Z Erziehungswiss (2014) (Suppl) 17:3–54
DOI 10.1007/s11618-014-0514-5

ZfE

Empirische Bildungsforschung und evidenzbasierte Bildungspolitik

Eine Analyse von Anforderungen an die Darstellung, Interpretation und Rezeption empirischer Befunde

Rainer Bromme · Manfred Prenzel · Michael Jäger

Zusammenfassung: In dem Beitrag geht es um die Möglichkeiten und Grenzen einer evidenzbasierten Bildungspolitik und evidenzbasierten pädagogischen Handelns. Die Erwartungen der unterschiedlichen Öffentlichkeiten (Bildungsadministration, Lehrkräfte, Eltern) an die Wissenschaft (Bildungsforschung) sind dabei eine Rahmenbedingung der Erkenntnisgewinnung und zugleich der Kommunikation zwischen Wissenschaft und Öffentlichkeit. Dieser Zusammenhang von Bildungsforschung, Evidenzbasierung und Wissenschaftskommunikation wird durch drei Fallstudien illustriert: Zur Wirkung gewalthaltiger Computerspiele, zum Zusammenhang von Klassengröße und Schülerleistungen und zu der Frage nach der Schulstruktur angesichts von PISA.

Schlüsselwörter: Evidenzbasierte Bildungspolitik · Pädagogisches Handeln · Wissenschaftskommunikation · Öffentlichkeit · Bildungsadministration · Gewalthaltige Computerspiele · Klassengröße · Schulstruktur

Prof. Dr. R. Bromme (✉)
Institut für Psychologie, Universität Münster,
Fliednerstr. 21,
48149 Münster, Deutschland
E-Mail: bromme@uni-muenster.de

Prof. Dr. M. Prenzel
Susanne Klatten-Stiftungslehrstuhl für Empirische Bildungsforschung,
TU München School of Education,
Schellingstr. 33,
80799 München, Deutschland
E-Mail: manfred.prenzel@tum.de

Dr. Dipl.-Psych. M. Jäger
Sebalder Forstweg 34,
91054 Buckenhof, Deutschland
E-Mail: michael.jaeger@unterrichtsevaluation.de

Educational research and evidence based educational policy – The challenge of exposing and of understanding educational research

Abstract: This paper discusses opportunities as well as constraints of evidence based policy and evidence based practice in education. Expectations held by different strands of the public (policy makers, teachers, parents) are discussed as constraint of the underlying educational research. Furthermore, such expectations constrain the communication between researchers and the public (a case of science communication). These relationships between educational research, evidence based policy and communication between research and the public are exemplified by case studies on three topics: The impact of violent computer games, the relationship between class size and students' learning outcomes and the issue of streaming in public schools.

Keywords: Evidence based policy · Educational practice · Science communication · The public · Violent computer games · Class size · Streaming in public schools

1 Einleitung

Das im November 2007 vom BMBF eingerichtete „Rahmenprogramm zur Förderung der empirischen Bildungsforschung" (Bundesministerium für Bildung und Forschung (BMBF) 2007) zeigt wie kaum ein anderes Dokument die „empirische Wende" in der Bildungspolitik an. Für eine evidenzbasierte Steuerung wird eine exzellente empirische Bildungsforschung benötigt, so lautet eine der zentralen Botschaften. Verlässliche wissenschaftliche Erkenntnisse über Bildungsprozesse und das Bildungssystem werden in dem Rahmenprogramm als Bedingungen für die Qualität eines Bildungswesens betrachtet. Das Programm soll unter anderem dazu beitragen, dass empirisch gesichertes Wissen für die Reform des Bildungssystems bereitgestellt wird.

Im Bildungsbereich zeichnet sich damit eine Entwicklung in Richtung einer daten- und evidenzgestützten Politik ab, die in anderen Bereichen (z. B. Gesundheit, Wirtschaft, Umwelt) bereits seit längerer Zeit begonnen hat. Die in der deutschen Öffentlichkeit stark beachteten internationalen Vergleichsstudien markieren einen Wendepunkt in der Bildungspolitik. Daten und deren Interpretation spielen eine wichtige Rolle im politischen Diskurs und in der Begründung von Entscheidungen. Die „empirische Wende", von der nach TIMSS und den „Konstanzer Beschlüssen" der KMK die Rede war, bleibt nicht nur auf das Messen beschränkt, sondern sie betrifft in der Konsequenz ebenso das bildungspolitische Handeln (Lange 2008).

Die Bildungspolitik ist auf Daten, Belege und theoretische Konzepte angewiesen, sei es zur Erfassung des Ist-Standes und von Problemlagen, zur Planung von Maßnahmen oder zur Bewertung von Perspektiven und Szenarien. Sie benötigt „Evidenz", um rational entscheiden zu können. Da Bildungsfragen auf vielen gesellschaftlichen Ebenen verhandelt werden (alle waren einmal Schüler, sehr viele Menschen haben Erfahrungen als Eltern), braucht sie die angesprochenen Wissensgrundlagen auch, um an dem gesellschaftlichen Diskurs zu Bildungsfragen mitzuwirken. Politische Entscheidungen müssen sich einem Diskurs mit unterschiedlichen (professionellen und nicht-professionellen) Öffentlichkeiten stellen und Entscheidungen rational und überzeugend begründen. Und sie setzen einen solchen Diskurs auch voraus – man denke zum Beispiel an den Wettbewerb zwi-

schen Bildungspolitik und anderen Politikfeldern um öffentliche Ressourcen. Auch da, wo die Politik die Bildungsforschung nicht zu der Lieferung von Daten zu bestimmten Problemlagen auffordert, muss sie sich mit den Diskursbeiträgen der bildungsbezogenen Forschung (ganz unterschiedlicher Provenienz) auseinandersetzen. Dabei kann sie heute in unterschiedlichem Ausmaß auf empirische Befunde zurückgreifen, die als einschlägig und gesichert gelten können.

Die Vorstellung einer bildungspolitischen Steuerung, die sich an empirischer Evidenz orientiert, überzeugt in demokratischen Gesellschaften durch den Anspruch, auf diese Weise Transparenz, Rationalität und sachliche Begründung politischer Entscheidungen sicherzustellen. Eine evidenzbasierte Steuerung wird jedoch zugleich mit großen Herausforderungen und Schwierigkeiten konfrontiert, die eben für Forschung typischerweise kennzeichnend sind. Empirische Befunde und ihre Belastbarkeit wie Aussagekraft hängen von den jeweiligen Fragestellungen, Theorien und Methoden ab. Gegenstandsbereiche werden durch Forschung keineswegs gleichmäßig abgedeckt. Ob Fragestellungen zufriedenstellend, ansatzweise oder noch gar nicht beantwortet werden können, hängt vom Stand der Forschung ab. Forschungsergebnisse können widersprüchlich sein und erst einmal weiterführende Fragen aufwerfen, die mit zusätzlichem Aufwand an Geld und Zeit untersucht werden müssten. Fast schon trivial erscheint hier die grundsätzliche Problematik, dass Daten und Befunde nicht einfach für sich sprechen, sondern unter Berücksichtigung von Theorien, Methoden und sonstigem Erkenntnisstand sorgfältig und kritisch interpretiert werden müssen, um Folgerungen ziehen zu können. Erst durch eine entsprechende Interpretation werden Daten zu „Evidenz“ für Einschätzungen oder Entscheidungen.

In Anbetracht einer relativ jungen empirischen Bildungsforschung und einer noch jüngeren Diskussion über Evidenz in der Bildungspolitik soll der vorliegende Beitrag Voraussetzungen und Bedingungen einer evidenzbasierten Steuerung in diesem besonderen Feld klären. Im Zentrum steht die Frage, wie die Rezeption von Befunden der empirischen Bildungsforschung und ihre Übersetzung in Evidenz verbessert werden können. Selbstverständlich steht dabei auch die Frage der Darstellung und Vermittlung von Befunden und Erkenntnissen aus der empirischen Bildungsforschung im Blickpunkt.

Unser Beitrag behandelt insbesondere folgende Fragestellungen:

- Wie kann den Akteuren in der Bildungspolitik und Bildungsadministration die Interpretation und Nutzung empirischer Forschung erleichtert werden? Welche Kriterien und Verfahren können sich dabei als hilfreich erweisen?
- Wer sind zentrale Adressaten zur Umsetzung der Erkenntnisse? Auf welchen Handlungsebenen agieren sie und welche Erkenntnisse erhoffen sie sich entsprechend?
- Wie wird in Politik und Administration mit der Vorläufigkeit und Widersprüchlichkeit empirischer Befunde der Bildungsforschung umgegangen? Was heißt „evidenzbasiert entscheiden“ in Anbetracht von Ungewissheit?
- Wie könnte die empirische Bildungsforschung den Erwartungen der Bildungspolitik besser Rechnung tragen, nützliche Erkenntnisse und Evidenz für relevante Entscheidungen zu liefern? Welche Programme sind dafür geeignet und wo sind die Grenzen einer „nutzenorientierten“ Forschung?

Um diese Fragen zu diskutieren, schaffen wir zunächst ein gemeinsames Verständnis der Begriffe und beschreiben die Nutzung von Evidenz als Entscheidungsgrundlage in unter-

schiedlichen Bereichen (Abschn. 2.2). Im Abschn. 2.2 wird auch damit begonnen, Rahmenbedingungen, Wirkungszusammenhänge und Konsequenzen einer evidenzbasierten Steuerung zu sammeln und zu verdichten.

Grundlage für die Bearbeitung der angesprochenen Fragestellungen sind drei inhaltliche Beispiele für Befunde aus der empirischen Bildungsforschung, die offensichtlich hohe bildungspolitische Relevanz besitzen. Die im dritten Abschnitt vorgestellten Beispiele repräsentieren unterschiedliche Konstellationen von mehr oder weniger abgesicherter Befundlage, Rezeption und Interpretation in Öffentlichkeit und Politik. Die Themen betreffen a) gewalthaltige Computerspiele, b) Auswirkungen von Klassengröße auf Schülerleistung und c) Konsequenzen aus PISA-Studie. Anhand dieser typischen Beispiele sollen Anforderungen und Probleme des Umgangs mit Evidenz aus der empirischen Bildungsforschung herausgearbeitet werden.

Der fünfte Abschnitt dieses Beitrags fasst die Ergebnisse der Beispielanalysen zusammen und verdichtet diese in Empfehlungen. Mit Blick auf Akteure und Handlungsebenen werden Handlungslogiken von Forschung, Politik, Administration und Öffentlichkeit beim Umgang mit Informationen der Bildungsforschung analysiert. Neben Empfehlungen zu zukünftigen Forschungsfeldern geben wir Hinweise zur Relevanz von Forschungsergebnissen in Hinblick auf die Steuerungsaufgaben von Bildungspolitik und Bildungsadministration.

2 Evidenz als Grundlage für Entscheidungen

Der Begriff der Evidenz scheint im alltäglichen Sprachgebrauch inzwischen selbstverständlich geworden zu sein. Im Wissenschaftsbereich wird der Begriff in der geisteswissenschaftlich-hermeneutischen Tradition, die zum Beispiel von „Evidenzerlebnissen“ spricht, allerdings deutlich anders gefasst als in den Naturwissenschaften. Dort handelt es sich um Befunde, die mit empirisch-wissenschaftlichen Methoden gewonnen wurden und mit denen Theorien (auch über Handlungsoptionen) bekräftigt oder widerlegt werden können (Schneider et al. 2007; Shavelson und Towne 2002). Bemerkenswert ist auch, dass der Begriff „Evidenz“ in romanischen Sprachen (Französisch, Italienisch) nicht vorkommt und typische Übersetzungsvorschläge (z. B. „les preuves“) deshalb missverständliche Konnotationen („Beweis“) transportieren. Evidenz im naturwissenschaftlichen Sinn dient eher als Beleg denn als Beweis für Vermutungen oder Theorien. Die neueren Diskussionen über Evidenzbasierung im Bildungsbereich werden von einem Evidenzbegriff geprägt, der dem naturwissenschaftlich-empirischen (und nicht dem geisteswissenschaftlich-hermeneutischen) Konzept entspricht.

Im Bildungskontext können zwei Stränge einer Diskussion über Evidenzbasierung unterschieden werden. Im Blickpunkt des ersten Stranges steht die Frage der Evidenzbasierung des Entscheidens und Handelns in der *pädagogischen Praxis*. Der zweite Strang betrifft die Evidenzbasierung des Entscheidens, Steuerns und Handelns *auf bildungspolitischer Ebene.* Die Unterschiede beziehen sich nicht nur auf den Aspekt einer „Mikro“- oder „Makro-Steuerung“. Vielmehr sind die institutionellen Randbedingungen, die Settings, die Aufgaben und Interaktionsfelder, in denen agiert wird, in beiden Strängen höchst unterschiedlich. Aus diesen Gründen konzentrieren sich die Überlegungen

in diesem Beitrag auf Fragen der Steuerung im Feld der Bildungsadministration und Bildungspolitik. Im Folgenden wird zunächst erläutert, wie aus Daten und Befunden gewissermaßen „Evidenz“ gewonnen wird. Der zweite Abschnitt behandelt Gegenstand und Fragestellungen der empirischen Bildungsforschung und bezieht diese auf Erkenntnisinteressen aus der Bildungsadministration und Bildungspolitik. Im dritten Abschnitt werden Möglichkeiten der Abstufung der Aussagekraft und Relevanz von Evidenz (auch unter Bezugnahme auf Beispiele aus anderen Disziplinen) diskutiert. Abschließend werden die Grenzen dieses Ansatzes und verschiedene Argumente gegen Evidenzbasierung erörtert.

2.1 Von Daten zur Evidenz

Empirisch ausgerichtete Forschungsansätze gewinnen Daten und Befunde im Kontext von Fragestellungen und Theorien. Wenn die Daten dazu dienen, Vermutungen, Hypothesen oder Theorien zu stützen – oder zu widerlegen – erhalten sie die Funktion von „Evidenz“. In diesem Sinne gibt es keine Evidenz „an sich“, sondern nur Evidenz „für“ oder „gegen“ Aussagen oder Vermutungen. Die enge Verflechtung von Evidenz mit Theorien und Methoden ist *innerhalb* der Wissenschaften selbstverständlich, und damit eben auch die genaue und sorgfältige Prüfung der Belastbarkeit und Aussagekraft von Daten mit Blick auf Aussagen. Einzelne Aussagen (z. B. Hypothesen) sind dabei normalerweise in größere Zusammenhänge (z. B. Theorien) eingebunden und werden entsprechend in diesem Kontext mit Daten, Belegen oder „Evidenz“ konfrontiert.

Für die Beurteilung von Evidenz spielen in der Wissenschaft methodologische Kriterien eine wichtige Rolle. Untersuchungsdesigns, Erhebungsmethoden und Auswertungsverfahren haben bestimmten Standards zu genügen; sie müssen dem aktuellen Stand in den wissenschaftlichen Disziplinen entsprechen. Es liegt auf der Hand, dass zum Beispiel die Messgenauigkeit oder Messfehler die Aussagekraft von Befunden beeinflussen und damit das Gewicht der „Evidenz“ reduzieren können. Von der Stichprobenziehung und von den Beteiligungsquoten hängt es ab, ob Befunde als repräsentativ gelten können. Sind sie nicht „repräsentativ“, dann können keine generalisierenden Aussagen getroffen werden; die Daten haben dann nur „lokale“ Gültigkeit und können als Hinweise oder vielleicht als „schwache“ Evidenz interpretiert werden. Die Aussagekraft von experimentell gewonnenen Befunden wiederum hängt von der Kontrolle möglicher Störfaktoren ab. Die Anwendung ungeeigneter statistischer Verfahren produziert Artefakte usw.

Ein wichtiges Prinzip wissenschaftlichen Arbeitens betrifft außerdem zum Beispiel die Replikation von Befunden. Replizierte Befunde stabilisieren die Aussagekraft, wenn unter anderen Randbedingungen (z. B. andere Forschergruppe, andere Stichprobe aus der Grundgesamtheit, mehr oder weniger variierende Durchführung) die gleichen Ergebnisse gewonnen werden. Umgekehrt wird ein erster (vor allem überraschender) Befund mit mehr Vorsicht betrachtet. Besonders schwierig wird die Beurteilung der Evidenz, wenn mehrere Studien zu gleichen Themen zu unterschiedlichen Befunden gelangen. Hier gilt es etwa, sehr sorgfältig die Unterschiede in den Methoden, Stichproben und Analyseverfahren zu prüfen, um die Qualität und das Gewicht der Befunde in Hinblick auf eine Hypothese/Aussage zu beurteilen.

Die Übersetzung von Daten/Befunden in Evidenz ist also ein aufwändiger und komplexer Vorgang, der insbesondere Expertise und kritische Reflexion verlangt – und gelegentlich auch in wissenschaftliche Kontroversen führt. Aus einer wissenschaftstheoretischen Sicht (z. B. des kritischen Rationalismus sensu Karl Popper 2007) sind das alles Merkmale normaler Wissenschaft. Auch empirisch gestützte wissenschaftliche Aussagen werden immer als „vorläufig" betrachtet; der wissenschaftliche Fortschritt erfolgt auf lange Sicht über das Ausschließen von Aussagen, die sich empirisch nicht bewährt haben und entsprechend nicht haltbar sind.

Bei einer Betrachtung der Wissenschaft „*von außen*" mögen die hier angesprochenen Kautelen und Vorbehalte irritierend erscheinen. Allerdings ist es so, dass grundsätzliche (wissenschaftstheoretisch begründete) Einschränkungen normalerweise nicht in der wissenschaftlichen Berichterstattung thematisiert werden – sie werden (in der Scientific Community) als bekannt vorausgesetzt und brauchen deshalb nicht ständig wiederholt zu werden. In wissenschaftlichen Artikeln behandelt die übliche „Diskussion" der Befunde jedoch (möglichst selbstkritisch) Grenzen der Methoden und der Interpretierbarkeit der dargestellten Studie. Diese Diskussion adressiert freilich üblicherweise nur die akademischen Kolleginnen und Kollegen. Sie bedient sich der Fachterminologie und unterstellt ein gemeinsames Bezugssystem für die Beurteilung der Tragfähigkeit von Befunden, das eine konzise Diskussion auf engem Raum erlaubt. Generell besteht in der Wissenschaft die Tendenz, in den Forschungsberichten den Erkenntnisfortschritt (gegenüber früheren Forschungsansätzen) herauszustellen. Dies entspricht durchaus der Logik des Forschungsprozesses. Dabei werden jedoch oft nicht die Einschränkungen wiederholt, die auch bei dem verbesserten Erkenntnisstand weiter bestehen könnten. Und es gibt eine Tendenz zur Betonung des jeweils Neuen, Andersartigen aktueller Forschungserträge gegenüber dem bisherigen Erkenntnisstand. Dies führt für den Außenstehenden leicht zu einer Unterschätzung der bereits kumulativ erzielten Kohärenz an Befunden und deren theoretischen Deutungen. Wissenschaftliche Forschungsberichte (insbesondere Zeitschriftenartikel) werden also in einem Wissenschaftssystem mit Transparenzregeln und internen Prüfungsmechanismen (z. B. Begutachtungen, Sekundäranalysen, Replikationen) veröffentlicht. Sie sind deshalb normalerweise nicht auf eine Rezeption durch eine fachfremde, wissenschaftsexterne Öffentlichkeit angelegt, für die eine deutlich andere und sehr viel ausführlichere Darstellung der Datenlage unter Evidenzgesichtspunkten erforderlich wäre.

Generell besteht die Erwartung, dass das Wissenschaftssystem selbst für die Qualitätssicherung zuständig ist und diese verantwortungsbewusst leistet. Aufgrund des Aufwands der methodischen Anstrengungen und der gegenseitigen Kontrolle im Forschungsprozess und bei der Veröffentlichung werden wissenschaftlichen Aussagen ein besonderes Gewicht und eine besondere Glaubwürdigkeit zugemessen. Wissenschaftliche Aussagen stehen nicht in Verdacht, durch subjektive Voreingenommenheiten verzerrt zu sein. Damit erhalten wissenschaftliche Aussagen eine hohe Attraktivität, um in Argumentationsketten als Beleg verwendet zu werden. Evidenz aus wissenschaftlichen Studien kann Gründe liefern, zum Beispiel Situationen auf bestimmte Weise zu beurteilen, weil sie auf Folgen aufmerksam macht, die mit hoher Wahrscheinlichkeit eintreten werden. Evidenz kann helfen, bestimmte Problemlagen zu erklären und Faktoren zu identifizieren, die kritisch sind. Evidenz kann Gründe liefern, bestimmte Maßnahmen zu ergreifen, um Probleme

zu lösen. Nicht zuletzt kann Evidenz zur Rechtfertigung von Entscheidungen beitragen, weil dadurch offensichtlich aktuelles und bestens abgesichertes wissenschaftliches Wissen berücksichtigt wurde.

Evidenz spielt also nicht nur innerhalb des Forschungssystems eine Rolle, wenn es darum geht, Hypothesen zu verwerfen oder beizubehalten. Evidenz erhält darüber hinaus überall dort eine Schlüsselstellung, wo Entscheidungen oder Handlungen unter Rationalitätsgesichtspunkten getroffen werden sollen. In fast allen Berufsfeldern folgen die Vorstellungen von professionellem Handeln einem Grundmodell rationalen Entscheidens und Handelns. Dies gilt auch für Entscheiden und Handeln in weiten Bereichen der Administration und Politik. Wissenschaftliche Evidenz hat herausragende Bedeutung für rationales Handeln, weil sie das beste verfügbare Wissen repräsentiert – zum Beispiel über Problemlagen und dort wirkende Einflussfaktoren, über mögliche Maßnahmen und deren kurz- und längerfristige Folgen und Nebenwirkungen. Selbstverständlich muss die wissenschaftliche Evidenz inhaltlich auf das jeweilige Entscheidungs- und Handlungsfeld passen. Der folgende Abschnitt erörtert deshalb Fragestellungen und Gegenstandsbereiche der empirischen Bildungsforschung und bezieht diese auf Anliegen der Bildungsadministration und Bildungspolitik sowie deren Bedarf an Wissen und Evidenz.

2.2 Gegenstand und Fragestellungen der empirischen Bildungsforschung

Gegenstand der empirischen Bildungsforschung sind Voraussetzungen, Prozesse, Ergebnisse sowie Ziele von Bildung über den Verlauf der Lebensspanne (vgl. Prenzel 2005). Die empirische Bildungsforschung untersucht diese Prozesse nicht nur auf individueller Ebene, sondern betrachtet diese ebenfalls systematisch im institutionellen oder gesellschaftlichen Kontext. Im Blickpunkt der Bildungsforschung stehen insbesondere *Relationen* (z. B. zwischen Voraussetzungen und Prozessen oder zwischen Prozessen und Ergebnissen; zeitliche Relationen im Lebenslauf, Relationen zwischen institutionellen Kontexten und Situationen etc.). „Empirische Bildungsforschung“ bezeichnet ein *Forschungsfeld*, in dem verschiedene Disziplinen problemorientiert wissenschaftlich arbeiten (ähnlich wie in der Klima-, Meeres- oder Krebsforschung). Zur empirischen Bildungsforschung tragen unter anderem die Erziehungswissenschaft, die verschiedenen Fachdidaktiken, die Psychologie und die Soziologie bei, aber zum Beispiel auch die Kommunikationswissenschaften, die Ökonomie oder die Politikwissenschaft.

