Enzyklopädie Erziehungswissenschaft Online

Nicola Döring

Qualitätskriterien für quantitative empirische Studien

39 Seiten
Aus: Enzyklopädie Erziehungswissenschaft Online; ISSN 2191-8325
Fachgebiet/Unterüberschrift: Methoden der empirischen erziehungswissenschaftlichen Forschung, Qualitätskriterien in der empirischen Forschung
hrsg. von Sabine Maschke und Ludwig Stecher

2015, DOI 10.3262/EEO07150345

Abstract: Wie lassen sich in der Erziehungswissenschaft seriöse Studien von nicht- oder pseudowissenschaftlichem Vorgehen abgrenzen? Der vorliegende Beitrag stellt dazu vier allgemeine Standards der Wissenschaftlichkeit vor (1. wissenschaftliches Forschungsproblem, 2. wissenschaftlicher Forschungsprozess, 3. Wissenschafts- und Forschungsethik sowie 4. Dokumentation des Forschungsprojekts), die jede Studie, die Wissenschaftlichkeit beansprucht, prinzipiell erfüllen muss. Diese vier Standards werden dann durch vier Kriterien der Wissenschaftlichkeit unterlegt, die graduelle Qualitätseinstufungen erlauben (1. Grad der inhaltlichen Relevanz des Forschungsproblems, 2. Grad der methodischen Strenge des Forschungsprozesses, 3. Grad der ethischen Strenge des Forschungsprozesses, 4. Grad der Dokumentations- und Präsentationsqualität). Anhand dieser Kriterien lässt sich abwägen, ob eine schwache, mittelmäßige oder hervorragende wissenschaftliche Studie vorliegt. Zur detaillierten Beurteilung speziell der methodischen Strenge einer empirisch-quantitativen Studie wird als zentrales Gütekriterium die Validität (d.h. die Gültigkeit von Schlussfolgerungen aus der Studie) herangezogen. Eine methodisch strenge Studie ist so angelegt, dass alle vier Unterformen der Validität (1. Konstruktvalidität, 2. interne Validität, 3. externe Validität und 4. statistische Validität) möglichst stark ausgeprägt sind, so dass aus der Studie tragfähige Schlussfolgerungen zu ziehen sind. Der Beitrag geht abschließend auf aktuelle Probleme sowie zukünftige Entwicklungen bei der Qualitätssicherung empirisch-quantitativer Studien in der Erziehungswissenschaft ein.

Schlüsselbegriffe: Standards der Wissenschaftlichkeit, methodische Strenge, Validität, Forschungsethik, wissenschaftliche Gütekriterien, Forschungsprozess

Inhalt

1. Einführung

Wissenschaftlicher Erkenntnisgewinn unterscheidet sich grundlegend von anderen Wissensformen. Gerade weil in der Erziehungswissenschaft oft alltagsnahe Fragestellungen bearbeitet werden, existieren zu diversen pädagogischen Themen auch zahlreiche vorwissenschaftliche sowie pseudowissenschaftliche Aussagen. Unter welchen Umständen kann man eine Studie also

als wissenschaftlich seriös einordnen? Dazu werden *vier Standards der Wissenschaftlichkeit* vorgestellt (Abschnitt 2).

Lässt sich eine empirische Studie im Feld der Erziehungswissenschaft begründet als wissenschaftlich einordnen, so stellt sich unter Qualitätsgesichtspunkten die Anschlussfrage, ob es sich um eine schwache, mittelmäße oder gar exzellente wissenschaftliche Studie handelt. Zur Einordnung werden in erster Linie die Kriterien der *inhaltlichen Relevanz* und der *methodischen Strenge* angelegt – im Englischen ist dies als Begriffspaar „relevance & rigour" etabliert. Zusätzlich werden der Grad der *ethischen Strenge* und der *Präsentationsqualität* in den letzten Jahren zunehmend differenzierter als Kriterien wissenschaftlicher Güte an empirische Studien angelegt. Somit kommt man auf insgesamt *vier Kriterien wissenschaftlicher Qualität*, welche die vier Standards der Wissenschaftlichkeit präzisieren und anhand derer Studien graduell bewertet werden können (Abschnitt 3).

Insbesondere bei der Beurteilung der methodischen Strenge einer empirischen Studie muss unterschieden werden, ob das Forschungsprojekt einer qualitativen Forschungslogik folgt oder im quantitativen Paradigma der empirischen Sozialforschung angesiedelt ist. Für quantitative empirische Studien, um die es in diesem Beitrag geht, hat sich ein relativ klares Inventar an *Detailkriterien zur Bewertung der methodischen Strenge* etabliert. Besonders einflussreich ist hier das Konzept der Validität (d.h. der Gültigkeit von Schlussfolgerungen aus der Studie), das maßgeblich von dem US-amerikanischen Psychologen und Methodiker Donald T. Campbell (1916-1996) ausgearbeitet wurde. Wir unterscheiden *vier Typen der Validität in der Campbell-Tradition*: Konstruktvalidität, interne Validität, externe Validität und statistische Validität, die durch methodisch strenges Vorgehen optimiert werden können und dadurch besonders aussagekräftige Resultate einer Studie sicherstellen (Abschnitt 4). Die im vorliegenden Kapitel behandelten Qualitätsaspekte sind Tabelle 1 zu entnehmen.

Standards der Wissenschaftlichkeit: *Liegt überhaupt eine wissenschaftliche Studie vor?*	**Kriterien der Wissenschaftlichkeit:** *Handelt es sich um eine schwache, durchschnittliche oder exzellente wissenschaftliche Studie?*	**Gütekriterien der methodischen Strenge:** *Ist die Studie insbesondere methodisch so streng angelegt, dass in Bezug auf das zu lösende Forschungsproblem wirklich tragfähige Schlussfolgerungen zu ziehen sind?*
1. Startet die Studie mit einem abgegrenzten und in den aktuellen Forschungsstand eingeordneten **wissenschaftlichen Forschungsproblem**?	1. Kann dem gewählten wissenschaftlichen Forschungsproblem ein **hoher Grad an theoretischer und/oder praktischer Relevanz** zugeschrieben werden?	
2. Folgt die Studie einem geordneten und reflektierten **wissenschaftlichen Forschungsprozess** unter Einsatz anerkannter Methodologien und Methoden?	2. Kann dem gewählten wissenschaftlichen Forschungsprozess in all seinen Phasen ein **hoher Grad an methodischer Strenge** zugeschrieben werden?	1. **Konstruktvalität** (tragfähiger Rückschluss von den erhobenen Messwerten auf die interessierenden theoretischen Konstrukte) 2. **Interne Validität** (tragfähiger Rückschluss auf die Existenz von Ursache-Wirkungs-Relationen zwischen den untersuchten Variablen) 3. **Externe Validität** (tragfähige Verallgemeinerbarkeit der gefundenen Effekte auf andere Orte, Zeiten, Personen und/oder Situationen) 4. **Statistische Validität** (tragfähige statistische Analyse zum zuverlässigen Nachweis der Größe und Richtung von Effekten in der Population)
3. Erfüllt die Studie die zentralen Anforderungen der **Wissenschafts- und Forschungsethik**, wie sie z.B. in den Ethik-Kodizes der wissenschaftlichen Fachgesellschaften festgeschrieben sind?	3. Kann dem gewählten wissenschaftlichen Forschungsprozess ein **hoher Grad an ethischer Strenge** zugeschrieben werden (sowohl hinsichtlich wissenschaftsinterner Fragen als auch im Umgang mit den Untersuchungspersonen)?	
4. Liegt eine umfassende **Dokumentation des Forschungsprojekts** (d.h. seines Ablaufs und seiner Ergebnisse) vor, so dass die Studie von Außenstehenden bewertet und bei Bedarf auch repliziert werden kann?	4. Kann der Dokumentation und ggf. auch Publikation der Studie ein **hoher Grad an Dokumentations- und Präsentationsqualität** zugeschrieben werden (etwa hinsichtlich Vollständigkeit und Anschaulichkeit der Darstellung statistischer Befunde oder Verständlichkeit und Diskriminierungsfreiheit des sprachlichen Ausdrucks)?	

Tabelle 1. Übersicht der Kernfragen zur Bewertung der Qualität einer quantitativen empirischen Studie in der Erziehungswissenschaft (vgl. Döring & Bortz, im Druck, Kap. 3)

Einschlägige Standards und Kriterien der Wissenschaftlichkeit – bis hin zu Details der Beurteilung der methodischen Strenge anhand unterschiedlicher Typen der Validität – zu kennen, ist zum einen notwendig, um vorliegende

quantitativ-empirische Studien in der Erziehungswissenschaft korrekt einordnen zu können – etwa bevor man sie als Grundlage eigener Forschungsarbeiten oder Praxismaßnahmen heranzieht. Zum anderen sollte man die Durchführung eigener Studien von vorne herein immer mit Blick auf die Qualitätskriterien ausrichten und ausgestalten. Denn Qualitätseinbußen – speziell auch hinsichtlich der Validität als zentralem Kriterium methodischer Strenge empirisch-quantitativer Studien – drohen in allen Phasen des Forschungsprozesses von der vorbereitenden Literaturrecherche bis zur abschließenden Interpretation statistischer Befunde. Eine empirisch-quantitative Studie, der es an methodischer Strenge fehlt bzw. deren Validität stark eingeschränkt ist, liefert Befunde begrenzter Gültigkeit und trägt deswegen nur sehr bedingt zum Erkenntnisfortschritt bei.

Insgesamt ist zu konstatieren, dass im Hinblick auf wissenschaftliche Qualitätskriterien im quantitativen Ansatz der empirischen Sozialforschung, in dem Forschungsprozesse hochgradig standardisiert organisiert sind, eine deutlich größere Einigkeit herrscht als in dem methodologisch und methodisch weitaus stärker ausdifferenzierten Feld der qualitativen Sozialforschung (für eine ausführliche Darstellung der Qualitätskriterien in der quantitativen wie qualitativen empirischen Sozialforschung siehe Döring & Bortz, im Druck, Kap. 3). Dennoch sind sowohl Festlegung als auch Prüfung von Qualitätskriterien auch im quantitativen Ansatz teilweise kontrovers und aus verschiedenen Gründen problematisch. Auf der Meta-Ebene werden deswegen auch *Grenzen der Qualitätsbewertung quantitativer empirischer Studien* angesprochen (Abschnitt 5). Insbesondere grobe Verletzungen von Qualitätskriterien – wie etwa Wissenschaftsfälschungen – fordern das Wissenschaftssystem heraus und verändern Prozesse der Qualitätssicherung. Der Beitrag endet mit einem *Ausblick auf die zukünftige Entwicklung* im Feld der Qualitätsbewertung empirisch-quantitativer Studien (Abschnitt 6).

2. Handelt es sich überhaupt um eine wissenschaftliche Studie? Vier Standards der Wissenschaftlichkeit

Von „Studien“, „Tests“ und „Experimenten“ rund um das Lehren und Lernen sowie andere erziehungswissenschaftliche Sachverhalte ist in vielen Zusammenhängen die Rede. Dabei handelt es sich teilweise um seriöse wissenschaftliche Studien, teilweise aber auch um vorwissenschaftliche, um pseudowissenschaftliche oder um parawissenschaftliche Beiträge. Zur Einordnung und Abgrenzung sind vier Standards der Wissenschaftlichkeit anzulegen: 1. Wissenschaftliches Forschungsproblem, 2. wissenschaftlicher Forschungsprozess, 3. Wissenschafts- und Forschungsethik sowie 4. Dokumentation des Forschungsprojekts. Sind diese vier Standards prinzipiell er-

füllt, hat man es nach gängigem Verständnis mit einer echten wissenschaftlichen Studie zu tun.

2.1 Wissenschaftliches Forschungsproblem

Jede seriöse erziehungswissenschaftliche Studie beginnt mit einem klar formulierten Forschungsproblem (meist ausdifferenziert in einzelne Forschungsfragen und/oder Forschungshypothesen), das es zu lösen gilt (Döring & Bortz, im Druck, Kap. 5). Das gewählte Forschungsproblem sollte grundsätzlich empirisch untersuchbar und auf dem aktuellen Kenntnisstand erklärbar sein, d.h. das Thema muss in den anerkannten erziehungswissenschaftlichen Forschungs- und Publikationskontext einzuordnen sein.

So stellen etwa Fragen rund um die Förderung von „Kompetenzen der interkulturellen Kommunikation" eindeutig wissenschaftliche Forschungsprobleme dar, nicht jedoch entsprechende Fragen zu „Kompetenzen der Kommunikation mit Außerirdischen". Denn hier würde man sich in wissenschaftliche Grenzbereiche (*Parawissenschaft*) bewegen, da über „Außerirdische" und die Kommunikation mit ihnen aktuell kein gesichertes Wissen vorliegt.

Auch wenn nur grob und vage ein Themengebiet benannt wird (z.B. „untersucht werden die Auswirkungen von Computerspielen"), kann eine Studie nicht als wissenschaftlich gelten, sondern muss eher als *pseudowissenschaftlich* eingestuft werden. Denn ohne strukturierendes Forschungsproblem können am Ende alle möglichen Ergebnisse beliebig interpretiert werden, der Erkenntniswert bliebe offen. Stattdessen muss für eine wissenschaftliche Studie ein konkretes und empirisch untersuchbares Forschungsproblem auf der Basis des aktuellen Forschungsstandes angemessen zugespitzt werden (z.B. „untersucht wird die Frage, ob – und wenn ja, wie stark und wie nachhaltig – computergestützte Bewegungsspiele die körperliche Fitness von Schulkindern steigern"). Mangelnde Klarheit und ungenügende Eingrenzung des bearbeiteten Forschungsproblems sind gravierende Schwächen, die dazu führen können, einem Projekt die Wissenschaftlichkeit abzusprechen.

2.2. Wissenschaftlicher Forschungsprozess

Neben dem eindeutig erkennbaren Forschungsproblem und dessen Einbettung in gesicherte wissenschaftliche Vorkenntnisse und Theorien, ist für eine empirische Studie in der Erziehungswissenschaft, die Wissenschaftlichkeit beansprucht, das Durchlaufen eines *geordneten und reflektierten Forschungsprozesses* entscheidend. In welche Phasen sich ein Forschungsprozess gliedert und welche wissenschaftlichen Methoden der Untersuchungsplanung, Datenerhebung und Datenanalyse jeweils indiziert sind, hängt vom konkreten Forschungsproblem sowie vom gewählten wissenschaftstheoretischen Forschungsansatz (z.B. qualitativer und/oder quantitativer Ansatz der

empirischen Sozialforschung) und von den forschungspraktischen Rahmenbedingungen (Zeit, Personal, Kosten) ab. In jedem Fall sollte der wissenschaftliche Forschungsprozess einer bekannten Methodologie (z.B. Evaluationsforschung; Experimentalforschung; ethnografische Feldforschung; Umfrageforschung) folgen und dabei passende wissenschaftliche Methoden der Datenerhebung und Datenanalyse einsetzen.

