Z Erziehungswiss (2016) (Suppl 1) 19:215–253
DOI 10.1007/s11618-016-0704-4

ZfE CrossMark

Leistungen, Leistungsfähigkeit und Leistungsgrenzen der empirischen Bildungsforschung

Das Beispiel von Large-Scale-Assessment-Studien zwischen Wissenschaft und Politik

Jürgen Baumert

Online publiziert: 22. September 2016

Zusammenfassung Der Beitrag untersucht Leistungen und Leistungsgrenzen empirischer Bildungsforschung im Spannungsfeld zwischen Wissenschaft und Politik am Beispiel von *Large-Scale-Assessment-Studies* (LSA). Im metatheoretischen Rahmen differenter Handlungslogiken von Wissenschaft und Politik und unter Bezugnahme auf Goldthorpe's (2001) Konzeption von Verursachung als generativem Prozess werden die Leistungen von LSA auf den Feldern der theoretischen und empirischen Rekonstruktion von bereichsspezifischen Leistungsdispositionen, der Beschreibung und Erklärung sozialer und ethnischer Disparitäten und der Identifikation von jungen Menschen mit einem besonders hohen Risiko der gesellschaftlichen Exklusion beschrieben. Dabei wird die Frage diskutiert, ob es in sozial-kommunikativen Kontexten theoretisch und empirisch sinnvoll ist, unterschiedliche Wissensformen – deskriptiv-analytische Rekonstruktion des Phänomens und unterschiedliche Modelle kausaler Erklärung – nach politischer Handlungsrelevanz zu unterscheiden. Der Beitrag thematisiert das Problem, wie die Kommunikation zwischen Wissenschaft und Politik trotz unterschiedlicher Funktionsrationalität auf Dauer gestellt werden kann, und macht auf die Gefahr von Grenzüberschreitungen und – damit verbunden – von nicht einlösbaren Leistungsversprechen seitens der Wissenschaft aufmerksam.

Prof. Dr. J. Baumert (✉)
Max-Planck-Institut für Bildungsforschung, Lentzeallee 94, 14195 Berlin, Deutschland
E-Mail: sekbaumert@mpib-berlin.mpg.de

Schlüsselwörter Empirische Bildungsforschung · Large-Scale-Assessment-Studien · Politikberatung · Evidenzbasierung · Erklärungswissen

Large Scale Assessment-Studies Between Science and Politics

The Contributions and Limitations of Empirical Educational Research

Abstract This article investigates contributions and limitations of empirical educational research at the interface between science and politics using the example of *large scale assessment-studies* (LSA). Within the meta-theoretical frame of the divergent rationalities of science and politics and with reference to Goldthorpe's (2001) concept of causation as a generative process, it describes the contributions of LSA in three fields: the theoretical and empirical conceptualization of domain-specific achievement dispositions, the description and explanation of social and ethnic disparities, and the identification of adolescents at risk of social exclusion at the transition to vocational training and the labor market. With reference to these examples, the article discusses whether it is theoretically and empirically advisable in social sciences to distinguish different forms of scientific knowledge according to their relevance for political decision-making. The article addresses the problem of how long-term communication can be established between science and politics despite the differences in their functional rationality and draws attention to the risk of not taking into account these differences and making promises that the social sciences cannot fulfill.

Keywords Empirical educational research · Large scale assessment-studies · Evidence based politics · Causal explanation

1 Einleitung und Arbeitsdefinitionen

Die nationale und internationale Diskussion über das Verhältnis von Bildungsforschung und Bildungspolitik einerseits und Bildungsforschung und Bildungspraxis andererseits ist kontrovers und insbesondere in Deutschland nicht einfach zu ordnen (vgl. Schrader 2014 und den Beitrag von Tillmann in diesem Band). Die Diskussion wird mit wissenschaftstheoretischen, methodischen, normativ-bildungstheoretischen, politischen, wissenschafts- und disziplinpolitischen und gelegentlich auch mit moralischen Argumenten geführt. Sie unterscheidet sich strukturell vom üblichen innerwissenschaftlichen Diskurs forschender Disziplinen, auch wenn es zurückhaltende und auf Konsens angelegte Versuche der Rückbindung gibt (z. B. Shavelson und Towne 2002 für das *National Research Council [NRC]*). Wenn man angesichts dieser Situation etwas über die Leistungen und die Leistungsgrenzen der empirischen Bildungsforschung – und zwar am Beispiel der *Large-Scale-Assessment*-Studien (LSA) – sagen und sich auf diese Weise auch mit Kritiken, die in diesem Band vorgetragen wurden, auseinandersetzen will , ist man gut beraten, zunächst eine Verständigung über die Sache, die verhandelt werden soll, herbeizuführen. Dies soll im Folgenden im ersten Schritt geschehen. Im zweiten Schritt soll versucht werden, im Rekurs auf die unterschiedlichen Handlungslogiken von Politik und Wissenschaft einen