Die skizzierte Gegenstandsbeschreibung der empirischen Bildungsforschung ist weit gefasst und deckt sich weitgehend mit dem Gegenstandsbereich von Bildungspolitik und Administration. Dieser kann unterschiedlichen Ressorts (z. B. Schule, tertiäre Bildung, Weiterbildung) zugeordnet werden beziehungsweise in unterschiedlichen ministeriellen Zuständigkeiten (Wissenschaft, Familie, Soziales, Arbeit) liegen. Dennoch bleiben die Wissensbedarfe vergleichbar – es spielt keine Rolle, ob Fragen nach der Ausstattung von Kindergärten, dem Qualifikationsniveau der Erzieherinnen oder den häufigsten Sprachproblemen und geeigneten Förderkonzepten von einem Sozial- oder Kultusministerium gestellt werden: Alle diese Fragen fallen prinzipiell in den Gegenstandsbereich der empirischen Bildungsforschung. Allerdings muss nicht immer die empirische Bildungsforschung konsultiert werden, um diese Fragen zu beantworten. Es kann durchaus sein, dass die Administration selbst über aussagekräftige Daten verfügt beziehungsweise

durch Abfragen erhalten kann. Freilich gilt es auch hier, die Qualität und Belastbarkeit der Daten abzuschätzen.

Die Aufgabe der empirischen Bildungsforschung kann grundsätzlich darin gesehen werden, empirisch gesichertes Wissen über ihren Gegenstandsbereich bereitzustellen (vgl. Prenzel 2012). Dieses Wissen umfasst zum Beispiel Aussagen über Voraussetzungen, Prozesse und Kontexte von Bildung sowie über deren Relationen. Dieses Wissen kann *Beschreibungen* enthalten (z. B. von erreichten Bildungsergebnissen), *Vorhersagen* erlauben (z. B. von Prozessen durch Voraussetzungen), *Erklärungen* gestatten (z. B. von Ergebnissen durch Prozesse, situative Bedingungen, Kontextfaktoren) oder Aussagen über Maßnahmen treffen, die bei gegebenen Voraussetzungen und Kontexten mit hoher Wahrscheinlichkeit bestimmte Ziele erreichen lassen (*Veränderungswissen*). Empirische Bildungsforschung kann im Übrigen Ziele beschreiben oder Relationen zwischen Zielen und Teilzielen klären oder Nebenwirkungen auf andere Ziele erfassen, aber sie setzt keine Ziele.

Die hier vorgenommene Differenzierung von Arten von Wissen ist mit Blick auf die Frage der Evidenzbasierung im Bildungsbereich von hoher Bedeutung. Aus wissenschaftlicher Perspektive ergeben sich zum Beispiel für das Bereitstellen von Erklärungswissen andere Anforderungen an die Untersuchungsdesigns, Stichproben und Methoden als für das Bereitstellen von Beschreibungswissen. Das heißt nicht (um einem Missverständnis entgegenzuwirken), dass es von vornherein methodisch schwieriger oder anspruchsvoller ist, Erklärungs- oder Veränderungswissen zu gewinnen als etwa Beschreibungswissen. Allerdings gibt es durchaus Abhängigkeiten zwischen den Wissensarten, denn ohne grundlegendes Beschreibungswissen wird es nicht möglich sein, Erklärungs- oder Veränderungswissen abzusichern.

2.3 Welche belastbare Evidenz kann von der empirischen Bildungsforschung erwartet werden?

Nach den Ausführungen über verschiedene Arten von Wissen und Erkenntnisinteressen der Bildungsforschung einerseits, der Bildungspolitik und Bildungsadministration andererseits, kehren wir zur Frage zurück, wie Evidenz aus der Forschung beurteilt und genutzt werden kann.

2.3.1 Analytische Bestimmung der Arten von Wissen, die die empirische Bildungsforschung liefern kann

Grundsätzlich hilfreich ist die Frage nach der *Art des Wissens*, das von der Forschung bereitgestellt – oder von der Bildungspolitik und -administration gewünscht wird. Geht es um empirisch gesichertes Wissen zum Zweck der Beschreibung, der Vorhersage, der Erklärung oder der Veränderung? Eine Verständigung über die Qualität des verfügbaren beziehungsweise gewünschten Wissens erleichtert eine Bestandsaufnahme und auch die Klärung der Frage, wo weiter Forschungsbedarf besteht.

Ohne hier im Einzelnen auf methodologische Fragen einzugehen, hilft die Klassifikation der Wissensarten auch Außenstehenden bei einer ersten Einschätzung der Belastbarkeit von Befunden. Überblickstudien, manchmal auch Surveys genannt, eignen sich

zum Beispiel sehr gut für die Bereitstellung von Beschreibungswissen. Sie erlauben auch statistische Analysen von Zusammenhängen (Korrelationen, Regressionen), aber sie reichen nicht aus, um Erklärungen zu prüfen und kausal relevante Faktoren zu identifizieren. Dafür braucht es Experimente. Experimentelle Designs werden auch benötigt, um Effekte von Maßnahmen solide zu prüfen und damit zu Veränderungswissen zu gelangen. Fortgeschrittene methodologische Fragen im Zusammenhang mit den Untersuchungsdesigns betreffen dann zum Beispiel die Stichproben (bzw. Populationen, über die Aussagen getroffen werden sollen), die Art der Kontrolle bei experimentellen Designs oder die Art der Kontrolle bei Feldversuchen oder Interventionsstudien.

Für die Verständigung hilfreich sind weitere der oben getroffenen Unterscheidungen. Bezieht sich das (Beschreibungs-, Erklärungs- oder Veränderungs-) Wissen auf Voraussetzungen bzw. Bedingungen, Prozesse oder Ergebnisse von Bildung (bzw. auf welche Relationen zwischen diesen drei Ebenen)? Und um welche geht es inhaltlich? Weiterführende, sozusagen fortgeschrittene methodologische Fragen könnten dann der Art der Erhebung und Messung dieser Konstrukte gelten. Hier greifen etwa die bekannten Gütekriterien der Validität (Gültigkeit), Reliabilität (Genauigkeit/Zuverlässigkeit) und Objektivität von empirischen Erhebungsverfahren. Ohne in Details zu gehen, ist es grundsätzlich hilfreich zu wissen, dass Tests mit bestimmten Bezeichnungen (z. B. Lese- oder Mathematiktest) sehr Unterschiedliches erfassen können (Validität), dass trotz größter Anstrengungen empirische Methoden nicht messfehlerfrei sind (wobei die Größe des Fehlers jedoch bestimmt werden kann) oder dass subjektive Einschätzungen in Fragebogen oder Ratingskalen von Bezugssystemen oder Erwünschtheitseffekten beeinflusst werden können.

Unter der Berücksichtigung der ebenfalls bereits angesprochenen Tatsache, dass Forschung in gewisser Weise immer vorläufige Erkenntnisse liefert, stellt sich die Frage, in wie vielen Studien bestimmte Sachverhalte auf vergleichbare/ähnliche Weise untersucht wurden, welche Stichproben (Umfang, Zusammensetzung, Zufallsstichprobe) dabei einbezogen waren und wie sich entsprechend das Gesamtbild darstellt. Wenn zu bestimmten Fragestellungen größere Zahlen von Studien vorliegen, können Metaanalysen und systematische Reviews dazu beitragen, nicht nur einen Überblick über die Befundlage zu gewinnen, sondern mit geeigneten statistischen Verfahren abzuschätzen, welche Faktoren erklärungsrelevant sind oder von welchen Maßnahmen stärkere Effekte zu erwarten sind. Auch im Bereich der empirischen Bildungsforschung sind solche Verfahren längst gebräuchlich (z. B. in der Unterrichtsforschung: Hattie 2009; Seidel und Shavelson 2007; Slavin et al. 2009a; Slavin et al. 2009b).

2.3.2 Hierarchien von Evidenz als rationale Grundlage von Forschungssynthesen?

In manchen Feldern, so zum Beispiel in der Medizin, wird seit einiger Zeit versucht, den Prozess der Begutachtung von Studien in Hinblick auf ihre Aussagekraft zu systematisieren. Als Gerüst für die Beurteilung der Relevanz von Studien (zum Beispiel durch praktizierende Ärzte) dient ein hierarchisches Modell der Evidenz, das von der Cochrane Collaboration postuliert wird. Das Regelsystem, das die Cochrane Collaboration für die Beurteilung der Evidenz vorschlägt und selbst anwendet, findet im Bereich der Medizin weitreichend Anerkennung, obschon zum Teil auch kritische Vorbehalte gemacht werden.

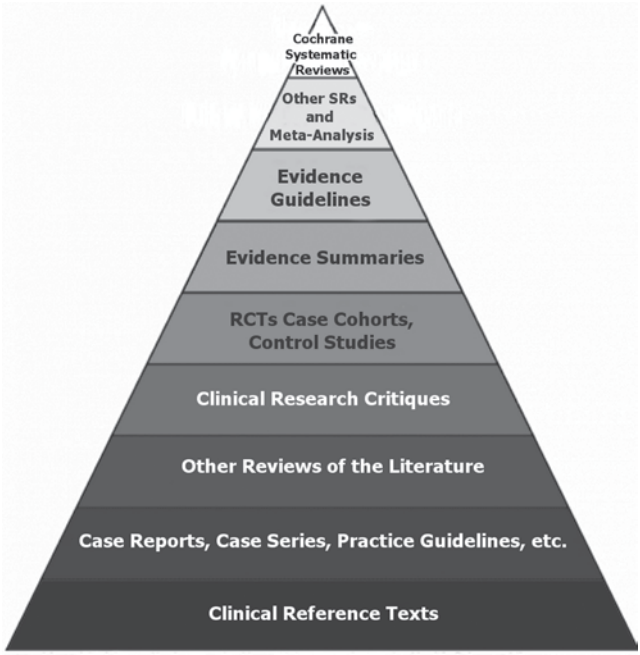

Abb. 1: Hierarchie der Evidenz im Kontext der Evidence-based-practice in der Medizin. (http://libguides.hsl.washington.edu/ebptools)

Im Folgenden soll das Modell vorgestellt und dann mit Blick auf den Bildungsbereich diskutiert werden (Abb. 1).

Um die Qualität von Studien einzuschätzen, werden die unterschiedlichen Informationsquellen in Form einer Pyramide dargestellt. An der Spitze der Pyramide stehen systematische Reviews nach den Regeln der Cochrane Collaboration (www.cochrane.org), einer Non-Profit-Organisation, die seit 1993 ein internationales Netzwerk aufbaut, um evidenzbasierte Entscheidungen im Gesundheitswesen zu unterstützen. Dazu wurde ein Prozess für die systematischen Reviews entwickelt, der wie auch die Begutachtung der medizinischen Studien, dem Peer-Review sowie einem kontinuierlichen Verbesserungsprozess unterworfen ist. Die aktuelle Version des Prozesses sowie die Anpassungen gegenüber den vorherigen Versionen sind unter www.cochrane-handbook.org dokumentiert.

Auf der nächsten Stufe stehen andere systematische Reviews oder Metaanalysen, sofern sie auf randomisierten Kontrollgruppenstudien aufgebaut werden. Im Unterschied zu den Cochrane Reviews fehlen hier zum Beispiel standardisierte und kontrollierte Bedingungen für den Reviewprozess. Die danach aufgeführten Evidence Guidelines sind konkrete Handlungsanweisungen und Verfahrensregeln, die ebenfalls auf kontrollierten und randomisierten Studienergebnissen aufsetzen.

Unsystematische Zusammenfassungen von randomisierten Kontrollgruppenstudien bilden die nächste Stufe der Bewertung der Evidenz, gefolgt von einzelnen randomisierten Kontrollgruppenstudien sowie echten Längsschnittanalysen. Mit dieser Stufe enden die streng kontrollierten Designs in der medizinischen Forschung.

Die nun folgenden Bewertungsstufen gehen von einer deutlich geringeren Aussagekraft der Studien aus, da die Forschungsdesigns alternative Erklärungen für die Effekte

abseits des jeweils kontrollierten Treatments nicht ausschließen können. Die Logik der Bewertung bleibt dabei bestehen: Zusammenfassende Berichte (Überblicksarbeiten), die übereinstimmend aus unterschiedlichen Perspektiven ähnliche Effekte berichten, werden höher gewichtet als ähnliche Ergebnisse aufgrund ähnlicher methodischer Ansätze. Einzelfallstudien sowie aus bekannten Effekten ohne weitere Überprüfung abgeleitete Zusammenhänge genießen die niedrigste Aussagekraft in dieser Bewertungshierarchie.

Bemerkenswert an diesem Konzept ist zunächst die Tatsache, dass in der Medizin weltweit Einrichtungen und Netzwerke etabliert wurden, die sich konsequent und systematisch mit konkreten Fragen der Evidenz für Behandlungsformen in der Medizin befassen. In Deutschland hat das 2004 gegründete *Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen* (IQWiG, Köln) begonnen, ähnliche Aufgaben zu übernehmen. Es gibt also im Medizinbereich in einzelnen Staaten unterschiedlich stark ausgebaute Unterstützungssysteme, die Evidenz für Behandlungen analysieren, bewerten und für professionelle Nutzer aufbereiten. „Evidenzbasierung" ist also kein Selbstläufer in Handlungsfeldern, sondern ein Ziel, das Ressourcen, Infrastruktur und Maßnahmen benötigt. Neben der strengen und sorgfältigen wissenschaftlichen Beurteilung von Evidenz müssen die Ergebnisse nachvollziehbar aufbereitet und den Gruppen zur Verfügung gestellt beziehungsweise kommuniziert werden, die auf der Basis der Evidenz Entscheidungen treffen sollen.

Die als Leitidee dienende Hierarchie von Evidenz je nach Art der Studien und Analysen folgt einer im Prinzip überzeugenden Logik: Für die Beurteilung der Evidenz für Behandlungsformen gewichtet sie systematische Reviews höher als randomisierte experimentelle Studien und diese wieder höher als Fallstudien, theoretische Papiere oder Praxisberichte. Diese hierarchische Betrachtung im Hinblick auf Behandlungen bildet jedoch – das sollte unterstrichen werden – nicht die *wissenschaftliche* Relevanz dieser Untersuchungszugänge und Forschungsergebnisse ab. Auch eine Fallstudie könnte aus Forschungsperspektive höchst innovativ und bahnbrechend für nachfolgende Studien sein.

Hervorzuheben ist schließlich die Relativierung, die bezogen auf das verfügbare Wissen aus unterschiedlichen Untersuchungsansätzen vorgenommen werden muss: Wenn keine systematischen Reviews zu einer Fragestellung vorliegen, dann steht möglicherweise eine kontrollierte Studie am Gipfel der Evidenzhierarchie oder sogar ein Praxisbericht.

2.4 Methodische Probleme einer Evidenzorientierung im Bildungsbereich

Überträgt man das Konzept der Evidenzhierarchie auf das Feld der empirischen Bildungsforschung, dann stellt sich dort die Situation in verschiedener Hinsicht anders dar: Das Forschungsfeld „Bildung" wird erst seit relativ kurzer Zeit im größeren Umfang mit empirischen Ansätzen bearbeitet und der Mitteleinsatz für Forschung ist (im Vergleich zur Medizin wie im Vergleich zu den Gesamtausgaben für das Bildungssystem) extrem gering (Burkhardt und Schoenfeld 2003; OECD 2003). Insgesamt gibt es also viele Teilbereiche und Fragestellungen, die bisher noch nicht Gegenstand von systematischer empirischer Forschung waren. Im Bildungsbereich steht die Forschung außerdem vor einigen fast unüberwindbar scheinenden Problemen. Zum Beispiel sind eine konsequente Randomisierung (z. B. zufällige Zuordnung von Schülerinnen und Schüler zu Klassenverbänden und Lehrkräften) oder die Kontrolle von potentiell relevanten Ein-

flussfaktoren über längere Zeiträume (z. B. in den Kontexten Elternhaus, Freundeskreis, Mediennutzung, Klasse, Schule) zu gewährleisten. Im Ergebnis führt das dazu, dass die „Spitze" der Evidenzhierarchie im Bildungsbereich sehr viel schwächer ausgeprägt oder besetzt ist als der „Rumpf".

Immerhin gibt es seit 2001 ein der Cochrane Collaboration in Aufbau und Ziel ähnliches Netzwerk, das unter der Bezeichnung Campbell Collaboration internationale Erkenntnisse in den Bereichen Bildung, Kriminologie, Sozialhilfe sowie Methoden und Statistik sammelt (http://www.campbellcollaboration.org) (vgl. auch Beelmann 2014 in diesem Band). Weitere Initiativen (eine Aufstellung findet man z. B. bei Fleischman 2009) verfolgen ähnliche Ziele, zum Beispiel die „Best Evidence Encyclopedia" an der Johns Hopkins University (http://www.bestevidence.org/). Allerdings zeigt sich hier überall immer noch der angesprochene große Bedarf an Metaanalysen und systematischen Reviews (Beelmann 2014; Pant 2014).

Betrachtet man die Möglichkeiten der Evidenzbasierung von Entscheidungen im Bildungsbereich auf der administrativen und politischen Ebene, werden weitere Besonderheiten erkennbar. So ist es auf der Ebene von Bildungssystemen fast unmöglich, durch experimentell angelegte Studien Evidenz für umfassende Strukturentscheidungen auf höheren Stufen der Hierarchie zu gewinnen. Zugespitzt ausgedrückt: Es verbietet sich, versuchsweise das (in vielen Leistungsvergleichen erfolgreiche) finnische Schulsystem in einer Experimentalgruppe von Staaten einzuführen und mit einer Kontrollgruppe zu vergleichen. Auch für andere Analysezugänge ist die Zahl von „Fällen" (Staaten) zu klein, um größere Mengen an unabhängigen und abhängigen Variablen statistisch kontrollieren zu können. Obwohl es weiter schwierig bleiben wird, Evidenz auf den höchsten Stufen der Evidenzhierarchie für weitreichende bildungspolitische Entscheidungen zu gewinnen, bleiben viele Möglichkeiten, für weniger weitreichende Entscheidungen solide Evidenz auf etwas niedriger angesiedelten Stufen der Hierarchie beizusteuern. In diesem Zusammenhang gewinnen andere Forschungszugänge (etwa vergleichende Surveys mit Trendanalysen, Längsschnitt- und Panelstudien) hohe Bedeutung (z. B. Blossfeld et al. 2011).

Deshalb ist es erforderlich, dass (wie in der Medizin) wissenschaftliche Methoden weiterentwickelt werden, die aus Forschungsergebnissen in einer kritischen Zusammenschau die Evidenzen destillieren, die entscheidungsrelevant werden. Sie weisen auch auf die Einschränkungen und Bedingungen der Gültigkeit und Anwendbarkeit hin.

Eine Verstärkung der Evidenzorientierung setzt also Infrastruktur voraus und bedeutet zusätzliche Kosten. Evidenzorientierung als erklärtes Prinzip hat zur Folge, dass fehlende Evidenz „erforscht" werden muss – und dies setzt Mittel und Anreize für eine entsprechend zugeschnittene Forschung voraus. Und sie hat zur Folge, dass neben der Evidenzbeurteilung, die sich an den randomized trials als dem „Goldstandard" ausrichtet, andere Formen der Evidenzsynthese und letztlich auch Beurteilungskriterien entwickelt werden müssen.

2.5 Die Akzeptanz der Evidenzorientierung in der erziehungswissenschaftlichen Forschung

Die bisherigen Ausführungen haben gezeigt, dass der empirischen Bildungsforschung seit einiger Zeit eine Schlüsselrolle für das Bereitstellen von Evidenz für Entscheidungen

im Bildungsbereich (auf der Ebene praktischen wie politischen Handelns) zugesprochen wird. Generell zeichnen sich empirische Zugänge durch eine hohe Realitätsorientierung und ausgeprägten Pragmatismus aus. Das unterscheidet sie etwa von vielen Vertretern einer geisteswissenschaftlich orientierten Pädagogik, die das Konzept der Evidenzorientierung – in Anbetracht der Komplexität von Bildungsprozessen – als viel zu eng gefasst wahrnehmen und ablehnen (z. B. Biesta 2010; Herzog 2010). Da Beiträge aus dem geisteswissenschaftlichen Bereich in hierarchischen Evidenzmodellen praktisch kein Gewicht haben, ist eine Tendenz zur Ablehnung des Ansatzes nachvollziehbar. Zum Teil werden auch ethische Bedenken gegen Vorstellungen experimenteller Herangehensweisen wie auch gegen das Modell einer Begründung pädagogischen Entscheidens und Handelns durch Evidenz vorgebracht.

Die empirisch orientierten Wissenschaftlerinnen und Wissenschaftler hingegen teilen das Selbstverständnis, mit ihrer empirischen Forschung zur Lösung von Problemen im Bildungsbereich beizutragen. Diese Beiträge können allerdings unterschiedlich gedacht sein. Zum Teil zielen sie darauf ab, Grundlagenwissen bereitzustellen, das auf längere Sicht ein anderes Verständnis von Bildungsprozessen erlaubt. Wichtige Beiträge werden von anderen darüber definiert, dass sie Probleme im Bildungsbereich identifizieren. Forschungsorientierungen unterscheiden sich auch oft darin, ob eingegrenzte und „bedingte" Theorien geprüft werden sollen oder stärker generalisierbare Erklärungsmodelle. Ähnliches gilt für das angestrebte Veränderungswissen, das als breiteres Handlungswissen oder spezieller technologisches Wissen verstanden werden kann. Unstrittig dürfte die Vorstellung sein, dass die empirische Bildungsforschung auf Erkenntnisse zielt, die sich kurz- oder längerfristig als nützlich erweisen. Ob sich die Forscherinnen und Forscher jedoch zu großen Teilen explizit mit einer Zulieferungsrolle für evidenzbasierte Entscheidungen identifizieren, kann bezweifelt werden. Im Vordergrund steht für viele eine epistemische Orientierung. Wenn die Erkenntnisse nützliche Evidenz für Entscheidungen beisteuern, ist das erfreulich, aber nicht das intrinsische Motiv für die Forschung. Aus diesen Gründen ist die primäre Öffentlichkeit, die von Wissenschaftlerinnen und Wissenschaftlern – auch in der empirischen Bildungsforschung – adressiert wird, die Scientific Community. Auf diese Kollegenkreise sind die Forschungsberichte zugeschnitten.

2.6 Funktionen wissenschaftlicher Evidenz für Bildungspolitik und Bildungsadministration

Aus der Sicht der Bildungsadministration und Bildungspolitik wie auch im öffentlichen Diskurs zu Bildungsfragen erfüllt wissenschaftliche Evidenz sehr unterschiedliche Funktionen. Anknüpfend an die oben beschriebene und letztlich wissenschaftslogisch begründete Unterscheidung von Wissensarten, die die empirische Bildungsforschung liefern kann, soll eine Unterscheidung solcher Funktionen eingeführt werden. Dabei geht es darum, wofür die wissenschaftliche Evidenz (Evidenzfunktionen) eigentlich gebraucht wird. Aus der Sicht der Bildungsadministration und Bildungspolitik entsprechen den Unterscheidungen der Evidenzfunktionen typische Erkenntnisinteressen. Um die Probleme der Rezeption empirischer Bildungsforschung zu verstehen, ist es notwendig, diese analytisch zu differenzieren.

Mit Blick auf Anforderungen einer evidenzbasierten Steuerung in der Bildungsadministration und Bildungspolitik können diese Funktionen mit exemplarischen Fragestellungen erläutert werden. Wir unterscheiden zwischen Beschreibungsfunktion, Erklärungsfunktion, Veränderungsfunktion und Evaluationsfunktion. Die voranstehend unterschiedenen Wissensarten (Beschreiben und Erklären) sind für diese Funktionen in unterschiedlichem Maße passend. Wenn die Bildungspolitik danach fragt, in welchem „Zustand" sich ein Ausschnitt des Bildungssystems (z. B. die Schule) befindet, so kann man diese Zustandsbeschreibungen in den meisten Fällen nicht ohne Annahmen darüber erfüllen, wie der jeweilige Ausschnitt des Bildungssystems funktioniert. Man benötigt also wenigstens Elemente von Kausalwissen, um zu bestimmen, was eigentlich beschrieben werden soll. Noch deutlicher wird das, wenn es darum geht, eine bestimmte Maßnahme, die (Veränderungsfunktion) entwickelt und dann umgesetzt wurde, zu evaluieren (Evaluationsfunktion). Dann will die Bildungspolitik nicht nur wissen, ob etwas funktioniert, sondern vor allem dann, wenn es nur partiell funktioniert, warum Ergebnisse nicht wie erwartet eintreten.