Ad hoc durchgeführte „Blitzumfragen" auf der Straße, „Online-Votings" auf Webseiten ohne klare Kenntnis des Teilnehmerkreises, sog. „Psycho-Tests" in Publikumszeitschriften oder „Selbstexperimente" einzelner Personen, wie sie zuweilen in der Presse und auch in der Fachliteratur kolportiert werden (z.B. eine Familie lebt eine Woche ohne Handy und Fernseher und beobachtet die Effekte auf das Familienleben) sind als *vorwissenschaftlich* einzuordnen. Sie haben rein anekdotischen Charakter und liefern keine generalisierbaren Erkenntnisse, auch wenn ihre Ergebnisse noch so plausibel wirken mögen. Hier fehlen nämlich zentrale Schritte jeden wissenschaftlichen Forschungsprozesses: Die kritische Reflexion des Untersuchungsdesigns, die begründete Auswahl von Einzelfällen oder die Ziehung einer Stichprobe, die gezielte Erhebung und vollständige Analyse aller Daten usw. Allerdings wird bei vorwissenschaftlichen sog. „Experimenten", „Tests" und „Umfragen" im Alltag meist gar nicht der Anspruch der Wissenschaftlichkeit erhoben, sondern steht der *Unterhaltungsaspekt* für das Publikum im Vordergrund.

Das ist bei *pseudowissenschaftlichen Studien* anders. Sie geben sich als seriöse Forschung aus, folgen aber nicht konsequent einem dem Forschungsproblem entsprechenden empirischen Forschungsprozess und dessen methodischen Anforderungen. Pseudowissenschaftliche Studien können aufgrund *mangelnder Methodenkompetenz* entstehen. So etwa wenn studentische Forschungsarbeiten vorgeben, wissenschaftlich fundierte Aussagen über Ursache-Wirkungs-Relationen treffen zu können, die jedoch gar nicht durch das dafür notwendige kausalanalytische Vorgehen gedeckt sind, sondern reine Überinterpretationen der Daten darstellen. Etwa wenn „Medienwirkungen auf Jugendliche" anhand nicht-experimenteller Querschnittstudien belegt werden sollen, obwohl Aussagen über Ursache-Wirkungs-Relationen eine experimentelle Variation und/oder Messwiederholungsdesigns erfordern würden. Derartigen Formen der *nicht-intendierten Pseudowissenschaft* (Döring & Bortz, im Druck, Kap. 3) gilt es durch verbesserte Methodenausbildung entgegen zu wirken.

Ein weiterer Grund für pseudowissenschaftliche Studien kann in *ökonomischen Interessen und/oder ideologischen Glaubenssystemen* liegen, auf deren Basis das inhaltliche Ergebnis der entsprechenden Studien von vorne herein feststeht. Unter dem Anschein der Wissenschaftlichkeit wird dieses vorher fest stehende Wunschergebnis als Resultat eines angeblich ergebnisoffenen Forschungsprozesses ausgegeben. Ein prominentes Beispiel sind die

über Jahrzehnte hinweg von der Zigarettenindustrie gesponserten Studien, die gezielt dafür sorgen sollten, längst bekannte Gesundheitsrisiken des Rauchens angeblich wissenschaftlich seriös zu widerlegen (Cummings, Brown & O'Connor, 2007). Heute wird von der Zigarettenindustrie die Gesundheitsschädlichkeit des Rauchens offiziell nicht mehr bestritten. Die Folgen der gezielten Fehlinformation der Öffentlichkeit durch pseudowissenschaftliche Studien zur vermeintlichen Harmlosigkeit des Rauchens erweisen sich aber nach wie vor als Herausforderung für die Gesundheitsbildung.

Die Grenze zwischen seriöser Forschung einerseits und interessengeleiteter Pseudowissenschaft andererseits ist indessen nicht immer so klar zu ziehen. Ein Beispiel hierfür ist die heftige wissenschaftsinterne und öffentliche Kontroverse um die Wirksamkeit des sog. „Dore Programme" zur Behandlung von Dyslexie und ADHS, das der britische Geschäftsmann Wynford Dore für seine Tochter Susie entwickelt und dann weltweit als „Wundertherapie" offensiv vermarktet hatte.[1] Das Programm basiert auf der Hypothese, dass Lese- und Aufmerksamkeits-Störungen durch ein *generelles Automatisierungsdefizit* entstehen, für welches das Kleinhirn verantwortlich sei, so dass sensomotorische Übungen, die das Kleinhirn trainieren (z.B. Balancieren, Ballwerfen), eine effektive Behandlung darstellen sollen. In der Fachzeitschrift „*Dyslexia. An International Journal of Research and Practice"* wurden zwei Studien veröffentlicht, welche die Wirksamkeit des Dore-Programms nachweisen sollen (Reynolds, Nicolson & Hambly, 2003; Reynolds & Nicolson, 2007). Die wissenschaftliche Seriosität dieser Studien wurde jedoch in Zweifel gezogen: Neun kritische Kommentare wurden publiziert und insgesamt sechs Mitglieder des Editorial Board der Zeitschrift traten aus Protest gegen die Veröffentlichung der von ihnen als fragwürdig angesehenen Studien zurück, wie die Fachzeitschrift *Nature Neuroscience* (2007) berichtet. Es wurde im Nachhinein aufgedeckt, dass die Studien von Dore gesponsert worden waren. Ob es sich hier um geplante und bezahlte Pseudowissenschaft handelt oder einfach nur um methodisch schwache Studien, die dann interessengeleitet überinterpretiert und zur Vermarktung des Programms genutzt wurden, ist schwer zu beurteilen. Die Kleinhirndefizit-Hypothese der Dyslexie, auf die sich das Dore-Programm beruft, wird in der Fachliteratur kritisch diskutiert (Zeffiro & Eden, 2001; Raberger & Wimmer, 2003). Zur endgültigen Klärung der Frage, ob das Dore-Programm wirkt, müssten strenge unabhängige Wirksamkeitsstudien in Form randomisierter Kontrollgruppenexperimente mit Kindern und Jugendlichen durchgeführt werden, bei denen Dyslexie bzw. ADHS diagnostiziert ist.

1 Siehe www.dore.co.uk

2.3 Wissenschafts- und Forschungsethik

Eine erziehungswissenschaftliche Studie kann nicht den Anspruch der Wissenschaftlichkeit erheben, wenn sie nicht den zentralen Regeln der Wissenschafts- und Forschungsethik folgt, wie sie u.a. im Ethik-Kodex der Deutschen Gesellschaft für Erziehungswissenschaft niedergelegt sind (DGfE, 2010). Die Wissenschaftsethik bezieht sich dabei auf wissenschaftsinterne Prozesse (z.B. Festlegung der Autorschaft), die Forschungsethik dagegen auf den Umgang mit Untersuchungspersonen (Israel & Hay, 2006; Mertens & Ginsberg, 2008; Döring & Bortz, im Druck, Kap. 4).

Verletzungen der *Wissenschaftsethik* liegen z.B. vor, wenn Daten manipuliert, Ideen gestohlen, Sponsoren oder Interessenskonflikte der Forschenden verschwiegen werden. Dies ist teilweise bei *pseudowissenschaftlichen Studien* der Fall. Dass etwa in den oben angeführten Dore-Studien das Sponsoring durch Dore nicht von Anfang an offen gelegt worden war (was den Verdacht der Pseudowissenschaft stützte), lag auch daran, dass die Zeitschrift „Dyslexia“ in ihren Artikeln keine Hinweise auf mögliche Interessenskonflikte verlangt, wie das in anderen Fachzeitschriften mittlerweile üblich ist. Auch fehlende Angaben zu Grenzen des Gültigkeitsbereichs der Befunde, das Hochstilisieren von Einzelergebnissen zu unumstößlichen Gewissheiten sowie die mangelnde Einbeziehung alternativer Erklärungen und kritischer Sichtweisen widerspricht jeweils den *Standards guter wissenschaftlicher Praxis*, wie sie z.B. von der Deutschen Forschungsgemeinschaft ausformuliert wurden (DFG, 2013).

Verletzungen der *Forschungsethik* entstehen dort, wo Menschen ohne ihr Einverständnis zu Untersuchungspersonen gemacht werden, wo sie durch die Tätigkeit der Forschenden ungerechtfertigt beeinträchtigt oder gar geschädigt werden. Forschungsfreiheit ist zwar in der Verfassung der Bundesrepublik Deutschland verankert (Art. 5 Abs. 3 Satz 1 des Grundgesetzes), darf aber nicht Persönlichkeits- und Menschenrechte beschneiden. International geht die Entwicklung in allen sozial- und humanwissenschaftlichen Disziplinen – so auch in der Erziehungswissenschaft – in die Richtung, dass die Einhaltung von Standards der Forschungsethik im Vorfeld der Datenerhebung durch *Ethik-Kommissionen* bescheinigt werden muss (Miethe, 2013). Zweifel an ethisch korrektem Vorgehen führen u.a. dazu, dass eine Studie nicht als wissenschaftlich anerkannt wird und z.B. nicht in Fachzeitschriften publiziert werden kann. Auch wird der Ruf seriöser Wissenschaft durch unethisches Verhalten beschädigt, und die Bereitschaft der Bevölkerung gemindert, an wissenschaftlichen Studien teilzunehmen.

2.4 Dokumentation des Forschungsprojekts

Zu den zentralen Standards der Wissenschaftlichkeit gehört immer auch die Transparenz, d.h. das gesamte Vorgehen im Forschungsprojekt muss in-

tersubjektiv nachvollziehbar dargelegt sein. Dies setzt eine umfassende Dokumentation der einzelnen Entscheidungen und Schritte im Forschungsprozess einschließlich der erhobenen Daten voraus (Döring & Bortz, im Druck, Kap. 13). Üblicherweise wird nicht nur ein *interner Forschungsbericht* verfasst, sondern die Studie auch *publiziert*. Erst durch die detaillierte Dokumentation und Publikation lässt sich von Außenstehenden überhaupt beurteilen, ob die Studie forschungsethisch unbedenklich ist, ob sie mit den richtigen Begründungen anerkannte wissenschaftliche Methoden korrekt einsetzt, und ob ein echtes Forschungsproblem am Anfang formuliert und am Ende auch angemessen gelöst wurde.

3. Handelt es sich um eine insgesamt besonders gute wissenschaftliche Studie? Vier Kriterien der wissenschaftlichen Qualität

Kann man sich anhand der oben ausgeführten vier elementaren Standards der Wissenschaftlichkeit darauf einigen, dass sich eine konkrete erziehungswissenschaftliche Studie auf dem Boden der Wissenschaftlichkeit bewegt und eben nicht als vorwissenschaftlich, parawissenschaftlich oder pseudowissenschaftlich einzustufen ist, dann gilt es in vielen Kontexten noch zu klären, wie hoch ihre wissenschaftliche Qualität ist. Entsprechend bewertet werden u.a. erziehungswissenschaftliche Studien, die zur Publikation in einer Fachzeitschrift, zur Präsentation auf einer Fachkonferenz oder zur Erlangung des Doktor-, Master-, oder Bachelorgrades eingereicht werden.

Dazu werden üblicherweise die vier oben ausgeführten Standards der Wissenschaftlichkeit mit detaillierteren Qualitätsanforderungen unterlegt, so dass graduelle Abstufungen der Qualitätsbewertung möglich sind.

3.1 Grad der inhaltlichen Relevanz des Forschungsproblems

Die Wissenschaftlichkeit verlangt die Wahl eines wissenschaftlichen Forschungsproblems als Startpunkt der Forschungsbemühungen. Eine *gute* erziehungswissenschaftliche Studie wird hierbei ein Forschungsproblem herausgreifen, das hohe inhaltliche Relevanz besitzt, d.h. dessen Bearbeitung einen großen Erkenntnisfortschritt bringt. Im Kontext der Grundlagenwissenschaft spricht man von *theoretischer Relevanz* (z.B. eine Studie trägt dazu bei, eine neue erziehungswissenschaftliche Theorie zu entwickeln oder eine etablierte Theorie kritisch zu prüfen). Im Kontext der Anwendungsforschung geht es vor allem um den Grad der *praktischen Relevanz* (z.B. eine Studie trägt dazu bei, eine Lehr- oder Lernmethode zu verbessern).

Relevanzbewertungen sind komplex und hängen stark von der Perspektivität der Urteilenden (etwa ihren eigenen Themenpräferenzen) sowie auch vom

„Zeitgeist“ ab. In jedem Fall muss für eine empirische Studie, die auf der Relevanzdimension hohe Qualität beanspruchen möchte, sorgfältig argumentiert werden, inwiefern sie einen wichtigen Erkenntnisbeitrag und nicht nur abseitige Befunde liefert und inwiefern sie für die jeweilige Zielgruppe der Konferenz oder Fachzeitschrift von Belang ist. Fachkonferenzen und Fachzeitschriften benennen üblicherweise konkrete Themenschwerpunkte, zu denen sie Einreichungen erwarten, während Studien außerhalb dieses Themenspektrums dann eher als irrelevant abgelehnt werden. Vordringlich ist die Begründung und Bewertung der inhaltlichen Relevanz einer Studie vor allem auch in der Planungsphase größerer Forschungsprojekte, für die Fördergelder akquiriert werden sollen. Hier legen wiederum unterschiedliche Forschungsfördereinrichtungen unterschiedliche Relevanzkriterien an (z.B. hinsichtlich der Frage, ob die theoretische oder die praktische Relevanz eines Forschungsproblems höher zu gewichten ist).

3.2 Grad der methodischen Strenge des Forschungsprozesses

Von einer wissenschaftlichen Studie verlangen wir, dass sie einen zum Forschungsproblem passenden, geordneten Forschungsprozess durchläuft und anerkannte wissenschaftliche Methoden korrekt anwendet. Von einer *guten* Studie erwarten wir, dass besonders aussagekräftige und anspruchsvolle Methoden eingesetzt werden, um möglichst gesicherte Erkenntnisse zu gewinnen. Aspekte der methodischen Strenge betreffen den gesamten Forschungsprozess, der je nach Forschungsparadigma (quantitativ versus qualitativ) und Methodologie (z.B. bevölkerungsrepräsentative Umfrageforschung versus ethnografische Feldforschung versus Experimentalforschung im Labor) ganz unterschiedlich beschaffen sein kann. Zur Beurteilung der methodischen Strenge müssen also zahlreiche Unteraspekte des methodischen Vorgehens einzeln auf den Prüfstand gestellt werden.

