Einführung zusammen mit einer Erläuterung an Lösungsbeispielen erheblich effektiver ist als eine offene Vorgehensweise, in welcher der Lerner die zu lernenden Inhalte auswählen darf. Diese Vorhersage wurde aus der Cognitive Load Theorie von Sweller abgeleitet.

Methode: Die Stichprobe bestand aus 32 Lehrerstudenten für das höhere Lehramt, die nach ihrem Vorwissen in Bezug auf das Datenbankprogramm FileMaker Pro in eine Gruppe mit und eine ohne Vorwissen aufgeteilt wurde. Diese beiden Gruppen wurden dann per Zufall auf zwei Gruppen aufgeteilt, eine, in der in vorstrukturierter Weise bestimmte Probleme analog zu vorgegebenen Lösungsbeispielen bearbeitet werden sollten, während die andere Gruppe ohne strukturierte Anleitung ähnliche Probleme wie in der Einführung mit Hilfe des Programms sowie der mitgelieferten Datenfiles zu lösen hatte. Die Trainingszeit wurde in allen Bedingungen konstant gehalten.

Ergebnisse: In der Gruppe ohne Vorwissen konnten diejenigen, die in der Trainingszeit konkrete Hinweise und durch Lösungsbeispiele angeleitet wurden, im anschließenden Test doppelt so viele Aufgaben richtig lösen wie Lehrerstudenten, die explorierend geübt hatten. Hatten die Studenten allerdings Vorwissen mit Datenbankprogrammen, war kein Unterschied mehr zwischen diesen Gruppen statistisch signifikant feststellbar.

Grenzen: In der Wissenschaft verlässt man sich nicht auf ein randomisiertes Experiment, sondern prüft, ob sich der gefundene Zusammenhang auch in anderen Kontexten bestätigen lässt (vgl. Kalyuga/Chandler/Sweller 2001; Kirschner/Sweller/Clark 2006).

In Experimenten sind folgende vier Gütekriterien wichtig:

(1) Interne Validität: Ein Experiment soll möglichst eindeutig nachweisen, dass die variierte Bedingung, und nicht irgendwelche anderen Faktoren, die erfasste Wirkung hervorgebracht hat. In der Forschung wird dazu das Verfahren der Randomisierung (Zufallsaufteilung auf die Behandlungsgruppen) verwendet. Als weniger günstig gelten Verfahren der Parallelisierung (vgl. Campbell/Stanley 1963). Die effektivste Kontrolle findet in Doppelblindversuchen statt. In diesen wissen weder die Versuchspersonen noch die Versuchsleiter, in welcher Bedingung eine Versuchsperson (bzw. eine Versuchseinheit) ist.

(2) Externe Validität: Diese bezieht sich auf die Frage der Gültigkeit der Theorie bzw. Hypothese in relevanten Anwendungsgebieten. So kann eine Hypothese durch Laborexperimente bestätigt werden, während bei einer Prüfung in Schulklassen keine Wirkung der »Behandlung« feststellbar ist. Deshalb ist zu fordern, dass Forschungen nicht nur in streng kontrollierten Laborexperimenten, sondern auch mit Schulklassen durchgeführt werden.

(3) Implementierungsgüte: Hier geht es um die gültige Umsetzung (Implementierung) der pädagogischen Methode (unabhängige Variable) in Versuchs- und Kontrollgruppe. Die Güte der Implementierung einer Methode kann z.B. durch Beobachtungen im Unterricht kontrolliert werden.

(4) Messvalidität: Die Messvalidität bezieht sich auf die gültige Erfassung der Wirkungen eines Versuchs durch gültige Messverfahren (z.B. Tests). Die Gültigkeit dieser Messungen wird durch zusätzliche Untersuchungen geprüft, wobei verschiedene Verfahren zur Verfügung stehen (z.B. Prüfung der Inhaltsvalidität und der Konstruktvalidität).

Messvalidität ist sowohl für deskriptive wie auch für hypothesenprüfende Forschungen wichtig. Für deskriptive Forschung ist zusätzlich Repräsentativität wichtig, die in der experimentellen Forschung keine Rolle spielt: Ein Experiment kann nicht repräsentativ sein, weil eine Zufallsstichprobe nicht aus allen potentiellen Mitgliedern der relevanten Grundgesamtheit gezogen werden kann (vgl. Gadenne 1976). In der hypothesenprüfenden Forschung ist hingegen interne und externe Validität wichtig, die für deskriptive Forschung keine Rolle spielt.