metatheoretischen Rahmen zu skizzieren, in dem die Leistungsfähigkeit, vor allem aber auch die Leistungsgrenzen empirischer Bildungsforschung beschrieben werden können. Im dritten Schritt soll die Ausdifferenzierung von unterschiedlichen Wissenstypen und ihre vermeintlich unterschiedliche Handlungsrelevanz im Anschluss an Goldthorpe's (2001) handlungstheoretischer Konzeption von Verursachung als eines generativen Prozesses diskutiert werden. Dies wird – hoffentlich – den Raum schaffen, um auch unterbewertete Leistungen der empirischen Bildungsforschung, insbesondere der *Large-Scale-Assessment*-Studien, an einigen typischen Beispielen würdigen zu können und die Übertragung von Erkenntnishierarchien (Higgins und Green 2008; vgl. Bromme et al. 2014; S. 12) in den Bereich der Erforschung sozialer Phänomene und Prozesse zumindest infrage zu stellen. Zum Abschluss soll noch einmal die Frage des Verhältnisses von Bildungsforschung und Bildungspolitik konstruktiv aufgenommen werden.

Als Arbeitsdefinition empirischer Bildungsforschung soll ein Vorschlag der Senatskommission „Impulse für die Bildungsforschung" der Deutschen Forschungsgemeinschaft (DFG) übernommen werden (Mandl und Kopp 2005). Danach ist empirische Bildungsforschung ein interdisziplinäres Forschungsfeld, das „die Voraussetzungen, Prozesse und Ergebnisse von Bildung über die Lebensspanne innerhalb und außerhalb von Institutionen" untersucht (Prenzel 2005). Bildungsforschung untersucht Bildungsprozesse – einschließlich ihrer Ziele und Ergebnisse – also nicht nur auf individueller Ebene, sondern auch in sozialen Zusammenhängen, die eine Mehrebenenstruktur aufweisen und von der sozialen Nahumwelt wie der Familie und dem Freundeskreis über institutionelle Kontexte bis zu gesamtgesellschaftlichen Zusammenhängen und ihren Veränderungen reichen. In einer diachronen Perspektive schließt empirische Bildungsforschung auch historische Fragestellungen ein. In der Forschungspraxis behandelt die empirische Bildungsforschung in der Regel spezifische Fragen, die im Anschluss an die einschlägige Forschungsliteratur und oftmals auch in Referenz zu politischen oder praktischen Problemlagen theoretisch entwickelt und begründet werden (vgl. Köller 2014). Diese Definition von Bildungsforschung ist bewusst breit gewählt und enthält weder thematische noch methodische Präferenzen.

Bildungsforschung in dieser Konzeption impliziert einen normativ offenen und für empirische Forschung anschlussfähigen Bildungsbegriff. Im Anschluss an Tenorth (1994, 2004; vgl. auch den Beitrag von Tenorth in diesem Band) soll im Folgenden unter Bildung der individuelle, aber sozial und gesellschaftlich gerahmte und im sozialen Austausch sich vollziehende Prozess der Sicherung der Voraussetzungen für gesellschaftliche Teilhabe, der Kultivierung von Lernfähigkeit und der Selbstkonstruktion der Identität im Lebenslauf verstanden werden. Diese Definition ist mit einem psychologischen Entwicklungsverständnis, nach der menschliche Entwicklung eine individuelle und soziale Ko-Konstruktionsleistung darstellt, kompatibel (Bronfenbrenner 1979; Lerner 1984; Baltes et al. 1998). Sie ist auch offen für nichtlineare und multidirektionale Entwicklungsverläufe. Die leitende Fragestellung empirischer Bildungsforschung ist also: „Wie ist Bildung in ihrer historischen Ausprägung rekonstruierbar und möglich?", nicht aber: „Wie soll sie sein?".