Für sieben Phasen des empirisch-quantitativen Forschungsprozesses sollen wichtige Aspekte der methodischen Strenge im Folgenden angesprochen werden. Ein methodisch strenges Vorgehen führt hierbei jeweils zu hoher Gültigkeit (Validität) der Befunde der Studie (Shadish, Cook & Campbell, 2002). Das zentrale Gütekriterium der Validität wird hier eingeführt und in Abschnitt 4 dann genauer ausgeführt.

3.2.1 Theorie und Forschungsstand

Um für eine empirisch-quantitative Studie den aktuellen Forschungsstand und den theoretischen Rahmen zu erarbeiten, ist eine Recherche und Aufarbeitung der für das Forschungsproblem einschlägigen wissenschaftlichen Fachliteratur notwendig (Döring & Bortz, im Druck, Kap. 6). Eine solche Literaturrecherche kann sehr streng bzw. systematisch erfolgen (z.B. vorherige Festlegung von deutschen und englischen Suchbegriffen, Verwendung mehrerer wissenschaftlicher Literaturdatenbanken) oder eher auf Zufalls-

funden basieren. Die Stringenz des Vorgehens bei der *Aufarbeitung des Theorie- und Forschungsstandes* entscheidet in der quantitativen Forschung vor allem auch darüber, wie gut die verwendeten theoretischen Konstrukte definiert sind, was wiederum die Voraussetzung für ihre möglichst unverzerrte Messung darstellt. Hier ist vor allem das Gütekriterium der *Konstruktvalidität* betroffen: Nur wenn vorab eindeutig geklärt wurde, was die interessierenden theoretischen Konzepte inhaltlich bedeuten, können passgenaue Messinstrumente ausgewählt bzw. entwickelt und dann aussagekräftige Messwerte erhoben werden.

3.2.2 Untersuchungsdesign

Mit dem Untersuchungsdesign ist die Gesamtkonzeption einer Studie gemeint. Das Design hat verschiedene Aspekte, wobei in der quantitativ-empirischen Forschung vor allem entscheidend ist, ob es sich um ein experimentelles, ein quasi-experimentelles oder ein nicht-experimentelles Studien-Design handelt (Döring & Bortz, im Druck, Kap. 7). Will man Ursache-Wirkungs-Relationen (Kausalität) prüfen, so ist die *experimentelle Kontrollgruppenstudie* der sog. Goldstandard des Erkenntnisgewinns. In einem echten Experiment wird nämlich der vermutete Kausaleffekt von den Forschenden in wiederholbarer und nachprüfbarer Weise selbst hergestellt. Dazu werden mindestens zwei Gruppen vergleichbarer Untersuchungspersonen gebildet (bei ausreichend großen Gruppen wird dies durch sog. Randomisierung = zufällige Zuordnung homogener Untersuchungspersonen zu den Gruppen erreicht). Beispiel: Ein Pool von freiwilligen Seniorinnen und Senioren ähnlicher Altersgruppe, kultureller Herkunft und gesundheitlicher Verfassung wird per Zufall (z.B. per Losverfahren) in zwei Gruppen eingeteilt. Durch diese Zufallsaufteilung ist davon auszugehen, dass sich im Durchschnitt beide Gruppen maximal ähneln und keine systematischen Unterschiede bestehen. Die Experimentalgruppe wird dem vermuteten Kausalfaktor ausgesetzt (z.B. Computerschulung mit einer neuen Unterrichtsmethode), die Kontrollgruppe dagegen nicht (z.B. Computerschulung gemäß der herkömmlichen Methode). Dann wird kurz-, mittel- und/oder langfristig beobachtet bzw. gemessen, ob sich der erwartete Kausaleffekt in der Weise zeigt, dass die Untersuchungspersonen in der Experimentalgruppe die erwarteten Wirkungen statistisch signifikant (überzufällig) und praktisch bedeutsam stärker ausprägen als die Kontrollgruppe (z.B. deutlich höherer Kompetenzzuwachs mit der neuen Unterrichtsmethode gegenüber der herkömmlichen Methode). Das Experiment stellt die strengste Kausalitätsprüfung dar, die Befunde weisen *hohe interne Validität* auf, d.h. sie erlauben sehr klare Rückschlüsse auf die Existenz eines Ursache-Wirkungs-Zusammenhangs, da Verfälschungen und Störeinflüsse weitgehend ausgeschlossen sind.

Beim *Quasi-Experiment* werden die Untersuchungsgruppen aus pragmatischen Gründen nicht randomisiert gebildet, sondern als natürliche Gruppen

vorgefunden und dann von den Forschenden gezielt unterschiedlich behandelt. Beispiel: Als Experimentalgruppe wird die Seniorengruppe des Freizeittreffs A eingesetzt, als Kontrollgruppe die Seniorengruppe des Freizeittreffs B. Wenn nun die Experimentalgruppe besser als die Kontrollgruppe abschneidet, könnte dies nicht nur an der Wirksamkeit der neuen Unterrichtsmethode (Experimentalbedingung), sondern auch daran liegen, dass sich Gruppe A bereits vor der Behandlung systematisch von Gruppe B unterschied (z.B. höherer Bildungsstand; bessere soziale Unterstützung, besserer Gesundheitszustand). Ein Quasi-Experiment ist weniger aufwändig (da man auf die Randomisierung verzichtet) als ein echtes Experiment, die Ergebnisse eines Quasi-Experiments sind dafür im Hinblick auf Ursache-Wirkungs-Aussagen aber auch weniger aussagekräftig, haben geringere interne Validität. Mit bestimmten Techniken der Untersuchungsplanung (z.B. Erhebung von Kontrollvariablen) kann die interne Validität eines Quasi-Experiments gesteigert werden (Cook & Campbell, 1979; Döring & Bortz, im Druck, Kap. 7).

Bei einer *nicht-experimentellen Studie* schließlich, in der nur vorgefundene Verhältnisse beobachtet bzw. gemessen werden und die Kausalbedingungen von den Forschenden überhaupt nicht beeinflusst werden, sind letztlich kaum klare Rückschlüsse auf Ursache-Wirkungs-Relationen möglich, da zu viele weitere unkontrollierte Einflussfaktoren im Spiel sind. Hier ist die interne Validität (also die Gültigkeit von Kausalaussagen auf der Basis der Studie) sehr gering, kann aber durch bestimmte Techniken des Designs und der statistischen Analyse leicht verbessert werden (z.B. Erfassung und statistische Kontrolle zumindest einiger Störvariablen). Beispiel: Vergleicht man in einem querschnittlichen nicht-experimentellen Design Schulkinder, die gar nicht oder selten digitale Spiele spielen, mit Schulkindern, die regelmäßig oder häufig spielen, und stellt signifikant geringere Schulleistungen bei den Vielspielenden fest, so ist kein valider Rückschluss in der Weise möglich, dass das Spielen die schlechten Schulleistungen verursacht hat. Denn Wenig- und Vielspielende unterscheiden sich von vorne herein in zahlreichen Hintergrundvariablen voneinander, die ebenfalls die Schulleistung beeinflussen. Eine Studie, die eine kausale Forschungsfrage beantworten will, muss also von vorne herein methodisch so streng angelegt werden, dass sie hohe interne Validität erreicht.

Ein strenger Nachweis der Kausalität unter experimentellen Bedingungen ist jedoch für tragfähige wissenschaftliche Schlussfolgerungen allein nicht ausreichend. Es muss zudem sichergestellt werden, dass die Befunde über die konkreten Bedingungen eines einzelnen Experiments hinaus generalisierbar sind – auf andere Orte, andere Zeiten, andere Situationen, andere Personengruppen etc. Man spricht von der *externen Validität*, wenn es darum geht, wie breit Befunde verallgemeinerbar sind. Hinsichtlich des Untersuchungsdesigns lassen sich unterschiedliche Maßnahmen ergreifen, um die externe Validität zu steigern. Etwa indem man die Experimentalsituation möglichst

eng an natürlichen Bedingungen orientiert, vielleicht sogar das Experiment nicht im Labor, sondern in Alltagssituationen durchführt (Feldexperiment). Auch steigt die Generalisierbarkeit, wenn man statt nur *einer* Experimental- und *einer* Kontrollgruppe mit mehreren Vergleichsgruppen arbeitet, d.h. die Dosierung oder Ausgestaltung des experimentellen Stimulus variiert: Bei der Prüfung der Wirkung einer neuen Unterrichtsmethode kann diese z.B. mehr oder minder häufig und/oder von verschiedenen Lehrkräften und/oder in verschiedenen Räumlichkeiten und/oder mit verschiedenen Gruppengrößen umgesetzt werden. Je mehr Variationen in der Studie abgedeckt werden, umso klarer sind Verallgemeinerungsmöglichkeiten der Befunde. Aus forschungsökonomischen Gründen (Zeitrahmen, Personal) ist man hierbei jedoch in der Regel bei der Durchführung einer Primärstudie (bei der eigene Daten erhoben und ausgewertet werden) limitiert.

Besondere Bedeutung haben deswegen Studiendesigns, die sich auf die Verarbeitung bereits vorliegender Daten stützen. Insbesondere die *Metaanalyse* ist eine Form der quantitativen Studie, die hohe externe Validität besitzt. Die Metaanalyse sucht und bündelt nämlich zu einem bestimmten Zeitpunkt alle qualitativ hochwertigen vorliegenden empirisch-quantitativen Studien über einen bestimmten Effekt, berechnet daraus den statistischen Gesamteffekt und arbeitet im Zuge einer Moderatorenanalyse auch heraus, unter welchen Bedingungen der Effekt stärker oder schwächer ausgeprägt ist (zu Forschungssynthese und Metaanalyse siehe Cooper, Hedges & Valentine, 2009; Döring & Bortz, im Druck, Kap. 16).

3.2.3 Operationalisierung

Die Operationalisierung legt fest, mit welchen Messinstrumenten (z.B. psychologischer oder pädagogischer Test; Einstellungs- oder Persönlichkeitsfragebogen; physiologische Messung; Beobachtungsplan etc.) die interessierenden Variablen bei einer Primärstudie erfasst werden sollen (Döring & Bortz, im Druck, Kap. 8). Die Operationalisierung kann dabei sehr streng erfolgen, indem mehrere verschiedene und jeweils *gut geprüfte Messinstrumente* – passend zu den interessierenden theoretischen Konstrukten – ausgewählt werden. Oder die Operationalisierung ist weniger streng und basiert z.B. größtenteils auf wenig oder gar nicht erprobten selbst konstruierten Skalen oder auf sehr einfachen Messinstrumenten (z.B. sog. *Single-Item-Measures*, die ein Konstrukt mit nur einer einzigen Frage oder Aufgabe abdecken, im Unterschied zu ganzen *psychometrischen Skalen*, die ein Konstrukt über ein Bündel ähnlicher Fragen bzw. Aufgaben erfassen). Als Gütekriterium der Operationalisierung ist hier wiederum vor allem die *Konstruktvalidität* angesprochen.

3.2.4 Stichprobenziehung

Bei empirisch-quantitativen Studien handelt es sich meist nicht um Einzelfallstudien und auch nicht um Vollerhebungen von Populationen, sondern

um Stichproben-Untersuchungen, in denen Ausschnitte der Zielpopulation untersucht werden (Döring & Bortz, im Druck, Kap. 9). Eine Studie ist im Hinblick auf die Stichprobenziehung besonders streng und aussagekräftig, wenn die Stichprobe ein möglichst unverzerrtes Miniaturabbild der Population darstellt. Dies wird am besten durch *zufallsgesteuerte (probabilistische) Stichproben* einer entsprechenden Mindestgröße erreicht. Sie können als global repräsentativ für die Population gelten. Weniger streng im Hinblick auf die Stichprobenziehung sind Studien, die mit *nicht-zufallsgesteuerten (nicht-probabilistischen) Stichproben* operieren, welche die Population nur sehr verzerrt widerspiegeln (wobei häufig Art und Umfang der Verzerrung gar nicht genau bekannt sind). Von Daten einer verzerrten Stichprobe lässt sich nur sehr bedingt auf die Verhältnisse in der Population rückschließen. Die externe Validität bzw. Generalisierbarkeit der Studienbefunde im Hinblick auf die Population ist bei nicht-zufälligen Stichproben (z.B. Gelegenheitsstichproben, Quotenstichproben oder Schneeball-Stichproben) somit eingeschränkt.

Beispiel: Zielsetzung der *PISA-Studie* ist es, die Lesekompetenz sowie die mathematische und naturwissenschaftliche Grundbildung der Population der 15-/16-jährigen Schüler/innen in Mitgliedstaaten der OECD zu erfassen und zu vergleichen. Hierzu wird eine zweistufige probabilistische Stichprobenauswahl realisiert[2]: Im ersten Schritt wird aus allen Schulen des jeweiligen Landes eine definierte Anzahl an Schulen per Zufall ausgewählt. Im zweiten Schritt wird von allen Schüler/innen der entsprechenden Altersgruppe der ausgewählten Schulen jeweils eine definierte Anzahl per Zufall in die Stichprobe gezogen. Zudem wird zur Sicherstellung der Repräsentativität der Stichprobe verlangt, dass sich die per Zufallsverfahren gezogenen Schulen und Schüler/innen auch mit großer Mehrheit (>85% bei Schulen und >80% bei Schüler/innen) an der Studie beteiligen, denn massenhafte Teilnahmeverweigerung würde zu Verzerrungen führen. Diese aufwändige zweistufige zufallsgesteuerte Stichprobenauswahl sichert hohe externe Validität der Befunde: Auf der Basis der untersuchten Stichprobe von Schulkindern kann mit hoher Zuverlässigkeit auf die Gesamtheit der Schulkinder des Landes geschlossen werden. Anders wäre es, würde man willkürlich nur einige für die Forschenden leicht erreichbare Schulen in geografischer Nähe von Forschungsinstituten ansprechen (nicht-zufällige Gelegenheitsstichprobe) oder nur Schulen einbeziehen, die sich auf einen öffentlichen Aufruf hin freiwillig melden (nicht-zufällige Selbstselektionsstichprobe).