In der Vergangenheit wurden anstelle des Experiments häufig Querschnitts- oder Längsschnittsuntersuchungen durchgeführt. In *Querschnittsuntersuchungen* werden z. B. durch einen Fragebogen zum gleichen Zeitpunkt sowohl unabhängige wie auch abhängige Variable erfasst. Wenn sich dann Zusammenhänge ergeben, wird leicht auf kausale Zusammenhänge geschlossen. Bei *Längsschnittuntersuchungen* erfasst man z. B. zuerst die Ausgangsleistung, danach, welche Methode von Lehrern im Unterricht angewendet wird, um dann am Ende erneut die Leistung zu messen. Wenn man dann Zusammenhänge findet, dann vermutet man auch hier einen kausalen Zusammenhang. In beiden Fällen handelt es sich um keine strengen Hypothesenprüfungen. Es kann hier – auch durch Einsatz komplexer statistischer Verfahren (vgl. Freedman 1991) – nicht eindeutig geklärt werden, ob die vermutete Ursache oder eine andere, damit korrelierende Variable für die Wirkung verantwortlich ist. Außerdem wird vorausgesetzt, dass die relevanten Merkmale in der theoretisch gewünschten Weise in der Stichprobe variieren. Wenn in einer Gruppe von Klassen Lehrer Hausaufgaben stellen, Lehrer der Kontrollgruppe dagegen nicht, und man keinen Effekt der Hausaufgaben finden kann, dann bedeutet dies nicht, dass theoretisch reflektierte Hausaufgabenpraxis keinen Effekt hat! (vgl. Elawar/ Corno 1985). Um Ursache-Wirkungszusammenhänge streng prüfen zu können, ist man somit auf experimentelle Forschungen angewiesen.

3.3. Entwicklungsforschung

Für die Entwicklungsforschung gelten ähnliche Gütemaßstäbe wie für experimentelle, hypothesenprüfende Forschung: Es muss möglichst eindeutig nachgewiesen werden, dass die spezifischen Programmelemente für die festgestellte Wirkung verantwortlich sind[9]; die Wirksamkeit des Programms sollte mindestens an den Personen und in dem Land streng überprüft sein, wo das Programm eingesetzt werden soll (externe Validität). Ferner sind auch hier Implementierungsgüte sowie Messgüte wichtig. Spezifische Relevanz hat die praktische Bedeutsamkeit des Programms (vgl. Wellenreuther 2000, Kap. 5), also das Verhältnis von Kosten und Nutzen.

Man kann in der Entwicklungsforschung folgende Phasen unterscheiden: Eine Entwicklungsphase (formative Phase), und eine abschließend überprüfende Phase (summative Phase). In der formativen Phase geht es um die Entwicklung und erste Erprobung des Programms. Hier spielt die kritische Diskussion unter Experten die entscheidende Rolle. In der summativen Phase werden u.a. randomisierte Experimente durchgeführt, um die Wirksamkeit im Vergleich zu alternativen Programmen zu prüfen (vgl. dazu Box 3). Zur individuellen Förderung des Lesenlernens bei größeren Defiziten gibt es z.B. das von M. Clay (vgl. Clay 1993) entwickelte Programm »Reading Recovery« (vgl. Box 3), dessen Wirksamkeit schon in mehreren Evaluationsstudien geprüft wurde.

Box 3: Entwicklungs- und Trainingsforschung

Wie leseschwache Kinder leichter lesen lernen (Iversen/Tunmer 1993)

Ziel: Die Wirksamkeit einer revidierten Fassung des Trainingsprogramms »Reading Recovery« (Clay 1993).

Besonderheit des Programms: Wichtigster Punkt war eine stärkere Gewichtung der Fertigkeit des phonologischen Dekodierens. Schüler sollten die Verknüpfungen zwischen visuellen Mustern und Lauten unterschiedlicher Wörter durch Nutzung von Phonogrammen einüben (z.B. »ight« (light, fight, might und sight). Die stärkere Gewichtung solcher Übungen wurde durch experimentelle Forschungen zum operativen Umgang mit Silben und Buchstabenklustern motiviert, der von entscheidender Bedeutung für das Lesenlernen ist (vgl. Bryant/Bradley 1985). Die Entwicklungsarbeit wurde hier also durch vorherige experimentelle Forschung geleistet.