Im vorliegenden Beitrag sollen Leistungen, Leistungsfähigkeit und Leistungsgrenzen von *Large-Scale-Assessment*-Studien als Teil der empirischen Bildungsfor-

schung diskutiert werden. Dies verlangt ein gemeinsames Verständnis über das, was unter LSA verstanden werden soll. Wir wählen eine Arbeitsdefinition, nach der unter *Large-Scale Assessments* Untersuchungen subsumiert werden, die (1) domänenspezifische kognitive Leistungen nach gültigen psychometrischen Standards erfassen, (2) auf Stichproben beruhen, die für ausgewählte Altersgruppen und größere Gebietseinheiten repräsentativ sind und ausreichend Testpower bieten, um Befunde praktischer Bedeutsamkeit zufallskritisch abzusichern, und die (3) Durchführungs- und Auswertungsobjektivität gewährleisten. LSA-Studien können als Querschnitt- oder Längsschnittuntersuchungen angelegt und als Beobachtungsstudien, quasi-experimentelle Untersuchungen oder als randomisierte Feldexperimente konzipiert sein. International vergleichende und auf Dauerbeobachtung angelegte Untersuchungen wie *Program for International Student Assessment* (PISA) der *Organisation for Economic Cooperation and Development* (OECD), *Progress in International Reading Literacy Study* (PIRLS) oder *Trends in International Mathematics and Science Study* (TIMSS) der *International Association for the Evaluation of Educational Achievement* (IEA) sind ebenso wie die regelmäßigen Überprüfungen der Bildungsstandards in Deutschland (BISTA) Spezialfälle von LSA. Die flächendeckenden Vergleichsarbeiten der Bundesländer (VERA) gehören mit Ausnahme der Erhebungen in Hamburg nicht zu den LSA.

Tab. 1 gibt einen Überblick über die wichtigsten in den vergangenen beiden Dekaden in Deutschland oder auch in Deutschland durchgeführten LSA. Die Übersicht zeigt in der Tat eine bemerkenswerte Aktivität auf diesem Forschungsfeld, die umso erstaunlicher ist, wenn man sich an die Forschungsabstinenz auf diesem Gebiet vor allem in den 1980er- und auch noch in den 1990er-Jahren erinnert. Auffällig ist ferner, dass die Zahl der auf Dauer gestellten Monitoring-Studien relativ gering ist. Es handelt sich um PISA, das internationale Grundschul-Monitoring PIRLS und TIMSS sowie die Überprüfung der Bildungsstandards in der Grundschule und der Sekundarstufe (BISTA). Gleichzeitig wurde vor allem in den vergangenen zehn Jahren eine erstaunliche Anzahl von Längsschnittstudien begonnen, die teils abgeschlossen wurden, teils noch fortgeführt werden, die eine bisher nicht verfügbare Datengrundlage für die Analysen von Bildungsprozessen bieten. Es steht außer Frage, dass es sich hier um ein hoch aktives und auch publikationsstarkes Forschungsgebiet handelt. Dennoch ist dieses Forschungsgebiet in der Bildungsforschung nicht dominant, andere Forschungsgebiete können einen ähnlichen Anstieg der Dynamik in den letzten zehn bis zwanzig Jahren verzeichnen. Dazu gehören u. a. die generische und vor allem auch die domänenspezifische Unterrichtsforschung, die ein Gebiet der Zusammenarbeit von Fachdidaktiken und der pädagogischen Psychologie ist, die Untersuchung von Entwicklungsprozessen im frühen Kindes- und Vorschulalter, die psychologische Altersforschung, die Professionsforschung, aber auch die qualitative und quantitative Institutionsforschung. In all diesen Bereichen werden auch *Large-Scale-Assessment*-Studien als Instrumente benutzt, sie machen jedoch nicht den Kern der Forschungstätigkeit aus. Die in Deutschland auf dem Gebiet der Bildungsforschung zu verzeichnende Steigerung der Forschungsintensität, der wissenschaftlichen Produktivität und des internationalen Einflusses ist kein Proprium von *Large-Scale Assessments* (Jones et al. 2010; Köller 2014; Botte et al. 2015; Schui und Krampen 2015; Hannah Greenbaum et al., 2016).