Beschreibungsfunktion. Die Bildungspolitik will wissen, in welchem „Zustand" sich ein Ausschnitt des Bildungssystems (z. B. die Schule) befindet. Allerdings ist bereits die Beschreibung oft nicht so einfach zu liefern. So kann sich herausstellen, dass sich die in der Administration (und auch der Forschung!) zu einem Zeitpunkt verfügbare Datenlage als völlig unzureichend erweist. Die Frage nach Bildungsergebnissen zum Beispiel wurde in Deutschland erst dann virulent, als durch internationale Vergleichsstudien Hinweise auf problematische Resultate nicht mehr zu übersehen waren. In der Folge hat sich die Aufmerksamkeit in der letzten Dekade zunehmend auf ein Wissen über erreichte Bildungsergebnisse verschoben. Zugleich wurden aber (neue, andere, verstärkte) Fragen zu Rahmenbedingungen, Voraussetzungen und Prozessen schulischer Bildung gestellt, die mit der gegebenen Datenlage ebenfalls noch nicht einmal auf der Ebene von Beschreibungswissen zufriedenstellend beantwortet werden konnten. Inzwischen wurden Anstrengungen erhöht, durch ein systematisches Bildungsmonitoring bessere Grunddaten zu erhalten und anhand wichtiger Indikatoren die Lage in bestimmten Zeitabständen abschätzen zu können. Allerdings kann immer die Frage gestellt werden, ob der Kranz an Indikatoren für ein Monitoring richtig ausgewählt und zusammengesetzt ist und ob die Balance zwischen Domänen, kognitiven und nicht kognitiven Merkmalen stimmt. Es ist also naheliegend anzunehmen, dass die Beschreibungsfunktion nur mit Beschreibungswissen erfüllt wird. Das ist jedoch nicht der Fall.

Bereits um Entscheidungen für eine Auswahl von Indikatoren für ein Monitoring treffen zu können, kommt Erklärungswissen (mindestens auf der Ebene theoretischer Modelle) ins Spiel. Dieses hilft abzuschätzen, wie Merkmale zusammenhängen und welche Nebenwirkungen und Folgen bestimmte Zustände haben können. Für die Beurteilung von Folgen und Nebenwirkungen braucht es zudem Wissen über Ziele, das nicht bei allgemeinen und abstrakten Statements über Bildungsziele stehen bleiben darf, sondern sich in Indikatoren übersetzen lässt. Je nachdem müssen mit erheblichem Aufwand Verfahren entwickelt werden, mit denen Facetten von Bildungsergebnissen oder Bildungsprozessen erfasst werden können, die relevant sind aber bisher nicht im Blickpunkt standen. Doch

auch aus der Forschung können Impulse resultieren, bestimmte neue Merkmale in einem Monitoring dauerhaft zu beobachten.

Erklärungsfunktion. Wenn die empirische Bildungsforschung das Beschreibungswissen liefert, das auf Probleme im Bildungsbereich aufmerksam macht, so wirft das schnell die Frage nach Erklärungen auf, nach Einflussfaktoren, die möglichst als Ursachen identifiziert werden können. Das Erklärungswissen liefert in der Bildungsforschung häufig die besonders spannenden theoretischen Ansätze; es ist aber auch in der Bildungsadministration und Bildungspolitik von hohem Interesse. Allerdings wird diese Funktion dort oft weniger durch theoretisches Erklärungswissen geliefert, als vielmehr durch Erklärungen, die leicht an die Öffentlichkeit zu kommunizieren sind und bei denen personalisiert werden kann, wer gegebenenfalls für Problemlagen Verantwortung trägt und wo denkbare Ansatzpunkte für eine Verbesserung liegen.

Veränderungsfunktion. Erklärungswissen kann tatsächlich der Schlüssel für Veränderungen sein. Das ist dann der Fall, wenn kausal relevante Faktoren identifiziert werden, die durch verfügbare Maßnahmen gestaltet werden können. Hilfreich ist hier eine Differenzierung von Handlungsebenen im Bildungssystem (vgl. Krapp 1979). Politische und administrative Maßnahmen können direkt die Struktur des Bildungssystems betreffen, sie können aber auch Rahmenbedingungen für Handlungen auf anderen Ebenen (Schulaufsicht, Schulleitung, Lehrkräfte) gestalten und auf diese Weise mittelbar wirken (z. B. durch Ausstattung, Curricula, Schulorganisation, Qualitätssicherung). Die empirische Bildungsfunktion kann die Funktion der Entwicklung von Maßnahmen erfüllen, weil sie durch das von ihr produzierte Erklärungswissen Ansatzpunkte für Veränderungen identifizieren kann und weil ihr Beschreibungswissen die Rahmenbedingungen von Interventionen aller Art erkennen lässt. Im Bildungsbereich hängen die Erfolgsaussichten fast aller Maßnahmen davon ab, ob bestimmte Randbedingungen und Voraussetzungen gegeben sind. Besonderes Gewicht erhält zudem empirisch gesichertes Wissen über die Implementation und Dissemination von Veränderungen im Bildungsbereich (Gräsel 2010; Jäger 2004; Prenzel 2010). Hier geht es um die Frage, wie die Veränderungsfunktion verbreitet und im Sinne von Interventionen praktisch umgesetzt wird.

Evaluationsfunktion. Interventionen aller Art geschehen unter Unsicherheit. Die realen Bedingungen des Bildungssystems unterscheiden sich in verschiedener Hinsicht von denen, die in der Forschung (und insbesondere in experimentellen Settings) untersucht wurden. Vor allem sind Interventionen mit Kosten (im wörtlichen und im übertragenen Sinne) verbunden. Deshalb verlangt die Bildungspolitik und Administration, dass die empirische Bildungsforschung auch eine Evaluationsfunktion erfüllt und die Wirkungen wie auch Folgen von Interventionen überprüft. Diese Funktion kann direkt durch Evaluationsstudien erfüllt werden, sie wird im Bildungssystem de facto aber auch häufig durch die Lieferung von eher beschreibenden Daten (z. B. dem kontinuierlichen Bildungsmonitoring) erfüllt. Es kann ja durchaus erforderlich sein, auf Problemlagen zu reagieren und Maßnahmen zu ergreifen, obwohl zu diesem Zeitpunkt noch keine wirklich überzeugende Evidenz für ein Maßnahmenpaket vorliegt. Beschreibungswissen in einem Feedbacksystem kann somit (unter Bedingungen eines unzureichenden Erklärungswissens oder bei

Ungewissheit über wirklich erfolgreiche Maßnahmen) einen wichtigen Beitrag zu einer evidenzbasierten Steuerung leisten.

3 Zwischenfazit

Die bisherigen Ausführungen haben Begriffe, Begründungen und Vorstellungen einer Evidenzbasierung im Bildungsbereich angesprochen und skizziert. Dabei wurde erkennbar, dass die empirische Bildungsforschung prinzipiell wichtige Beiträge für eine evidenzbasierte Steuerung in Bildungspolitik und Bildungsadministration beisteuern kann. Ein Schlüsselproblem besteht in der „Übersetzung" von Befunden in Evidenz: Dies ist ein Prozess der Forschungssynthese, der zugleich eine Bewertung der Belastbarkeit unterschiedlicher Arten von Forschungsergebnissen umfasst. Wir sprechen also von evaluativer Forschungssynthese, um hervorzuheben, dass es nicht bloß um eine Befundaggregation gehen kann. Die Bewertung von Forschungsergebnissen muss dann je nach bildungspolitischer Fragestellung und je nach Art des erzeugten Wissens unterschiedlich erfolgen. So sind z. B. Fallstudien nicht schon deshalb auszusortieren, weil sie nicht dem „Goldstandard" der „randomized clinical trials" genügen können. Es kommt vielmehr darauf an, wie man die Erkenntnisse aus unterschiedlichen Studien und Forschungszugängen im Rahmen einer evaluativen Forschungssynthese nutzt. Wir haben ebenfalls deutlich gemacht, dass es zwar möglich ist (Beispiel Campell und Cochrane-Ansätze), die Forschungssynthese ihrerseits methodisch elaboriert vorzunehmen, dass hier jedoch auch ein ganz eigenständiger Forschungs- und Entwicklungsbedarf zum Thema evaluative Forschungssynthesen besteht. Weder die normativen Standards (z. B. randomized clinical trials) noch die forschungspraktischen Möglichkeiten der Disziplinen, in denen diese Standards und Synthesemethoden entwickelt wurden, sind unmittelbar auf die Bildungsforschung zu übertragen. Daraus resultieren jedoch keine unüberwindbaren Probleme, denn die Evidenz, die die empirische Bildungsforschung produziert, kann unterschiedliche Funktionen erfüllen und in diesem Kontext auch an spezifischen Standards beurteilt werden. So kann z. B. eine experimentelle Studie zu Unterrichtsmethoden im Kontext einer Modellschule rigorosen Standards der internen Validität genügen. Sie muss deshalb nicht notwendigerweise auch repräsentative Befunde (gültig für alle Schulen des gleichen Schultyps) liefern, sondern kann als Modell auch dann bildungspolitisch relevantes Wissen im Sinne der „Veränderungsfunktion" liefern, wenn ihre Befunde nicht unmittelbar übertragbar auf alle Schulen sind. Wichtig ist dann allerdings, dass diese Einschränkungen der externen Validität den beteiligten Akteuren im Bildungsdiskurs bewusst bleiben. Die voranstehenden analytischen Unterscheidungen sollten dazu beitragen.

4 Evidenzbasierte Bildungspolitik erfordert Wissenschaftskommunikation: Beispiele

Eingangs wurde erläutert, dass sich mit der evidenzbasierten Steuerung die Erwartung verbindet, dass die empirische Bildungsforschung wissenschaftliche Grundlagen zu einer rationalen Entscheidungsfindung liefert. Diese Erwartung kann die empirische Bildungs-

forschung oft nur zum Teil oder auf längere Sicht erfüllen. Die voranstehende Übersicht zu den Problemen einer evaluativen Forschungssynthese hat bereits gezeigt, dass es auf viele Fragen keine einfachen Antworten geben kann. Manchmal ist dies dem noch mangelhaften Stand der Forschung geschuldet, oft aber auch der prinzipiellen Unabgeschlossenheit wissenschaftlicher Erkenntnisproduktion.

Die angemessene Darstellung und Interpretation der Evidenz bezogen auf Fragen und Entscheidungsbedarfe aus der Sicht von Bildungspolitik und Bildungsadministration erfordert also immer auch eine ihrerseits rationale (und das heißt wiederum: wissenschaftlich begründete) Wissenschaftskommunikation. Dies betrifft die Ergebnisse selbst, es betrifft aber auch die methodischen Grundlagen und die in der Öffentlichkeit vorhandenen Vorkenntnisse und Erwartungen über die Möglichkeiten und die Methoden empirischer Forschung. Die voranstehende Überlegung können wir so zusammenfassen:

Evidenzbasierte Bildungspolitik ist nur möglich als Teil von Wissenschaftskommunikation. Sie setzt die wechselseitige Verständigung zwischen denen, die Bildungsforschung betreiben, denen, die sie interpretieren und denen, die dann an Bildungspolitik und Bildungsadministration beteiligt sind, voraus.

Wir schlagen deshalb vor, Überlegungen und Maßnahmen zur Verbesserung und Systematisierung von evaluativen Forschungssynthesen mit Maßnahmen (Forschungs- und Entwicklungsaufgaben) zur Wissenschaftskommunikation zu verbinden, weil bereits die Erstellung, vor allem aber die Nutzung von evaluativen Forschungssynthesen zur empirischen Bildungsforschung nur immer im Kontext des öffentlichen Diskurses zu Bildungsfragen zu sehen sind.

Bei Wissenschaftskommunikation geht es um die Wechselbeziehung von *Wissenschaft und Öffentlichkeit*. Dabei gibt es natürlich nicht „die Öffentlichkeit"; es gibt nur ganz unterschiedliche „Öffentlichkeiten". Im Zusammenhang mit diesem Beitrag schlagen wir folgende (noch immer vereinfachende) Unterscheidungen von Öffentlichkeiten vor:

- die Akteure in Bildungsadministration und Bildungspolitik,
- die professionellen Akteure im Bildungssystem (insbesondere Lehrkräfte und Personen in Lehrerausbildung und Lehrerfortbildung),
- die Subjekte des Bildungssystems (d. h. Schülerinnen/Schüler, Studierende und ihre Eltern),
- die mediale Öffentlichkeit (d. h. also die Berichterstattung zu Bildungsthemen und zu Wissenschaftsthemen) in Massenmedien (Zeitungen, TV & Radio, Internet),
- die Erziehungswissenschaft sowie die anderen Wissenschaftsdisziplinen (z. B. Fachdidaktiken), soweit sie empirische Bildungsforschung vornehmlich rezipiert und diskutiert und nicht selbst dazu beiträgt (z. B. weil sie andere Aufgaben oder ein anderes methodisches Selbstverständnis hat, siehe oben Teil 2.5).

Der Zusammenhang zwischen den Möglichkeiten – und, wie sich auch zeigen wird, den Grenzen – evaluativer Forschungssynthesen und Wissenschaftskommunikation soll nachfolgend an Hand dreier Fallstudien aufgezeigt werden. Drei (nach bestimmten Gesichtspunkten) ausgewählte Beispiele für die Präsentation, Rezeption und Interpretation von Evidenz aus der empirischen Bildungsforschung werden – gewissermaßen im Sinne einer situierten und multiperspektivischen Herangehensweise – vorgestellt und diskutiert. Die Beispiele illustrieren den derzeitigen Umgang mit Evidenz im Spannungs-

feld der Interessen von Forschung, Politik, Administration und Öffentlichkeit. An diesen konkreten Problemstellungen versuchen wir, den praktischen Nutzen von wissenschaftlicher Evidenz im politischen Entscheidungsprozess zu beleuchten, widerstreitende Interessen in diesem Prozess aufzuzeigen und exemplarisch weiterführende Forschungsfragen zu entwerfen. Die Präsentation der Beispiele ersetzt keine systematischen Reviews zu den jeweiligen Themengebieten, greift jedoch – wo möglich – auf Überblicksarbeiten und Metaanalysen zurück und argumentiert an diesen Beispielen für evidenzgestützte politische Entscheidungsprozesse und eine an Zielgruppen orientierte Publikation von Forschungsergebnissen.

Als erstes Beispiel dient die Diskussion um Gewalt und Computerspiele; ein Zusammenhang, dessen Auswirkungen vor allem in einer politischen Wertediskussion sowie möglichen Regeln und Zugangsbeschränkungen für Computerspiele zum Ausdruck kommen. Diese Diskussion wird sowohl in Fachkreisen als auch in der Öffentlichkeit mit stellenweise großer Entschiedenheit (und nicht immer sachlich begründet) geführt. Innerhalb der Wissenschaft wird der kausale Zusammenhang zwischen Aggressivität und Computerspielnutzung unterschiedlich bewertet. In der öffentlichen Diskussion hingegen steht das Verbot bzw. die Ächtung von Gewaltspielen im Mittelpunkt und produziert politischen Handlungsdruck. Dieses Beispiel repräsentiert somit die Problematik einer Nutzung von Evidenz unter Bedingungen einer nicht eindeutigen Forschungslage und hoher öffentlicher Empfindsamkeit.

Das zweite Beispiel betrifft den Zusammenhang zwischen Klassengröße und Schülerleistung. Hier stehen auf der einen Seite Ergebnisse der Bildungsforschung und Bildungsökonomie, die mit hoher Übereinstimmung keinen systematischen Zusammenhang zwischen kleineren Klassen und besseren Schülerleistungen feststellen. Diese Befunde rechtfertigen also keine pauschalen Maßnahmen einer Reduzierung der Klassengröße zur Verbesserung von Bildungsergebnissen. Auf der anderen Seite fordern Lehrergewerkschaften und Eltern auf der Basis ihrer Wahrnehmungen und Überzeugungen kleinere Klassen, um die Belastung der Lehrkräfte zu verringern und die Betreuung der Kinder zu verbessern. Welche Implikationen hat diese Konstellation für die Präsentation von Ergebnissen, für die öffentliche Diskussion und für politische bzw. administrative Entscheidungen?

Das dritte Beispiel schließlich betrifft Erkenntnisse aus PISA sowie daraus verschiedentlich abgeleitete Schlussfolgerungen für das Schulsystem. Hier betrachten wir unterschiedliche Erkenntnisse z. B. zur Lesekompetenz der Schülerinnen und Schüler sowie zur Diagnosekompetenz von Lehrkräften und zur Gestaltung des Schulsystems. Hier interessiert vor allem die Frage, inwieweit die politischen und schulorganisatorischen Forderungen durch die methodische Anlage der PISA-Studien überhaupt gestützt werden können und inwieweit die Einschränkungen kommuniziert und wahrgenommen werden.

Diese Fallbeispiele illustrieren den Umgang mit Evidenz aus der empirischen Bildungsforschung. Sie machen aber auch auf Kommunikationsprobleme zwischen Wissenschaft und den unterschiedlichen „Öffentlichkeiten“ aufmerksam, die letztlich durch die oben skizzierten Grenzen der Generierung belastbarerer Evidenz (durch originäre Forschung und dann durch evaluative Synthesen) in der empirischen Bildungsforschung bedingt sind.

Im fünften Abschnitt dieses Beitrags werden dann Schlussfolgerungen aus diesen Beispielen zusammengefasst, systematisch analysiert und diskutiert. Im Kern geht es darum zu klären, welche Forschungs- und Entwicklungsmaßnahmen zur Verbesserung von evaluativen Forschungssynthesen und zur Wissenschaftskommunikation durchgeführt werden können.

4.1 Evidenz für oder gegen ein Verbot gewalthaltiger Computerspiele

In Deutschland wurden insbesondere die Gewalttaten in Erfurt im Jahr 2002, Emsdetten im Jahr 2006, Winnenden oder Ansbach im Jahr 2009 zum Anlass genommen, den Einfluss gewalthaltiger Computerspiele auf Kinder und Jugendliche in Öffentlichkeit und Politik zu diskutieren. In allen Fällen war über die Gewalttäter in Erfahrung gebracht worden, dass sie einen erheblichen Teil ihrer Freizeit mit gewalthaltigen Computerspielen verbracht hatten. In gewisser Weise wurde diese Gemeinsamkeit als fallbasierte Evidenz für das Gefährdungspotential von gewalthaltigen Computerspielen verstanden. Wenn die Nutzung von gewalthaltigen Computerspielen tatsächlich für gewalttätiges Handeln in der Realität kausal relevant ist, liegt es nahe, den Zugang zu solchen Computerspielen zu verhindern.

Dieses Beispiel der Wirkung von gewalthaltigen Computerspielen steht sicher inhaltlich nicht im Zentrum der empirischen Bildungsforschung, obwohl es auch mit Lernen und nicht gelungenen Sozialisationsprozessen zu tun hat. Obwohl die Forschung zur Mediennutzung ja nicht das Ziel hat, die individuelle Entwicklung in Richtung unerklärlicher Gewalttaten von jungen Menschen aufzuklären, wird sie in der öffentlichen Rezeption doch zu diesem Zweck herangezogen. Ein beträchtlicher Teil der Forschung in diesem Bereich bemüht sich um die Aufdeckung mehr oder weniger grundlegender Erkenntnisse und weniger darum, Evidenz für aktuelle und relevante gesellschaftliche, institutionelle oder politische Entscheidungen zu erarbeiten.

Eine Option, die in den Medien und in der Politik seit Jahren leidenschaftlich diskutiert wird, ist das Verbot von gewalthaltigen Computerspielen. Hier handelt es sich also um ein klares Beispiel einer auf politischer Ebene zu treffenden Entscheidung. Entsprechend wurde in den öffentlichen Debatten, die auf die dramatischen Gewalttaten folgten, politisch Stellung bezogen. Im Folgenden sollen politische Positionen in dieser Debatte skizziert werden, ohne dabei die Äußerungen einzelner Persönlichkeiten aus der Politik zu zitieren.

4.1.1 Die öffentliche Debatte

Vertreter, die sich *für ein Verbot gewalthaltiger Computerspiele* aussprechen, unterstellen einen kausalen Zusammenhang zwischen der Nutzung solcher Computerspiele und der Tendenz, im realen Leben Gewalt auszuüben. Meistens wird folgendermaßen argumentiert: Gewalthaltige Computerspiele, die in den Medien dann gerne auch als „Killerspiele" bezeichnet werden, lassen Gewalt als selbstverständlich beziehungsweise das Morden von Menschen als etwas Lustvolles erscheinen. Realität und Fiktion verschwimmen, die Schwelle zur Gewalt wird abgesenkt. Es besteht die Gefahr, dass insbesondere Kinder und Jugendliche die in diesen Computerspielen geforderten Aktivitäten im Alltag nachvollziehen wollen. Neben diesen Argumenten, die auf Evidenz geprüft werden können,

werden auch Wertpositionen geltend gemacht. Sie betreffen die Frage, ob die Darstellung von exzessiver Gewalt in Medienprodukten und Spielen moralisch verantwortet werden kann. Verwiesen wird zudem auf wirtschaftliche Interessen einer Unterhaltungsindustrie, die mit gewalthaltigen Inhalten zum Kauf und Konsum anregt.

In den Positionen, die sich *gegen ein Verbot gewalthaltiger Computerspiele* aussprechen, findet man andere Argumente. Hier wird betont, dass ein Zusammenhang zwischen dem spielerischen Ausüben virtueller Gewalt und der Gewaltausübung in der Realität bisher nicht belegt ist. Weiter wird darauf gesagt, dass Verbote die Befassung mit gewalthaltigen Inhalten nicht verhindern könnten. Vielmehr wird empfohlen, die Medienkompetenz (und damit das Nutzungsverhalten) von Kindern und Jugendlichen zu stärken und damit die Auswahl und die Art der Nutzung von Computerspielen zu beeinflussen. Auch Gegner eines Verbots verweisen auf Wertpositionen, etwa auf die Kunstfreiheit, die für die Gestaltung von Computerspielen und für das Ausüben kultureller Praktiken geltend gemacht werden kann (Bisky 2008; Otto 2008).

Die Debatte in der Öffentlichkeit über die Verbotsfrage wird – bedingt durch Interview- und Talk-Show-Formate – schlagwortartig und zugespitzt geführt. Mehrdeutige bildhafte Bezeichnungen („Killerspiel“) dienen als rhetorisches Mittel, es werden Betroffenheit und emotionales Commitment ausgedrückt sowie grundsätzliche Wertpositionen artikuliert. Auf konkrete Evidenz wird in solchen Zusammenhängen sehr selten verwiesen. Gelegentlich werden Personen zitiert, die als Experten für ein breites Fragenspektrum Bekanntheit erlangt haben. In der Argumentation für oder gegen ein Verbot werden vorwiegend Kasuistiken und Plausibilitäten bemüht. Verweise auf den Forschungsstand erfolgen allenfalls pauschal, etwa dergestalt, dass der Zusammenhang zwischen Nutzung von gewalthaltigen Computerspielen und Gewaltbereitschaft nicht belegt sei.

4.1.2 Der Stand der Forschung

In der öffentlichen Debatte über Computerspiele und Gewalt im Zusammenhang mit der Frage eines Verbots spielt wissenschaftliche Evidenz keine zentrale Rolle. Das ist im Kontext der Evidenzthematik deshalb erstaunlich, weil es zum Thema „Computerspiele und Gewalt“ einen durchaus beachtenswerten Forschungsstand inklusive Überblicksarbeiten und Metaanalysen gibt.