Methode: Die Studie wurde an 23 Schulen in Rhode Island mit 26 speziell für »Reading Recovery« ausgebildeten Lehrern durchgeführt. 64 leseschwache Schüler aus 34 Klassen nahmen an der Untersuchung teil. Zwei Gruppen von Lehrern, die voneinander keine Kenntnis hatten, wurde an verschiedenen Orten für die Durchführung des alten bzw. des revidierten Programms »Reading Recovery« trainiert. Danach wurden Kinder ausgesucht, die für eine individuelle Leseförderung in Betracht kamen. Diese wurden per Zufall drei Behandlungsgruppen zugeteilt: Eine mit dem alten Reading Recovery Programm, eine mit dem neuen, modifizierten Programm und eine dritte mit dem Standard-Förderprogramm für leseschwache Schüler.

Ergebnisse: Alle mit Reading Recovery geförderten Schüler konnten nach Beendigung der individuellen Förderung im normalen Klassenunterricht erfolgreich mitarbeiten. Beim modifizierten Programm benötigte man dafür jedoch nur noch 42 halbstündige Sitzungen, beim Standardprogramm 57 Sitzungen.

Offene Fragen: Durch weitere Forschung sollte geklärt werden, wie die hohen Kosten reduziert werden können. (1) Man sollte das Programm so modifizieren, dass man gleichzeitig mit zwei oder drei Schülern arbeiten kann (vgl. Iversen, Tunmer/Chapman 2005). (2) sollte man mehr mit freiwilligen Erwachsenen arbeiten, die von Lesespezialisten in Schulen angeleitet werden (Farkas 1998; Johnston/Invenizzi & Juel 1998).

4. Ausblick: Zur Bedeutsamkeit des randomisierten Experiments

Die geringe Glaubwürdigkeit pädagogischer Forschung im Vergleich zur medizinischen Forschung hat vor allem mit der Aufweichung strenger Gütekriterien zu tun. Staaten wie z.B. die USA haben deshalb durch Gesetze (No Child Left Behind Act; Education Sciences Reform Act, vgl. Reyna 2005) festgelegt, in Zukunft nur noch strenge experimentelle Forschung staatlich zu fördern. Boruch (2005) schreibt dazu: »Randomisierte Versuche produzieren, wenn sie gut durchgeführt werden, statistisch unverzerrte Schätzungen der relativen Wirkungen von ökonomischen, medizinischen, verhaltensartigen und sozialen Einflüssen. Dennoch vertreten einige Leute die Auffassung, solche Versuche seien überflüssig, weil andere Versuche wie Meinungsumfragen und Quasi-Experimente[10] faire Schätzungen der Wirkungen liefern könnten.

Demgegenüber steht fest: Analysen von Daten aus passiven Meinungsumfragen, von Daten aus den Verwaltungen oder aus Quasi-Experimenten können nicht vergleichbare unverzerrte Schätzungen der Wirkungen von Interventionen sicherstellen« (Boruch 2005, S. 181).

Boruch belegt seine Aussage anhand vieler Untersuchungen, in denen randomisierte Experimente zu anderen Ergebnissen führen als andere Untersuchungen. So deuteten viele Berichte in den Medien darauf hin, dass dissoziale Jugendliche durch den Besuch von Haftanstalten geläutert werden. Randomisierte Experimente zeigten hingegen, dass solche Besuche keinen positiven Effekt hatten (vgl. Petrosino/Turpin-Petrosino/Buehler 2002). Ähnlich verhält es sich mit Forschungen über Lerntypen (auditiver Typ, visueller Typ). Auch hier zeigen experimentelle Forschungen, dass die von vielen Lehrern vertretene Theorie falsch ist (vgl. Kavale/Forness 1987).

Für die Güte eines Experiments ist die gezielte, theoretisch gut begründete Implementierung pädagogisch relevanter Bedingungen wichtig. So wurden in einem der berühmtesten pädagogischen Experimente zum Einfluss der Klassengröße auf schulisches Lernen nicht nur die Klassengröße variiert (15 Schüler – 24 Schüler), sondern ob bei normaler Klassengröße durch Einbeziehung eines Lehrerassistenten mehr gelernt wird. Innerhalb einer Schule wurden die Klassen in jedem Schuljahr erneut per Zufall einer dieser Bedingungen zugeordnet. Die Ergebnisse besagten, dass eine Reduktion der Klassengröße auf 15 Schüler in den unteren Klassen (Vorschule bis Klasse 3) beträchtliche Wirkungen hatte, in späteren Klassen waren die Wirkungen dagegen gering. Die im Kindergarten und in den ersten Klassenstufen erzielten Effekte erwiesen sich über mehrere Jahre stabil (vgl. Mosteller 1995).