2 Innerwissenschaftliche Evidenz als Kriterium der Beurteilung von Forschungsleistungen

Empirische Forschung erzeugt Daten im Rahmen von Theorien, Modellen, theoretischen Fragestellungen und Hypothesen. Zu innerwissenschaftlicher „Evidenz" werden – wie Bromme et al. (2014) richtig herausstellen – Befunde dann, wenn sie im Licht der theoretischen Fragestellungen für oder gegen Annahmen interpretiert werden. „In diesem Sinne gibt es keine Evidenz ‚an sich', sondern nur Evidenz ‚für' oder ‚gegen' Aussagen oder Vermutungen." (Bromme et al. 2014, S. 7). Dies gilt auch für rein deskriptive Befunde, die Sinn und Bedeutung erst durch ihre Interpretation in einem wie auch immer elaborierten konzeptuellen Rahmen erhalten. Innerwissenschaftliche Evidenzen sind Geltungsbehauptungen. Dies impliziert auch, dass nur Fragestellungen Gegenstand empirischer Forschung sein können, die prinzipiell an der Realität scheitern können. Wie belastbar eine Evidenz oder Geltungsbehauptung ist, entscheidet sich im innerwissenschaftlichen, kommunikativen Validierungsprozess. Der Validierungsprozess beginnt mit der Publikation und der Überwindung unterschiedlich hoher Zugangshürden zu Publikationsorganen, deren Abstufung und Unterschiedlichkeit in der Regel auch Außenseiterpositionen Chancen auf Veröffentlichungen eröffnen, wenn auch die Rezeptionschancen variieren. Der eigentliche Validierungsprozess vollzieht sich dann in der Rezeption, die im Modus des Anschlusses und der Kritik – in der Regel in Kombination von beidem – oder auch der Nichtbeachtung erfolgt. Die dem wissenschaftlichen Betrieb zugrunde liegende Handlungslogik lässt sich im Anschluss an Luhmann (1990) knapp als Suche nach Wahrheit unter den Bedingungen institutionalisierten Zweifels beschreiben. Im praktischen Forschungsprozess stehen Innovativität und damit der Zweifel besonders hoch im Kurs. Das kann für Außenstehende den Eindruck erwecken, dass Wissenschaft im Wesentlichen eine Sammlung konkurrierender Geltungsbehauptungen sei (und Politik und mediale Öffentlichkeit haben gelernt, damit in kritischer Attitüde gegenüber der Wissenschaft in ihrer eigenen Handlungslogik zu spielen). Für Kohärenz und Kumulativität sorgen jedoch im Hintergrund der Anschluss an und die Auseinandersetzung mit der vorgängigen Forschungslage (vgl. Bromme et al. 2014). Besonders in den Sozial- und Humanwissenschaften stellen Forschungsreviews Haltepunkte dar, an denen Kumulativität sichtbar wird.

Maßstab der innerwissenschaftlichen Bewertung von Untersuchungen ist ihr Beitrag zur Erkenntnisproduktion, der diskursiv und in Referenz auf wissenschaftliche Regeln und Methoden im Rezeptionsprozess ermittelt wird. Dies gilt im Prinzip auch für angewandte Forschung oder nutzeninspirierte Grundlagenforschung *(use inspired basic research)* (Stokes 1997; Mandl und Kopp 2005; Schrader 2014). Hier erfolgt die Selektion der Fragestellung nicht allein und oftmals auch nicht primär an der innerwissenschaftlichen Forschungslage, sondern auch unter politischen, sozialen oder praktischen Relevanzgesichtspunkten. Dennoch wird über die Qualität der Forschung auch in diesen Fällen im innerwissenschaftlichen Diskurs des Validierungsprozesses entschieden.

Wählt man diese innerwissenschaftliche Perspektive für die Beurteilung der Leistungen empirischer Bildungsforschung, lassen sich leicht Forschungsgebiete identifizieren, in denen in den letzten beiden Jahrzehnten systematische und kumulative

Tab. 1 Wichtige *Large-Scale-Assessment*-Studien[1]

Querschnitte Deutschland: Programme	Querschnitte Deutschland: Einzelstudien	Längsschnitte Deutschland: Erweiterung von Querschnitten	Längsschnitte Deutschland: Einzelstudien und Programme	Längsschnitte Regionen: Einzelstudien
BISTA	Civic Education	COACTIV	DESI	BIJU
PIRLS/IGLU	LEO-Level One	PIAAC-L	NEPS	BIKS
PISA	PIAAC	PISA-Plus	–	BILWISS
TIMSS (Grundschule)	Reading Literacy	TIMSS-II Deutschland	–	COACTIV-R
–	TEDS-M	–	–	ELEMENT
–	TIMSS-II (Mittelstufe)	–	–	KESS
–	TIMSS-III (Oberstufe)	–	–	KOALA
–	UEBERGANG	–	–	LAU
–	–	–	–	LEK
–	–	–	–	LEK-R
–	–	–	–	LISA
–	–	–	–	PALMA
–	–	–	–	Sommercamp Bremen
–	–	–	–	TOSCA
–	–	–	–	TRAIN