3.2.5 Datenerhebung

Im Zuge der Datenerhebung werden die in der Phase der Operationalisierung ausgewählten bzw. entwickelten Messinstrumente tatsächlich einge-

2 Siehe www.oecd.org/de/pisa

setzt, um Daten von allen Personen der Stichprobe zu erlangen (z.B. die Tests werden durchgeführt; Fragebögen ausgeteilt und eingesammelt; physiologische Messungen im Labor vorgenommen; Döring & Bortz, im Druck, Kap. 10). Die Datenerhebung zeichnet sich durch methodische Strenge aus, wenn sie plangemäß und standardisiert erfolgt, etwa durch geschultes Testpersonal, so dass möglichst wenig Messfehler, fehlende Werte und Antwortverweigerungen entstehen. Weniger streng ist eine Datenerhebung in der quantitativen Forschung, die nicht unter vergleichbaren Umständen stattfindet, bei der die Befragungspersonen z.B. unterschiedliche Zusatzinformationen über die Studie erhalten oder den Testpersonen unterschiedlich lange Bearbeitungszeiten zugebilligt werden. Nachlässigkeiten bei der Datenerhebung gehen wiederum auf Kosten der *Konstruktvalidität*, da Verzerrungen, Fehler und Lücken in den Messwerten entstehen, die somit nicht genau die theoretisch angezielten Konstrukte erfassen.

3.2.6 Datenaufbereitung

Fast jeder empirische Datensatz, der am Ende der Datenerhebungsphase vorliegt, ist mehr oder minder stark „verschmutzt“: Er enthält unvollständige Daten, Tippfehler, unplausible Spaßantworten, Dopplungen usw. Eine methodisch strenge Studie führt deswegen eine sorgfältige und systematische Datenbereinigung durch, um die Datenqualität zu erhöhen (Schendera, 2007; Döring & Bortz, im Druck, Kap. 11). Dazu gehört neben dem Eliminieren von Fehlern auch das Hinzufügen von Meta-Informationen (z.B. genaue Angaben zu Orten und Zeiten der Datenerhebung; schlüssige namentliche Kennzeichnung aller Variablen und Werte) sowie ggf. das Sicherstellen der Anonymität der Untersuchungspersonen durch Herausnahme oder Veränderung identifizierender Informationen. Dabei wird das Vorgehen detailliert dokumentiert und begründet. Weniger methodisch strenge Studien adressieren Fragen der Datenqualität nicht ausdrücklich und schließen z.B. unvollständige Fälle mehr oder minder willkürlich ein oder aus.

3.2.7 Datenanalyse und Interpretation

Im quantitativen Paradigma der empirischen Sozialforschung werden mit standardisierten Verfahren numerische Messwerte erhoben und im Zuge der Datenanalyse statistisch ausgewertet (Eid, Gollwitzer & Schmitt, 2011; Döring & Bortz, im Druck, Kap. 12). Eine methodisch strenge Studie nutzt dabei jeweils genau diejenigen statistischen Verfahren, die zur Beantwortung der Forschungsfragen bzw. Prüfung der Forschungshypothesen am besten geeignet sind und auf die jeweilige Datenlage anwendbar sind (z.B. Prüfung von statistischen Voraussetzungen vor Einsatz eines bestimmten statistischen Signifikanztests). Ebenso wird eine methodisch strenge empirisch-quantitative Studie sich nicht allein auf das Berichten der *statistischen Signifikanzen* beschränken, sondern auch die *Teststärke* (d.h. war der Stichprobenumfang der Studie groß genug, damit ein in der Population vorhandener

Effekt bestimmter Größe überhaupt signifikant werden kann?) sowie die *Effektgröß*e (ist ein statistisch signifikanter Effekt groß genug, um ihm – im jeweiligen Forschungsfeld – praktische Bedeutsamkeit zuzusprechen?) berechnen und diskutieren. Der korrekte Umgang mit dem statistischen Instrumentarium der Datenanalyse ist Voraussetzung dafür, dass eine Studie sinnvoll interpretierbare Befunde hervorbringt und somit das studienleitende Forschungsproblem gelöst werden kann. Man spricht hier auch von *statistischer Validität*. Eine weniger strenge Studie geht bei der statistischen Analyse nicht gleichermaßen sorgfältig vor, wodurch sich mehr oder minder große Berechnungsfehler und inhaltliche Fehlschlüsse einschleichen können.

Wenn etwa ein Signifikanztest berechnet wird, ohne die statistischen Voraussetzungen der Daten zu prüfen oder ohne Ausreißerwerte korrekt zu behandeln, können stark verzerrte und ggf. *falsch-positive Befunde* resultieren (sog. Alpha-Fehler oder Fehler erster Art: fälschliche Annahme der Forschungshypothese). Problematisch sind auch statistische Analysen, die nur die statistische Signifikanz betrachten und die Frage ausblenden, wie groß und praktisch bedeutsam die gefundenen Effekte eigentlich sind. Gerade bei großen Stichproben werden schon winzige Effekte signifikant, die vielfach kaum praktische Bedeutung haben.

Ein anderes Problem besteht darin, dass ein nicht-signifikanter Befund fälschlich als Beleg gegen die Forschungshypothese ausgelegt wird (sog. Beta-Fehler oder Fehler zweiter Art: *falsch-negativer Befund*), ohne dass geprüft wurde, ob die Teststärke überhaupt ausreichend war, um einen Effekt nachzuweisen. Aktuell geht man davon aus, dass aufgrund mangelnder Betrachtung der statistischen Teststärke (die sich oft als viel zu gering erweist) ein Großteil sozialwissenschaftlicher Studien falsch-negative Befunde berichtet (Ellis, 2010, S. 76). Schwächen in der statistischen Validität einer Studie können somit zu eklatanten Fehlschlüssen führen. Dass es sich bis heute nicht überall eingebürgert hat, die Befunde einer quantitativen Studie in der Gesamtschau zu würdigen und dabei statistische Signifikanz, Effektgröße und Teststärke ausdrücklich einzubeziehen, zeigt auf, dass Qualitätsprobleme empirisch-quantitativer Studien nicht nur das Vorgehen einzelner Forschender betreffen, sondern das gesamte Wissenschaftssystem (z.B. Publikationswesen und dessen Anforderungen; Art und Umfang der Methodenausbildung in einzelnen Studiengängen).

3.3 Grad der ethischen Strenge des Forschungsprozesses

Eine wissenschaftliche Studie muss Regeln der Wissenschafts- und Forschungsethik einhalten. Eine *gute* Studie weist einen besonders hohen Grad an ethischer Strenge auf, d.h. sie erfüllt Standards der Wissenschafts- und Forschungsethik besonders umfassend und nachhaltig. Diese Qualität ist insbesondere in Forschungskontexten gefragt, in denen ethische Dilemmata

und Konflikte gehäuft auftreten oder auftreten können. Bei diversen ethischen Fragen gibt es im Detail keine „Patentlösung", deswegen ist hohe ethische Strenge daran erkennbar, wie intensiv nach überzeugenden Lösungen gesucht wird, möglichst unter Einbeziehung aller betroffenen Anspruchsgruppen (Israel & Hay, 2006; Mertens & Ginsberg, 2008).

Das bezieht sich im Rahmen der Wissenschaftsethik z.B. auf faire Regelungen rund um die *Autorschaft von wissenschaftlichen Publikationen*, die umso komplizierter werden, wenn man mit großen, interdisziplinären und ggf. auch interkulturellen Teams arbeitet. Angesichts der Komplexität allein dieses Einzelaspekts der Wissenschaftsethik hat die in Fragen des wissenschaftlichen Publizierens sehr aktive American Psychological Association umfangreiches Informationsmaterial im Web bereitgestellt, um Forschende darin zu unterstützen, ethisch korrekt mit wissenschaftlicher Autorschaft umzugehen (sog. „Responsible Authorship").[3] Zu beachten ist dabei etwa, welche Beiträge zu einer Studie (z.B. Idee zum Studiendesign; Analyse der Daten; Korrektur oder Übersetzung des Manuskripts) dazu berechtigen (oder eben auch nicht dazu ausreichen), Mitautorschaft zu beanspruchen. Hier unterscheiden sich aber auch teilweise die Gepflogenheiten je nach Wissenschaftsdisziplin und Forschungseinrichtung deutlich.

Im Bereich der Forschungsethik, die sich auf den Umgang mit Untersuchungspersonen und ihren Rechten bezieht, stellen sich zahlreiche *neue Herausforderungen* u.a. durch innovative Methoden der Datenerhebung im Internet: Unter welchen Umständen etwa dürfen Beiträge in Online-Diskussionsforen, in virtuellen Welten und Online-Spielen, auf Foto- oder Videoplattformen, in Blogging- und Microblogging-Diensten, auf Social-Networking- oder E-Learning-Plattformen frei für Forschungszwecke verwendet werden? Unter welchen Umständen muss das ausdrückliche Einverständnis der jeweils beteiligten Mitgliederkreise bzw. der einzelnen Beitrags-Autorinnen und -Autoren eingeholt werden? Derartige kontextspezifische ethische Fragen der Abwägung der wissenschaftlichen Interessen der Forschenden einerseits und der Wahrung der Rechte der Beforschten andererseits sind nicht trivial (da sich z.B. herkömmliche Grenzen zwischen Privatheit und Öffentlichkeit im Internet verschieben) und werden uns zukünftig noch stärker beschäftigen (siehe Abschnitt 6.3).

Zudem bleiben bereits lange diskutierte Ethikfragen weiterhin aktuell: Ethisch hochrelevant sind u.a. *Kompetenzmessungen im Bildungs- und Berufswesen*, da sich für die Testpersonen – insbesondere jenseits anonymisierter grundlagenwissenschaftlicher Forschung – in der Praxis oft Entscheidungen großer Tragweite ergeben können. Hier sind in Forschung und Praxis Standards der *Testethik* und *Testfairness* stets zu beachten, etwa um

3 Siehe http://apa.org/research/responsible/publication/

bestimmte Gruppen von Testpersonen nicht ungewollt zu benachteiligen. Maßgebliche Richtschnur bilden hier die von drei führenden internationalen Fachgesellschaften gemeinsam herausgegebenen „Standards for Educational and Psychological Testing“.[4]

3.4 Grad der Dokumentations- und Präsentationsqualität

Jede wissenschaftliche Studie muss in ihrem Ablauf und ihren Ergebnissen für Dritte nachvollziehbar dokumentiert werden. Eine *gute* Studie zeichnet sich darüber hinaus dadurch aus, dass sie die Dokumentation besonders gründlich und die Präsentation der Ergebnisse besonders zielgruppengerecht, ansprechend und ausgewogen gestaltet. Dies kann z.B. auch beinhalten, dass die statistischen Befunde übersichtlich und vollständig in Tabellen und Grafiken aufbereitet sind. Angesichts wachsender Bedeutung von *Forschungssynthese* (Zusammenfassung von Einzelstudien z.B. im Rahmen von Metaanalysen und systematischen Forschungsreviews; Cooper, Hedges & Valentine, 2009) gewinnt Präsentationsqualität in der Weise an Bedeutung, dass für jede statistische Analyse alle statistischen Kennwerte vollständig und in standardisierter Form berichtet werden müssen – nur so ist es möglich, die Studie später problemlos in einer Metaanalyse zu verrechnen.

Wie die Ergebnispräsentation im Detail vorzunehmen ist (z.B. welche statistischen Kennwerte für eine Varianzanalyse oder ein Strukturgleichungsmodell in welcher Weise im Fließtext, in einer Tabelle und/oder Grafik zu präsentieren sind) wird vom APA Publication Manual vorgegeben, das in vielen sozialwissenschaftlichen Disziplinen als bindend gilt (APA, 2010). Fachzeitschriften formulieren darüber hinaus teilweise leicht abweichende oder ergänzende Anforderungen an die Aufbereitung von Studien in Fachartikeln. Zur Präsentationsqualität empirisch-quantitativer Studien gehört indessen nicht nur die Darstellung der statistischen Ergebnisse, sondern auch der Sprachstil. So sind laut APA Publication Manual z.B. diskriminierende Bezeichnungen für Personengruppen oder auch ein *Gender Bias* in der Weise, dass sprachlich nur die männlichen Formen verwendet werden (sog. Generisches Maskulinum), aber eigentlich Menschen aller Geschlechter gemeint sind, als Mängel in der wissenschaftlichen Präsentationsqualität zu werten (APA, 2010, S. 73f.). Denn bei Verwendung des generischen Maskulinums werden z.B. weibliche Personen nachweislich eben nicht gleichermaßen „mitgedacht“ und fühlen sich auch weniger angesprochen (Irmen & Köncke, 1996; Miller & James, 2009).

4 Siehe www.teststandards.org (AERA, APA, & NCME, 2014).

4. Handelt es sich speziell um eine methodisch besonders strenge Studie? Vier Typen der Validität in der Campbell-Tradition

Die methodische Strenge haben wir bereits als ein in sich stark ausdifferenziertes Qualitätskriterium wissenschaftlicher Studien kennengelernt, das in diversen Unteraspekten in allen Phasen des empirischen Forschungsprozesses eine Rolle spielt (siehe oben Abschnitt 3.2). Wenn es darum geht, die methodische Strenge einer quantitativ-empirischen Studie zu bewerten, ist das zentrale Gütekriterium die *Validität* (d.h. die Gültigkeit der aus der Studie abgeleiteten Schlussfolgerungen; Shadish, Cook & Campbell, 2002, S. 35). Obwohl die Validität die Gültigkeit der Schlussfolgerungen bzw. Ergebnisse einer methodisch strengen Studie meint, wird verkürzend und vereinfachend auch oft die Studie selbst, ihr Design, ihre Stichprobe oder das verwendete Datenerhebungsinstrument als „valide“ bezeichnet. Diese verkürzende Sprechweise ist üblich, aber streng genommen eben ungenau.

In der Tradition des Methodikers Donald T. Campbell werden vier bereits angesprochene Formen der Validität differenziert (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002): Konstruktvalidität, interne Validität, externe Validität sowie statistische Validität. Eine wichtige Leistung von Campbell und Kollegen bestand darin, sehr präzise herauszuarbeiten, auf welche Weise das mehr oder weniger strenge methodische Vorgehen einer Studie die jeweilige Validität steigert (oder eben mindert) und damit die Aussagekraft der Befunde der Studie beeinflusst. Genaue Kenntnisse über die Mechanismen der Validitätsbedrohung in den verschiedenen Phasen des Forschungsprozesses helfen wiederum dabei, a) vorliegende empirisch-quantitative Studien und ihre Befunde differenziert zu bewerten sowie b) eigene Studien möglichst optimal aussagekräftig zu planen.