Der Durchführung kostspieliger Experimente sollten theoretische Überlegungen über eine optimale Gestaltung der zu prüfenden Methoden vorausgehen. Experimente über die Wirkung von Hausaufgaben, wie sie mehr oder weniger gedankenlos von Lehrern aufgegeben werden (vgl. Cooper 1989), können nicht darüber informieren, wie unter theoretischen Gesichtspunkten optimierte Hausaufgaben wirken. So ergab eine experimentelle Studie von Elawar/ Corno (1985), dass Hausaufgaben pädagogisch wirksam waren, wenn der Lehrer dreimal in der Woche in Mathematik die Hefte genau durchsah[11] und konkrete Kommentare unter die Hausaufgaben schrieb wie »das … hast du gut gemacht, aber in den Punkten … solltest du dich noch verbessern.«. In der Pädagogik muss deshalb zwischen der Evaluierung einer häufig auftretenden Praxis (»normale« Hausaufgabenpraxis) und einer nach theoretischen Gesichtspunkten optimierten Praxis unterschieden werden.[12]

Randomisierte Experimente mit Schulklassen als Untersuchungseinheit verursachen einen sehr großen Aufwand, weil je nach Größe des zu erwartenden Effekts mindestens jeweils 20 Schulklassen per Zufall auf die Versuchs- und Kontrollklassen aufgeteilt werden müssen. Deshalb sollten Voruntersuchungen und Experimente mit Zufallsaufteilung von Schülern durchgeführt werden, mit denen möglichst viele der anstehenden Fragen schon geklärt werden können, bevor aufwendige Experimente mit Schulklassen als Untersuchungseinheit durchgeführt werden. Levin/O'Donnel (1999) haben dazu ein Vierstufenmodell entwickelt, auf das ich nun kurz eingehen möchte.

Graham, Harris und Zito (2005) erläutern dieses Stufenmodell an ihrem SRSD-Projekt (**S**elf **R**egulated **S**trategy **D**evelopment), in dem die Kompetenz von leistungsschwachen Schülern beim Schreiben von Berichten verbessert werden sollte.

Stufe 1: Hier wird die einschlägige theoretische Literatur gesichtet und geprüft, welche empirischen Untersuchungen zu dem gewählten Forschungsproblem vorhanden sind. In dem von Graham, Harris und Zito (2005) durchgeführten Projekt gab es schon eine Reihe von Studien mit älteren Schülern, aber noch keine Studien zu Kindern der zweiten und dritten Klassenstufe. Beobachtungen von leistungsschwachen Schreibern zeigten, dass diese beim Verfassen einer Berichts bzw. einer Geschichte ohne irgendeinen Plan alles, was ihnen gerade einfiel, nie-

derschrieben. Vermutlich beansprucht die Schreibtätigkeit und das Richtigschreiben soviel Arbeitskapazität, dass sie für ein sinnvolles Planen keine Kapazitäten mehr frei haben. Ein strukturiertes Vorgehen nach einem schriftlichen Plan, nach dem Geschichten zu entwickeln sind, kann auch das Arbeitsgedächtnis entlasten.

Stufe 2: Um Hinweise auf mögliche Wirkfaktoren zu erhalten, sind hier »best practice« Studien sinnvoll (z.B. »best practice« Klassen in einem Land mit relativ guten Ergebnissen beim Lesenlernen; vgl. Wilkinson/Townsend 2000[13]). Solche Studien können dann Ausgangspunkt für Trainingsstudien bzw. für Unterrichtsexperimente sein. Man kann dann in der Schule neue Trainingsmethoden an einzelnen Schülern oder an kleinen Gruppen von Schülern experimentell erproben, die von geschulten Studenten implementiert werden.

Stufe 3: Hier im Idealfall ein randomisiertes Experiment mit Zufallsaufteilung von Klassen bzw. von ganzen Schulen auf Versuchs- und Kontrollgruppen durchgeführt werden. Eine Möglichkeit besteht darin, Lehrer beide Methoden in verschiedenen Klassen unterrichten zu lassen, wobei eine Zufallszuordnung der Klassen zu den Methoden erfolgt (vgl. z.B. Copeland 1991). Lehrer werden zuerst in der neuen Methode geschult, wobei sich dieses Training über einen längeren Zeitraum (z.B. ein Jahr) hinziehen kann (vgl. Brown/Pressley/Van Meter/Schuder 1996). Wichtig ist diese experimentelle Prüfung unter realen Bedingungen, weil nur dadurch die pädagogische und praktische Bedeutung der neuen Methode evaluiert werden kann.