[1] Auflösung der Akronyme s. Abkürzungen

Erkenntnisgewinne zu verzeichnen sind, an denen auch die Forschung in Deutschland Anteil hat (vgl. Köller 2014). Es sollen drei Felder beispielhaft genannt werden. Das erste Gebiet sind die LSA selbst, die an dieser Stelle nur gestreift werden. Hier geht erhöhte Sensibilität für die theoretische Fundierung der erfassten Konstrukte mit wachsender Kenntnis der methodischen Probleme von Trendmessungen mit komplexen Rotationsdesigns einher. Die wichtigsten Leistungen sind wohl die fachdidaktische Fundierung insbesondere der neueren Tests (z. B. BISTA), der Transport von Papier-und-Bleistift-Instrumenten auf Computerplattformen und die Entwicklung intelligenter Aufgaben, die das Potenzial der Rechner nutzen, sowie eine gezielte und begründete Reduktion der Komplexität von Testdesigns. Im Bereich von Lehren und Lernen und der Unterrichtsforschung – dem zweiten Beispiel – sind mit der Differenzierung von Sicht- und Tiefenstrukturen, der Berücksichtigung der Multikriterialität von Instruktionsprozessen, der Identifikation von Basis-Dimensionen qualitätsvollen Unterrichts und der Ausarbeitung des Konzepts der domänenspezifischen kognitiven Aktivierung sowie dem Lernen in interaktiven Lernumgebungen sichtbare Erkenntnisfortschritte erreicht worden (Klieme et al. 2001; Seidel und Shavelson 2007; Helmke 2009; Kunter und Trautwein 2013; Kunter und Voss 2013; Seidel und Reiss 2014; Leutner et al. 2014; Seidel 2014). Dazu haben auch die Fachdidaktiken, insbesondere die Mathematik- und Naturwissenschaftsdidaktik beigetragen. Aber auch auf der Ebene der mikrogenetischen Analyse von Lernprozessen sind z. B. mit den Untersuchungen zum Verständnis naturwissenschaftlicher

Konzepte, dem mathematischen Modellieren, den kognitiven Integrationsleistungen beim Lesen, dem Lernen mit ausgearbeiteten Beispielen oder der Nutzung von Lernstrategien kumulative Erkenntnisgewinne zu verzeichnen, die auch zu einem revidierten Gesamtverständnis von Lehren und Lernen beigetragen haben (Sawyer 2006; Mayer 2008; Renkl 2008; Nückles und Wittwer 2014). Als drittes Gebiet soll die pädagogische Professionsforschung erwähnt werden, die sich in Deutschland zu einem aktiven Forschungsfeld mit internationaler Ausstrahlung entwickelt hat. Im Rahmen eines generischen Modells professioneller Kompetenz hat mittlerweile eine ganze Serie von Untersuchungen zu einem neuen und fundierteren Verständnis der Bedeutung professioneller Kompetenzen für die Qualität des Unterrichts und den Wissenserwerb und die Persönlichkeitsentwicklung von Schülerinnen und Schülern geführt (Baumert und Kunter 2006; Blömeke et al. 2010; Kunter et al. 2013; Kunter und Voss 2013; Blömeke und Delaney 2014; König et al. 2015; Lohse-Bossenz, Kunina-Habenicht, Dicke, Leutner & Kunter 2015; Tillmann 2015; Voss et al. 2015; Blömeke et al. 2016; König und Kramer 2016; König et al. 2016). Gleichzeitig haben diese Forschungsergebnisse aber auch darauf aufmerksam gemacht, wie wenig wir über die Mechanismen wissen, die professionelles Wissen, Überzeugungen und motivationale Orientierungen mit der mentalen Organisation des praktischen Könnens und dem praktischen Handeln verbinden. In der Erziehungswissenschaft gibt es – wie oft – kluge und treffende Kritik (z. B. Neuweg 2015a, 2015b; vgl. auch Herzog in diesem Band), aber wenig Forschung (vgl. aber Oser et al. 2012; Holzberger et al. 2016). Beiträge kommen eher aus der Psychologie und betreffen oft andere Gebiete professionellen Handelns.

3 Wissenschaftliche Befunde und die Logik bildungspolitischen Handelns

Ob Erkenntnisse der empirischen Forschung politische oder praktische Relevanz haben, wird nicht innerwissenschaftlich entschieden, sondern hängt davon ab, ob eine „Evidenz“ in das Aufmerksamkeitsraster der Referenzsysteme fällt und in der Logik politisch-administrativen bzw. pädagogischen Handelns interpretiert und reformuliert und letztlich in die politische Agenda bzw. das praktische professionelle Handeln integriert werden kann. Wichtige Vermittler zwischen Wissenschaft und Politik bzw. Wissenschaft und professioneller Praxis sind die mediale und zivilgesellschaftliche Öffentlichkeit bzw. die akademische und staatliche Aus- und Fortbildung, die politische und professionelle Aufmerksamkeit vorstrukturieren. Wer meint, dass innerwissenschaftliche Evidenz direkte handlungsanleitende Funktion für Politik und Praxis haben könne oder gar die Voraussetzung rationalen Handelns sei, übersieht, dass aus einer wissenschaftlichen Beschreibung oder Erklärung keine eindeutige Konstruktionsregel für praktisches Handeln folgt, sondern bestenfalls Handlungsoptionen vorgezeichnet werden, die je nach den normativen Vorstellungen über das Wünschenswerte und der Beurteilung des Möglichen ganz unterschiedlich bewertet werden können (Bromme und Kienhues 2014). Gleichzeitig übergeneralisiert er/sie die Logik wissenschaftlichen Handelns und verkennt, dass andere gesellschaftliche Subsysteme unterschiedlichen Handlungsrationalitäten folgen.