Unter anderem wurde zum Beispiel eine Überblicksarbeit von Kunczik und Zipfel im Auftrag des Bundesministeriums für Familie, Senioren, Frauen und Jugend erstellt (Kunczik und Zipfel 2005), die das Thema differenziert aus unterschiedlichen Perspektiven betrachtet und eine vorsichtige Entwarnung ausspricht. Zu einer deutlich kritischeren Bewertung gelangen die Autoren einer Metaanalyse, die im Psychological Bulletin erschien und Anlass zu einer lebhaften Fachdiskussion gab (Metaanalyse von Anderson et al. 2010; Kommentare von Bushman et al. 2010; Ferguson und Kilburn 2010; Huesmann 2010). Zu einem kritischen Gesamtergebnis kommen auch Greitemeyer und Mügge (2014) in ihrer Metaanalyse, die sich auch mit dem Verhältnis von aggressionsfördernden und prosozialen Elementen in Spielen befassen.

Das Forscherteam um Craig A. Anderson gelangt auf der Basis einer Metaanalyse zu einem klaren Fazit: Es gibt klare Evidenz für einen ursächlichen Zusammenhang zwischen gewalthaltigen Computerspielen und vermehrten aggressiven Gedanken, Gefühlen

und vermehrtem aggressivem Verhalten. Gleichzeitig reduziert der Umgang mit gewalthaltigen Computerspielen die Empathiefähigkeit sowie prosoziales Verhalten (Anderson et al. 2010). Studien, die von den Autoren als methodisch besser konzipiert betrachtet werden, berichten auch die deutlicheren Effekte. Eine Verzerrung der Aussagen durch die Bevorzugung von in renommierten Zeitschriften veröffentlichten Arbeiten wird durch die Autoren nicht festgestellt. Diese Ergebnisse stimmen gut mit früheren (Überblicks-) Arbeiten überein, die von Craig A. Anderson und seinen Kolleginnen und Kollegen im Verlauf der vergangenen zehn Jahre publiziert wurden.

Ist damit – wie es in einem Kommentar (Huesmann 2010) zur publizierten Studie heißt – jeglicher Zweifel über die Schädlichkeit von gewalthaltigen Computerspielen entkräftet?

Dies ist wohl noch nicht der Fall, denn ein kritischer Kommentar zu dieser Metaanalyse (Ferguson und Kilburn 2010) sowie einige weitere Arbeiten dieser Forschergruppe (u. a. auch metaanalytische Arbeiten Ferguson 2007a, b, 2009; Ferguson und Rueda 2010) weisen darauf hin, dass die in der Metaanalyse präsentierte Evidenz nicht ohne Probleme ist. Auch die Antwort der Autoren (Bushman et al. 2010) auf diese Kritik kann nicht alle angesprochenen Punkte entkräften.

So wird diese Debatte auch aktuell noch weiter geführt. Ein aktueller Überblicksartikel fasst den Forschungsstand eher im Sinne einer Entwarnung zusammen (Elson und Ferguson 2014b), erfährt aber gleichzeitig deutliche Kritik (Bushman und Huesmann 2014; Krahé 2014; Warburton 2014). Auch bei diesem erneuten Disput zum Thema bleiben Fragen offen. Sowohl die warnenden Stimmen als auch diejenigen, die die Auswirkungen gewalthaltiger Computerspiele deutlich harmloser einschätzen, bleiben bei ihren Argumentationslinien und interpretieren Forschungsergebnisse unterschiedlich (Elson und Ferguson 2014a).

Die Diskussion über die widersprüchlichen Ergebnisse der beiden Metaanalysen dreht sich letztlich um methodische Fragen. Sie betreffen vor allem die Qualität der ausgewählten Studien, die Genauigkeit der Operationalisierung von Gewalt sowie die begriffliche Abgrenzung zwischen Aggression (im Sinne einer Gewaltbereitschaft) und (manifester) Gewalt und die Vergleichbarkeit der methodischen Verfahren. Dies führt dazu, dass die Kontroverse ohne fundierte Fachkenntnis kaum mehr zu durchdringen und zu interpretieren ist. Sowohl Fürsprecher als auch Gegner eines Verbots von gewalthaltigen Computerspielen können sich somit bei den verschiedenen wissenschaftlichen Lagern mit Evidenz bedienen, auch wenn ihnen die (begrenzte) Belastbarkeit der Befunde möglicherweise nicht bewusst ist.

4.1.3 Bessere Evidenz durch eine Ausweitung der Forschungsperspektive?

Die zum Forschungsstand berichteten Zusammenhänge stammen aus Fragebogenstudien mit Selbstauskünften oder (quasi-)experimentellen Studien, die zum Teil auch Reaktionsmessungen verwendeten. Viele der Studien gelangen jedoch an Grenzen, wenn genauere Aussagen darüber getroffen werden sollen, wie das Nutzen gewalthaltiger Computerspiele das Verhalten im Alltag beeinflusst. Für die Einschätzung der kausalen Relevanz ist ein Wissen über die Prozessabläufe wichtig. Die Beobachtung von Kunczik und Zipfel (2005) scheint auch heute noch zu gelten: „Die bisherige Forschung hat sich stärker auf

Wirkungseffekte denn auf ihr Zustandekommen konzentriert. Im Forschungsdesign wäre eine bessere Berücksichtigung der Spezifika von Computerspielen wünschenswert (z. B. Interaktivität). Darüber hinaus sollte die genauere Untersuchung der Einflussvariablen sowie der Risikogruppenansatz im Mittelpunkt weiterer Forschungsbemühungen stehen" (Kunczik und Zipfel 2005, S. 238).

Insbesondere zeichnen sich in der vorliegenden Forschung zu diesem Thema folgende Schwachpunkte und Desiderata ab:

- Das „Allgemeine Aggressionsmodell", das in der Medienwirkungsforschung insbesondere im Zusammenhang mit gewalthaltigen Filmen als theoretisches Modell angenommen wurde (Generalized Aggression Model GAM, erklärt bei Kunczik und Zipfel 2005, S. 111 ff.), postuliert eine kausale Wirkrichtung (Verstärkung von Präferenzen durch Mediennutzung). Ebenso plausibel wäre eine Umkehr dieser Wirkrichtung (Selektion durch individuelle Präferenzen), wobei das konkurrierende theoretische Modell (Ferguson 2008) bislang deutlich weniger Aufmerksamkeit in der Forschungslandschaft erhält.
- Die Katharsis-Hypothese ist für passive Mediennutzung klar widerlegt. Möglicherweise gilt diese aber für die aktive Beschäftigung mit Computerspielen (Ferguson und Rueda 2010; Kunczik und Zipfel 2005, S. 210).
- Die Habitualisierung, d. h. eine Abstumpfung gegenüber Gewalt, kann nicht eindeutig belegt werden. Problematisch ist hier aber vor allem, dass das Konzept der Messung des Gewöhnungseffekts nicht klar definiert ist.
- Der Transfer zwischen virtueller Gewalt und realer Gewalt ist ebenfalls offen. Auch hier sind die Methoden der Erfassung und damit die Ergebnisse sehr unterschiedlich.
- Die Kultivierung, d. h. langfristige Übernahme und Internalisierung von Verhaltensweisen, kann für Computerspiele bislang nicht nachgewiesen werden. Auch hier fehlen aussagekräftige Studien (Kunczik und Zipfel 2005, S. 222).
- Schließlich wird die wissenschaftliche Debatte nicht mehr alleine auf der Basis der Evidenz, sondern auch vor dem Hintergrund individuell unterschiedlicher normativer Positionen und Überzeugungen geführt, was die Appelle zur Versachlichung der Debatte von unterschiedlichen Beteiligten an dieser Diskussion (Elson und Ferguson 2014a; Krahé 2014) deutlich machen.

Folgt man den angesprochenen Reviews und Meta-Analysen, dann findet man einerseits Evidenz *für* und andererseits *gegen* die Annahme, dass gewalthaltige Computerspiele sich auf die Gewaltbereitschaft auswirken. Dennoch lässt der (insgesamt kontroverse) Forschungsstand einige stabile Ergebnisse erkennen, auf die aufgebaut werden kann (vgl. auch Kunczik und Zipfel 2005):

- Gewalthaltige Computerspiele können kurzfristige aggressive Gedanken und Haltungen auslösen. Langfristige Effekte sind ebenfalls nicht ausgeschlossen. Maßgeblich für die Effekte sind Umfeld- und Personenvariablen.
- Gewalthaltige Computerspiele können motivieren und das Selbstwertgefühl steigern.
- Gewalthaltige Computerspiele können (wie andere Spiele auch) Stresserleben reduzieren und damit positive Auswirkungen auf den gesamten Organismus haben.

- Die Katharsis-Hypothese, nach der Medienkonsum ein Ventil für Aggressionen bilden und damit entlastend wirken soll, findet keine Unterstützung. Nicht eindeutig geklärt ist, ob die aktive Beschäftigung im Computerspiel kathartische Wirkung entfalten kann.
- Personen, die insgesamt aggressiver sind, spielen auch eher Gewaltspiele. Der Umkehrschluss gilt jedoch nicht.
- Es gibt zahlreiche weitere Einflussfaktoren, die aggressive Gedanken und Handlungen begünstigen. Dazu zählen neben einer persönlichen Disposition äußere Faktoren wie z. B. Verhalten der Eltern, Freundeskreis und kulturelle Hintergründe.

Insgesamt muss nach derzeitigem Stand der Forschung davon ausgegangen werden, dass es Risikofaktoren für die schädliche Wirkung von gewalthaltigen Computerspielen gibt, die erstens in der Person selbst, also deren Veranlagung und Lerngeschichte, begründet sind, zweitens durch die soziale Situation beeinflusst werden und die drittens durch die Darstellungsform und Interaktionsmöglichkeit des Spiels befördert werden. Gewalthaltige Computerspiele dürften somit individuell unterschiedlich (problematisch) wirken.

Zweifellos bleiben aber beim aktuellen Stand noch viele Forschungsfragen offen, die wissenschaftlich wie auch praktisch und politisch hoch interessant und wichtig sind, zum Beispiel:

- Wie kann die Anreizstärke unterschiedlicher Gewaltspiele standardisiert gemessen werden? Welches Maß eignet sich zur standardisierten Erfassung der Effekte?
- Wie können Risikofaktoren zuverlässig identifiziert werden? Wie wirken diese Risikofaktoren auf unterschiedliche Personen?
- Wie können statistische Analysen aus der Soziologie und der Kriminologie helfen, die Effekte besser einzuordnen?
- Welche Schutzmechanismen gegen die möglichen schädlichen Effekte sind wirksam? Warum wirken diese Schutzmechanismen? Was ist ein Maß für Medienkompetenz?
- Wie wirksam sind Trainings zur Medienkompetenz (z. B. Möller und Krahé 2013)?
- Wie können die positiven Effekte der Computerspiele genutzt werden? Welche weiteren positiven Effekte neben der verbesserten visuellen Wahrnehmung und Orientierung sowie der möglichen Motivations- und Selbstwirksamkeitssteigerung gibt es?
- Könnten prosoziale Spielinhalte die negativen Effekte der aggressiven Inhalte kompensieren oder sogar übertreffen (Greitemeyer und Mügge 2014).

Schließlich sind weitere Metaanalysen sinnvoll, um innerhalb dieses Forschungsgebiets Evidenz zusammenzuführen und zu klären. Freilich sind Metaanalysen und systematische Reviews hochgradig abhängig von der konzeptionellen und methodischen Qualität der Primärstudien und deren Reichweite.

4.1.4 Die Wirkung gewalthaltiger Computerspiele: Enttäuschte Erwartungen

Das Beispiel der Forschung zur Wirkung gewalthaltiger Computerspiele steht für die Vorläufigkeit wissenschaftlicher Evidenz, die den Normalfall darstellt und als Herausforderung für die Kommunikation zwischen Wissenschaft und Öffentlichkeit betrachtet werden muss. Gewalthaltige Computerspiele und ihre möglichen Auswirkungen auf

aggressives Verhalten sind ein gutes Beispiel dafür, dass wissenschaftliche Forschung einerseits durchaus belastbare Ergebnisse erbringen kann und dennoch gleichzeitig diese Ergebnisse vorläufig, ergänzungswürdig und im Prozess der Forschung auch innerwissenschaftlich kontinuierlich umstritten bleiben. Dieser vorläufige Charakter besteht grundsätzlich, er kann nicht einfach durch Verweis auf die methodische Sorgfalt von Untersuchungen aufgelöst werden, obwohl Unterschiede in den methodischen Zugängen vermutlich für die Unterschiede in den Befunden verantwortlich sind. Mit Hilfe der oben angeführten Funktionen von wissenschaftlicher Evidenz für die (Bildungs)-Politik kann man festhalten, dass es durchaus Kausalwissen und auch Beschreibungswissen zum Zusammenhang von Spieleerfahrungen und bestimmten Erlebens- und Verhaltensweisen gibt, die ihrerseits aggressionsrelevant sind. Weil jedoch Amokläufe (glücklicherweise) seltene Ereignisse sind, und weil diese auf multiple Bedingungen in der Person des Täters zurückgehen, kann man aus diesem Kausalwissen keine *Veränderungsfunktion* ableiten. In anderen Worten: Die in Metaanalysen gezeigten Kausalstärken zwischen Spielen, Gewöhnung an Gewalt und dann tatsächlichem extremem aggressivem Verhalten sind zu gering, um daraus ein Verbot solcher Spiele zwingend ableiten zu können. Wie oben ausgeführt ist durch weitere Forschung zu erwarten, dass die genannten Zusammenhänge noch weiter aufgeklärt werden können. Möglicherweise wird auch zukünftige Forschung keine stärkeren Zusammenhänge zwischen derartiger Spielerfahrung und extremen Verhaltensweisen wie Amokläufen aufzeigen können. Sie kann aber möglicherweise Zusammenhänge mit weniger extremen, jedoch häufigeren und ebenfalls problematischen Verhaltens- und Erlebensweisen im alltäglichen Umgang mit Aggression herausarbeiten, aus denen dann wiederum auch veränderungsbezogene Vorschläge abgeleitet und empirisch überprüft werden können. Die oben angesprochenen Überlegungen in Bezug auf medienpädagogische Interventionen gehen in diese Richtung.

Es ist jedoch durchaus denkbar, dass „die Öffentlichkeit“ (also z. B. die Bildungspolitik) das in dieser Übersicht skizzierte bereits verfügbare Kausalwissen als Heuristik für normative Diskussionen nimmt. Man kann aus dem skizzierten Zusammenhang zwischen Spielerfahrungen und gewissen Gewöhnungseffekten die normative Schlussfolgerung ziehen, dass es gesellschaftlich wünschenswert ist, die öffentliche Erlebbarkeit von Gewalt zu reduzieren. Es wäre nicht gerechtfertigt, dies als vermeintliches Veränderungswissen aus den genannten Ergebnissen abzuleiten, aber die theoretischen Annahmen, die dem Kausalwissen zu Grunde liegen, können als Heuristik für normative Diskussionen (was ist moralisch, was ist politisch gewollt?) dienen. Betrachtet man die Diskurse zu dieser Thematik in den unterschiedlichen Öffentlichkeiten, so finden sich durchaus Beispiele für eine solche heuristische Nutzung. Aus der derzeitigen Forschungslage kann wenigsten ein Schluss zu solchen heuristischen Nutzungen gezogen werden: Die vorliegende wissenschaftliche Evidenz reicht offensichtlich nicht aus, um von vornherein eine politische Position (für oder gegen ein Verbot) als unbegründet auszuschließen. Diese (Meta-) Erkenntnis ist durchaus relevant. In anderen Worten: Sie macht den normativen (politischen) Gehalt von Verbotsforderungen und Verbotsablehnungen deutlich und das kann zur Transparenz und Versachlichung von Debatten beitragen.

4.2 Kleinere Schulklassen – bessere Schülerleistungen?

Bei dem zweiten Fallbeispiel handelt es sich um eine Problemstellung, die weniger dramatisch ist, aber viele Gemüter erhitzt und ebenfalls immer wieder Gegenstand öffentlicher und bildungspolitischer Debatten ist. Es geht um die Frage, welche Rolle die Klassengröße für die Qualität des Lernens im Unterricht spielt. Die Klassengrößen variieren in Deutschland je nach Schulart, Schulstufe und Einzelschule (Statistische Ämter des Bundes und der Länder 2013, S. 110).

Die Klassengrößen in Deutschland unterscheiden sich nicht von internationalen Kennwerten wie etwa dem OECD-Durchschnitt (z. B. Sälzer et al. 2013a). Sie sind im Verlauf der letzten Jahrzehnte gesunken. Dennoch wird von verschiedenen Seiten immer wieder die Erwartung formuliert, dass kleinere Klassen eine bessere Betreuung, Unterstützung und Förderung der Schülerinnen und Schüler gestatten. Vor diesem Hintergrund ist die Forderung nach Verringerung der Klassengröße und Absenkung des Klassenteilers (des Wertes, ab dem Klassen geteilt werden müssen) ein ständiges Thema in den Bildungsdiskussionen und Gegenstand von Forderungen unterschiedlicher Lehrer- wie Elternverbände und Parteien. Vor dem Hintergrund von Bildungsergebnissen, die nicht den Hoffnungen und Erwartungen entsprechen, wird die Frage der Klassengröße auch als Maßnahme zur Verbesserung des Leistungsstandes ins Gespräch gebracht. Im Hinblick auf politische und ggf. administrative Entscheidungen über die Reduzierung von Klassengrößen besteht die Frage nach der Evidenz für Effekte solcher Maßnahmen. Im Blickpunkt stehen meist die Wirkungen für Lernfortschritte und Leistungsstand der Schülerinnen und Schüler.

4.2.1 Die öffentliche Debatte

Die öffentliche Diskussion über die erforderliche oder wünschenswerte Größe von Schulklassen wird von mehreren Interessengruppen geprägt, die man nach den oben eingeführten „Öffentlichkeiten" unterscheiden kann. Neben politischen Akteuren tragen vor allem Lehrerverbände und Organisationen der Elternvertretung regelmäßig zu der Debatte bei.

Aus der *Perspektive der Eltern* ist der Wunsch nach kleineren Klassen leicht nachzuvollziehen. Die Aufmerksamkeit und Unterstützung, die ihr Kind in der Klasse erfährt, hängt rechnerisch von der Klassengröße ab. Aus der subjektiven Erfahrung von Eltern erscheint es einfacher zu sein, eine kleinere Anzahl von Schülerinnen und Schülern gut und persönlich kennen zu lernen, ihren Lernstand und Förderbedarf festzustellen und sie individuell zu betreuen. In einer kleineren Klasse haben die Schülerinnen und Schüler mehr Zeit, sich einzubringen und in Interaktionen mit der Lehrkraft oder im Gruppenzusammenhang einbezogen zu werden. Außerdem wird vermutet, dass bei kleineren Klassen die Lehrkräfte weniger belastet sind und damit bessere Voraussetzungen für einen pädagogisch qualitätsvollen oder innovativen Unterricht vorliegen. Diese und ähnliche Argumente werden typischerweise von Elternvertretern in öffentlichen und politischen Debatten vorgebracht, um kleinere Klassengrößen zu begründen. Auf wissenschaftliche Evidenz zu dem Themenkreis wird normalerweise nicht Bezug genommen.

Auch aus der *Perspektive von Lehrkräften* ist es ebenfalls offensichtlich, dass kleinere Klassengrößen bessere Lernbedingungen für Schülerinnen und Schüler bedeuten und zu

besseren Lernergebnissen führen. Kleinere Klassen erlauben einen besseren Kontakt zu den Schülerinnen und Schülern und verbessern die Möglichkeit, individuell auf die Schülerinnen und Schüler einzugehen und diese gezielt zu fördern. Die Belastung durch Lärmpegel und Störungen ist geringer; der Arbeitsaufwand (z. B. für Korrekturen) kann eher bewältigt werden. Hinweise auf problematische Ergebnisse und auf erforderliche Verbesserungen des Unterrichts werden oft mit der Klassengröße verbunden beziehungsweise von dieser abhängig gemacht. Die Klassengröße wird aus der Sicht von Lehrerverbänden als eine wichtige, wenn nicht die zentrale Stellschraube für die Qualität von Unterricht und Schule betrachtet. In dieser Wahrnehmung und Argumentation findet man einen breiten Konsens zwischen Lehrerverbänden, unabhängig von deren sonstigen schulpolitischen Orientierungen. Die Klassengröße repräsentiert eine zentrale Bedingung der alltäglichen Berufsausübung aus der Lehrerperspektive; es ist nur konsequent, dass Lehrerverbände dieser Bedingung größte Aufmerksamkeit schenken. Die von Lehrerverbänden vorgebrachte Argumentation stützt sich stark auf generelle und verbreitete (in der Zunft wohl völlig unstrittige) Erfahrungen aus der alltäglichen Praxis. Auf wissenschaftliche Evidenz wird in der Argumentation normalerweise nicht Bezug genommen. Gelegentlich werden jedoch Forschungsergebnisse, die Unterschieden in der Klassengröße keine Effekte zuschreiben, abgelehnt (siehe z. B. Gewerkschaft Erziehung und Wissenschaft 2009). Ihnen wird abgesprochen, die tatsächlichen Verhältnisse in Klassen erfasst zu haben.

Insgesamt ist die Sicht der Lehrkräfte ähnlich wie die der Eltern vor allem durch persönliche Erfahrungen geprägt. Im Zentrum der Begründung für kleinere Klassengrößen steht ein Argument, dass sachlogisch zwingend zu sein scheint: Die Klassengröße determiniert letztlich die Anteile von Zeit, Aufmerksamkeit und Förderung, die den Schülerinnen und Schülern pro Kopf gewidmet werden können. Dieses Argument stützt die tiefe Überzeugung, die mit der alltäglichen Erfahrung zu korrespondieren scheint. Dieses Argument wird auch sehr viel stärker und häufiger bemüht als das Argument der persönlichen Belastung durch Korrekturaufwand, Lärm und Stress. Die Interessenvertretungen zielen in der öffentlichen Debatte darauf ab, hohe Klassenstärken als Problem zu klassifizieren, das sich negativ auf die Qualität des Unterrichts auswirkt.

Aus der *Perspektive der Bildungspolitik* werden die Wünsche und Forderungen der Lehrer- und Elternverbände sehr wohl wahrgenommen. Es wird nachvollzogen, dass die Klassengröße grundsätzlich eine Rolle für die Unterrichtsqualität und für die Belastung der Lehrkräfte spielt. Die letztlich entscheidende Frage ist, wo unter den Restriktionen des Finanzbudgets die durchschnittliche Klassengröße und die Richtlinien für Klassenteilungen kritisch werden. Grundsätzlich lässt sich seit Jahren keine Tendenz beobachten, die Klassengrößen zu erhöhen. Im Gegenteil werden häufig die Absicht und das Bemühen ausgedrückt, die Klassengrößen weiter zu reduzieren, auch um Realisierungsbedingungen und Akzeptanz für anspruchsvollere pädagogische Konzepte zu schaffen. In diesem Zusammenhang wird in den letzten Jahren zunehmend auf Evidenz aus der empirischen Bildungsforschung hingewiesen. Sie wird oft so dargestellt und interpretiert, dass eine Reduzierung der Klassengröße vielleicht das Wohlbefinden von Lehrkräften und Eltern fördert, aber für sich genommen keine bessere Unterrichtsqualität inklusive individueller Förderung und Leistungssteigerung garantiert.

4.2.2 Der Stand der Forschung

Im Unterschied zum ersten Fallbeispiel – den gewalthaltigen Computerspielen – zeichnet sich bei der Frage der Wirkung der Klassengröße eine deutlich größere Übereinstimmung der Befunde ab. Sie besagt, kurz gefasst: Kleinere Klassen bringen nicht automatisch bessere Unterrichtsergebnisse. Allerdings findet man auch zu diesem Thema aus wissenschaftlicher Sicht unterschiedliche methodische Zugänge, Interpretationen und insbesondere Schlussfolgerungen.