Stufe 4: Hier geht es um die Übertragung und Anwendung eines empirisch streng geprüften Programms in die Schulpraxis. Lehrer müssen dabei zunächst über das Programm informiert werden, z.B. durch entsprechende Veröffentlichungen in Zeitschriften und in der Presse.

Häufig ist es schwierig, Schulen für eine Mitarbeit an Experimenten zu gewinnen, vor allem wenn diese als Kontrollgruppen »nur« Vergleichsdaten zur Verfügung stellen sollen. Um solche Probleme zu lösen, haben Borman, Slavin, Cheung, Chamberlain und Chambers (2005) Schulen sowohl als Versuchs- wie auch als Kontrollklassen verwendet. Allerdings wird hier der Nachweis von Wirkungen erschwert, weil Lehrer der Kontrollklassen durch Gespräche mit Versuchsklassenlehrern etwas für ihren Unterricht lernen können. Dennoch kann man die relative Wirksamkeit einer neuen Methode letztlich nur durch sorgfältig geplante Experimente mit Zufallsaufteilung von Klassen bzw. Schulen feststellen. Praxisnähe ist dabei durchaus mit methodischer Strenge zu vereinbaren. Es ist keineswegs notwendig, Designexperimente[14] durchzuführen, wenn man praxisnahe Schulforschung betreiben will.

5. Literatur

Adorno, Theodor W./Dahrendorf, Ralf/Pilot, Harald/Albert, Hans/Habermas, Jürgen/Popper, Karl R.: Der Positivismusstreit in der deutschen Soziologie (1969). Ulm: Luchterhand Verlag 1972

Anderson, John R./Reder, Lynne M./Simon, Herbert A. (1999): Applications and Misapplications of Cognitive Psychology to Mathematics Education. Texas Educational Review. http://act-r.psy.cmu.edu/papers/misapplied.html

Anderson, John R./Greeno, James G./Reder, Lynne M./Simon, Herbert A. (2000): Perspectives on Learning, Thinking, and Activity. Educational Researcher 29, H. 4, 11–13

Borman, Geoffrey D./Slavin, Robert E./Cheung, Alan/Chamberlain, Anne/Madden, Nancy/ Chambers, Betty (2005): The National Randomized Field Trial of Succes for All: Second Year Outcomes. American Educational Research Journal 42, H. 4, 673–696

Boruch, Robert (2005): Beyond the Laboratory or Classroom: The Empirical Basis of Educational Policy. In: Gary D. Phye/Daniel H. Robinson/Joel Levin (Hrsg.): Empirical Methods for Evaluating Educational Interventions., 177–191

Bryant, Peter E./Bradley, Lynette (1985): Children's reading problems. Oxford; England: Basil Blackwell.

Brown, Ann L. (1992): Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. Journal of the Learning Sciences 2, H. 2, 141–178

Brown, Rachel/Pressley, Michael/Van Meter, Peggy/Schuder, Ted (1996): A Quasi-Experimental Validation of Transactional Strategies Instruction With Low-Achieving Second-Grade Readers. In: Journal of Educational Psychology 88, H. 1, 18–37

Bühner, Markus (2006): Einführung in die Test- und Fragebogenkonstruktion. 2. Aufl., München: Pearson Studium

Campbell, Donald T./Stanley, Julian C. (1963): Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally

Clay, Marie M.: Reading Recovery (1993): A Guidebook for Teachers in Training. Portsmouth: Heinemann

Cooper, Harris M. (1989): Synthesis of research in homework. Educational Leadership 47, H. 3, 85–91

Copeland, Willis D. (1991): Microcomputers and teaching actions in the context of historical inquiry. Journal of Educational Computing Research 7, 421–454

Cronbach, Lee J./Ambron, Sueann R./Dornbusch, Sanford M./Hess, Robert D./Hornik, Robert C./Phillips, D.C./Walker, Decker F./Weiner, Stephen S. (1980): Toward Reform of Program Evaluation. San Francisco: Jossey-Bass 1980

Elawar, Maria C./Corno, Lyn (1985): A Factorial Experiment in Teacher´s Written Feedback on Student Homework: Changing Teacher Behavior a Little rather Than a Lot. Journal of Educational Psychology 77, H. 2, 162–173

Farkas, George (1998): Reading One-To-One: An Intensive Program Serving a Great Many Students While Still Achieving large Effects. In: Jonathan Crane (Hrsg.): Social Programs That Work. New York: Russell Sage Foundation, 75–109