Folgt man wiederum Luhmann (2000), lässt sich die Handlungslogik der Bildungspolitik als Herstellung kollektiv bindender Entscheidungen beschreiben, und zwar – so muss man hinzufügen – unter den Bedingungen einer volatilen Öffentlichkeit und einer professionellen Praxis, die auf operativer Ebene einer eigenen autonomen Handlungslogik folgt. Dies bedeutet, dass politische Gestaltungsprogramme, auch wenn es sich nur um die Aufrechterhaltung des Status quo handelt, immer auch die Bedingungen des Machterhalts zu berücksichtigen haben, damit nicht als Folge der Durchsetzung von Entscheidungen die demokratische Legitimationsbasis des Handelns selbst in Gefahr gerät (Luhmann 1975). Und selbst bei Umsetzung der Entscheidung ist ungewiss, ob die erwünschte Wirkung auch tatsächlich eintritt. Gleichzeitig stehen alle Gestaltungsmaßnahmen, die nicht nur die institutionellen Rahmenbedingungen pädagogischen Handelns verändern, sondern auch das Handlungsprogramm auf operativer Ebene optimieren wollen, unter der Restriktion strukturell begrenzter Durchsetzbarkeit von bürokratischen Entscheidungen in professionellen Handlungskontexten. Kommunikatives Handeln lässt sich nicht anordnen. Politisches Handeln ist – mit oder ohne Wissenschaft – Handeln unter Unsicherheit. Dabei ist die Bildungspolitik doppelter Beobachtung ausgesetzt: extern durch die mediale und zivilgesellschaftliche Öffentlichkeit, die Themen in schwer berechenbarer Weise politisieren kann, und intern durch eine professionelle Lehrerschaft, die selbst organisiert und kollektiv sprechfähig ist.

Es steht außer Frage, dass Politik und Verwaltung, um überhaupt handlungsfähig zu sein, auf kontinuierliche und vor allem auch auf vorausschauende Information über die Funktion und die Funktionsfähigkeit des Bildungssystems angewiesen sind. Dazu gehören auch wissenschaftliche Informationen, vor allem dann, wenn im Wissenschaftssystem selbst Forschungsergebnisse zur Leistungsfähigkeit des Bildungssystems erzeugt werden, die öffentliches Interesse finden und politisiert werden können (vgl. Tillmann et al. 2008; Kuhlmann 2012). In den vergangenen Jahrzehnten haben in Deutschland die Bildungsverwaltungen aller Länder Systeme der quantitativen und qualitativen Dauerbeobachtung und damit verbunden der Qualitätssicherung institutionalisiert. Dass Bildungsverwaltungen von demografischen Schwankungen oder Veränderungen der Bildungsnachfrage überrascht werden, wie dies noch in den 1960er-Jahren geschehen konnte, ist heute schwer vorstellbar. Die Einheiten können als Abteilungen oder Referate in die Ministerialverwaltung selbst integriert, in nachgeordnete Dienststellen ausgelagert oder in selbstständigen wissenschaftlichen Einrichtungen, die aber staatlicher Kontrolle unterliegen – wie z. B. das IQB in Berlin oder einige Staatsinstitute der Länder –, institutionalisiert sein. Dabei kann die Bildungsverwaltung ihre Informationen auch selbst im Modus der Wissenschaft erzeugen oder erzeugen lassen. Die Kommunikation mit Wissenschaft ist unterschiedlich organisiert, in jedem Fall aber auf Dauer gestellt. In den Planungs- und Qualitätssicherungseinheiten werden potenziell steuerungsrelevante Informationen aufbereitet, fortgeschrieben und *politisch* interpretiert (vgl. Tillmann et al. 2008). Damit verfügen die Länderverwaltungen über Schnittstellen zur Wissenschaft, die in ihrer Selektion von Themen und der Rekontextualisierung von Befunden an die Funktionslogik des politischen Systems gebunden sind, ohne auf die Durchführungsstandards qualitätsvoller Forschung verzichten zu müssen.