4.1 Konstruktvalidität

Eine verbreitete Grundsatzkritik am quantitativen Paradigma der empirischen Sozialforschung lautet, dass komplexe soziale Sachverhalte hier von den Forschenden recht beliebig und simplifizierend auf einzelne Variablen und deren Messwerte heruntergebrochen werden und die daraus errechneten Statistiken letztlich weitgehend aussagelose und beliebige „Erbsenzählerei“ darstellten.

Tatsächlich herrscht jedoch im quantitativen Paradigma der empirischen Sozialforschung gerade keine naive „Zahlengläubigkeit“ vor. Vielmehr wird ein Großteil der Forschungsaktivitäten genau darauf gelenkt, genau zu klären, welche Ausschnitte der sozialen Wirklichkeit für das Forschungsproblem relevant sind. Die einzelnen theoretischen Konstrukte und ihre Relationen untereinander werden auf der Basis des gesicherten Forschungsstandes

und etablierter Theorien definiert. Im Zuge der Operationalisierung wird auf dieser Basis begründet festgelegt, wie die Variablenausprägungen zu ermitteln sind. Entsprechende Messinstrumente durchlaufen dabei in der Regel mehrere Überprüfungs- und Revisionsdurchgänge (Döring & Bortz, im Druck, Kap. 8).

All diese Fragen werden in der Campbell-Tradition unter dem Gütekriterium der *Konstruktvalidität* zusammengefasst. Bei hoher Konstruktvalidität können die erhobenen Daten dann tatsächlich als Indikatoren für die Ausprägung der theoretisch interessierenden Konstrukte dienen (Shadish, Cook & Campbell, 2002, S. 73). Umgekehrt lässt sich formulieren: Die Frage, ob ein Messinstrument tatsächlich genau das Merkmal erfasst, das es zu messen vorgibt (Misst ein konkreter Intelligenztest tatsächlich Intelligenz – oder vor allem Schulbildung? Misst ein konkreter Persönlichkeitstest tatsächlich die Persönlichkeitseigenschaft Schüchternheit – oder doch eher Ängstlichkeit?) ist eine Frage der Konstruktvalidität. Längere Zeit wurde im Kontext der Testtheorie die *Güte von Messinstrumenten* mit einer Fülle von Validitätskonzepten beschrieben (z.B. neben der Konstruktvalidität auch die *Kriteriumsvalidität*, die *Augenschein-Validität*, die *logische Validität* etc.). Dies setzt sich bis heute teilweise fort, obwohl in der Methodenliteratur inzwischen die Konstruktvalidität als Oberbegriff bevorzugt wird (Messick, 1995; AERA, APA, & NCME, 2014). Noch einmal ist zu betonen, dass die Vorstellung und Formulierung, ein Messinstrument als solches sei „(konstrukt)valide“, verkürzend ist, denn konstruktvalide können immer nur Testergebnisse sein, die in einem spezifischen Studiendesign an bestimmten Personen erhoben und im Lichte von Theorien in bestimmter Weise von den Forschenden im Hinblick auf das Forschungsproblem interpretiert werden. Ein Test kann nicht universell beanspruchen, immer und unter allen Umständen zu gültigen Befunden zu führen, vielmehr müssen die Einsatzbedingungen und die aus den Testscores gezogenen Schlussfolgerungen kritisch bewertet werden. Des Weiteren ist zu betonen, dass die *Reliabilität* (Messgenauigkeit) ein der Validität untergeordnetes Kriterium ist: Dass ein Test wenig Messfehler erzeugt, ist notwendig für inhaltlich valide Schlussfolgerungen, aber eben nicht hinreichend (allein hohe Messgenauigkeit kann trotzdem bedeuten, dass der Test das inhaltlich falsche Konstrukt erfasst).

Die Konstruktvalidität steigt, je besser bei der Erarbeitung des Theorie- und Forschungsstandes die interessierenden theoretischen Konstrukte spezifiziert und definiert wurden. Sie steigt auch, wenn ein gut geprüftes Messinstrument genutzt wird, anstelle eines selbstkonstruieren Ad-hoc-Instruments. Die Konstruktvalidität steigt weiter, wenn das interessierende Merkmal nicht nur mit einem einzelnen Messinstrument, sondern mit mehreren möglichst unterschiedlichen Messinstrumenten erfasst wird. Eine gründliche Mehrfach-Messung ist vor allem für die zentralen Konstrukte einer Studie empfehlenswert, nicht für alle Variablen (insbesondere nicht für einfache sozialstatistische Hintergrundvariablen wie z.B. das Alter). Wenn

es aber beispielsweise darauf ankommt, in einer Studie den Lernerfolg in einem bestimmten Lehr-Lern-Setting zu untersuchen, so sollte dieser im Sinne der Konstruktvalidität nicht einfach durch ein von den Forschenden spontan entworfenes Einzelitem (etwa „Ich habe viel gelernt." – „stimmt gar nicht – stimmt wenig – stimmt teils-teils – stimmt ziemlich – stimmt völlig") erhoben werden, sondern möglichst bereits erprobte Instrumente aus der Literatur einbeziehen und zudem neben Selbstauskunftsdaten der Lernenden möglichst auch objektive Leistungsindikatoren (Kompetenztests; Beurteilungen durch Lehrkräfte etc.) einbeziehen. Zu beachten ist auch, dass vermeintlich einfache Hintergrundvariablen (z.B. Geschlecht, kultureller Hintergrund) theoretisch durchaus anspruchsvoll sind und einer reflektierten Operationalisierung bedürfen (z.B. Döring, 2013), um konstruktvalide Messwerte zu erzeugen.

4.2 Interne Validität

Im quantitativen Paradigma der empirischen Sozialforschung strebt man in der Tradition der Wissenschaftstheorie des *Kritischen Rationalismus* (Popper, 1934/2002; Döring & Bortz, im Druck, Kap. 2) danach, Theorien und Hypothesen über Ursache-Wirkungs-Relationen kritisch auf den Prüfstand zu stellen, um per Falsifikationsprinzip falsche Hypothesen auszusondern und die bewährten Hypothesen und Theorien vorläufig zu behalten.

Doch wenn sich in den empirischen Daten statistisch überzufällige Gruppenunterschiede, Messwertveränderungen über die Zeit oder Merkmalszusammenhänge zeigen, ist damit nicht automatisch die Frage nach der Kausalität beantwortet. Das Gütekriterium der internen Validität bezieht sich nun darauf, wie eindeutig die Befunde einer Studie als Hinweise auf Ursache-Wirkungs-Verhältnisse gelten können (Shadish, Cook & Campbell, 2002, S. 55).

Die höchste interne Validität weisen echte Experimentalstudien bzw. randomisierte Kontrollgruppenstudien auf, die deswegen auch als *Goldstandard* gelten, wenn es darum geht, *evidenzbasiert* zu argumentieren, welche pädagogischen Maßnahmen wirksam oder unwirksam sind. Doch um ausreichende interne Validität für eine schlüssige Kausalaussage zu beanspruchen, reicht es nicht, dass ein Experiment oder zumindest ein Quasi-Experiment durchgeführt wurde. Es muss sich auch um ein *gutes* Experiment handeln. Denn bei der Durchführung einer experimentellen oder quasi-experimentellen Studie können zahlreiche Probleme auftreten, die ihre Aussagekraft dramatisch senken. In der Campbell-Tradition (Shadish, Cook & Campbell, 2002, S. 55ff.) wird hier z.B. auf Einbußen der internen Validität durch experimentelle Mortalität (z.B. bestimmte Personen fallen aus einer Studie mit Messwiederholungen heraus) oder durch einen Testübungseffekt hingewiesen. Entsprechende Probleme gilt es umfassend im Zuge der Untersuchungsplanung zu antizipieren und bestmögliche Lösungen zu erarbei-

ten (z.B. durch Kontrollgruppen), wobei ein Höchstmaß an interner Validität meist aus forschungsökonomischen Gründen nicht erreichbar ist. Die verbleibenden Validitätseinbußen müssen dann offen gelegt und in die Ergebnisdiskussion einbezogen werden.

4.3 Externe Validität

Ein strikter Kausalitätsnachweis im Experiment erfordert oft eine Untersuchungsdurchführung unter hochkontrollierten, „künstlichen" Bedingungen (teilweise im Labor), zunächst auch meist mit wenigen Untersuchungspersonen (in der Experimentalforschung sind Gruppengrößen von 20-40 Personen üblich). Hier stellt sich dann die Frage, ob und inwieweit ein unter so speziellen Bedingungen nachgewiesener Kausal-Effekt auf andere Orte, Zeiten, Situationen und Personen generalisierbar ist (Shadish, Cook & Campbell, 2002, S. 87).

Hohe *externe Validität* im Sinne hoher Verallgemeinerbarkeit von Untersuchungsergebnissen erreicht man durch eine *größer angelegte Studie*, in der z.B. mehr Untersuchungsbedingungen durchvariiert und/oder mehr verschiedene Personengruppen untersucht werden. Weiterhin lässt sich die externe Validität eines Befundes vergrößern, wenn dafür Sorge getragen wird, dass die ursprüngliche Studie von den Forschenden selbst und/oder von anderen Forschungsgruppen unter mehr oder minder variierten Bedingungen wiederholt (repliziert) wird. Je vielfältiger die Kontexte, unter denen *Replikationsstudien* gelingen, umso besser ist die externe Validität des Effekts gesichert.

Wenn genügend Studien zu einem bestimmten Effekt vorliegen, kann und sollte der gesamte Forschungsstand in einem *systematischen Forschungsreview* zusammengefasst werden. Dabei sollten dann die vorliegenden statistischen Einzelbefunde *metaanalytisch* zu einem Gesamtbefund verrechnet werden – inklusive Moderatorenanalyse, die aufzeigt, unter welchen Bedingungen der Effekt stärker oder schwächer ausgeprägt ist (Cooper, Hedges & Valentine, 2009; Döring & Bortz, im Druck, Kap. 16). Metaanalysen, die sich zuerst in der Medizin etabliert haben, gewinnen in allen empirischen Sozialwissenschaften an Bedeutung, vor allem wenn es darum geht, gut gesicherte (*evidenzbasierte*) Aussagen über die Wirksamkeit von (quasi-)experimentell geprüften Interventionen zu treffen. (Man beachte, dass außerhalb der Experimentalforschung – etwa in der Umfrage- und Meinungsforschung sowie in der Kompetenzdiagnostik – hohe externe Validität oft durch große und repräsentative Stichproben erzielt wird, die jedoch in der Experimentalforschung selten zum Einsatz kommen.)

4.4 Statistische Validität

Eine quantitative Studie weist hohe statistische Validität auf, wenn die statistischen Datenanalysen passend zu den konkreten Forschungsfragen bzw. Forschungshypothesen sowie passend zum vorliegenden Datenmaterial korrekt und umfassend ausgeführt und interpretiert werden (Shadish, Cook & Campbell, 2002, S. 45).

Zu vermeiden sind sowohl falsch-positive Befunde (also das Begehen eines Alpha-Fehlers im Sinne fälschlicher Annahme der Forschungshypothese) als auch falsch-negative Befunde (also das Begehen eines Beta-Fehlers im Sinne fälschlicher Ablehnung der Forschungshypothese). Neben der *statistischen Signifikanz* (hier wird üblicherweise ein Signifikanzniveau von $\alpha = 5\%$ oder $\alpha = 1\%$ angelegt) sind die *Teststärke* (hier wird üblicherweise eine Teststärke $1\text{-}\beta > 80\%$ gefordert) sowie die Effektgröße (hier wird eine möglichst hohe Effektgröße erwartet) zu bestimmen und angemessen zu interpretieren. Zur Einordnung von standardisierten Effektgrößenmaßen hat sich in der Tradition des Statistikers Jacob Cohen (1988) die Klassifikation von Effekten in klein ($d = 0{,}20$) mittel ($d = 0{,}50$) und groß ($d = 0{,}80$) weithin eingebürgert, wobei „d“ als standardisiertes Effektgrößenmaß eine an der Standardabweichung relativierte Mittelwertsdifferenz zwischen zwei Gruppen beschreibt. Vereinfacht gesagt: Je größer diese Differenz, umso stärker zeigt sich ein erwarteter Effekt in der Experimentalgruppe im Vergleich zur Kontrollgruppe. Je nach Untersuchungsdesign und Datenlage sind jedoch auch andere Effektgrößenmaße zu bestimmen (vgl. Olejnik & Algina, 2000).

Es ist sehr wichtig zu betonen, dass empirisch gefundene Effektgrößen niemals allein anhand von numerischen Grenzen, sondern immer im Hinblick auf die im konkreten Forschungsfeld üblichen sowie theoretisch als bedeutsam zu erachtenden Größenordnungen inhaltlich einzuschätzen sind (Cortina & Landis, 2009): Wo es um Gesundheit und Menschenleben geht, sind auch numerisch sehr kleine Effekte praktisch hochgradig bedeutsam. Bei Bildungsmaßnahmen, die teuer und aufwändig sind, können dagegen oft nur sehr große Lerneffekte als praktisch bedeutsam angesehen werden, da man den finanziellen und personellen Aufwand für kleine Effekte meist nicht aufbringen oder rechtfertigen kann. Weiterhin ist zu beachten, dass ein in der Stichprobe gefundener Effekt nur deskriptivstatistisch einzuordnen ist und die Basis bilden muss, um den *Populationseffekt* zu schätzen, etwa über entsprechende Konfidenzintervalle (Döring & Bortz, im Druck, Kap. 14). Eine Analyse der Fachliteratur hat ergeben, dass das Berechnen, Berichten und korrekte Interpretieren von Effektgrößen sich in der erziehungswissenschaftlichen Literatur langsam etabliert, aber immer noch Schwächen aufweist (Peng, Chen, Chiang & Chiang, 2013).

Wenn sich nach Abschluss einer Studie herausstellt, dass die Teststärke zu gering ist, um überhaupt aussagekräftige statistische Datenanalysen durch-

zuführen, zeigt dies einen Planungsfehler an. Die erzielte Teststärke einer Untersuchung sollte sich nicht willkürlich ergeben, sondern vor allem durch die Wahl des sog. *optimalen Stichprobenumfangs* sowie durch weitere Maßnahmen (z.B. bessere Kontrolle von Störvariablen; ausreichende Dosierung der unabhängigen Variablen) gezielt beeinflusst werden. Für welche nachzuweisende Effektgröße bei welchem Signifikanztest welcher Stichprobenumfang groß genug ist, um ausreichende Teststärke zu sichern, lässt sich den Tabellenwerken des Statistikers Cohen (1988, 1992) entnehmen oder auch mit entsprechenden Software-Tools komfortabel bestimmten (z.B. mit dem kostenlosen Programm g*power: www.gpower.hhu.de; Faul, Erdfelder, Lang & Buchner, 2007; Döring & Bortz, im Druck, Kap. 14). Während bei statistischen Analysen in den Sozialwissenschaften inzwischen neben der Signifikanz auch die Effektgröße verstärkt betrachtet wird, erfolgt eine Bestimmung der Teststärke nach wie vor äußerst selten (Fritz, Scherndl & Kühlberger, 2013).