Freedman, D. A. (1991): Statistical Models and Shoe Leather. In: Mardsen, P. V.: Sociological Methodology, Vol.21, Washington, 291-358

Feuer, Michael J./Towne, Lisa/Shavelson, Richard J. (2002): Scientific Culture and Educational Research. Educational Researcher 31, H. 8, 4–14

Gadenne, Volker (1976): Die Gültigkeit psychologischer Untersuchungen. Stuttgart: Kohlhammer

Gersten, Russel/Baker, Scott K./Smith-Johnson, Joyce/Flojo, Jonathan R./Hagan-Burke, Shanna (2004): A Tale of Two Decades: Trends in Support for Federally Funded Experimental Research in Special Education. Exceptional Children 70, H. 3, 323–332

Glass, Gene V.: Meta-Analysis at 25. 2000. http://glass/ed.asu.edu/gene/papers/meta25.html

Graham, Steven/Harris, Karen H./Zito, Jennifer (2005): Promoting internal and external validity: A Synergism of Laboratory-like experiments and classroom-based self-regulated strategy development research. In: Gary D. Phye/Daniel H. Robinson/Joel R. Levin (Hrsg.): Empirical Methods for Evaluating Educational Interventions. San Diego: Elsevier Academic Press, 235–265

Hager, Willi (1995): Programme zur Förderung des Denkens bei Kindern. Konstruktion, Evaluation und Metaevaluation. Göttingen: Hogrefe

Herbart, Johann Friedrich: Pädagogische Schriften. 3 Bände. Walter Asmus (Hrsg.; 1982). Stuttgart: Klett

Hsieh, Peggy/Acee, Taylor/Chung, Wen-Hung/Hsieh, Ya-Ping/Kim, Hyunjin/Thomas, Greg D./You, Ji-in/Levin, Joel R./Robinson, Daniel H. (2005): Is Educational Intervention Research on Decline? Journal of Educational Psychology 97, H. 4, 523–529

Ingenkamp, Karlheinz (1992): Ausbreitung und Akzeptanz der empirisch-orientierten Pädagogik. In: Karlheinz Ingenkamp/Reinhold S. Jäger/Hanns Petillon/Bernhard Wolf (Hrsg.): Empirische Pädagogik 1970–1990. Band 1, Weinheim: Deutscher Studien Verlag, S. 4–15

Iversen, Sandra/Tunmer, William E. (1993): Phonological Processing Skills and the Reading Recovery Program. Journal of Educational Psychology 85, H. 1, 112–126

Iversen, Sandra/Tunmer, William E./Chapman, James W. (2005): The Effects of Varying Group Size on the Reading Recovery Approach to Preventive Early Intervention. Journal of Learning Disabilities 38, H. 5, S. 456–472

Johnston, Francine. R./Invernizzi, Marcia/Juel, Connie: Book Buddies: guidelines for volunteer tutors of emergent and early readers. New York: Guilford Press 1998

Kalyuga, S., Chandler, P., Sweller, J. (2001): Learner Experience and Efficiency of Instructional Guidance. Educational Psychology, Vol. 21, No. 1, 5-23

Kavale, Kenneth A./Forness, Steven R. (1987): Substance Over Style: Assessing the Efficacy of Modality Testing and Teaching. Exceptional Child 54, H. 3, 228–239

Kirschner, Paul A./Sweller, John/Clark, Richard E. (2006): Why Minimal Guidance During Instruction Does not Work: An Analysis of the failure of Constructivist, Discovery, Problem-Based, Experimental, and Inquiry-Based Teaching. Educational Psychologist 41, H. 2, 75–86

Leinhardt, Gaea (1989): Math Lessons: A Contrast of Novice and Expert Competence. Journal for Research in Mathematics Education 20, H. 1, S. 52–75

Leschinsky, Achim (2004): Die Ausdifferenzierung und Weiterentwicklung der Schulforschung seit den 1970er Jahren. In: Werner Helsper/Jeanette Böhme (Hrsg.): Handbuch der Schulforschung. Wiesbaden: Verlag für Sozialwissenschaften

Levin, Joel R. (2005): Randomized Classroom Trials on Trial. In: Gary D. Phye/Daniel H. Robinson/Joel R. Levin (Hrsg.): Empirical Methods for Evaluating Educational Interventions. San Diego: Elsevier Academic Press, 3–27

Levin, Joel R./O'Donnell, Angela M.(1999): What to do about educational Research's Credibility Gaps. Issues in Education 5, H. 2, 177–229; 279–293