Den Zusammenhängen zwischen Klassengröße und Leistungsniveau oder Belastungserleben kann auf unterschiedliche Weise nachgegangen werden (vgl. Arnhold 2005). Im Rahmen von groß angelegten Vergleichsstudien wird analysiert, inwieweit Input-Merkmale (Klassengröße) unter Kontrolle anderer relevanter Bedingen (z. B. Zusammensetzung der Schülerschaft nach Herkunft oder Leistungsniveau) mit den Ergebnissen von Leistungstests kovariieren (z. B. Hanushek 1998). Bei einer differenzierteren Betrachtung werden (im Rahmen von Überblicksstudien oder Ex-post-facto Studien) auch Prozessmerkmale des Unterrichts bei unterschiedlichen Klassengrößen erfasst sowie weitere abhängige Merkmale (z. B. fächerübergreifende Kompetenzen, Belastungserleben der Lehrkräfte). Beim interessantesten (aber sehr aufwändigen) Forschungszugang wird die Klassengröße in einem experimentellen Design systematisch variiert, um dann kurz- und längerfristige Effekte zu messen.

Alle jüngeren internationalen Vergleichsstudien (egal ob IGLU, TIMSS oder PISA) gelangten bisher zu vergleichbaren Befunden: Die Klassengröße korreliert nicht mit der Leistung, die von den Schülerinnen und Schülern in Tests erzielt wird. Tendenziell zeichnet sich sogar eine negative Korrelation ab, die besagt, dass das Leistungsniveau bei großen Klassen besser ist als bei kleineren. Die großen Klassen in leistungsstarken asiatischen Staaten könnten für diesen Effekt verantwortlich sein. Doch findet man auch bei einer Analyse der Daten nur für Deutschland keine Korrelation zwischen Klassengröße und Leistungsniveau.

Seltener untersucht sind Beziehungen zwischen der Klassengröße und der Art des Unterrichts. Insgesamt ergibt sich hier ein weniger klares Bild (Schrader et al. 2001). In manchen Studien finden sich Hinweise darauf, dass in großen Klassen ein effektives Klassenmanagement weniger gut gewährleistet werden kann und deshalb Lernzeit verloren geht (z. B. Hargreaves et al. 1998). Andererseits gibt es auch Hinweise, dass der Unterricht in großen Klassen im Durchschnitt besser und klarer strukturiert ist als in kleinen Klassen (z. B. Helmke und Weinert 1997). Insgesamt verdichten sich jedoch die Hinweise dahin gehend, dass Lehrkräfte die Möglichkeiten kleinerer Klassen bei der Unterrichtsgestaltung oft nicht ausnutzen.

Bezogen auf das Belastungserleben als Lehrkraft wird immer wieder die Klassengröße als belastender Faktor genannt (z. B. Schaarschmidt 2005). Dies gilt zumindest für Lehrerbefragungen, die das Belastungserleben generell erheben und die mit Fragen der Unterrichtsforschung verknüpft sind. Schwache Korrelationen zwischen Klassengröße und subjektivem Belastungserleben (die zum Teil wiederum durch das Alter der Lehrkräfte erklärt werden konnten) fanden Hosenfeld et al. (2002) in der sogenannten MARKUS-Studie.

Eine aktuelle Analyse von IGLU-Daten erhellt die Situation (Lankes und Carstensen 2010), weil dort verschiedene Zusatzdaten zum Unterricht und aus Lehrerbefragungen berücksichtigt werden. In dieser Studie verwenden die Autoren als Inputparameter die Klassengröße und setzen diese in Beziehung zu den Outputvariablen (a) IGLU-Leistungsdaten der Schülerinnen und Schüler und (b) Belastungsempfinden der Lehrkräfte (Fragebogendaten). Aus der Gesamtstichprobe von 357 Klassen wurden in einem weiteren Schritt die Klassen herausgenommen, die aufgrund spezieller Lernbenachteiligungen wie z. B. Lernbehinderung oder mehr als einem Drittel der Kinder mit Sprachschwierigkeiten absichtlich klein gehalten wurden. Damit ergibt sich eine Analysestichprobe von 318 Klassen, deren Klassengröße zwischen 15 und 30 Schülerinnen und Schülern liegt.

Der Einfluss der Klassengröße auf die Lesekompetenz entspricht in der Gesamtstichprobe nicht dem von Lehrkräften und Eltern erwarteten Wert: Je größer die Klassen, desto besser ist die Leseleistung. Zu berücksichtigen ist jedoch, dass sich in der Gesamtstichprobe Klassen befinden, die aufgrund ihrer Zusammensetzung (z. B. größere Anteile von Schülerinnen und Schülern mit besonderem Förderbedarf) mit gutem Grund weniger Schülerinnen und Schüler umfassen. Werden diese Klassen aus den Analysen ausgeschlossen, dann reduziert sich der Zusammenhang zwischen Klassengröße und Leseleistung. Zur Erklärung der Leistungsunterschiede erweisen sich dann Unterrichtsmerkmale, zum Beispiel die Nutzung von Lesestrategien, als sehr viel relevanter.

Zur Frage des Einflusses der Klassengröße auf die Belastung bringen die Analysen von Lankes und Carstensen (2010) ein ähnlich eindeutiges Ergebnis. Die Klassengröße eignet sich nicht als Prädiktor für das subjektive Belastungserleben. Stattdessen treten Merkmale des Schulklimas sowie der Lehrerpersönlichkeit und der beruflichen Historie in den Vordergrund. Lange Berufstätigkeit als Lehrkraft wirkt sich auf die Belastung eher erhöhend aus, gutes Schulklima und gute Fortbildungsmöglichkeiten dagegen eher reduzierend. Außerdem zeigen Lehrkräfte, die Verantwortung für die Lernergebnisse der Schülerinnen und Schüler aktiv übernehmen, geringere Belastung. Wenn dagegen Misserfolge (z. B. durch die Schulleitung) den Lehrkräften angelastet werden, dann erhöht dies die Belastung.

Diese Befunde wären allerdings missverstanden, wenn sie – weil keine Zusammenhänge gefunden wurden – nun zur Begründung von Vergrößerungen der Klassen herangezogen würden. Um dies am Beispiel der IGLU-Studie zu erläutern: Die Daten lassen keine Vorhersagen darüber zu, was passieren würde, wenn alle Klassen um mehrere Schüler vergrößert oder wenn bisher unterdurchschnittlich großen Klassen zusätzliche Schüler zugewiesen würden. Basis der Aussagen von Lankes und Carstensen ist die vorgefundene Unterschiedlichkeit in Klassengrößen. Denn Lehrkräften würden durch solche Maßnahmen tatsächlich zusätzliche Belastungen zugemutet und viele von ihnen hätten vermutlich nicht die Ressourcen, um mit einer so veränderten Situation produktiv umzugehen.

Auf der anderen Seite gibt es Evidenz, dass eine Verringerung der Klassengrößen durchaus positive Effekte haben kann. Das Projekt Student/Teacher Achievement Ratio (STAR), das von 1985 bis 1989 in den USA im Bundesstaat Tennessee lief, gilt als ein bahnbrechendes und mutiges Forschungsvorhaben, da im Schulumfeld in großem Rahmen ein randomisiertes und längsschnittlich angelegtes Experiment durchgeführt wurde. Im Jahr 2008 wurden schließlich die Daten und Dokumente aus diesem Projekt auf einer

Internetseite veröffentlicht, sodass die Überprüfung der Zusammenhänge trotz der inzwischen vergangenen 20 Jahre leicht möglich ist (Achilles et al. 2008).

Im Projekt STAR wurde der Einfluss kleinerer Klassen im Zusammenhang mit zahlreichen weiteren Einflussfaktoren untersucht. Was die Ergebnisse aus der Masse der übrigen Studien heraushebt, ist die methodische Konzeption und Vorgehensweise. In 79 Grundschulen, die sich freiwillig für die Teilnahme an der Studie gemeldet hatten, wurden zufällig Schülerinnen und Schüler auf drei unterschiedliche Bedingungen verteilt: kleine Klassen (13–17 Schülerinnen und Schüler), normale Klassen (22–25 Schülerinnen und Schüler) sowie normale Klassen mit einer zusätzlichen Person zur Unterstützung der Lehrkraft.

Nach vier Jahren endete das Projekt und die Klassen wurden wieder alle auf die normale Größe gebracht. Über die Projektlaufzeit und am Ende der Highschool 1998 wurden Leistungsdaten erhoben. Die Ergebnisse zeigen, dass in den kleineren Klassen höhere Leistungen erzielt werden, und der Leistungsvorsprung nach der Grundschulzeit (und bei „normalen" Klassengrößen) nicht verloren geht. Von den kleinen Klassen profitieren zwei Teilgruppen besonders: Sozial schwächer gestellte sowie schwarze Schülerinnen und Schüler. Sie haben letztlich höhere Chancen ein Studium zu beginnen als Schülerinnen und Schüler, die normale Klassen (mit oder ohne Assistenzkraft) besucht hatten.

Diese methodisch bemerkenswerte Studie liefert starke Hinweise dafür, dass eine deutliche Senkung der Klassengröße (13–17 versus 22–26 Schülerinnen und Schüler pro Klasse) einen Effekt auf den Unterricht und das Lernen haben kann. Freilich ist zu berücksichtigen, dass dieses Experiment für die teilnehmenden Schulen höchst willkommen war, weil zumindest in einem Teil ihrer Klassen deutlich günstigere Bedingungen geschaffen wurden. Natürlich war auch den Lehrkräften bei diesem Design bewusst, ob sie der Experimental- oder der Kontrollbedingung zugeordnet waren. Dadurch wurde möglicherweise der Ehrgeiz stimuliert, die günstigen Bedingungen wirklich produktiv zu nutzen. Bemerkenswert ist der Befund, dass die normalen Klassen mit einer zusätzlichen Vollzeitkraft nicht bessere Ergebnisse erzielten, denn im Ende hatten sie zahlenmäßig die günstigste Lehrer-Schüler-Relation.

Im Zusammenhang mit den hier berichteten anderen Befunden unterstützt diese Studie jedenfalls die Auffassung, dass kleine Klassengrößen als Gelegenheitsstruktur für einen stärker individualisierenden, fördernden und innovativen Unterricht verstanden werden können, die freilich genutzt werden muss.

4.2.3 Bessere Evidenz durch eine Ausweitung der Forschungsperspektive?

In der Grundrichtung konvergieren die Ergebnisse der unterschiedlichen methodischen Ansätze bei der Untersuchung von Zusammenhängen zwischen Klassengröße, Unterricht und Leistung. Umfangreiche Daten zu diesen Fragen liegen aus den großen internationalen und nationalen Vergleichsstudien vor. Allerdings handelt es sich hier eher um ein Nebenprodukt dieser Studien. Die Klassengröße ist ein relativ einfach und zuverlässig zu erhebender Input-Indikator, der aus bildungsökonomischer, pädagogischer und bildungspolitischer Perspektive großes Interesse findet und deshalb bei den Analysen gerne berücksichtigt wird. Diesem Indikator wird auch aus bildungspolitischer Sicht Steue-

rungsrelevanz zugesprochen. Es ist allerdings nicht so, dass die Anlage dieser großen Vergleichsstudien detaillierte Fragen zu Effekten der Klassengröße beantworten ließe.

Zum Beispiel wird PISA in vielen Staaten (und anders als in Deutschland) nur mit Stichproben auf Schulebene durchgeführt. Bei diesem eingeschränkten Design ist es nicht möglich, Klassengrößen, Unterrichtsmerkmale und Leistungen in der gleichen Untersuchungseinheit zu analysieren. Deshalb sind manche Befunde aus diesen Analysen sorgfältig und kritisch zu beleuchten (Schümer und Weiß 2008). Die Erweiterungen jedoch, die zum Beispiel in Deutschland bei der vorhin zitierten IGLU-Studie vorgenommen wurden, gestatten es, stärkere Evidenz für das Wirkungsgeschehen und Effekte der Klassengrößen zu gewinnen. Eine Erweiterung des Studiendesigns durch längsschnittliche Erhebungen würde einen weiteren Qualitätsgewinn bedeuten. Bei der an PISA 2003 angekoppelten Follow-up Erhebung fanden sich allerdings keine Belege für Effekte der Klassengröße auf den Leistungszuwachs in Mathematik im Verlauf eines Schuljahrs (Schöps et al. 2006).

Das Beispiel des STAR-Projekts weist in die Richtung, dass gezielt spezifische Studien zur Frage der Klassengröße durchgeführt werden müssten. Das gilt vor allem dann, wenn (z. B. im Kontext des demographischen Wandels) aus politischer Sicht ernsthaft Verringerungen der Klassengröße in Betracht gezogen werden. Die entscheidende Frage geht dann in die Richtung, *wie* Verringerungen der Klassengröße produktiv umgesetzt werden. Für solche Fragestellungen liegt bisher wenig gesichertes Wissen vor. Unbeantwortet sind aber auch Fragen, inwieweit sich Lehrkräfte unter Bedingungen größerer Klassen mehr herausgefordert (und am Ende wirksamer) erleben. Die Herausforderung könnte von Lehrkräften auch so verstanden werden, durch bestimmte Maßnahmen (z. B. Gruppenarbeiten, gegenseitiges Lehren und Lernen, individualisierende Aufträge) oder durch gezieltes Vorbereiten, die schwierige Situation pädagogisch zu meistern. Aus solchen Untersuchungen könnte gelernt werden, welche Unterstützungs- und Fortbildungsmaßnahmen geeignet sind, um Lehrkräfte mit unterschiedlichen Klassengrößen ertragreich und zu ihrer eigenen Zufriedenheit unterrichten zu lassen. Weitgehend ungeklärt (auch trotz des STAR-Projekts) bleibt die Frage, ob und wie Konstellationen größerer Klassen mit Team-Teaching oder durch (zeitweise) Unterstützung durch pädagogisches Personal – eventuell mit besonderen Qualifikationen (Sonderpädagoge, Unterrichtsassistent, externer Experte) – produktiv unterrichtet werden können. Durch solche breiter angelegten Fragestellungen bewegt sich das Thema Klassengröße weg von einer bloßen Budgetfrage hin zu konzeptionellen Problemen.

Entsprechende weitergehende Fragestellungen sollten auch aus der Perspektive von Lehrkräften und Eltern interessanter sein als Studien, die den Eindruck erwecken, ihr primärer Zweck sei es, Ressourcen zu rechtfertigen oder vielleicht sogar zu reduzieren. Befunde über Null-Korrelationen zwischen Klassengröße und Unterricht sowie Leistung und sogar Belastung werden von Lehrkräften und ihren Verbänden in dieser Zuspitzung selbstverständlich nicht gerne rezipiert. Die öffentliche Darstellung muss erkennen lassen, dass diese Befunde auch produktiv interpretiert werden können: Kleine Klassengrößen erfordern anderen Unterricht als große. In der Passung des Unterrichtsansatzes mit der Klassengröße liegt ein Schlüsselproblem, es gibt Lehrkräfte, die dieses Problem vorbildlich meistern und andere, die daran scheitern. Betrachtet man den Zeitraum der letzten zehn Jahre, dann ergibt sich also folgendes Bild: Die durchschnittlichen Klassen-

größen in Deutschland unterscheiden sich nicht vom internationalen Mittelwert; sie sind in der letzten Dekade leicht verringert worden. Die Evidenz aus Vergleichsstudien weist der Klassengröße kein nennenswertes Gewicht für die Vorhersage von Leistungsunterschieden zwischen Klassen zu. Aus einer pädagogischen und didaktischen Sicht sollten unterschiedliche Klassengrößen mit unterschiedlichen Unterrichtskonzepten beantwortet werden. Die Untersuchungen mit Stichproben aus Deutschland zeigen jedoch, dass in großen Klassen normalerweise nicht anders unterrichtet wird als in kleinen Klassen. Nicht nur Interaktionsmuster sondern auch die wahrgenommene Unterrichtsqualität variieren unabhängig von der Klassengröße. Selbst die Annahme einer stärkeren (subjektiven) Belastung durch größere Klassen lässt sich in den Schulvergleichsstudien nicht substantiell bekräftigen. Und es stellt sich die Frage, ob Klassengröße überhaupt ein geeigneter Ansatzpunkt für die Verbesserung von Unterrichtsqualität ist:

> Die Menge an Studien, die sich immer wieder aufs Neue und ohne überzeugendes Ergebnis darum bemühen, einen generellen Effekt der Klassengröße auf die Leistung aufzuzeigen, lassen Zweifel an der Sinnhaftigkeit dieser Fragestellung aufkommen. Neuere Studien betrachten deshalb Maßnahmen zur Klassengrößenreduktion aus der Perspektive von Kosten-Nutzen-Überlegungen und im Vergleich mit anderen Maßnahmen zur Steigerung der Unterrichtseffektivität“ (Gundlach 2006; OECD 2008; Yeh 2009). Dabei zeichnet sich ab, dass „Maßnahmen zur Steigerung der Unterrichtsqualität, die sich direkt auf den Unterricht richten, etwa die Unterstützung der Diagnosefähigkeit von Lehrkräften oder die Implementierung von Fördermaßnahmen oder auch einfach ein Mehr an Unterricht, der Reduzierung der Klassengröße sowohl in Bezug auf die Kosteneffektivität als auch in Bezug auf den Erfolg überlegen sind. Hier öffnet sich ein breites und noch wenig erforschtes Gebiet. (Lankes und Carstensen 2010, S. 140 f.)

4.2.4 Der Zusammenhang von Klassengröße und Unterrichtsqualität: Stabile, aber unzutreffende Erwartungen der Öffentlichkeit

Die Frage der Klassengröße ist ein „Dauerbrenner“ in der öffentlichen und politischen Diskussion. Aus der Wahrnehmung von Eltern und Lehrkräften (und insbesondere ihren Verbänden) ist die Klassengröße ein wesentlicher Indikator für Bildungsqualität. Diese Annahme stützt sich auf durchaus plausible Argumente – kleinere Klassen erleichtern individualisierende pädagogische Zugänge sowie einen persönlichen Umgang und bedeuten weniger Korrekturaufwand. Diese Befunde schließen selbstverständlich nicht aus, dass Eltern und Lehrkräfte in einzelnen Fällen von ganz überdurchschnittlich großen Klassen die berechtigte Sorge haben, die Schülerinnen und Schüler würden dort schlechter betreut und unterstützt. Tatsächlich streuen die durchschnittlichen Klassengrößen über eine gewisse Spannweite. Die großen Klassen am Rande der Verteilung ziehen die kritische Aufmerksamkeit auf sich, nicht jedoch die besonders kleinen Klassen am anderen Ende der Verteilung.

Während Vertretungen der Lehrkräfte und der Eltern weiter Handlungsbedarf zur Verringerung der Klassengrößen sehen, wird das Thema von Seiten der empirischen Bildungsforschung sehr viel gelassener betrachtet. Eine bloße Absenkung der Klassengrößen

wird als unwirksam betrachtet, weil aufgrund der Befundlage in Deutschland davon ausgegangen werden muss, dass sich die Art des Unterrichts und dessen Qualität damit nicht (von alleine) verbessern würden. Aus dieser Perspektive wird Investitionen, die direkt zur Verbesserung der Unterrichtsqualität dienen (z. B. durch schulnahe Qualitätsentwicklungsprojekte), ein größerer Nutzen zugesprochen. Auch die Einstellung zusätzlichen Personals mit spezifischen Qualifikationen (sonder- und förderpädagogisch oder sozialpädagogisch) könnte bei bestimmten Klassenzusammensetzungen einen stärkeren Förderungseffekt haben und ebenfalls mit einer Entlastung der Lehrkräfte verbunden sein.

Das Beispiel des STAR-Projekts, das vielleicht von Vertretern der Lehrer- und Elternschaft als Beleg für ihre Forderung nach kleineren Klassen herangezogen wird, zeigt eine besondere Konstellation. Hier wurden Klassengrößen in einem so beträchtlichen Ausmaß reduziert (13–17 versus 22–26), dass die Lehrkräfte tatsächlich in dem Modellprojekt mit einer spürbar neuen Klassenkonstellation herausgefordert und stimuliert waren, diese Situation zu nutzen (auch im Kontrast zu den Kontrollgruppen). Hier handelt es sich also nicht um die Größenordnung einer durchschnittlichen Absenkung der Klassengröße um ein oder zwei Schüler, sondern um eine gewaltige Umstellung der Unterrichtsorganisation mit erheblichen Kosten. Letztlich stellt sich aber auch hier die Frage, wie ein solcher Ansatz in der Breite unterstützt und implementiert werden müsste, um eine verbesserte Unterrichtsqualität zu erzielen. Dazu besteht weiter großer Forschungsbedarf.

Das Prinzip, das dem STAR-Projekt zu Grunde liegt, ist in der Bildungsforschung jedoch längst gut belegt: Die Qualität von Input-Faktoren stellt erst einmal nur Gelegenheiten bereit, die z. B. für eine genauere individuelle Diagnostik, Lernbegleitung und Rückmeldung genutzt werden müssten. Deshalb interessiert in der Forschung nicht nur der Input, sondern auch der Output. Die Befundlage für die letzte Dekade zeigt nun, dass die Unterschiede im Input – im Sinne kleinerer Klassengrößen – nicht systematisch mit dem Output (Lernzuwachs, Lernfreude) zusammenhängen. Vor diesem Hintergrund würde die alleinige Verringerung von Klassengrößen zwar zeitweise den Druck auf die Politik reduzieren, aber voraussichtlich nicht zu besseren Lernprozessen und Lernergebnissen führen. Die Kosten für eine Verringerung der Klassengröße ließen vermutlich nur kleinere Absenkungen der Klassenteiler und Durchschnittsgrößen zu. Wenn – wie aufgrund der Evidenz zu vermuten ist – diese Reduzierungen nicht zu einer besseren Lern- und Ergebnisqualität führen, sind dann wieder neue und weitergehende Forderungen nach kleineren Klassen zu erwarten. Das Thema Klassengröße bliebe somit auch dann ein Dauerbrenner, wenn nennenswerte Reduzierungen politisch umgesetzt würden.

Für die politische Entscheidungsfindung bietet die empirische Bildungsforschung durchaus relevante Wissensgrundlagen. Allerdings sind Ergebnisse über die (fehlenden) Zusammenhänge zwischen Klassengröße und Lernerfolg (und sogar Belastung der Lehrkräfte) nicht die Befunde, die in der öffentlichen Diskussion die große Beachtung erfahren haben. Dies führt zu der Frage, ob und wie die empirische Bildungsforschung zu einer Aufklärung der Öffentlichkeit und rationaleren Diskussionen über Bildungsfragen beitragen kann. Selbstverständlich muss die empirische Bildungsforschung gewissermaßen als „Zeuge“ zur Verfügung stehen, wenn über Klassengrößen diskutiert wird. Aber kann sie mehr beitragen? Das relativ geringe Interesse der Medien an dieser Thematik weist auch darauf hin, dass wenig Interesse darin besteht, Eltern (als große Teile der Leserschaft) mit Befunden über fehlende Evidenz für ihre Forderung zu ernüchtern.

Allerdings muss man sich fragen, was denn unmittelbare Konsequenzen einer überzeugenden Vermittlung des empirisch gesicherten Wissens in der öffentlichen und politischen Debatte wären. Die Befundlage muss letztlich so interpretiert werden, dass offensichtlich ein beachtlicher Teil von Lehrkräften unter der Bedingung „große Klasse" sehr erfolgreich arbeitet – und ein vielleicht ähnlich großer Anteil unter der Bedingung „kleine Klasse" deutlich weniger gut. Das führt wiederum zu der grundsätzlicheren Frage, worin die Unterschiede in der Qualität der pädagogischen Arbeit unter den skizzierten Konstellationen bestehen.