Dass eine empirisch-quantitative Studie methodisch streng angelegt worden ist, erkennt man im Hinblick auf statistische Validität also unter anderem auch daran, dass schon bei der Untersuchungsplanung begründet berechnet wurde, welcher Stichprobenumfang für eine aussagekräftige inferenzstatistische Analyse notwendig ist.

5. Grenzen der Qualitätsbewertung quantitativer empirischer Studien

Wer im Kontext erziehungswissenschaftlicher Forschung tätig ist, hat in unterschiedlichen Rollen mit der Qualitätsbewertung quantitativer empirischer Studien zu tun: Es geht darum, Studien, die man liest und auf die man sich in Forschung und Praxis stützen will, hinsichtlich Seriosität und Güte einzuordnen. Es geht aber auch darum, eigene Forschungsarbeiten qualitätsbewusst zu planen und umzusetzen, um sie in hochrangigen Publikationsorganen zu platzieren oder Fördergelder zu akquirieren. Nicht zuletzt sind viele Forschende auch selbst in den Peer-Review-Prozess eingebunden und beurteilen offiziell Qualifikationsarbeiten, Projektanträge, Vortrags- und Publikationseinreichungen.

Die Frage nach wissenschaftlichen Qualitätskriterien muss in der Wissenschaft, die den Anspruch erhebt, nicht nur Erkenntnisse zu liefern, sondern immer auch den Erkenntnisprozess als solchen zu reflektieren, auch auf der Meta-Ebene betrachtet werden. Wie werden die wissenschaftlichen Qualitätskriterien festgelegt und welche Probleme gibt es dabei? Wie wird die Einhaltung der wissenschaftlichen Qualitätskriterien geprüft und welche Schwierigkeiten sind damit verbunden?

5.1 Probleme bei der Festlegung der Qualitätskriterien und der Bewertungsmaßstäbe

Im Kern besteht über die Qualitätsanforderungen an quantitativ-empirische Studien weithin Einigkeit. Dennoch gibt es im Detail der Definition von Validitätstypen oder der Auflistung von Validitätsbedrohungen unterschiedliche Sichtweisen in der Fachliteratur. Die hier ausgeführten Validitäts-Konzeptionen der Campbell-Tradition sind nicht unkritisiert geblieben. Campbell und Kollegen selbst haben die Abgrenzungen ihrer Unterformen der Validität im Laufe der Jahre immer wieder deutlich verändert (Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002). Und man kann sich z.B. streiten, ob es sinnvoll ist, die Konstruktvalidität als Teilaspekt der externen Validität und die statistische Validität als Teilaspekt der internen Validität einzuordnen, wie die Autoren es tun.

Eine große Schwierigkeit liegt zudem darin, jeweils adäquate *Bewertungsmaßstäbe* anzulegen. Insbesondere wenn man in Rechnung stellt, dass Studien je nach institutionellem Umfeld und Position der Forschenden, aber auch in Abhängigkeit von Eigenschaften des Forschungsfeldes, unter sehr unterschiedlichen *forschungsökonomischen Bedingungen* realisiert werden (z.B. Ausstattung mit Untersuchungsräumen, Personal, Mitteln für Reisekosten, Messgeräte). Keine reale Studie kann alle methodischen und sonstigen Qualitätskriterien gleichermaßen und ideal erfüllen. Es sind immer Abstriche zu machen. Welche Abstriche unter welchen Bedingungen akzeptabel oder inakzeptabel sind, also welche Bewertungs*maßstäbe* jeweils gelten sollen, ist jedoch nicht verbindlich festgelegt und vom jeweiligen Forschungskontext abhängig. So mag im einen Fall eine erziehungswissenschaftliche Doktorarbeit zur Kompetenzmessung als wissenschaftlich anspruchsvoll genug akzeptiert werden, obwohl sie mit einer nicht-repräsentativen Stichprobe operiert (z.B. wenn die Arbeit als externe Promotion im Alleingang bewältigt wird), während im anderen Fall die Verwendung von Daten aus repräsentativen Samples verlangt wird (z.B. wenn die Arbeit an ein entsprechend großes Forschungsprojekt mit Datenerhebungsmöglichkeiten angebunden ist).

Einzelne Qualitätskriterien stehen nicht selten auch in einem *Spannungsverhältnis* zueinander, so dass die Verbesserung des einen Qualitätsaspekts automatisch eine Verschlechterung des anderen nach sich zieht. So wurden im Bereich der Evaluationsforschung, die sich der wissenschaftlichen Bewertung von Produkten, Personen, Organisationen und Maßnahmen widmet, internationale „*Standards für Evaluation*“ formuliert, die auf 25 Einzelstandards angeben, was eine gute Evaluationsstudie ausmacht (DeGEval, 2008). Da Evaluationsstudien meist Auftragsforschung darstellen und im Feld stattfinden, ist *Effizienz (Evaluationsstandard D3)* ein wichtiges Qualitätskriterium (d.h. es soll im Rahmen der Evaluationsstudie kein unnötiger Aufwand betrieben werden, der überflüssige Kosten und zeitliche Belastungen oder

Störungen im Praxisfeld nach sich zieht). Gleichzeitig ist natürlich *Genauigkeit* ein wichtiger Qualitätsstandard (um diesen zu sichern ist jedoch Mehraufwand u.a. bei der Informationsbeschaffung – Evaluationsstandard G5 – im Sinne von zusätzlichen Erhebungsorten, Erhebungszeiten, weiteren Messinstrumenten etc. erforderlich, wobei der Zusatznutzen manchmal vorab nicht genau prognostiziert werden kann). Maximal genau und maximal ökonomisch bzw. effizient kann eine Evaluationsstudie nicht gleichzeitig sein. Das rechte Verhältnis ist diskussions- und begründungspflichtig und damit sind entsprechende Methodenentscheidungen auch immer angreifbar.

5.2 Probleme bei der Überprüfung der Qualitätskriterien

Angesichts der oben genannten Probleme bei der Festlegung von Kriterien und Beurteilungsmaßstäben sowie angesichts der Komplexität konkreter Studien ist nachvollziehbar, dass einzelne Studierende und Forschende in ihren Beurteilungen vorliegender Studien sehr unsicher sein können. Neben der *Übung*, die sich durch Rezeption einer wachsenden Zahl von Studien einstellt, hilft hier der Austausch im Kreis der Mitstudierenden bzw. Mitforschenden, um die *eigene Urteilsfähigkeit* auszubilden. Auch ist es in vielen Instituten üblich, Beurteilungsmaßstäbe z.B. für Studien, die im Rahmen akademischer Qualifikationsarbeiten angefertigt werden, relativ konkret auszuformulieren und in Form von Checklisten oder Handreichungen bereitzustellen. Dadurch sollte es einfacher sein, intersubjektiv nachvollziehbar z.B. die Qualität einer empirisch-quantitativen Masterarbeit oder Doktorarbeit einzustufen.

Das *Peer-Review-Verfahren*, bei dem Studien durch Fachkollegen begutachtet werden und entweder nur die Reviewer (einfachblindes Verfahren) oder Reviewer und Autoren (doppelblindes Verfahren) wechselseitig füreinander anonym bleiben, ist der zentrale Mechanismus der Qualitätssicherung im Wissenschaftssystem. Er soll eine gründliche, fachkundige und unvoreingenommene Qualitätsbeurteilung sicherstellen und dafür sorgen, dass keine mangelhaften Studien veröffentlicht werden. Dies geschieht, indem Beiträge mit großen Mängeln abgelehnt werden und Beiträgen mit kleineren Mängeln die Möglichkeit zur Nachbesserung gegeben wird. Weichen die vorliegenden Gutachten in ihren Einschätzungen zu weit voneinander ab, werden weitere Gutachten eingeholt. Das Peer-Review-Verfahren ist weithin etabliert und anerkannt, kämpft jedoch auch mit einer Reihe systematischer Probleme: Da die Begutachtung ehrenamtlich erfolgt und Forschende ohnehin meist zeitlich überlastet sind, kann oft nur eine *grobe Prüfung* von Manuskripten vorgenommen werden. Das ist eine Erklärung dafür, warum manchmal auch gravierende Mängel bis hin zu offensichtlichen Wissenschaftsfälschungen von Gutachtenden unentdeckt geblieben sind. Abgesehen von Zeitaufwand und Gründlichkeit wird vermutet, dass eine Reihe von *Urteilsverzerrungen* die Bewertung beeinträchtigen, dass etwa Gutachtende

dazu neigen, Studien von renommierten Hochschulen oder Studien mit Ergebnissen, die ihren eigenen Theorien entsprechen, besser zu bewerten. Auch ist die formale Anonymisierung der Autorinnen und Autoren meist wirkungslos, da Gutachtende die Autorschaft anhand von Studienthema, Literaturliste und sonstigen Kennzeichen oft erschließen können, insbesondere in kleinen Forschungsfeldern. Die bisherige Forschung zu Verzerrungen im Peer-Review-Prozess liefert widersprüchliche Ergebnisse (Lee, Sugimoto, Zhang & Cronin, 2013).

Eine Reihe von Maßnahmen wird diskutiert und erprobt, um den *Peer-Review-Prozess zu verbessern*. Generell sind Gutachtende gefordert, sich im Sinne g**ut**er wissenschaftlicher Praxis (DFG, 2013) ethisch zu verhalten und im Zweifelsfall eine Begutachtung aus Befangenheit abzulehnen. Zudem sind diejenigen, die Peer-Review-Prozesse z.B. für Konferenzen, Fachzeitschriften oder Forschungsfördereinrichtungen organisieren, gefordert die Gutachtenden sorgfältig auszuwählen und anzuleiten, bei fragwürdigen Gutachten entsprechend zu intervenieren und die Begutachtungsprozesse transparent zu halten (z.B. sollten Gutachtende jeweils die anderen Gutachten und die Begutachtungsentscheidung erhalten, um ihr eigenes Urteil einordnen zu können). Auch sollten Peer-Review-Prozesse evaluiert werden (z.B. um Verzerrungen im Review-Prozess zu erkennen). Die in Abschnitt 2.2 vorgestellte Kontroverse um die Evaluation des Dore-Programms zur Behandlung von Dyslexie zeigt Grenzen des Peer-Reviews, etwa wenn Review-Entscheidungen so kontrovers ausfallen, dass Beiratsmitglieder einer Zeitschrift unter Protest zurücktreten und offenbar kein Konsens über die fachliche Einschätzung der Studie zu erreichen war.

Um zu prüfen, ob eine Evaluationsstudie den Qualitätsstandards der Evaluationsforschung genügt (DeGEval, 2008), kann – insbesondere bei aufwändigen und folgenreichen Studien – eine sog. *Meta-Evaluation* durchgeführt werden, also eine qualitätsbeurteilende Evaluation der Evaluation. Dabei wird die Evaluationsstudie prozessbegleitend oder abschließend von externen Fachkollegen analysiert (etwa anhand von Interviews mit den Forschenden; Sitzungsprotokollen; Forschungsberichten) und mit Blick auf ihre wissenschaftliche Qualität beurteilt. Dies ist jedoch wiederum mit nicht unbeträchtlichem Aufwand und zusätzlichen Kosten verbunden.

Es mehren sich inzwischen auch Studien, welche die *Qualität wissenschaftlicher Publikationen wissenschaftlich untersuchen*, etwa indem sie empirisch-quantitative Zeitschriftenartikel dahingehend inhaltsanalytisch auswerten, wie viele und welche Indikatoren für unterschiedliche Typen der Validität sie enthalten (Wester, Borders, Boul & Horton, 2013). Mit Meta-Evaluationen und wissenschaftlichen Studien zur Qualität von Publikationen kann u.a. überprüft werden, ob und inwiefern von den wissenschaftlichen Fachgesellschaften vorgegebene methodische Standards aktuell und/oder im historischen Zeitverlauf tatsächlich umgesetzt werden.

6. Ausblick zur Qualitätssicherung empirisch-quantitativer Forschung

Die Qualitätssicherung empirisch-quantitativer Forschung unterliegt fortlaufendem Wandel, sie reagiert beispielsweise auf Wissenschaftskrisen und technologische Innovationen. Einige Trends für die vier großen Bereiche der Qualitätsbewertung (1. wissenschaftliches Forschungsproblem, 2. wissenschaftlicher Forschungsprozess, 3. Wissenschafts- und Forschungsethik, 4. Dokumentations- und Präsentationsqualität) werden im Folgenden skizziert.

6.1 Zukunft der Relevanzbewertung von Forschung

Welche wissenschaftlichen Forschungsprobleme innerhalb der Erziehungswissenschaft als besonders wichtig anzusehen sind, ist begründungspflichtig und kontrovers (siehe Abschnitt 3.1). Es zeichnet sich der Trend ab, die Bedeutung von Studien zunehmend unabhängig von ihren konkreten Fragestellungen und Befunden und stattdessen mit Bezug auf quantitative Indikatoren ihrer Rezeption zu bewerten: Je häufiger die Studie zitiert wird, umso höhere Relevanz und somit umso höhere Qualität wird ihr zugeschrieben. Die Digitalisierung des wissenschaftlichen Publikationswesens erlaubt eine fortlaufende Zählung der Abrufe und Zitationen digitaler Artikel. Entsprechende szientometrische Indikatoren wie der *Journal Impact Factor* (eine Maßzahl, die angibt, wie häufig Artikel der betreffenden Zeitschrift zitiert werden) dienen verstärkt als Beurteilungsgrundlage, wenn es darum geht, einzelne Forschende, konkrete Studien oder ganze Forschungsfelder in ihrer Bedeutung einzuordnen.

Dabei darf hohe Resonanz nicht automatisch mit besonderer theoretischer oder praktischer Relevanz einer Studie (siehe Abschnitt 3.1) gleichgesetzt werden. So sind Zitationszahlen in größeren Fachdisziplinen bzw. breiteren Forschungsfeldern von vorne herein deutlich höher als in kleineren Gebieten. Auch ist zu beachten, dass Zitations-Indizes, die sich allein auf wissenschaftliche Literaturdatenbanken stützen, wachsende Teile des Publikationswesens nicht erfassen (z.B. Konferenzbände, Wissenschafts-Blogs etc.; Larsen & Ins, 2010).