Mayer, Richard E. (2005): The Failure of Educational Research to impact Educational Practice. In: Gary D. Phye/ Daniel H. Robinson/Joel R. Levin (Hrsg.): Empirical Methods for Evaluating Educational Interventions. San Diego: Elsevier Academic Press, 67–81

Mosteller, Frederick (1995): The Tennessee Study of Class Size in the Early School Grades. The Future of Children. Critical Issues for Children and Youths 5, H. 2, 113–127

Oelkers, Jürgen (1998): Pädagogische Reform und Wandel der Erziehungswissenschaft. In: Christoph Führ/Carl-Ludwig Furck (Hrsg.): Handbuch der deutschen Bildungsgeschichte. Bd. IV, 1945 bis zur Gegenwart – Zweiter Teilband: Deutsche Demokratische Republik und neue Bundesländer. München: C. H. Beck, 217–243

Phye, Gary D./Robinson, Daniel H./Levin, Joel R (Hrsg.; 2005): Empirical Methods for Evaluating Educational Interventions. San Diego: Elsevier Academic Press

Pressley, Michael/Gaskins, Irene W./Solic, Katie/Collins, Stephanie (2006): A Portrait of Benchmark School: How a School Produces High Achievement in Students Who Previously Failed. Journal of Educational Psychology 98, H. 2, 282–306

Petrosino, Anthony/Turpin-Petrosino, Carolyn/Buehler, John: Scared Straight. A Campbell Collaboration systematic review: Campbell Library 2002. http://campbellcollaboration.org

Raudenbush, Stephen W. (2004): Learning from Attempts to Improve Schooling: The Contribution of Methodological Diversity. http://www7.nationalacademies.org/cfe/Multiple_Methods_Raudenbush_Paper.pdf

Reyna, Valerie F. (2005): The No Child Left Behind Act, Scientific Research and Federal Educational Policy: A View from Washington, D.C. In: Gary D. Phye/ Daniel H. Robinson/Joel R. Levin (Hrsg.): Empirical Methods for Evaluating Educational Interventions. San Diego: Elsevier Academic Press 2005, S. 29–52

Roth, Heinrich (1962): Die realistische Wendung in der Pädagogischen Forschung. Neue Sammlung 2, 481–490

Schümer, Gundel. (2004): Zur doppelten Benachteiligung von Schülern aus unterprivilegierten Gesellschaftsschichten im deutschen Schulwesen. In: Gundel Schümer, Klaus-Jürgen Tillmann/Manfred Weiß (Hrsg.): Die Institution Schule und die Lebenswelt der Schüler. Vertiefende Analysen der PISA-2000-Daten zum Kontext von Schülerleistungen. Wiesbaden: VS Verlag für Sozialwissenschaften

Stevenson, Harold W./Stigler, James W. (1992): The Learning Gap. Why our schools are failing and what we can learn from Japanese and Chinese Education. New York: Summit Books

Stigler, James W./Hiebert, James (1998): Teaching Is A Cultural Activity. American Educator 22, H. 4, S. 1–10

Terhart, Ewald (1997): Entwicklung und Situation des qualitativen Forschungsansatzes in der Erziehungswissenschaft. In: Barbara Friebertshäuser/Annedore Prengel (Hrsg.): Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft. Weinheim: Juventa-Verlag, 27–42

Tuovinen, Juhani E./Sweller, John (1999): A Comparison of Cognitive Load Associated With Discovery Learning and Worked Examples. Journal of Educational Psychology 91, H. 2, 334–341

Valentine, Jeffrey C./Cooper, Harris M. (2005): Can we measure the quality of causal research in Education? In: Gary D. Phye/ Daniel H. Robinson/Joel R. Levin (Hrsg.): Empirical Methods for Evaluating Educational Interventions. San Diego: Elsevier Academic Press. 85–111

Webb, Eugene J./Campbell, Donald T./Schwartz, Richard D./Seechrest, Lee. (1966): Unobstrusive Measures: Nonreactive Research in the social sciences. Chicago: Rand McNally

Wellenreuther, Martin (2000): Quantitative Forschungsmethoden in der Erziehungswissenschaft. Eine Einführung. Weinheim: Juventa-Verlag.

Wellenreuther, Martin (2004): Lehren und Lernen – aber wie? Empirisch-experimentelle Forschungen zum Lehren und Lernen im Unterricht. Baltmannsweiler: Schneider Verlag Hohengehren.