Das Thema „Klassengröße" ist also ein Beispiel für ein Thema, bei dem hinreichendes Kausal- und Beschreibungswissen vorliegt, um zumindest die Richtung von Veränderungen anzugeben; nämlich eher auf der Ebene der Unterrichtsqualität als auf der Ebene der numerischen Klassengröße. Und es ist ein Beispiel dafür, dass die bildungspolitische Diskussion innerhalb der oben skizzierten „Öffentlichkeiten" weitgehend unter Ausblendung der verfügbaren empirischen Evidenz geschieht. Schließlich ist sie auch ein Beispiel dafür, dass Wissenschaftler in der Wissenschaftskommunikation ihre Erkenntnisse immer in Auseinandersetzung mit bereits existierenden alltagsweltlichen (subjektiven) Theorien kommunizieren müssen. Dafür wäre es hilfreich, wenn die Struktur und auch die Aufrechterhaltung solcher unzutreffender, aber verbreiteter subjektiver Vorstellungen zu Bildungsthemen selbst besser untersucht und verstanden würde. So stellt sich z. B. die Frage, wodurch die breite Akzeptanz der vermuteten Kausalannahme über „kleinere Klassen = besserer Unterricht" eigentlich bedingt ist. Ist sie auf die Übergeneralisierung von Erfahrungen mit sehr großen Klassen zurückzuführen? Dies wäre dann anzunehmen, wenn diese Kausalannahme vor allem unter älteren Erwachsenen verbreitet wäre, die erheblich größere Klassen erlebt haben. Oder aber die Kausalannahme ist so stark, weil sie einen intuitiv nachvollziehbaren, allgemeinen Mechanismus beschreibt (kleinere Systeme sind besser beherrschbar als große Systeme). Denkbar ist auch, dass es insgesamt zu wenige alltagsweltliche Kausaltheorien zur Erklärung der subjektiv (in der persönlichen Schulzeit, bei den eigenen Kindern und medial vermittelt) erlebten Varianz von Schulerfolg und Schulerleben in unterschiedlichen Schulklassen gibt, sodass man auf die Erklärungen zurückgreift, die eben verfügbar sind.

4.3 PISA und die Frage nach der Schulstruktur

Das dritte Fallbeispiel bezieht sich auf die Frage, welche Rolle die Struktur des Schulsystems für die dort erzielten Ergebnisse spielt. Die Frage nach dem optimalen Schulsystem wird in Deutschland seit Jahrzehnten kontrovers diskutiert. Dabei lassen sich verschiedene Höhepunkte der Debatten identifizieren. Ein erster Höhepunkt waren Diskussionen, die durch Publikationen und Vorschläge (z. B. für Schulversuche mit Gesamtschulen) des Deutschen Bildungsrates stimuliert wurden. Eine Neubelebung der Debatten erfolgte mit den Ergebnisberichten zum „Programme for International Student Assessment" (PISA), die von der OECD (2001, 2004, 2007) und den für PISA in Deutschland verantwortlichen Wissenschaftlergruppen (Baumert et al. 2001; Prenzel et al. 2007; Prenzel et al. 2004) veröffentlicht worden waren. Im Kontext dieser Publikationen wurde über unterschiedliche Maßnahmen zur Verbesserung der Qualität des Lehrens und Lernen an den Schulen in Deutschland gesprochen. Die Debatten gelangen bis heute immer wieder zu der

Frage zurück, ob ein mehrgliedriges oder ein nicht differenzierendes Schulsystem bessere Bedingungen für ein hohes Leistungsniveau, eine geringere Zahl von Schulversagern und für eine höhere Bildungsgerechtigkeit schafft. In diesen Diskussionen wird meistens auf Erkenntnisse aus PISA Bezug genommen, um für oder gegen eine bestimmte Schulorganisation zu argumentieren, vgl. auch Bieber et al. (2014), die im internationalen Vergleich der Frage nachgehen, in welchem Maße und abhängig von welchen Randbedingungen bildungspolitische Effekte von einem Programm wie PISA ausgehen.

4.3.1 Die öffentliche Debatte

Die Ergebnisse der ersten PISA-Erhebung aus dem Jahr 2000 schockierten Deutschland (Baumert et al. 2001; OECD 2001). In allen drei Domänen (Lesen, Mathematik, Naturwissenschaften) lagen die Leistungen der Fünfzehnjährigen signifikant unter dem Mittelwert der OECD. Deutschland lag damit im unteren Drittel der Rangliste der OECD-Staaten. Fast ein Viertel der Schülerinnen und Schüler erreichte ein Kompetenzniveau, das nicht über Grundschulanforderungen hinausreichte und sehr schlechte Chancen für die weitere Ausbildung und eine Berufskarriere bedeutete. Im internationalen Vergleich wurden ein sehr enger Zusammenhang zwischen sozialer Herkunft und Kompetenz sowie große Disparitäten nach Migrationshintergrund beobachtet. Die Lage in Deutschland stellte sich damit in mehrfacher Hinsicht als problematisch dar. Der internationale Vergleich legte eine Gegenüberstellung mit Staaten nahe, die in diesen Belangen sehr gute Ergebnisse erzielten. Damit rückten Finnland, Japan und Korea, aber auch Kanada oder Schweden als Beispiele „guter" Praxis in den Blickpunkt. Im Wesentlichen waren es diese Ergebnisse, die in den Medien als Botschaften transportiert wurden und die Diskussion prägten.

Die Diskussion wurde durch zahlreiche Kommentare in den Medien stimuliert. Neben Stellungnahmen aus Expertenkreisen waren politische Statements zur Erklärung der Befunde und zu möglichen Handlungsoptionen erwünscht. Unter dem Eindruck der insgesamt problematischen Ergebnisse schienen kleine Nachbesserungen wenig Erfolgsaussichten zu versprechen. Grundlegende Reformen des Schulwesens in Deutschland waren in Betracht zu ziehen. Eine nicht unwichtige Rolle in den Debatten spielte die Interpretation der Ergebnisse. In Berichten und Stellungnahmen der OECD war unter anderem auch auf Zusammenhänge zwischen der Schulstruktur und der Qualität der Ergebnisse (Leistungsniveau, Kopplung mit Herkunft) aufmerksam gemacht worden. Ein wesentlicher Strang der Diskussion richtet sich damit auf die Systemfrage. Perspektiven mit hoher Anschaulichkeit ergaben sich aus dem „Benchmarking", dem Vergleich mit offensichtlich erfolgreichen Schulsystemen. Staaten an der Spitze der OECD-Rangliste überzeugten nicht nur durch ein hohes Leistungsniveau und eine geringe Leistungsstreuung, sondern auch durch relativ schwache Zusammenhänge zwischen Merkmalen der Herkunft und Leistungskennwerten. Und: In diesen Staaten erfolgte eine Aufgliederung der Schülerinnen und Schüler in unterschiedliche Schulformen erst am Ende der ersten Sekundarstufe. In den Blickpunkt der Beispielbetrachtung gerieten freilich noch weitere Aspekte (Qualitätssicherungssystem, Autonomie von Schulen, Wertschätzung von Schule, Förderungs- versus Selektionsorientierung). In vielen Stellungnahmen wurde der „finnische Weg" als Perspektive zur Lösung vieler Probleme im deutschen Bildungssystem präsentiert.

Die Ergebnisse des Vergleichs der Bundesländer in Deutschland (Baumert et al. 2002) brachte allerdings Erkenntnisse, die sowohl die Interpretation der Situation in Deutschland als auch das Nachdenken über Maßnahmen zur Verbesserung neu anregten. Der Ländervergleich in Deutschland zeigte überraschend große Unterschiede im durchschnittlichen Leistungsniveau, lieferte aber keine Belege dafür, dass in Ländern mit hohen Anteilen von Gesamtschulen oder mit einer sechsjährigen Grundschule der Zusammenhang zwischen Herkunft und Kompetenz schwächer ausgeprägt wäre. Im innerdeutschen Vergleich schienen Länder mit einem klaren Bekenntnis zu einem leistungsorientierten dreigliedrigen Schulsystems in verschiedener Hinsicht deutlich besser abzuschneiden als Länder, die sich zum Beispiel mit vermehrten Gesamtschulangeboten um bessere Bildungschancen unabhängig von der sozialen Herkunft bemüht hatten. Damit wurde die kontroverse Diskussion darüber, ob die Schulstruktur in Deutschland verändert werden solle, wiederbelebt. Die internationalen und nationalen Vergleiche im Rahmen von PISA lieferten nicht nur Anlass, sondern auch Stoff (Daten, Beispiele, Interpretationen) für die Debatte. In dieser Kontroverse positionierten sich neben den politischen Parteien (auf Landes- und Bundesebene) wiederum Lehrerverbände (je nach vertretener Schulart). Nebenbei bemerkt gab es weniger politische Kontroversen bezogen auf Maßnahmen zur Qualitätssicherung (Bildungsstandards, Vergleichsarbeiten, Schulevaluation), die ihre Begründung letztlich auch aus dem Benchmarking mit erfolgreichen PISA-Staaten bezogen. Vermutlich wurde bei der Frage der Schulstruktur an die bereits älteren und seit Jahrzehnten vertretenen Überzeugungen angeschlossen.

4.3.2 Der Stand der Forschung

Ob Gesamtschulen zu mehr Chancengerechtigkeit beitragen, ist in Deutschland seit den Diskussionen des Deutschen Bildungsrates Gegenstand von Untersuchungen (z. B. Fend 1982, 2008; Fend et al. 1976). Allerdings bezogen sich die Studien vorwiegend auf Modellversuche beziehungsweise Gesamtschulen als weiteres Schulangebot neben Gymnasien, Real- und Hauptschulen. Es gab in Deutschland somit keinen „Feldversuch" einer flächendeckenden Einführung von Gesamtschulen im Sinne eines Regelsystems. Die Frage, ob die Einführung eines übergreifenden Grundschulsystems bis zum Ende der ersten Sekundarstufe (wie zum Beispiel in Skandinavien) bessere Aussichten auf ein hohes Leistungsniveau und gerechte Bildungschancen als das bisherige Schulsystem in Deutschland bietet, kann mit der Evidenz aus den früheren Gesamtschulstudien nicht überzeugend beantwortet werden.

Die im Zusammenhang mit der Debatte häufig zitierte PISA-Studie selbst liefert zwar anschauliche Beispiele für erfolgreiche Schulsysteme, gelangt aber aufgrund ihrer Anlage als Querschnittstudie dann an Grenzen, wenn Zusammenhänge (z. B. zwischen Strukturmerkmalen und Bildungsergebnissen) kausal erklärt werden sollen. Allerdings beginnt bereits bei der einfachen Frage, ob es einen Zusammenhang zwischen dem Strukturmerkmal eingliedriges versus mehrgliederiges Schulsystem und dem Leistungsniveau gibt, die Evidenz zu „zerfallen". Da zahlreiche Staaten mit einem nicht differenzierenden System (z. B. USA, Norwegen, Italien) schwache Leistungen erbringen, lässt sich kein bedeutsamer Zusammenhang finden. Ein solcher Befund kann so verstanden werden, dass man sich von einer bloßen Umgestaltung der Schulstruktur keine Leistungsverbes-

serung erhoffen darf. Etwas anders sieht das Bild beim Zusammenhang zwischen dem Alter der Aufgliederung in unterschiedliche Schularten und Indikatoren für die soziale Kopplung zwischen Herkunft und Leistung aus. Aber obwohl hier ein (schwacher) statistischer Zusammenhang besteht, darf dieser nicht kausal interpretiert werden. Nimmt man zum Beispiel die jüngsten PISA-Daten (Klieme et al. 2010b), dann findet man für Staaten wie Neuseeland (mit einem nicht differenzierten System) seit der ersten PISA-Erhebung eine ständige Verstärkung des Zusammenhangs, für Deutschland aber eine Reduzierung, sodass inzwischen in beiden Staaten das gleiche Niveau der Kopplung zwischen Herkunft und Kompetenz beobachtet wird. Sehr wenige (oder gar nur einzelne) solcher Beispiele stellen vermutete kausale Effekte der Schulstruktur auf die Förderung oder Benachteiligung von Kindern unterschiedlicher sozialer Lagen massiv in Frage. Letztlich zeigen solche Beispiele, dass unter den Bedingungen „ein- versus mehrgliedriges Schulsystem" unterschiedliche Zustände und Entwicklungen möglich sind. Wie die Befunde belegen, kann in einem mehrgliedrigen System (Beispiel Deutschland) innerhalb von zehn Jahren die Kopplung mit sozialer Herkunft signifikant abgesenkt werden, während bei einem eingliedrigen System dieser Zusammenhang deutlich zunehmen kann. Offensichtlich üben hier andere (und bisher nicht identifizierte bzw. in der Wirkung nicht ausreichend belegte) Faktoren einen wichtigen Einfluss aus. Mit einer ähnlichen Argumentation haben Kobarg und Prenzel (2009) Entwicklungen in den nordischen Staaten analysiert und am Beispiel der Unterschiede zwischen skandinavischen Staaten argumentiert, dass die Grundstruktur des dortigen Schulsystems keine positive Leistungsentwicklung garantiert. Allerdings fehlt derzeit auch gut gesichertes Wissen darüber, welche Faktoren innerhalb eines in bestimmter Weise strukturierten Schulsystems zu einer Verbesserung des Leistungsstandes und der Bildungschancen substantiell beitragen. Im Rahmen einer Bilanz der bisherigen PISA-Runden wurden einige solcher Faktoren (z. B. Qualitätssicherung) ausführlicher diskutiert. (Klieme et al. 2010a). Betrachtet man die Veränderungen im Kompetenzniveau sowie in der Kopplung zwischen Leistung und Herkunftsmerkmalen in den OECD-Staaten von PISA 2000 bis PISA 2012, dann zeichnet sich ab, dass die Schulstruktur keineswegs Erfolg garantiert oder positiven Leistungsentwicklungen im Wege steht (Müller und Ehmke 2013; Sälzer et al. 2013b).

4.3.3 Bessere Evidenz durch eine Ausweitung der Forschungsperspektive?

Bei Entscheidungen, die eine Ausrichtung der nationalen Schulstruktur betreffen, ist es praktisch kaum möglich, Evidenz aus experimentell angelegten Studien zu erhalten. Experimentelle Designs könnten zwar helfen, die Wirkung von Faktoren (z. B. Strukturmerkmale) besser zu verstehen, aber ein solches Experiment hätte weitreichende Voraussetzungen, die z. B. Curricula, Lehrerbildung, Räumlichkeiten etc. beträfen. Die mit dem Experiment verbundenen Auswirkungen auf Schülerinnen und Schüler (worst case) müssten abgeschätzt und abgewogen werden. Auch für eine solche Abschätzung benötigte man bereits im Vorfeld einigermaßen überzeugende Evidenz. Noch unrealistischer scheinen experimentell angelegte Schulversuche mit Struktureingriffen. Zwar könnten solche Experimente in bestimmten Regionen in kleiner Zahl durchgeführt werden, doch ist es unmöglich, die unterschiedlichen Randbedingungen und intervenierenden Faktoren in einer größeren Zahl zu kontrollieren. Eine Veränderung der Schulstruktur, die vor Jahr-

zehnten in Finnland mit Erfolg durchgeführt wurde, liefert ein Fallbeispiel, aber keine ausreichende Evidenz, dass in Deutschland – unter deutlich anderen Ausgangslagen und Bedingungen – eine analoge Veränderung zum Erfolg führen würde. Nicht zuletzt müssen solche Veränderungen (z. B. von den Eltern und Lehrkräften) mitgetragen werden. Spätestens hier wird sichtbar, dass Entscheidungen über Strukturmerkmale eines Bildungssystems wesentliche politische Komponenten enthalten. Eine Begründung durch wissenschaftliche Evidenz reicht vor diesem Hintergrund nicht aus.

Wissenschaftliche Evidenz kann jedoch auf dringenden politischen Handlungs- und Entscheidungsbedarf hinweisen. So zeigten etwa die Befunde der Ländervergleiche in Deutschland, die seit PISA 2003 auch nach Schularten differenzierten (Prenzel et al. 2005, 2008), äußerst problematische Ergebnisse für die Hauptschulen in den Stadtstaaten, die nur noch von einem ca. zehnprozentigen Anteil des Altersjahrgangs besucht wurden. Mit der Verringerung des Anteils von Schülerinnen und Schülern, die zum Teil auch durch das alternative Angebot von Gesamtschulen befördert wurde, hatte sich über die letzten Jahre auch die soziale und kulturelle Komposition der Klassen in den Stadtstaaten verändert. Damit standen dort die Hauptschulen vor riesigen Herausforderungen, zielführend und erfolgreich zu unterrichten. Die PISA-Befunde halfen, diese Problemlage zu identifizieren und in ihrer Reichweite politisch sichtbar zu machen. Politische Reaktionen folgten dann. Dieses Beispiel mag auch dafür stehen, dass die bei PISA 2000 noch vorherrschende politische Ablehnung von Ländervergleichen auf der Ebene aller Schularten auf zugängliche Erkenntnisse (und nützliche Evidenz) verzichtet hatte.

Wie im letzten Abschnitt ersichtlich, helfen inzwischen Zeitreihen von internationalen Vergleichsstudien und längsschnittliche Designs (z. B. Blossfeld et al. 2011; Prenzel et al. 2006), um besser abschätzen zu können, inwieweit Merkmale der Schulstruktur kausal relevant und damit wichtige Stellschrauben für Qualitätsverbesserungen sind. Für ein besseres Verständnis der Wirkungen werden freilich elaborierte Theorien benötigt, die Prozessabläufe von distalen Merkmalen der Schulstruktur hin zu proximalen Merkmalen der Unterrichtsqualität und des Lehrens und Lernens modellieren. Um dies an einem Beispiel zu erläutern: Mehrgliedrige Systeme treffen zu einem mehr oder weniger frühen Zeitpunkt Entscheidungen über die Zuordnung von Schülerinnen und Schülern zu Bildungsgängen. In einem mehrgliedrigen System könnten Lehrkräfte dazu tendieren, auf der Basis von Annahmen über stabile Begabungen weniger Anstrengungen für eine individuelle Förderung zu unternehmen. Sie könnten auch zu eher uniformen Unterrichtskonzepten neigen und weniger Anstrengungen in Richtung Differenzierung unternehmen. Maaz et al. (2011) haben kürzlich erneut zeigen können, dass die Zensurengebung durch die soziale Herkunft der Schüler beeinflusst wird und dass u. a. auf diese Weise die Kopplung von sozialer Herkunft und Übergangsentscheidungen zwischen unterschiedlichen Schulformen erzeugt wird. Aus derartigen Befunden ergeben sich wiederum weiterführende Fragen: Können solche Tendenzen von Lehrkräften in einem mehrgliedrigen System durch bestimmte Maßnahmen ausgeglichen und wie kann eine Durchlässigkeit von „unten nach oben“ gestützt werden (z. B. Lehrerbildung, Qualitätssicherung, Anreize)? Man könnte auch fragen, ob Lehrkräfte durch eine Veränderung der Schulstruktur veranlasst werden, entsprechende Tendenzen aufzugeben.

Die Frage nach Evidenz für eine Entscheidung über die Schulstruktur führt damit sehr schnell in breitere Forschungsfragen, zum Beispiel zu Überzeugungen und Kompeten-

zen von Lehrkräften und zur Unterrichtsqualität, die Untersuchungsprogramme nach sich ziehen müssten. Zusätzlich wären Fragen zu den Bedingungen einer erfolgreichen Implementation zu stellen, die ebenfalls empirisch zu beantworten wären. Die zuletzt angesprochenen Erkenntnisinteressen müssten nicht unbedingt durch neue Projekte bedient werden, wenn Ergebnisse aus anderen, vielleicht stärker auf Grundlagenwissen bezogenen Projekten vorliegen.

4.3.4 Der Zusammenhang von Schulstruktur und Unterrichtserfolg: Erwartungen der Öffentlichkeit an die empirische Bildungsforschung

In diesem letzten Beispiel ging es um die Frage, ob die aus PISA oder anderen Studien vorliegende Evidenz bildungspolitische Entscheidungen über die Schulstruktur begründen kann. Aufgrund des Designs kann PISA Hinweise im Rahmen eines Benchmarking geben, die politisch höchst anregend sein können. PISA kann auf gravierende Problemlagen und Handlungsbedarf hinweisen (z. B. Hauptschulen in Stadtstaaten). Doch bedingt die Anlage der Studie, dass beobachtete Zusammenhänge zwischen Schulstrukturen und Output-Indikatoren nicht als kausal interpretiert werden dürfen.

Die Diskussion der Befunde aus PISA zeigt vielmehr, dass – nimmt man Staaten oder Länder als Fälle – Strukturmerkmale mit unterschiedlichen Ergebnissen (Leistungsniveau, Kopplung zwischen Herkunft und Kompetenz) verbunden sein können. Auch zeichnet sich bei der Betrachtung der Daten über mehrere Erhebungsrunden ab, dass die Schulstrukturen für sich noch keine Vorhersage gestatten, ob Indikatoren für Bildungsgerechtigkeit sich negativ oder positiv verändern (Beispiele Neuseeland und Deutschland). Entsprechende Befunde sind eher als Evidenz dafür zu verstehen, mit Entscheidungen über Schulstrukturen höchst vorsichtig umzugehen. Sie sind mit hohen Kosten verbunden und verlangen große Anstrengungen zur Vorbereitung und Implementation.

Deshalb ist es auf jeden Fall sinnvoll, theoriegeleitet nach Ansatzpunkten für Qualitätsverbesserungen zu suchen, die weniger aufwändig sind und für die abgesichert ist, dass sie am Ende tatsächlich die alltägliche Qualität des Lernens im Unterricht befördern. Auch Geld, das zur genaueren Erforschung dieser Wirkungszusammenhänge ausgegeben wird, ist gut anlegt.

Freilich gilt auch in diesem Bereich, dass Untersuchungsergebnisse von Wissenschaftlerinnen und Wissenschaftlern zum Teil sehr unterschiedlich interpretiert oder methodische Einschränkungen hinsichtlich der (kausalen oder technologischen) Interpretierbarkeit von Befunden nicht immer gleich beurteilt werden. In politischen Kontroversen besteht dann die Tendenz, dass die politischen Seiten sich jeweils auf kongruente wissenschaftliche Auffassungen berufen. Das Beispiel „PISA und die Schulstruktur“ steht für eine heftige politische Kontroverse in Deutschland. Der Stand der Forschung zu dieser Frage hat sich im Verlauf dieser Debatte zwar so weit entwickelt, dass die Kontroverse heute sehr viel sachlicher geführt werden kann, es bleibt aber festzuhalten, dass es im Kern nicht möglich ist, die Frage nach der Eignung einer bestimmten Schulstruktur alleine durch Berufung auf Ergebnisse der empirischen Bildungsforschung zu beantworten. Sie bleibt eine Frage, die nur unter Bezug auf normative, man könnte auch sagen, politische Wertungen zu beantworten ist. Das hat im Wesentlichen zwei Gründe:

Wie oben ausgeführt, ist es praktisch unmöglich, die breite Einführung grundlegender neuer Schulstrukturen sozusagen vorab empirisch zu testen. Eine solche Erwartung ist durch die empirische Bildungsforschung nicht zu erfüllen. Das oben erwähnte Beispiel der Gesamtschulstudien aus den 70er Jahren hat das gezeigt. Es waren aufschlussreiche Studien zu den Effekten von Gesamtschulen innerhalb eines Bildungssystems mit ganz unterschiedlichen Schulstrukturen, es waren jedoch keine empirischen Untersuchungen, in denen ein homogenes und ein heterogenes Schulsystem bei gleichzeitiger experimenteller Kontrolle weiterer Rahmenbedingungen verglichen wurden. Solche Vergleiche sind, wenn überhaupt, nur nationenübergreifend möglich. Die damit aber wiederum verbundenen Begrenzungen der Aussagekraft solcher Vergleiche wurden bereits oben diskutiert.