Szientometrische Maßzahlen wie Zitationszahlen von Zeitschriften, einzelnen Fachartikeln oder Forschenden können in Qualitätsbewertungen der Forschung einfließen, müssen aber korrekt interpretiert werden und ersetzen keine inhaltliche Debatte darüber, welche Forschungsprobleme innerhalb der Erziehungswissenschaft aus welchen Gründen zu einem bestimmten historischen Zeitpunkt und in einem bestimmten kulturellen Umfeld als vordringlich oder unwichtig einzustufen sind.

6.2 Zukunft der methodischen Strenge von Forschung

Im Hinblick auf die methodische Strenge setzt sich bei der *statistischen Datenanalyse* die Entwicklung zu mathematisch immer komplexeren und anspruchsvolleren computergestützten Methoden fort. Einfache statistische Signifikanztests, die man prinzipiell noch per Hand ausrechnen könnte, sind teilweise durch *Strukturgleichungsmodellierung* abgelöst worden, die es erlaubt, die Messmodelle der Konstrukte und gleichzeitig ganze Hypothesensysteme zu prüfen. Die Vorteile anspruchsvoller statistischer Verfahren gehen jedoch mit dem Nachteil einher, dass Details der Analyse und Interpretation für größere Teile der Forschungscommunity immer schwerer nachzuvollziehen sind. Damit durch die Verfügbarkeit komplexer Datenanalyseverfahren die Qualität der Forschung steigt (etwa im Sinne statistischer Validität; siehe Abschnitt 4.4), müssen die Methoden sachgerecht eingesetzt und interpretiert werden, also muss entsprechende Aus- und Weiterbildung in quantitativen Forschungsmethoden angeboten und genutzt werden.

Das gilt auch im Zusammenhang mit *innovativen Methoden der Datenerhebung* und der Erschließung neuer Datenquellen, die aktuell etwa unter dem Stichwort „*Big Data*“ diskutiert werden (Boyd & Crawford, 2012): Die im Digitalzeitalter massenhaft protokollierten Verhaltensdaten von Menschen (z.B. Suchanfragen in Suchmaschinen, Beiträge in Online-Foren, Aktivitäten auf E-Learning-Servern) werden von Unternehmen für die kommerzielle Marktforschung genutzt. Können und sollen diese Datenmassen nicht auch dem wissenschaftlichen Erkenntnisfortschritt dienen? Wie können Big Data für die Erziehungswissenschaft erschlossen und sinnvoll theoriebildend oder theorieprüfend statistisch ausgewertet werden, v.a. wenn dazu eine Rechnerkapazität notwendig ist, die den in der erziehungswissenschaftlichen Forschung üblichen Arbeitsplatzrechner überfordert?

Als Herausforderung zu betrachten ist auch die in den letzten Jahren zu beobachtende stärkere Hinwendung zu *neurowissenschaftlichen Ansätzen in Pädagogik und Psychologie*. Die neurowissenschaftliche Forschung ist hinsichtlich Theorien, Datenerhebungsverfahren, Datenstrukturen und Datenanalysetechniken ausgesprochen komplex. Fachfremden fehlen hier meist die notwendigen Grundlagen, um überhaupt Primärstudien zu verstehen und einordnen zu können, gleichzeitig beeindrucken derartige Studien durch ihre aufwändige Messmethodik. Es besteht die Gefahr pseudowissenschaftlicher Argumentation, wenn man sich auf „neurowissenschaftliche Studien“ stützt, weil dieser naturwissenschaftliche Zugang zum menschlichen Geist besonders „objektiv“ wirkt, ohne die methodische Qualität und Aussagekraft der Arbeiten ernsthaft selbst beurteilen zu können. Dem regelrechten Hype um die sog. „Spiegelneuronen“, die uns angeblich automatisch die Intentionen anderer Menschen erschließen und für Empathie zuständig sind, steht beispielsweise ein aktueller Forschungsstand gegenüber, demgemäß nicht nur

die möglichen Funktionen, sondern sogar die Existenz von Spiegelneuronen beim Menschen bis heute umstritten sind (Kilner & Lemon, 2013).

Chancen und Risiken gleichermaßen birgt auch der deutliche Aufschwung der *Forschungssynthese* und das wachsende Verständnis dafür, dass Einzelstudien nur Bausteine sind und erst aus der systematischen Zusammenfassung der Gesamtheit der Studien zu einem Effekt tragfähige Schlüsse zu ziehen sind (Cooper, Hedges & Valentine, 2009; Döring & Bortz, im Druck, Kap. 16). Die Medizin hat mit der 1993 gegründeten, nach dem schottischen Arzt Archibald Leman Cochrane (1909–1988) benannten *Cochrane-Collaboration*[5] und der von ihr betriebenen Online-Datenbank *Cochrane Library* (Verlag Wiley) ein weltweites System der Forschungssynthese etabliert. In systematischen Cochrane-Reviews, die definierten Qualitätskriterien entsprechen müssen, wird immer wieder von Fachleuten der jeweils aktuelle Forschungsstand zu unterschiedlichen medizinischen Fragestellungen objektiv zusammengefasst und Wissenschaft und Öffentlichkeit bereitgestellt. Damit soll *evidenzbasierte Medizin* ermöglicht werden, also medizinische Praxis, die auf dem aktuellen Forschungsstand basiert, der jederzeit online über die Cochrane-Library abrufbar ist.[6]

Für *evidenzbasierte Bildung, Politik und Justiz* durch Erarbeitung und Bereitstellung von Forschungsreviews setzt sich die *Campbell-Collaboration*[7] ein, benannt nach dem Psychologen und Methodiker Donald T. Campbell, dessen Validitäts-Konzeption in diesem Beitrag ausführlich dargelegt wurde. Der meistabgerufene Forschungsüberblick der Campbell-Datenbank (mehr als 25.500 Abrufe; Stand: August 2014) geht der Forschungsfrage nach, ob die Anwendung von Techniken der Klassenführung (classroom management) durch Lehrkräfte dazu führt, dass unpassendes, störendes und aggressives Schülerverhalten zurückgeht (Oliver, Wehby & Reschli, 2011). Dazu wurden 12 aussagekräftige Studien in der Fachliteratur identifiziert und zusammenfassend betrachtet mit dem Ergebnis, dass Techniken der Klassenführung (Routinen und Regeln für die Abläufe in der Klasse etablieren, konstruktives Schülerverhalten verstärken etc.) wirksam sind. Die Qualität der Forschung in einem Themenfeld steigt, wenn sich mehr Forschende für systematische Forschungssynthese engagieren, d.h., wenn sie a) ihre Primärstudien so publizieren, dass diese metaanalytisch problemlos verarbeitet werden können, und wenn sie b) gute Forschungssynthesen erstellen, vielleicht sogar regelmäßig in der Campbell-Collaboration mitarbeiten.

Was verursacht gute Schülerleistungen in der Schule? Dieser Frage ging der neuseeländische Pädagoge John Hattie in einer legendären „Meta-Metaanalyse“ nach: Er fasste statistisch mehr als 800 Metaanalysen zusam-

5 Siehe www.cochrane.org

6 Siehe www.thecochranelibrary.com

7 Siehe www.campbellcollaboration.org

men, die ihrerseits auf mehr als 52.000 Einzelstudien mit mehr als 80 Millionen Untersuchungspersonen basieren. Daraus berechnete er zentrale Erfolgsfaktoren für das Lehren und Lernen und empfiehlt evidenzbasiert einen Ansatz, bei dem „Lehren und Lernen sichtbar gemacht werden", indem Lehrende und Lernende sich intensives Feedback geben und aktiv die Perspektive des jeweils anderen einnehmen (Hattie, 2008). Die Hattie-Studie wurde weithin regelrecht begeistert aufgenommen und teilweise als „heiliger Gral" gefeiert. Gleichzeitig wurde aber auch deutlich, dass sowohl Forschende als auch Lehrkräfte nicht selten überfordert waren, die Studie selbst zu verstehen und kritisch einzuordnen. Es musste erst Vermittlungsarbeit geleistet werden, um das methodische Vorgehen bei einer „Meta-Metaanalyse" und die Bedeutung der berichteten „Effektgrößen" (siehe Abschnitt 4.4) verstehbar zu machen. Betrachtet man die Studie mit entsprechender Methoden-Expertise, so lassen sich ihre Stärken, aber auch ihre Schwächen aufzeigen (z.B. Terhart, 2011). Nicht allein die Tatsache, dass es sich um eine Metaanalyse handelt und auch nicht der schiere Umfang der Studie (die in dieser Form ohne computergestützte Recherche- und Analysetechniken nicht möglich gewesen wäre) sorgen automatisch für große methodische Strenge. Es sind wiederum viele Einzelentscheidungen im Forschungsprozess, die es sachkundig zu bewerten gilt. Damit Metaanalysen heute und in Zukunft sachgerecht durchgeführt und interpretiert werden können, ist spezifische Methodenkompetenz vonnöten, die im Studium der Erziehungswissenschaft bislang meist nicht ausreichend entwickelt wird.

6.3 Zukunft der ethischen Strenge von Forschung

Forschungsethik wird zunehmend ernster genommen und ihre Einhaltung formalisiert. International haben sich *Ethik-Kommissionen* an den Forschungseinrichtungen etabliert, die von den Forschenden im Vorfeld jeder einzelnen wissenschaftlichen Untersuchung zu konsultieren sind, um sich ethische Unbedenklichkeit des Studienkonzepts formal bescheinigen zu lassen. Diese Bescheinigungen wiederum werden von immer mehr internationalen Fachzeitschriften verlangt, damit die Studie veröffentlicht werden kann. Deutschland hat in der Erziehungswissenschaft (ebenso wie in anderen Sozialwissenschaften) das System der Ethik-Kommissionen (IRB: Institutional Review Boards) bislang nicht aufgebaut, steht hier aber unter Handlungsdruck, damit hiesige Forschung international konkurrenzfähig bleibt. Prüfungen durch Ethik-Kommissionen steigern indessen nicht automatisch die ethische Strenge der Forschung (diese sollte bei entsprechend ausgebildeten und ethisch verantwortungsvollen Forschenden ohnehin gegeben sein), sondern können auch Negativfolgen haben: Eine weitere Bürokratisierung des Wissenschaftsalltags droht. So muss teilweise mit monatelangen Wartezeiten auf einen Ethik-Bescheid gerechnet werden, was z.B. Forschungsseminare mit Studierenden, die in einem Semester realisiert werden müssen, blockieren würde. Auch können Studien in bestimmten Themenfel-

dern behindert werden, etwa wenn sexualbezogene Fragestellungen oder Methoden der Online-Datenerhebung von manchen Ethik-Kommissionen vorurteilsbehaftet per se als „ethisch bedenklich“ eingestuft werden (zur Institutionalisierung forschungsethischer Standards in der Erziehungswissenschaft siehe Miethe, 2013).

Insbesondere die *ethischen Kriterien der boomenden Online-Forschung* sind nach wie vor im Detail unklar und strittig: Sind öffentliche Online-Beiträge von Privatpersonen (z.B. Blog-Einträge, Posts in Online-Foren, Profile auf Social-Networking-Plattformen) genau wie Zeitungsartikel zu betrachten und somit für die Forschung frei verwendbar? Oder muss hier vor einer Nutzung für Forschungszwecke ausdrückliches Einverständnis eingeholt werden? Antworten sind nicht so leicht, wie es vielleicht scheint (Zimmer, 2010). Die Deutsche Gesellschaft für Online-Forschung hat inzwischen Standesregeln für Marktforschung in Sozialen Medien verabschiedet und unterscheidet dabei zwischen offenen Bereichen (die auch der Forschung frei zur Verfügung stehen) und geschlossenen Bereichen, in denen Datenerhebung nur nach ausdrücklicher Einwilligung der Plattformbetreiber und Mitglieder erfolgen darf (DGOF, 2014).

Auch im Hinblick auf *Wissenschaftsethik* werden neue formalisierte Methoden der Prüfung und Kontrolle etabliert. Ein Beispiel sind *computergestützte Plagiats-Checks* akademischer Qualifikationsarbeiten, die an einigen Fakultäten bereits eingeführt sind. Auch hier gilt, dass computergestützte Verfahren als Hilfestellungen dienen können, dass die eigentliche Qualitätssicherung im Bereich Wissenschaftsethik heute und in Zukunft aber bei der besseren Ausbildung der Studierenden und ihrem ethischen Selbstverständnis ansetzen muss.

6.4 Zukunft der Dokumentations- und Präsentationsqualität von Forschung

Die Bedingungen der Dokumentation und Präsentation empirischer Studien verändern sich dramatisch mit der Digitalisierung des wissenschaftlichen Publikationswesens. Fachzeitschriften, aber auch Fachbücher, erscheinen zunehmend nur noch digital und bieten neue Darstellungsformate: Das Einbinden von Fotos, 3D-Animationen und Videos in Fachartikel ist heute ebenso bereits Realität wie die Bereitstellung aller Erhebungsinstrumente und Originaldatensätze einer Studie zusammen mit der Publikation. Damit können sich Interessierte noch viel detaillierter ein Bild der Untersuchung verschaffen. Gleichzeitig sind Forschende gefordert, ihre Befunde auf innovative Weise aufzubereiten.

Der Wissenschaftsverlag Elsevier hat beispielsweise mit „*Article of the Future*“ ein Projekt zur Erprobung innovativer Präsentationsformen in Fach-

zeitschriften gestartet.[8] Bislang sind hier Natur- und Technikwissenschaften Vorreiter, aber die Erziehungswissenschaft und andere Sozialwissenschaften werden sicher folgen.

„Eine neue Zeitschrift für eine neue Ära" heißt das programmatische Editorial der 2013 gegründeten Fachzeitschrift „*Archives of Psychological Science*"[9], in der u.a. auch für die Erziehungswissenschaft relevante entwicklungs-, kognitions- und schulpsychologische Arbeiten publiziert werden können. Das Journal verfolgt einen „open method, open data, open access"-Ansatz, d.h. bei jeder Artikel-Einreichung müssen in einem Formular detaillierte Angaben zu allen methodischen Aspekten gemacht, zudem alle Messinstrumente und alle Datensätze öffentlich bereitgestellt werden. Die Zahl der publizierten Beiträge ist bislang noch sehr gering, so dass eine Evaluation dieses neuen, besonders transparenten Publikationsmodells abzuwarten bleibt.