Wilkinson, Ian A. G./Townsend, Michael A. R. (2000): From Rata to Rimu: Grouping for instruction in best practice New Zealand classrooms. The Reading Teacher 53, H. 6, 460–471

Anmerkungen

1 Für die kritischen Anmerkungen von J. Düring, K. Nölle und Th. Petzel möchte ich mich an dieser Stelle bedanken.

2 Vgl. auch Herbart 1982, Bd. 1, S. 125.

3 Ingenkamp (1992, S. 15) schreibt dazu: »Die Bevorzugung qualitativer Methoden und subjektiver Wahrheitskriterien machte die Handlungsforschung für einen Teil der Geisteswissenschaftler attraktiv, die den dornigen Weg der quantitativ-statistischen Methodenausbildung scheuten…. Die Handlungsforschung [hat] in ihren Anfängen eine kontinuierliche Ausweitung der empirischen Pädagogik sehr behindert und die Akzeptanz einzelner Methoden weitgehend aufgehoben…«

4 Nach Ingenkamp (1992, S. 13) bildete sich damals eine »bildungsphilosophische Koalition«, deren Argumentation »seltsam irreal und provinziell« wirke, »da in fast allen Stellungnahmen die internationale Methodologiediskussion völlig übergangen wird. Die empirische Pädagogik ist aber vor 1945 nur in Deutschland ins Abseits gedrängt worden.«

5 Quantitative, hypothesenprüfende Forschung sei »positivistisch, mechanisch, dehumanisierend, sinnlos, technokratisch etc.« (Terhart 1997, S. 37). »Während empirisch-quantitative Forschung auf eine streng Theorie- und hypothesengeleitete Quantifizierung von Ereignissen, Abläufen und Zusammenhängen in der sozialen Wirklichkeit ausgerichtet ist, wobei dies Zergliederung, Dimensionierung, Messung bedeutet, orientiert sich qualitativ-empirische Forschung am Ziel einer möglichst gegenstandsnahen Erfassung der ganzheitlichen, kontextgebunden Eigenschaften sozialer Felder« (Terhart 1997, S. 27).

6 TIMSS ist die Abkürzung für «Third International Mathematics and Science Study«.

7 PISA ist die Abkürzung für »Program for International Student Assessment«.

8 Bei der Darstellung stütze ich mich auf die Zusammenfassung des Max Planck Instituts für Bildungsforschung in Berlin, die im Internet veröffentlicht ist (vgl. http://www.mpib-berlin.mpg.de/TIMSS-Video/TIMSS_homepage/html/method.htm).

9 Es macht z.B. nur begrenzt Sinn, ein kostspieliges individuelles Förderprogramm mit dem »normalen« Klassenunterricht zu vergleichen. Wichtiger ist der Nachweis, dass das zu prüfende Förderprogramm mit einem vergleichbaren individuellen Förderprogramm verglichen wird: Der Forscher muss sich somit den Nachweis der Wirksamkeit möglichst schwer machen.

10 Quasi-Experimente sind z.B. Experimente ohne Zufallsaufteilung. Häufig werden z.B. zu Schulen, in denen ein Programm eingesetzt wird, möglichst vergleichbare Schulen für die Kontrollgruppe gesucht, in denen nach der traditionellen Methode verfahren wird. Auch wenn die Gruppen der Versuchsschulen und der Kontrollschulen bezüglich der sozialen Herkunft parallelisiert sind, kann nicht ausgeschlossen werden, dass die Schulen, die bereit sind, das neue Programm einzuführen, auch insgesamt engagierter sind.

11 Diese Hausaufgabenpraxis erstreckte sich über 10 Wochen. Das Experiment wurde in der sechsten Klassenstufe durchgeführt.

12 Dies ist völlig analog zur medizinischen Forschung. Auch dort interessiert vor allem die Wirksamkeit eines Medikaments, das in verschiedenen Hinsichten (Nebenwirkungen, Wirksamkeit) optimiert wurde.

13 Bei solchen »best practice« Studien muss präzise anhand von Lernergebnissen der Schüler belegt sein, dass es sich tatsächlich um »best practice«-Lehrer handelt!

14 Im »Designexperiment« werden die Behandlungen nicht starr festlegt, sondern bei Bedarf abgeändert. Auch verzichtet man auf die Zufallszuteilung von Lehrern zu Klassen. Dadurch ist es sehr schwer möglich, eindeutige Folgerungen bezüglich der relevanten wirksamen Faktoren zu ziehen. Solche Designexperimente können in Stufe 2 sinnvoll sein, sie eignen sich jedoch nicht für Stufe 3.