Darüber hinaus haben Bund und Länder nach der Föderalismusreform I im Jahre 2006 auf der Grundlage des neuen Artikels 91b Abs. 2 GG ein Bundesländer übergreifendes und international anschlussfähiges Instrument der Dauerbeobachtung des Bildungssystems geschaffen, das gleichzeitig die Kommunikation zwischen Bund, Ländern und der Bildungsforschung in einer weit wirksameren Weise verstetigt, als es die Bund-Länder-Kommission (BLK) im Rahmen der Gemeinschaftsaufgaben jemals zuvor geleistet hatte (Verwaltungsabkommen vom 27. Mai 2007; BAnz. S. 5863). Mit der Einrichtung einer gemeinsamen „Steuerungsgruppe" wurde ein organisatorischer Ort für eine relativ systematische Verständigung zwischen Politik und Wissenschaft über die Selektion wichtiger Probleme und Fragestellungen, die Ordnung und Integration von Wissen und die Identifikation von Wissenslücken gefunden. Diese Verständigung findet ihren Niederschlag auch in der Auswahl und Fortschreibung der Indikatoren und den thematischen Schwerpunktsetzungen des Nationalen Bildungsberichts (Baumert und Füssel 2012).

Insgesamt ist hier ein Kommunikationssystem zwischen Politik und Wissenschaft entstanden, das, wie Tenorth (2014, 2015) in seiner kleinen Geschichte der politischen Beratung feststellt, die fallbezogene Beratung durch in der Regel gemischte Kommissionen nicht ersetzt, sondern in ein neues, weitaus komplexeres Netzwerk einfügt. Die bemerkenswerte Funktionsfähigkeit dieses Systems beruht im Kern auf der beiderseitigen Anerkennung der Differenz der Handlungslogiken von Politik und Wissenschaft und damit der Akzeptanz unterschiedlicher Kontextualisierung von Problemen und Befunden. Politik und Wissenschaft können sich über die Relevanz von Problemstellungen verständigen, prozedieren jedoch für die Erkenntnisgewinnung und die Erkenntnisnutzung im jeweils eigenen Rationalitätsmodus. Belehrung – auch in Gestalt der in der Erziehungswissenschaft beliebten „Kritik" (vgl. Bellmann 2015; Heinrich 2015) – ist hier keine Form erfolgreicher, d. h. anschlussfähiger Kommunikation.

Die Akzeptanz unterschiedlicher Handlungslogiken impliziert aber auch die Anerkennung der prinzipiellen Öffentlichkeit wissenschaftlicher Erkenntnis. Dies bedeutet, dass immer mit der medialen Öffentlichkeit als einem dritten Mitspieler zu rechnen ist, der Befunde nach eigenen Relevanzgesichtpunkten selektiv wahrnimmt und interpretiert und auch interpretieren darf. Dies ist eine dauerhafte Quelle potenzieller politischer Dynamik, die je nach Betroffenheit nicht nur Freude bei politisch Handelnden auslöst. Die Situation wird komplexer. Dies legt die Versuchung nahe, der Wissenschaft, die zur Steigerung der Komplexität beigetragen hat, zumindest rhetorisch auch (Mit-)Verantwortung für die Lösung des Problems zuzuweisen z. B. in der Erwartung, endlich „abgesichertes und anwendbares Steuerungswissen" (Meyer-Hesemann 2008) zur „rationalen Begründung … bildungspolitischer Entscheidungen" (BMBF 2007, S. 4) zu erzeugen.

Wissenschaftliche Befunde ersetzen aber keine politischen Entscheidungen und geben keine linearen Handlungsanleitungen. Auf jede innerwissenschaftliche Evidenz gibt es mindestens zwei und in der Regel mehrere politische oder praktische Antworten. Sozialwissenschaftliche Befunde können Aufmerksamkeit vorstrukturieren, möglicherweise auch orientieren und bestenfalls Optionen unter der Perspektive von Umsetzbarkeit, möglichen Folgen und Nebenfolgen beschreiben. Darüber hinaus sichert empirische Evidenz in keinem Fall die praktische Umsetzung von politi-