Der zweite Grund liegt darin, dass Veränderungen der Schulstruktur, insbesondere dann, wenn sie über lokale und/oder modellhafte Versuche hinausgehen, immer eingebettet sind in breite politische und gesellschaftliche Debatten. Man kann also eine solche Veränderung nicht als Experiment denken, das die Rückwirkungen solcher politischen Debatten auf die Akteure ausblenden könnte. Die unterschiedlichen Öffentlichkeiten sind als Politiker, Lehrer, Eltern, Schüler sowohl Akteure als auch Rezipienten dieser Debatten und sie wären zugleich in einem solchen „Experiment“ involviert.

Die empirische Bildungsforschung und – damit eng verbunden – die pädagogische Psychologie haben vielfach zeigen können, dass die Erwartungen, Wahrnehmungen und alltagsweltlichen Interpretationen von Lehr-Lernprozessen in erheblichem Maße das Lehren und Lernen bestimmen. Oben wurden Beispiele dafür genannt: Wenn die Übergangsempfehlungen von Lehrern u. a. von den subjektiven Vorstellungen zur individuellen Stabilität von Begabungen abhängen, dann ist dies ein Beispiel dafür, dass Schulstrukturen nicht per se, sondern in Abhängigkeit von deren Interpretation durch die Beteiligten wirksam werden. Ein anderes Beispiel sind die inzwischen gut untersuchten Auswirkungen der Erwartungen von Lehrern, Eltern und Schülern zu geschlechtsspezifischen Leistungsstärken und Schwächen in den MINT-Fächern einerseits und sprachlichen Fächern andererseits. Obwohl solche Erwartungen eigentlich unrealistisch sind (es gibt keine biologischen Gründe für derartige Leistungsunterschiede) haben sie nachweisbare Effekte auf Motivation, Selbstkonzept und schließlich sogar die tatsächliche Leistungsausprägung von Mädchen und Jungen.

Diese Befunde betreffen vor allem die (insgesamt gut untersuchten) Effekte der alltagsweltlichen Theorien zur Erklärung von interindividueller Varianz von Schulleistungen auf das Handeln von Lehrkräften, Eltern und Schülern. Weniger gut untersucht sind bislang die Effekte der alltagsweltlichen Theorien zur Rolle von Konstanz und Veränderung von Rahmenbedingungen des schulischen Lernens. Die empirische Bildungsforschung weiß bislang wenig darüber, wie sich die Konstanz oder aber die Veränderung organisatorischer, curricularer unterrichtsmethodischer Rahmenbedingungen schulischen Lehrens und Lernens auf die verschiedenen Akteure auswirkt. Gerade weil Veränderungen der Schulstruktur ein heftiger Gegenstand öffentlicher Debatten sind, muss die Rückwirkung dieser Debatten auf die Akteure berücksichtigt werden. Man kann annehmen, dass z. B. die Haltung von Eltern zu Übergangsempfehlungen ihrer Kinder auch von dem beeinflusst wird, was sie wiederum in der öffentlichen bildungspolitischen Debatte über Ergebnisse der empirischen Bildungsforschung gehört und verstanden haben. In anderen Worten: Die öffentliche Debatte zu Bildungsfragen ist ein Teil der Rahmenbedingun-

gen von Bildungsprozessen und insofern sollte sie auch zum Forschungsgegenstand der empirischen Bildungsforschung werden.

Vor diesem Hintergrund bleibt die Einschätzung, dass Entscheidungen über Schulstrukturen nur zu einem kleinen Teil durch wissenschaftliche Evidenz getragen werden können. Zu einem großen Teil handelt es sich um politische Entscheidungen, die von Überzeugungen in gesellschaftlichen Gruppen und von Wählerwünschen beeinflusst werden und für die Konsens gesucht werden muss. Hier bestehen allerdings bislang auch erhebliche Forschungslücken. Sowohl die Wirkung von strukturellen Veränderungen auf die beteiligten Akteure als auch das Verständnis und die Diskussion von Befunden der empirischen Bildungsforschung in den unterschiedlichen Öffentlichkeiten sollten umfassender empirisch untersucht werden. Basierend auf derartiger Forschung könnte die empirische Bildungsforschung besser dazu beitragen, die Evidenzlage, die Erfolgsaussichten und die Risiken von Veränderungen sichtbar zu machen.

5 Schlussfolgerungen aus den Fallbetrachtungen: Zwei Forschungsprogramme und ihr Zusammenhang

An drei Fallbeispielen haben wir exemplarisch die Prozesse der Rezeption von Forschungsergebnissen mit dem Schwerpunkt Bildung und Erziehung dargestellt. Wir verweisen auf die voranstehenden Abschnitte und fassen die dortigen Vorschläge und Empfehlungen zur Ausweitung der jeweiligen Forschungsperspektiven (Computerspiele, siehe S. 33, Klassengröße, siehe S. 46, Schulstruktur, siehe S. 58) hier nicht erneut zusammen.

Am Beispiel der gewalthaltigen Computerspiele haben wir zu zeigen versucht, dass empirische Forschung empirische Evidenz (Kausalwissen) liefern kann, die gleichzeitig belastbar und dennoch unabgeschlossen ist. Die Forschung zu diesem Thema illustriert, dass Vorläufigkeit wissenschaftlicher Evidenz der Normalfall wissenschaftlicher Forschung ist (Nowotny 1999), der jedoch einer Nutzung der Erkenntnisse in Bildungspolitik und Bildungsadministration nicht entgegensteht. Im Gegenteil, gerade wenn man versteht, dass die Unabgeschlossenheit der Ergebnisse und zugleich ihre Aussagekraft bereits bei dem vorliegenden Erkenntnisstand zusammen zu sehen sind, kann die Forschung zur Versachlichung der jeweiligen Debatten beitragen. Allerdings setzt dies methodisch angemessene Forschungssynthesen (die zu diesem Thema durchaus vorliegen) und ein sozialwissenschaftlich fundiertes Verständnis der öffentlichen Debatten und Auffassungen zu dieser Thematik voraus.

Auch die beiden anderen Beispiele der Wirkungen von Klassengröße und von Schulformen haben gezeigt, dass die Evidenz, wie sie empirische Bildungsforschung liefern kann, zwar heuristische Anregungen zur Gestaltung bestimmter bildungspolitischer Maßnahmen liefert, dass diese aber letztlich auch eine politische, normative Begründung benötigen. Die drei Beispiele haben aber auch gemeinsam, dass viele Annahmen, die in die dann notwendige politische Diskussion eingehen, durchaus empirisch und konzeptuell mit den Methoden der Bildungsforschung überprüft werden können.

Auf die Dauer wird sich ein solcher versachlichender Effekt wissenschaftlicher Evidenz auf die Debatte nur dann erreichen lassen, wenn die Beteiligten und das heißt insbe-

sondere die Wissenschaftler, die solche Evidenz liefern und in diese öffentliche Debatte einbringen, sich dabei ihrerseits auf sozialwissenschaftlich fundiertes Wissen über die Erwartungen und das Vorverständnis der „Öffentlichkeiten" stützen können; in anderen Worten: Eine nachhaltige und ihrerseits rational geplante Beteiligung der Bildungsforschung (d. h. hier also der Wissenschaftlerinnen und Wissenschaftler, die Bildungsforschung betreiben) setzt sozialwissenschaftliche Forschung zum Verhältnis Wissenschaft und Öffentlichkeit im Bereich der Bildungsforschung voraus. Im Ergebnis empfehlen wir zwei Maßnahmen im Bereich der empirischen Bildungsforschung: 1)Verstärkung von Forschung und Entwicklung zur Erstellung von evaluativen Forschungssynthesen. 2) Verstärkung von Forschung und Entwicklung zur Kommunikation und dem Verhältnis zwischen Wissenschaft und Öffentlichkeit zu Themen der empirischen Bildungsforschung. Wie die Fallstudien zeigen, hängen beide Themenbereiche zusammen; die Gestaltung von Forschungsprogrammen sollte also diesen Zusammenhang reflektieren.

5.1 Verstärkung von Forschung und Entwicklung zur Erstellung von evaluativen Forschungssynthesen im Bereich der empirischen Bildungsforschung

Oben wurde bereits auf die elaborierte metanalytische Methodik hingewiesen, die im Kontext der Campbell Foundation (http://www.campbellcollaboration.org) entwickelt wurde. Dabei wurde aber auch deutlich, dass die mit diesem Ansatz implizierte Hierarchie von Evidenz unterschiedlicher Stärke als alleiniges normatives Modell für evaluative Forschungssynthesen nicht realistisch anwendbar ist. Hier sind also weitere Methodenentwicklungen notwendig, die den oben skizzierten Besonderheiten auf dem Gebiet der empirischen Bildungsforschung gerecht werden. Die Beiträge von Beelmann (2014) und Pant (2014) in diesem Heft beleuchten die Desiderate solcher Methodenentwicklung. Wie bereits angedeutet wurde, führt nicht alleine eine Verstärkung von Sammel- und Review-Aktivitäten zu einer besseren Abschätzung, wie belastbar der Forschungsstand zu bestimmten Fragen und Entscheidungsproblemen ist. Generell käme es darauf an, nicht nur mehr, sondern vor allem bessere Metaanalysen über Befundlagen zu Fragestellungen der empirischen Bildungsforschung zu erstellen. So verdienstvoll umfassende Metaanalysen (z. B. Hattie 2009) auch sind, so besteht doch die Gefahr, dass der Komplexität der Datenlage nicht immer kritisch genug Rechnung getragen wird (Pant 2014). Dies zeigt sich z. B. im Vergleich zu der Metaanalyse von Seidel und Shavelson (2007), die thematisch stärker fokussiert war und sehr viel weiterreichende Ansprüche verfolgt und umgesetzt hat: Die Berücksichtigung mehrfacher Zielbezüge (z. B. Effekte von Unterricht auf kognitive *und* motivationale Outcomes) und die Aufschlüsselung von Methodeneffekten (z. B. experimentelle versus nicht experimentelle Designs, Selbstauskünfte versus Beobachtungsmethoden). Erst mit entsprechenden analytischen und kritischen Herangehensweisen besteht die Chance, substantielle Erkenntnisse aus einer unübersichtlichen Forschungslage herauszuarbeiten – und es besteht damit auch die Chance, auf das Forschungsfeld zurückzuwirken und bestimmte anspruchsvollere Forschungszugänge hervorzuheben und anzuerkennen.

Verfahren einer „Best-Evidence Synthesis", die seit Jahren insbesondere um R.E. Slavin (z. B. Slavin et al. 2009a) entwickelt und umgesetzt wurden, bedeuten im Bereich der empirischen Bildungsforschung beträchtliche Fortschritte, indem Stück für Stück

der aktuelle Erkenntnisstand zu bestimmten Fragen aufbereitet und zugänglich gemacht wird. Entsprechende umfangreiche und systematisch angelegte Initiativen finden wir im deutschsprachigen Raum bisher nicht. Das ist auch deshalb schade, weil Problemlagen im Bildungsbereich immer auch durch Besonderheiten nationaler Bildungssysteme und Traditionen bestimmt werden und nicht allein auf der Basis von Studien aus dem englischsprachigen Raum angemessen beurteilt werden können.

So verdienstvoll Evidenzsynthesen im Bildungsbereich auch sind, so stellt sich doch die Frage, wer entsprechende umfangreiche Arbeiten tatsächlich konsequent leisten kann. Anzumerken bleibt, dass es sich hier ja nicht um das übliche Geschäft wissenschaftlichen Arbeitens und Publizierens handelt. Ob zum Beispiel ein starkes Engagement in der Synthetisierung und Bewertung von Evidenz für die Forscherkarriere förderlich ist, lässt sich derzeit kaum abschätzen. Daneben ist unklar, wie das wissenschaftliche Engagement für solche Vorhaben im Rahmen der üblichen Drittmittelförderung in Deutschland unterstützt werden könnte oder ob bestimmte wissenschaftliche Einrichtungen mit Aufgaben der Evidenzsynthese betraut werden sollten. Derzeit gibt es jedenfalls keine entsprechenden Förderprogramme für die aufwändigen Arbeiten einer systematischen Evidenzbeurteilung im Bildungsbereich.

Dennoch gab und gibt es in Deutschland immer wieder Initiativen zu einer kritischen Beurteilung von Evidenz, die durch die Wahrnehmung akuter Probleme im Bildungsbereich (z. B. Mathematik- und Naturwissenschaftsdefizite nach TIMSS, Leseprobleme in Folge von PISA) ausgelöst und dann in Form von Expertisen dargelegt und interpretiert wurden. Die dabei erstellten Gutachten waren von sehr hoher wissenschaftlicher Qualität, entstanden aber immer problemgetrieben und unter hohem Zeitdruck. Sie folgten auch nicht einem expliziten Verfahren zur Evidenzabschätzung, sondern der Erfahrung und dem Ideenreichtum von Expertinnen und Experten. Offensichtlich gibt es gute Beispiele, aber bisher noch keine etablierten Verfahren oder Heuristiken. Für die Erstellung von Expertisen in Anspruch genommen wurden und werden oft Wissenschaftlerinnen und Wissenschaftler von einschlägig ausgewiesenen Instituten, die zwar das Forschungsfeld ständig im Blick haben, aber bisher wenig Grund hatten, Einschätzungen der Forschungslandschaft mit einer bestimmten Regelmäßigkeit vorzunehmen und zu publizieren. Deshalb wäre es höchst wünschenswert, auf der Basis von Erfahrungen aus entsprechenden und erfolgreichen Gutachten einerseits und international erkennbaren Heuristiken für die Synthese bester Evidenz andererseits eine Art Leitfaden oder Kriterienkatalog für die Evidenzbeurteilung im Bildungsbereich zu erstellen. Dieser Leitfaden könnte zugleich als Rahmen für die Ausschreibung eines Förderprogramms zur Evidenzsynthese dienen, das sich thematisch auf eine kleine Gruppe von relevanten Fragen bzw. Entscheidungsproblemen bezieht. Gegenstand des Förderprogramms müsste gleichzeitig die Entwicklung geeigneter Verfahren und Standards für die Evidenzbeurteilung sein. Auf der Basis einer solchen Förderinitiative könnte dann entschieden werden, ob und wie einzelne Wissenschaftlerinnen und Wissenschaftler, einschlägige Institute oder Verbünde von Forschungseinrichtungen für Aufgaben einer systematischen Evidenzbeurteilung im Kontext der Bildungsforschung gewonnen werden könnten.

In den einleitenden Abschnitten dieses Beitrags wurde gezeigt, dass bereits die evaluative Forschungssynthese ein Prozess ist, der in den öffentlichen bildungsbezogenen

Diskurs eingebettet ist und der auch auf diesen reagiert (etwa bei der Auswahl der Fragestellungen). Auch die drei Fallstudien haben solche Zusammenhänge gezeigt.

Außerdem zeigen jüngere Studien zum Verhältnis von Wissenschaft und Öffentlichkeit in den Naturwissenschaften, dass es unter Wissenschaftlern selbst ein zunehmend strategisches Verhalten gegenüber der Öffentlichkeit gibt, weil die Zuteilung von Ressourcen (öffentliche Mittel, aber auch Renommee) an Wissenschaftler selbst wiederum davon abhängig ist, wie ihre Arbeit in der Öffentlichkeit wahrgenommen wird (Peters 2012; Weingart 2005). Es ist bislang aber nicht empirisch untersucht, ob diese Beobachtungen auch für die empirische Bildungsforschung und deren Akteure gelten. Wenn das so sein sollte, stellt sich die Frage, wie sich dies wiederum auf mögliche Forschungssynthesen auswirkt. Auch diese Entwicklung veranlasst uns zu der zweiten Empfehlung, das Verhältnis von Wissenschaft und Öffentlichkeit im Bereich der Bildungsforschung empirisch zu untersuchen. Nur durch eine solche Untersuchung können die konzeptuellen Grundlagen geschaffen werden, um zu bewerten, wie sich mögliche Veränderungen der Haltungen vieler Wissenschaftler zum öffentlichen Diskurs auf die Wissenschaft selbst und auf ihre Rolle in bildungspolitischen Diskursen auswirken.

5.2 Verstärkung von Forschung und Entwicklung zum Verhältnis von Wissenschaft und Öffentlichkeit im Bereich der empirischen Bildungsforschung

Im Kern besteht unser zweiter Vorschlag darin, die öffentliche Diskussion und auch die Implementation von Ergebnissen der empirischen Bildungsforschung als eine Instanz von Wissenschaftskommunikation aufzufassen und selbst zum Gegenstand eines Forschungsprogramms der empirischen Bildungsforschung zu machen. Mit Wissenschaftskommunikation bezeichnet man üblicherweise die Kommunikation zwischen „Wissenschaft & Öffentlichkeit“, wie sie z. B. im Bereich der Naturwissenschaft im Kontext von Initiativen für Public Understanding of Science, im Kontext von Ingenieurwissenschaften und im Kontext der Risikokommunikation betrieben wird. Wissenschaftskommunikation ist ihrerseits auch ein Gegenstand sozialwissenschaftlicher Forschung (Bromme & Kienhues, 2014; Beispiele: Die Fachzeitschrift Public Understanding of Science (http://pus.sagepub.com) oder das aktuelle DFG Schwerpunktprogramm „Wissenschaft und Öffentlichkeit“ (http://wissenschaftundoeffentlichkeit.de/DFG-SPP1409). Bislang gibt es aber erst wenige empirische Studien zur Wissenschaftskommunikation im Kontext der empirischen Bildungsforschung (Böttcher et al. 2009; Killus und Tillmann 2011)).

Ein solches Forschungs- und Entwicklungsprogramm sollte interdisziplinär angelegt sein und könnte durch Wissenschaftler aus der empirischen Bildungsforschung, der Psychologie (Pädagogische Psychologie, Medienpsychologie, Sozialpsychologie), der Kommunikationswissenschaft und der Wissenschaftssoziologie betrieben werden. Die genaue thematische Struktur eines solchen Forschungsprogramms wäre abhängig von den verfügbaren Forschungsressourcen wie auch von der Beteiligung aus den genannten Disziplinen. Das oben genannte DFG-Forschungsprogramm kann dazu ebenfalls Anregungen liefern, es bezieht sich jedoch überwiegend auf naturwissenschaftliches Wissen (z. B. zu Medizin, Klimawandel und Nanotechnologie). Allerdings wird im Rahmen dieses Programms das Verhältnis von Wissenschaft & Öffentlichkeit am Beispiel der Debatte um gewalthaltige Computerspiele untersucht. Zu diesem unmittelbar bildungs-

bezogenen Thema konnte der Zusammenhang zwischen persönlicher Bewertung derartiger Computerspiele und der (verzerrenden) Darstellung und Rezeption der Ergebnislage empirisch gezeigt werden (vgl. Gollwitzer et al. 2014 in diesem Heft). Die Ergebnisse (z. B. Bromme & Goldmann, 2014) des DFG Schwerpunktprogramms „Wissenschaft und Öffentlichkeit“ (das noch bis zum Jahr 2015 läuft) sind auch für die Bildungsforschung interessant, sie erübrigen aber nicht ein Forschungsprogramm, das den Besonderheiten der empirischen Bildungsforschung Rechnung trägt, von denen wir einige in diesem Beitrag skizziert haben.

5.3 Beispiele für Fragestellungen und Adressaten eines Forschungsprogramms zum Verhältnis von Wissenschaft und Öffentlichkeit:

a. Erwartungen und alltagsweltliche Theorien zu Bildungsprozessen bei Eltern, Lehrern und Schülern sowie bei Entscheidungsträgern im Bereich von Bildungspolitik und Bildungsadministration.
   - Im Zusammenhang mit den Fallstudien wurde auf die Rolle solcher Erwartungen und auch auf die Bedeutung alltagsweltlicher Theorien für Lehr-Lernprozesse, aber auch für die Implementation von Neuerungen aller Art im Bildungssystem bereits hingewiesen.
   - Diese Erwartungen und alltagsweltlichen Theorien sind, wie oben erwähnt, bereits recht gut untersucht, wenn es um die Erklärung interindividueller Leistungsunterschiede geht. Sie sind deutlich weniger gut untersucht, wenn man die Erwartungen an die empirische Bildungsforschung betrachtet. Unrealistische Ergebniserwartungen (z. B. Wissenschaft soll „abgeschlossene“ Ergebnisse liefern), problematische subjektive Erklärungsmuster für konfligierende Evidenz (z. B. wissenschaftliche Kontroversen entstehen, weil Wissenschaftler sich profilieren wollen und unnötig streiten) und problematische methodische Erwartungen (z. B. wenn man Befunde zu ganzen Schülerpopulationen hat, dann müsste man doch auch etwas über einzelne Schüler aus der Stichprobe sagen können) in der Öffentlichkeit sind Teil des Kontextes, in dem Wissenschaftskommunikation geschieht. Sie sind eine Herausforderung für die aktive Aufbereitung und Kommentierung von Ergebnissen der empirischen Bildungsforschung. Aus derartigen Studien könnten Empfehlungen abgeleitet werden, wie die Ergebnisse empirischer Bildungsforschung für unterschiedliche Öffentlichkeiten angemessen aufbereitet werden können. Andere Wissenschaftsbereiche bieten interessante „best practice“ – Beispiele und Empfehlungen für die öffentliche Darstellung komplexer wissenschaftlicher Sachverhalte. Beispielhaft sind hier die Bemühungen um die Kommunikation von Klimawandel-Information des Center for Research on Environmental Decisions der Columbia University (http://www.cred.columbia.edu/guide/) anzuführen.

b. Welche „statistische und forschungsmethodische Grundbildung“ ist notwendig, welche ist möglich und auch von welchen Öffentlichkeiten zu erwarten bzw. dort zu fördern?