Bislang ist die Medizin Vorreiterin darin, das Problem des sog. *Publication Bias* durch veränderte Publikationsanforderungen systematisch zu bekämpfen. Der Publication Bias besteht darin, dass bevorzugt Studien publiziert werden, die hypothesenkonforme (signifikante) Ergebnisse erbracht haben. Studien mit hypothesenkonträren Befunden bleiben dagegen häufiger unveröffentlicht (sei es, weil Forschende sie seltener zur Publikation einreichen, sei es, weil Zeitschriften sie wegen vermeintlich geringerer Relevanz seltener zur Publikation akzeptieren), wodurch sich ein insgesamt verzerrtes Bild des Forschungsstandes ergibt. Als Gegenmaßnahme ist es in der Medizin inzwischen bei Medikamententests (in denen typischerweise ein neues Medikament gegen ein altes Medikament oder ein Placebo geprüft wird) erforderlich, jede Studie *vor* ihrer Durchführung offiziell mit ihrem geplanten Versuchsablauf anzumelden. Die geplante Studie wird in einer Datenbank registriert. Im Falle hypothesenkonträrer Ergebnisse (d.h. das neue Medikament schneidet nicht besser ab) lässt sich die Studie somit im Nachhinein nicht einfach ignorieren oder verheimlichen, sondern geht in den Gesamtforschungsstand ein. Die Weltgesundheitsorganisation hat dazu die *internationale Studien-Registrierungs-Plattform ICTRP* (International Clinical Trials Registry Platform) eingerichtet.[10]

Ob und inwiefern eine öffentliche Studien-Registrierung sinnvoll von medizinischen Medikamententests auf empirische Sozialforschung übertragbar ist – etwa auf die Wirksamkeitsforschung zu pädagogischen Interventionen – wäre zu diskutieren. Dass im herkömmlichen Publikationswesen gar keine Studien mit Nulleffekten publiziert werden, ist indessen auch nicht der Fall. So wird beispielsweise in der Forschung zur Effektivität von technologiege-

8 Siehe www.articleofthefuture.com

9 Siehe www.apa.org/pubs/journals/arc/

10 Siehe www.who.int/ictrp/en/

stütztem bzw. Fernunterricht einerseits im Vergleich zum Präsenzunterricht andererseits ein sog. „*No Significant Difference* Effect“ postuliert, weil zahlreiche Studie vorliegen, die keine Gruppenunterschiede zeigen (Russell, 1999).[11] Die pauschale Schlussfolgerung, dass somit belegt sei, dass Fernunterricht genauso gut wie Präsenzunterricht ist, muss angesichts der methodischen Probleme vieler dieser Vergleichsstudien jedoch hinterfragt werden (Lockee, Burton & Cross, 1999; Bernard et al., 2004). Dieses Beispiel unterstreicht erneut die Bedeutung hoher Präsentationsqualität im Sinne *ausführlicher und anschaulicher Methodendarstellungen* in Publikationen, um diese kritisch beurteilen zu können, sowohl wenn signifikante als auch nicht-signifikante Befunde präsentiert werden.

So zeigte eine aktuelle randomisierte Kontrollgruppenstudie, dass Medizinstudierende anatomische Details anhand eines Computermodells (sowohl 2D- als auch 3D-Abbildungen) signifikant schlechter lernen als anhand eines greifbaren Plastik-Modells (Khot, Quinlan, Norman & Wainman, 2013). Zur Beurteilung der Studie ist u.a. die Vergleichbarkeit der anatomischen Lernmaterialien (d.h. der drei Ausprägungen der unabhängigen Variable) entscheidend. Eine digitale Publikation könnte der Leserschaft mit der Einbettung der Computermodelle und einem Video des Plastik-Modells genau demonstrieren, wie die Untersuchungsbedingungen ausgesehen haben und somit eine fundiertere Einschätzung der Studie erlauben.

7. Fazit

Der vorliegende Beitrag hat aufgezeigt, dass Qualitätsdiskussionen rund um empirisch-quantitative Studien vor allem vier Dimensionen beachten müssen, 1) die Relevanz des wissenschaftlichen Forschungsproblems, 2) die methodische Strenge des Forschungsprozesses, 3) wissenschafts- und forschungsethische Aspekte sowie 4) die umfängliche und nachvollziehbare Dokumentation des Studienablaufs und der Befunde (siehe Tabelle 1 in Abschnitt 1).

Hier sind die *einzelnen Forschenden* gefragt, ihre Studien mit Blick auf Qualität zu planen und zu realisieren und vorliegende Studien sachgerecht anhand von Qualitätskriterien zu beurteilen. Beides erfordert entsprechende Methodenkompetenz und Übung.

Gleichzeitig ist Qualitätssicherung auch eine Aufgabe für das *Wissenschaftssystem als Ganzes*. Das umfasst u.a. eine zeitgemäße und gründliche Ausbildung in den quantitativen Methoden der empirischen Sozialforschung in erziehungswissenschaftlichen Studiengängen, eine fortlaufende Evaluation und Verbesserung von Qualitätssicherungsmaßnahmen wie dem Peer-

11 Siehe www.nosignificantdifference.org

Review-Verfahren, eine Weiterentwicklung des Publikationswesen und eine Stärkung von Initiativen der Forschungssynthese wie etwa der Campbell-Collaboration.

Auch konzertierte Aktionen wie die sog. *Reproducability-Initiative* in der Psychologie, die sich zum Ziel gesetzt hat, zentrale psychologische Studien systematisch von unabhängigen Forschergruppen weltweit replizieren zu lassen, um den Bestand gesicherten psychologischen Wissens aktuell zu prüfen[12], können für Forschungsfelder innerhalb der Erziehungswissenschaft inspirierend sein (Cook, 2014). Auslöser für diese Initiative war die Aufdeckung mehrerer Wissenschaftsfälschungen in der Psychologie und die Erkenntnis, dass Replikationsstudien, die die Existenz und Generalisierbarkeit von Befunden sichern, viel zu selten durchgeführt werden (Pashler & Wagenmakers, 2012). Viele Fachzeitschriften verlangen in erster Linie „Originalität" und belohnen damit teilweise effekthascherische Studien ohne fundierte Theoriebasis, lassen aber zu wenig Raum für eine differenzierte Ausarbeitung und Weiterentwicklung gesicherter Theorien.

Die *Durchführung von Replikationen guter und einschlägiger Studien* des eigenen Fachgebiets sollte *im Studium der Erziehungswissenschaft* (wie jeder anderen empirischen Sozialwissenschaft) fest verankert werden. Erst eine Replikationsstudie lehrt, eine vorliegende quantitativ-empirische Studie detailliert zu durchdringen, enthüllt alle nebulösen Angaben und fehlenden Details zum methodischen Vorgehen, das man nachstellen möchte, macht den hinter einer publizierten Studie steckenden tatsächlichen Arbeitsaufwand erlebbar und gibt zugleich eine realistische Richtung vor, wie seriöse wissenschaftliche Studien im jeweiligen Forschungsfeld anzulegen sind. Gleichzeitig sind unabhängige, methodisch saubere Replikationen der beste Nachweis für die Existenz und Generalisierbarkeit theoretisch vorhergesagter Effekte im Feld der Erziehungswissenschaft.

Literatur

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education) (2014). Standards for Educational and Psychological Testing (5th edition). Washington: AERA.

APA (American Psychological Association) (2010). Publication Manual of the American Psychological Association (6th edition). Washington, DC: APA.

Bernard, R., Abrami, P., Lou, Y., Brokhovski, E. Wade, A., Wozney, L., Wallet, P., Fiset, M., Huang, B. (2004). How Does Distance Education Compare With Classroom Instruction? A Meta-Analysis of the Empirical Literature. Review of Educational Research, 74 (3), 379–439.

Boyd, D. & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. Information, Communication & Society 15(5): 662–679.

12 Siehe https://osf.io/ezcuj/wiki/home/

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd Edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis for field settings. Chicago: Rand McNally.

Cook, B. (2014). A Call for Examining Replication and Bias in Special Education Research. Remedial and Special Education, online first. doi: 10.1177/0741932514528995

Cooper, H. Hedges, L.V. & Valentine, J.C. (Eds.) (2009). The Handbook of Research Synthesis and Meta-Analysis (2nd edition). New York: Sage.

Cortina, J.M., & Landis, R.S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences (pp. 287–308). New York: Routledge.

Cummings, K. M., Brown, A. & O'Connor, R. (2007). The Cigarette Controversy. Cancer Epidemiology, Biomarkers & Prevention, *16,* 1070–1076.

DeGEval (Gesellschaft für Evaluation) (2008). Standards für Evaluation. (4. Aufl.). Köln: DeGEval.

DFG (Deutsche Forschungsgemeinschaft) (2013). Sicherung guter wissenschaftlicher Praxis (2. Aufl.). Weinheim: Wiley-VCH.
http://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/

DGfE (Deutsche Gesellschaft für Erziehungswissenschaft) (2010). Ethik-Kodex der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE). Online-Dokument:
http://www.dgfe.de/fileadmin/OrdnerRedakteure/Service/Satzung/Ethikkodex_2010.pdf

DGOF (Deutsche Gesellschaft für Online-Forschung) (2014). Richtlinie für Untersuchungen in den und mittels der Sozialen Medien (Soziale Medien Richtlinie). Online-Dokument: http://rat-marktforschung.de/fileadmin/user_upload/pdf/R11_RDMS_D.pdf.

Döring, N. (2013). Zur Operationalisierung von Geschlecht im Fragebogen: Probleme und Lösungsansätze aus Sicht von Mess-, Umfrage-, Gender- und Queer-Theorie. Gender - Zeitschrift für Geschlecht, Kultur und Gesellschaft, 2/2013, 94–113.

Döring, N. & Bortz, J. (im Druck). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften (5. Aufl.). Heidelberg: Springer.

Eid, M., Gollwitzer, M. & Schmitt, M. (2011). Statistik und Forschungsmethoden (2. Aufl.). Weinheim: Beltz.

Ellis, P.D. (2010). The essential guide to effect sizes: an introduction to statistical power, meta-analysis and the interpretation of research results. Cambridge: Cambridge University Press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175–191.

Fritz, A., Scherndl, T. & Kühlberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? Theory & Psychology 23 (1), 98-122.

Hattie, J. A. C. (2008). Visible Learning. A Synthesis of over 800 Meta-Analyses relating to Achievement. London: Routledge.

Irmen, L. & Astrid Köncke, A. (1996). Zur Psychologie des „generischen" Maskulinums. Sprache & Kognition 15 (3), 152–166.

Israel, M. & Hay, I. (2006). Research Ethics for Social Scientists. Between ethical conduct and regulatory compliance. London: Sage.

Khot, Z., Quinlan, K., Norman, G.R. & Wainman, B. (2013). The relative effectiveness of computer-based and traditional resources for education in anatomy. Anatomical Sciences Education 6 (4), 211–215.

Kilner, J.M. & Lemon, R. N. (2013). What We Know Currently About Mirror Neurons. Current Biology, 23 (23), 1057–1062.

Larsen, P. & Ins, M. von (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84 (3), 575–603.

Lee, C. J., Sugimoto, C. R., Zhang, G. & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology 64 (1), 2–17.

Lockee, B., Burton, J. & Cross, L. (1999). No comparison: Distance education finds a new use for "No significant difference". Educational Technology Research and Development, 47 (3), 33–42.

Mertens, D. M. & Ginsberg, P. E. (Eds.) (2008). The Handbook of Social Research Ethics. Thousand Oaks, CA: Sage.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist 50, 741–749.

Miethe, I. (2013). Institutionalisierung forschungsethischer Standards – Welchen Weg geht die Erziehungswissenschaft? Erziehungswissenschaft 24 (47), 13–21.

Miller, M. & James, L. (2009): Is the generic pronoun he still comprehended as excluding women? The American Journal of Psychology, 122 (4), 483–49.

Nature Neuroscience (2007). Editorial: A Cure for Dyslexia. *Nature Neuroscience* 10, 135.

Olejnik, S. & Algina, J. (2000). Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations. Contemporary Educational Psychology, 25 (3), 241–286.

Oliver, R., Wehby, J. & Reschly, D. (2011). Teacher classroom management practices: effects on disruptive or aggressive student behavior. Campbell Systematic Reviews 2011: 4. Campbell Collaboration. http://www.campbellcollaboration.org/lib/project/164/

Pashler, H. & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Reproducability in Psychological Science. A Crisis of Confidence? Perspectives on Psychological Science 7 (6), 528–530.

Peng, C.-Y., J. Chen, L.-T., Chiang, H.-M. & Chiang, Y-C. (2013). The Impact of APA and AERA Guidelines on Effect Size Reporting. Educational Psychology Review, 25 (2), 157–209.

Popper, K. (1934/2002). Logik der Forschung. Tübingen: Mohr Siebeck.

Raberger, T. & Wimmer, H. (2003). On the automaticity/cerebellar deficit hypothesis of dyslexia: balancing and continuous rapid naming in dyslexic and ADHD children. Neuropsychologia 41(11), 1493–1497.

Reynolds, D. & Nicolson, R.I. (2007). Follow-up of an exercise-based treatment for children with reading difficulties. Dyslexia, 13 (2), 78–96.

Reynolds, D., Nicolson, R.I. & Hambly, H. (2003). Evaluation of an exercise-based treatment for children with reading difficulties. *Dyslexia* 9 (1), 48–71.

Russell, T. (1999). The No Significant Difference Phenomenon. Raleigh: North Carolina State University.

Schendera, C. (2007). Datenqualität mit SPSS. München: Oldenbourg.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. Journal of Curriculum Studies, 43 (3), 425–438.

Wester, K., Borders, L.D., Boul, S. & Horton, E. (2013). Research Quality: Critique of Quantitative Articles in the Journal of Counseling & Development. Journal of Counseling & Development 91 (3), 280–290.

Zeffiro, T. & Eden, G. (2001). The cerebellum and dyslexia: perpetrator or innocent bystander?: Comment from Thomas Zeffiro and Guinevere Eden to Nicolson *et al. Trends Neurosci.* 24 (9), 512–513.

Zimmer,M. (2010). "But the data is already public": on the ethics of research in Facebook. Ethics in Information Technology 12, 313–325.

Nicola Doering, Prof. Dr. phil. habil., leitet das Fachgebiet für Medienpsychologie und Medienkonzeption am Institut für Medien und Kommunikationswissenschaft der Technischen Universität Ilmenau. Arbeitsschwerpunkte: Psychologische und soziale Aspekte der Online-, Mobil- und Mensch-Roboter-Kommunikation, Technikpsychologie und Medienkonzeption, Geschlechter- und Sexualforschung, Lernen und Lehren mit neuen Medien, Medienpädagogik, sozialwissenschaftliche Forschungsmethoden und Evaluation.

E-Mail-Adresse: nicola.doering@tu-ilmenau.de