- Umgang mit Evidenz bedeutet immer noch Umgang mit Unsicherheit. Jedoch bietet Evidenz eine Möglichkeit zur Risikoabschätzung und damit zu einer besseren Entscheidungsgrundlage. Dazu braucht es eine verlässliche Einordnung der vorhandenen Evidenz und die Übersetzung der unter Berücksichtigung wissenschaftlicher Standards erstellten Forschungsergebnisse in eine aus Anwenderperspektive gültigen Evidenzhierarchie. Es bedarf zum Beispiel eines bestimmten Grundverständnisses von Wahrscheinlichkeitsrechnung (General statistical literacy im Sinne von Gigerenzer 2002), um nachvollziehen zu können, wieso es angesichts der Vielfalt von individuellen und organisatorischen Unterschieden im schulischen Lehren und Lernen möglich ist, im Rahmen großer Schulleistungsstudien die Wirkung recht spezifischer Lernervoraussetzungen auf die Kriteriumsleistungen (z. B. Leistungstest) vorherzusagen. Dieses Grundverständnis ist durchaus auch im Rahmen von Wissenschaftskommunikation an interessierte Personen zu vermitteln. Es bedarf dazu jedoch der Entwicklung und Erprobung geeigneten Materials. Solche Entwicklungsarbeiten könnten anknüpfen an analoge Arbeiten zur Verbesserung des Verständnisses von Wahrscheinlichkeitsaussagen im Kontext von Public Health (Gigerenzer et al. 2007). Auch ein grundlegendes öffentliches Verständnis von Kompetenzmessung und des Kompetenzbegriffs wird dann sehr wichtig, wenn dieses Konzept bei Bestandsaufnahmen wie auch bei Veränderungen des Bildungssystems grundlegend ist.

c. Empirische Bildungsforschung im Internet.
   - Die eingangs geschilderte Problematik der Erzeugung von Forschungssynthesen wie auch die Vielfalt von Wissensarten, die empirische Bildungsforschung liefern kann, haben gezeigt, dass Wissenschaftskommunikation nicht nach dem Sender – Empfänger Modell der Kommunikation verläuft. In anderen Worten: Empirische Bildungsforschung kann Evidenz „liefern“ für evidenzbasierte Bildungspolitik, aber diese Metapher des „Lieferns“ kann nicht als Problem einer lediglich adressatengerechten Verpackung und dann Überbringung von Ergebnispaketen verstanden werden. Vielmehr findet dieses „Liefern“ auf vielen Ebenen des bildungsbezogenen gesellschaftlichen Diskurses statt und die Kommunikation ist auch nicht so einseitig, wie es die Metapher des „Lieferns“ nahelegt. Ein zunehmend wichtiges Medium dieser Diskussion ist das Internet. Es dient sowohl als Informationsquelle für die Öffentlichkeit wie auch als Diskussionsforum (z. B. in Blogs zu bildungspolitischen Themen). Ausgehend von einer spezifischen Problemstellung kann man sehr einfach vielfältige wissenschaftliche oder wissenschaftsbasierte Informationen erhalten. Dieser durch moderne Informationstechnologie mögliche weitreichende Zugriff auf wissenschaftsbasierte Informationen ermöglicht eine deutlich leichtere Teilhabe an wissenschaftlichen Diskursen. Gleichzeitig verlangt diese leichte Zugänglichkeit vom Rezipienten eine Einordnung und Interpretation der rezipierten Information, was eine erhebliche Herausforderung darstellt (Stadtler et al. 2014). Die kognitiven und kommunikativen Prozesse der Informationsselektion und der Informations- sowie Quellenbewertung werden zwar gegenwärtig in dem oben erwähnten DFG Schwerpunktprogramm Wissenschaft & Öffentlichkeit untersucht, dort stehen jedoch überwiegend keine bildungsbezogenen Inhalte

der Wissenschaftsrezeption durch Laien im Internet im Mittelpunkt. Der Umgang mit bildungsbezogenen Wissenschaftsinformationen im Internet durch die unterschiedlichen Öffentlichkeiten, insbesondere durch Beteiligte wie Eltern oder Lehrer, ist bislang sehr wenig untersucht. Hier ergeben sich sowohl grundlagenorientierte Forschungsfragestellungen als auch Fragen einer empirisch evaluierten Gestaltung von Informationsbereitstellungen aus der Bildungsforschung. Ein Beispiel für eine solche grundlagenorientierte Frage ist die Problemstellung, wie Personen ohne ausreichendes Hintergrundwissen Urteile darüber treffen, welche Informationen aus dem unüberschaubaren Angebot als vertrauenswürdig gelten und bei welchen Fragestellungen es sinnvoll wäre, weiteren Expertenrat einzuholen, um über deren Akzeptanz zu entscheiden (Bromme et al. 2014; Thomm und Bromme 2012).

- Ein anderes Beispiel (nicht nur im Internet) ist die Unterscheidung von normativen Kontroversen und innerwissenschaftlichen, z. B. methodisch bedingten, Kontroversen. Die oben skizzierten Fallstudien haben die Wichtigkeit dieser Unterscheidung für den Bildungsdiskurs gezeigt. Es ist aber derzeit nicht bekannt, wie wissenschaftliche Kontroversen so dargestellt werden können, dass die Rezipienten diese Unterscheidung wahrnehmen. Ebenso ist bislang unbekannt, welche Rolle es für die öffentliche Wahrnehmung von wissenschaftlicher Evidenz spielt, wenn Wissenschaftler, die in der Öffentlichkeit auftreten, solche Unterscheidungen ignorieren.

d. Die Rolle von Medien und Journalismus für die Rezeption und Diskussion der empirischen Bildungsforschung.
   - Aus der empirischen Forschung zur Wissenschaftskommunikation ist bekannt, dass man das Handeln von Wissenschaftsjournalisten nicht einfach als „Übersetzung“ wissenschaftlicher Ergebnisse in verständliche Sprache begreifen kann. Die Tätigkeit von Wissenschaftsjournalisten folgt vielmehr einer eigenen journalistischen Logik, die durch den speziellen medialen Kontext (z. B. Zielpublikum, Art des Mediums) und die Entscheidungs- und Selektionsprogramme des Journalismus bestimmt ist (Blöbaum 2008; Kessler et al. 2014, in diesem Heft). Auch hierüber weiß man bislang zu wenig. Bildungsthemen und damit auch die Fragestellungen und Befunde der empirischen Bildungsforschung werden in unterschiedlichen Medien sehr breit und nicht nur im Kontext der Wissenschaftsberichterstattung abgehandelt. Bislang ist aber wenig darüber bekannt, welche Rolle welche Medien (die ja ihrerseits ganz unterschiedlich mit Bildungsthemen umgehen) für welche Öffentlichkeiten spielen. Gerade weil die unterschiedlichen Akteure ganz unterschiedliche Medien nutzen und weil sie zugleich auch Teil mehrerer Öffentlichkeiten sind (z. B. als Bildungspolitiker und als Eltern), wäre es gut, die Wechselwirkungen zwischen den Veröffentlichungen von Wissenschaftlern, bestimmten Medien (z. B. solchen, die Wissenschaftsjournalismus zu Bildungsthemen betreiben) und dem Verständnis von Akteuren (z. B. Elterninitiativen im Kontext von Schulreformen) empirisch begründet besser zu verstehen.

Literatur

Achilles, C. M., Bain, H. P., Bellot, F., Boyd-Zaharias, J., Finn, J., Folger, J., et al. (2008). *Tennessee's student teacher achievement ratio (STAR) project V1*. http://hdl.handle.net/1902.1/10766. Zugegriffen: 18. März 2014.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., et al. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin, 136,* 151–173. doi:10.1037/a0018251.supp.

Arnhold, G. (2005). *Kleine Klassen – große Klasse? Eine empirische Studie zur Bedeutung der Klassengröße für Schule und Unterricht*. Bad Heilbrunn: Klinckhardt.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., et al. (Hrsg.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske & Budrich.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., et al. (Hrsg.). (2002). *PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske & Budrich.

Beelmann, A. (2014). Möglichkeiten und Grenzen systematischer Evidenzkumulation durch Forschungssynthesen in der Bildungsforschung. In R. Bromme & M. Prenzel (Hrsg.), *Von der Forschung zur evidenzbasierten Entscheidung: Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung* (in diesem Heft). Sonderheft der *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer VS.

Bieber, T., Martens, K., Niemann, D., & Windzio, M. (2014). Grenzenlose Bildungspolitik? Empirische Evidenz für PISA als weltweites Leitbild für nationale Bildungsreformen. In R. Bromme & M. Prenzel (Hrsg.), *Von der Forschung zur evidenzbasierten Entscheidung: Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung* (in diesem Heft). Sonderheft der *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer VS.

Biesta, G. J. J. (2010). Why „What Works" still won't work: From evidence-based education to value-based education. *Studies in Philosophy and Education, 29,* 491–503. doi:10.1007/s11217-010-9191-x.

Bisky, L. (2008). Im Fokus von Kulturkritik und Marktinteresse – Computerspiele als massenmedialesProdukt der Populär- und Alltagskultur. In O. Zimmermann & T. Geißler (Hrsg.), *Streitfall Computerspiele: Computerspiele zwischen kultureller Bildung, Kunstfreiheit und Jugendschutz* (S. 44–46). Berlin: Deutscher Kulturrat e. V.

Blöbaum, B. (2008). Wissenschaftsjournalisten in Deutschland. Profil, Tätigkeiten und Rollenverständnis. In H. Hettwer, M. Lehmkuhl, H. Wormer, & F. Zotta (Hrsg.), *WissensWelten. Wissenschaftsjournalismus in Theorie und Praxis* (S. 245–260). Gütersloh: Bertelsmann Stiftung.

Blossfeld, H.-P., Maurice, J. v., & Schneider, T. (2011). The national educational panel study: need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, Sonderheft, 14,* 5–17. doi:10.1007/s11618-011-0178-3.

Böttcher, W., Dicke, J. N., & Ziegler, H. (Hrsg.). (2009). *Evidenzbasierte Bildung. Wirkungsevaluation in Bildungspolitik und pädagogischer Praxis*. Münster: Waxmann.

Bromme, R., & Goldman, S. (Guest Eds.). (2014). Understanding the Public Understanding of Science: Psychological Approaches (Special Issue). *Educational Psychologist*. (im Druck).

Bromme, R., & Kienhues, D. (2014). Wissenschaftsverständnis und Wissenschaftskommunikation (S. 55–81). In T. Seidel & A. Krapp (Hrsg.), *Pädagogische Psychologie* (6 Aufl.). Weinheim: Beltz.

Bromme, R., Thomm, E., & Wolf, V. (2014). From understanding to deference: Laypersons' and medical students' views on conflicts within medicine. *International Journal of Science Education, Part B: Communication and Public Engagement*. doi: 10.1080/21548455.2013.849017.

Bundesministerium für Bildung und Forschung (BMBF). (2007). *Rahmenprogramm zur Förderung der empirischen Bildungsforschung*. Bonn: BMBF.

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research:Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32,* 3–14. doi:10.3102/0013189X032009003.

Bushman, B. J., & Huesmann, L. R. (2014). Twenty-five years of research on violence in digital games and aggression revisited: A reply to Elson & Ferguson (2013). *European Psychologist, 19*(1), 47–55. doi:10.1027/1016-9040/a000164.

Bushman, B. J., Rothstein, H. R., & Anderson, C. A. (2010). Much ado about something: Violent video game effects and a school of red herring: Reply to Ferguson and Kilburn (2010). *Psychological Bulletin, 136,* 182–187. doi:10.1037/a0018718.

Elson, M., & Ferguson, C. J. (2014a). Does doing media violence research make one aggressive? The ideological rigidity of social cognitive theories of media violence and response to Bushman and Huesmann (2013), Krahé (2013), and Warburton (2013). *European Psychologist, 19*(1), 68–75. doi:10.1027/1016-9040/a000185.

Elson, M., & Ferguson, C. J. (2014b). Twenty-five years of research on violence in digital games and aggression: Empirical evidence, perspectives, and a debate gone astray. *European Psychologist, 19*(1), 33–46. doi:10.1027/1016-9040/a000147.

Fend, H. (1982). *Gesamtschule im Vergleich. Bilanz der Ergebnisse des Gesamtschulversuchs*. Weinheim: Beltz.

Fend, H. (4. Januar 2008). Schwerer Weg nach oben. *Die Zeit*. http://www.zeit.de/2008/02/C-Enttaeuschung. Zugegriffen: 18. Februar 2014.

Fend, H., Knörzer, W., Nagl, W., Specht, W., & Väth-Szusdziara, R. (1976). *Gesamtschule und dreigliedriges Schulsystem – eine Vergleichsstudie über Chancengleichheit und Durchlässigkeit*. Stuttgart: Klett.

Ferguson, C. J. (2007a). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior, 12*(4), 470–482. doi:10.1016/j.avb.2007.01.001.

Ferguson, C. J. (2007b). The Good, The bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly, 78,* 309–316. doi:10.1007/s11126-007-9056-9.

Ferguson, C. J. (2008). An evolutionary approach to understanding violent antisocial behavior: Diagnostic implications for a dual-process etiology. *Journal of Forensic Psychology Practice, 8*(4), 321–343. doi:10.1080/15228930802199168.

Ferguson, C. J. (2009). Media violence effects: Confirmed truth or just another X-file. *Journal of Forensic Psychology Practice, 9*(2), 103–126. doi:10.1080/15228930802572059.

Ferguson, C. J., & Kilburn, J. (2010). Much Ado about nothing: The misestimation and overinterpretation of violent video game effects in eastern and western nations: Comment on Anderson et al. (2010). *Psychological Bulletin, 136*(2), 174–178. doi:10.1037/a0018566.

Ferguson, C. J., & Rueda, S. M. (2010). The Hitman studiy. Violent video game exposure effects on aggressive behavior, hostile feelings, and depression. *European Psychologist, 15*(2), 99–108. doi:10.1027/1016-9040/a000010.

Fleischman, S. (2009). User-driven Research in Education: A key element promoting evidence-based education. In W. Böttcher, J. N. Dicke, & H. Ziegler (Hrsg.), *Evidenzbasierte Bildung* (S. 69–82). Münster: Waxmann.

Gewerkschaft Erziehung und Wissenschaft. (2009). *Kleine Klassen Große Klasse! Argumente für einen sachlichen Umgang mit einer umstrittenen Frage*. Frankfurt: Gewerkschaft Erziehung und Wissenschaft.

Gigerenzer, G. (2002). *Das Einmaleins der Skepsis*. Berlin: Berlin Verlag.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*(2), 53–96. doi:10.1111/j.1539-6053.2008.00033.x.

Gollwitzer, M., Rothmund, T., Klimmt, C., Nauroth, P., & Bender, J. (2014).Gründe und Konsequenzen einer verzerrten Darstellung und Wahrnehmung sozialwissenschaftlicher Forschungsbefunde: Das Beispiel der „Killerspiele-Debatte". In R. Bromme & M. Prenzel (Hrsg.), *Von der Forschung zur evidenzbasierten Entscheidung: Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung* (in diesem Heft). Sonderheft der *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer VS.

Gräsel, C. (2010). Stichwort Transfer und Transferforschung. Geheimnisvoller Transfer? *Zeitschrift fuer Erziehungswissenschaft, 13,* 7–20. doi:10.1007/s11618-010-0109-8.

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin, 40*(5), 578–589 (published online 23 January 2014). doi:10.1177/0146167213520459.

Gundlach, E. (2006). *Bildungspolitik im Zeitalter der Globalisierung. Zukunft der Sozialen Marktwirtschaft*. Stuttgart: Lucius & Lucius.

Hanushek, E. A. (1998). *The Evidence on Class Size*. Occasional Paper (98–1), W. Allen Wallis Institute of Political Economy, Universtity of Rochester. http://www.wallis.rochester.edu/WallisPapers/wallis_10.pdf. Zugegriffen: 18. Februar 2014.

Hargreaves, L., Galton, M., & Pell, A. (1998). The effects of changes in class size on teacher–pupil interaction. *International Journal of Educational Research, 29*(8), 779–795. doi:10.1016/S0883-0355(98)00063-9.

Hattie, J. (2009). *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Helmke, A., & Weinert, F. E. (1997). Unterrichtsqualität und Leistungsentwicklung – Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 241–252). Weinheim: Beltz.

Herzog, W. (2010). Die Erziehungswissenschaft am Gängelband der Bildungspolitik. *Zeitschrift für pädagogische Historiographie, 16,* 103–105.

Hosenfeld, I., Helmke, A., Ridder, A., & Schrader, F.-W. (2002). Die Rolle des Kontextes. In A. Helmke & R. Jäger (Hrsg.), *Das Projekt MARKUS – Mathematik-Gesamterhebung Rheinland-Pfalz. Kompetenzen, Unterrichtsmerkmale, Schulkontext* (S. 155–256). Landau: Verlag Empirische Pädagogik.

Huesmann, L. R. (2010). Nailing the coffin shut on doubts that violent video games stimulate aggression: Comment on Anderson et al. (2010). *Psychological Bulletin, 136*(2), 179–181. doi:10.1037/a0018567.

Jäger, M. (2004). *Transfer in Schulentwicklungsprojekten*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kessler, S. H., Guenther, L., & Ruhrmann, G. (2014). Die Darstellung epistemologischer Dimensionen von evidenzbasiertem Wissen in TV-Wissenschaftsmagazinen. Ein Lehrstück für die Bildungsforschung. In R. Bromme & M. Prenzel (Hrsg.), *Von der Forschung zur evidenzbasierten Entscheidung: Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung* (in diesem Heft). Sonderheft der *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer VS.

Killus, D., & Tillmann, K.-J. (2011). *Der Blick der Eltern auf das deutsche Schulsystem. Die 1. JAKO-O-Bildungsstudie*. Münster: Waxmann.

Klieme, E., Jude, N., Baumert, J., & Prenzel, M. (2010a). PISA 2000–2009: Bilanz der Veränderungen im Schulsystem. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 277–300). Münster: Waxmann.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., et al. (Hrsg.). (2010b). *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster: Waxmann.

Kobarg, M., & Prenzel, M. (2009). Der Mythos der nordischen Bildungssysteme. *Zeitschrift für Erziehungswissenschaft, 12,* 597–615. doi:10.1007/s11618-009-0098-7.

Krahé, B. (2014). Restoring the spirit of fair play in the debate about violent video games: A comment on Elson and Ferguson (2013). *European Psychologist, 19*(1), 56–59. doi:10.1027/1016-9040/a000165.

Krapp, A. (1979). *Prognose und Entscheidung*. Weinheim: Beltz.

Kunczik, M., & Zipfel, A. (2005). *Medien und Gewalt. Befunde der Forschung seit 1998*. Berlin: Bundesministerium für Familie, Senioren, Frauen und Jugend.

Lange, H. (2008). Vom Messen zum Handeln: „empirische Wende" der Bildungspolitik? *Recht der Jugend und des Bildungswesens, 56,* 7–15.

Lankes, E.-M., & Carstensen, C. H. (2010). Klassengrößen. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert, & R. Valtin (Hrsg.), *IGLU 2006 – die Grundschule auf dem Prüfstand. Vertiefende Analysen zu Rahmenbedingungen schulischen Lernens* (S. 121–142). Münster: Waxmann.

Maaz, K., Baeriswyl, F., & Trautwein, U. (2011). *Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Eine Studie im Auftrag der Vodafone Stiftung Deutschland*. http://www.vodafone-stiftung.de/pages/publikationen/index.php3?ACTION=MENUEPUNKT&ID=10640&displayText=23407. Zugegriffen: 18. Februar 2014.

Möller, I., & Krahé, B. (2013). *Mediengewalt als pädagogische Herausforderung – Ein Programm zur Förderung der Medienkompetenz im Jugendalter*. Göttingen: Hogrefe.

Müller, K., & Ehmke, T. (2013). Soziale Herkunft als Bedingung der Kompetenzentwicklung. In M. Prenzel, Ch. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012: Fortschritte und Herausforderungen in Deutschland* (S. 245–274). Münster: Waxmann.

Nowotny, H. (1999). *Es ist so. Es könnte auch anders sein*. Frankfurt: Suhrkamp.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD.

OECD. (2003). *New challenges for educational research*. Paris: OECD.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

OECD. (2007). *PISA 2006 Science competencies for tomorrow's world*. Paris: OECD.

OECD. (2008). *Education at a glance 2008 – OECD indicators*. Paris: OECD.

Otto, H.-J. (2008). Deutscher Verbotsaktionismus schadet der kulturellen Vielfalt – Das Beispiel Computerspiele. In O. Zimmermann & T. Geißler (Hrsg.), *Streitfall Computerspiele: Computerspiele zwischen kultureller Bildung, Kunstfreiheit und Jugendschutz* (S. 41–43). Berlin: Deutscher Kulturrat e. V.

Pant, H. A. (2014). Aufbereitung von Evidenz für bildungspolitische und pädagogische Entscheidungen: Metaanalysen in der Bildungsforschung. In R. Bromme & M. Prenzel (Hrsg.), *Von der Forschung zur evidenzbasierten Entscheidung: Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung* (in diesem Heft). Sonderheft der *Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer VS.

Peters, H. P. (2012). Scientific sources and the mass media: Forms and consequences of medialization. In S. Rödder, M. Franzen & P. Weingart (Hrsg.), *The sciences' media connection – public communication and its repercussions* (S. 217–240). Dordrecht, NL: Springer.

Popper, K. (2007). *Logik der Forschung* (3 Aufl.). Berlin: Akademie Verlag.

Prenzel, M. (2005). Zur Situation der Empirischen Bildungsforschung. In H. Mandl & B. v. Kopp (Hrsg.), *Impulse für die Bildungsforschung. Stand und Perspektiven. Dokumentation eines Expertengesprächs. Deutsche Forschungsgemeinschaft* (S. 7–21). Berlin: Akademie Verlag.

Prenzel, M. (2010). Geheimnisvoller Transfer? Wie Forschung der Bildungspraxis nützen kann. *Zeitschrift für Erziehungswissenschaft, 13,* 21–37.

Prenzel, M. (2012). Empirische Bildungsforschung morgen: Reichen unsere bisherigen Forschungsansätze aus? In M. Gläser-Zikuda, T. Seidel, C. Rohlfs, A. Gröschner, & S. Ziegelbauer. (Hrsg.), *Mixed Methods in der empirischen Bildungsforschung* (S. 273–286). Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., et al. (Hrsg.). (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland -Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., & Schiefele, U., et al. (Hrsg.). (2005). *PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche*? Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Schiefele, U., et al. (Hrsg.). (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Münster: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (Hrsg.). (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (Hrsg.). (2008). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich*. Münster: Waxmann.

Sälzer, Ch., Prenzel, M., & Klieme, E. (2013a). Schulische Rahmenbedingungen der Kompetenzentwicklung. In M. Prenzel, Ch. Sälzer, E. Klieme, & O. Köller (Hrsg.), *PISA 2012: Fortschritte und Herausforderungen in Deutschland* (S. 155–187). Münster: Waxmann.

Sälzer, Ch., Reiss, K., Schiepe-Tiska, A., Prenzel, M., & Heinze, A. (2013b). Zwischen Grundlagenwissen und Anwendungsbezug: Mathematische Kompetenz im internationalen Vergleich. In M. Prenzel, Ch. Sälzer, E. Klieme, & O. Köller (Hrsg.), *PISA 2012: Fortschritte und Herausforderungen in Deutschland* (S. 47–97). Münster: Waxmann.

Schaarschmidt, U. (Hrsg.). (2005). *Halbtagsjobber? Psychische Gesundheit im Lehrerberuf – Analyse eines veränderungsbedürftigen Zustandes* (2 Aufl.). Weinheim: Beltz.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects: Using experimental and observational methods*. Washington, D.C.: AERA.

Schöps, K., Walter, O., Zimmer, K., & Prenzel, M. (2006). Disparitäten zwischen Jungen und Mädchen in der mathematischen Kompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 209–224). Münster: Waxmann.

Schrader, F.-W., Helmke, A., Hosenfeld, I., & Ridder, A. (2001). Klassengröße und Mathematikleistung. *Empirische Pädagogik, 15*(4), 601–625.

Schümer, G., & Weiß, M. (2008). *Bildungsökonomie und Qualität der Schulbildung. Kommentar zur bildungsökonomischen Auswertung von Daten aus internationelen Schulleistungsstudien*. Frankfurt: Gewerkschaft Erziehung und Wissenschaft.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. doi:10.3102/0034654307310317.

Shavelson, R. J., & Towne, L. (2002). *Scientific research in education*. Washington, D.C.: NationalAcademy Press.

Slavin, R. E., Groff, C., & Lake, C. (2009a). Effective programs in middle and high school mathematics: A best evidence synthesis. *Review of Educational Research, 79*(2), 839–911. doi:10.3102/0034654308330968.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009b). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79,* 1391–1466.

Stadtler, M., Bromme, R., & Rouet, J.-F. (2014). „Science meets Reading“: Worin bestehen die Kompetenzen zum Lesen multipler Dokumente zu Wissenschaftsthemen und wie fördert man sie? *Unterrichtswissenschaft, 42,* 55–68.

Statistische Ämter des Bundes und der Länder. (Hrsg.). (2013). *Internationale Bildungsindikatoren im Ländervergleich*. Wiesbaden: Statistisches Bundesamt.

Thomm, E., & Bromme, R. (2012). „It should at least seem scientific!“ Textual features of „scientificness“ and their impact on lay assessments of online information. *Science Education, 96*(2), 187–211. doi:10.1002/sce.20480.

Warburton, W. (2014). Apples, oranges and the burden of proof: Putting media violence findings in context. *European Psychologist, 19*(1), 60–67. doi:10.1027/1016-9040/a000166.

Weingart, P. (2005). *Die Wissenschaft der Öffentlichkeit*. Weilerswist: Velbrück.

Yeh, S. S. (2009). Class size reduction or rapid formative assessment? A comparison of cost effectiveness. *Educational Research Review, 4*(1), 7–15. doi:10.1016/j.edurev.2008.09.001.