scher Entscheidung in einem professionellen Handlungssystem. Im Grunde ist allen Akteuren im politischen System dieser Sachverhalt klar, auch wenn in politischen Handlungsprogrammen der empirischen Bildungsforschung mehr zugemutet wird. So formuliert etwa Meyer-Hesemann, der sich für die Bundesländer Handlungswissen wünscht, auch: „‚Wissen für Handeln' darf nicht die falsche Erwartung wecken, wissenschaftlich abgesichertes Steuerungswissen ließe sich unmittelbar kraft Rationalität der Erkenntnis in den Beschluss von bildungspolitischen Maßnahmen umsetzen. Diese Erwartung ist naiv, denn sie verkennt die vollkommen unterschiedlichen handlungsbestimmenden Koordinaten politischen Handelns und wissenschaftlichen Arbeitens" (Meyer-Hesemann 2008, S. 13). Für den Hamburger Staatsrat Lange, der im Rahmen der KMK die Arbeit der Amtschefs der Länder lange Jahre koordinierte und den Begriff der „empirischen Wende der Bildungspolitik" (Lange 2008) erfunden hatte, war immer klar, dass man wissenschaftliche Befunde wie Seekarten oder Seewetterberichte nutzen kann, um das „Staatsschiff zu segeln", diese aber weder das Ziel noch den aktuellen Kurs bestimmen und bei schwerer See auch nur begrenzt helfen. Mit der Veröffentlichung seiner Ergebnisse verliert die Wissenschaftlerin/der Wissenschaftler die ausschließlichen Nutzungsrechte an seinen nicht patentierten Ergebnissen, auch wenn er die Urheberrechte behält, wie Tillmann et al. (2008) für die politische PISA-Rezeption sehr schön gezeigt haben. Umgekehrt kann die politische Seite von der Bildungsforschung Sensitivität für gesellschaftliche Problemlagen, kommunikative Verständigung über politisch relevante und wissenschaftlich untersuchbare Fragestellungen, die dann nach den wissenschaftlichen Regeln bearbeitet werden, oder auch die systematische Integration von Forschungsbefunden zu politisch bedeutsamen Handlungsfeldern erwarten. Alles was darüber hinausgeht, ist Zumutung.

Wie kommt es aber dann zu politischen Programmen, die von der empirischen Bildungsforschung verlangen, „... belastbare wissenschaftliche Informationen zu generieren, die eine rationale Begründung bildungspraktischer und bildungspolitischer Entscheidungen ermöglichen" (BMBF 2007, S. 4), oder „... abgesichertes und anwendbares Steuerungswissen zur zentralen Herausforderung, wie eine erhöhte Bildungsqualität bei gleichzeitiger Verbesserung der Bildungschancen erreicht werden kann" (Meyer-Hesemann 2008), zu erzeugen – also Programme, die in verschiedenen Beiträgen dieses Bandes systematisch kritisiert werden (vgl. die Beiträge von Bellmann und Herzog; Bellmann 2006; Bellmann und Müller 2011)? Wissenschaftliche Förderungsprogramme des Bundes haben wie auch Ressortforschung im engeren Sinne politische Funktionen. Aus der Perspektive des Bundes werden mit dem „Rahmenprogramm zur Förderung empirischer Bildungsforschung" (BMBF 2007) Felder erhöhter politischer Aufmerksamkeit definiert, die der Bund für die Qualitätsentwicklung des Bildungssystems auf gesamtstaatlicher Ebene für relevant hält, auch wenn er für die politische Gestaltung im Rahmen der föderalen Kompetenzverteilung nicht oder nicht allein zuständig ist. Insofern ist das Förderprogramm ein Instrument zur indirekten Dynamisierung bildungspolitischer Prozesse, die bei unterschiedlichen föderalen Interessen und Prioritäten gesamtstaatlich nicht vorankommen. Spitz könnte man formulieren, der Bund erwartet mit seinen Förderprogrammen auch Argumente, die helfen können, die Länder dazu zu bringen, das zu tun, was der Bund wünscht und die Länder nicht tun wollen oder nicht tun

können. Wenn von Länderseite nach bislang nicht verfügbarem Handlungswissen gefragt wird – gelegentlich auch zu Problemen, die man wissenschaftlich nicht bearbeiten kann (z. B. Rabe 2013) –, ist dies auch ein politisches Argument, um zu begründen, dass man das nicht tut, was man politisch nicht tun kann. Problematisch wird dies erst, wenn seitens der Wissenschaft darauf mit Versprechen geantwortet wird, durch Änderung der Forschungspraxis und insbesondere durch die Privilegierung bestimmter Forschungstypen „Evidenzen" für politisches Handeln liefern zu wollen, die strukturelle Differenzen zwischen Wissenschaft und Politik überbrücken können.

4 Wissensformen und ihre politische Handlungsrelevanz

Aufseiten der Wissenschaft werden gelegentlich Wissensformen nach unterschiedlicher Affinität zum politischen und praktischen Handeln unterschieden. Klieme (2013, 2014) etwa unterscheidet Diagnose-, Erklärungs- und Handlungswissen. Bromme et al. (2014) differenzieren ganz ähnlich Beschreibungs-, Erklärungs- und Veränderungswissen. Erklärungswissen liefert die Gründe für festgestellte Effekte, während Handlungs- oder Veränderungswissen auf dem Nachweis der Effekte von Manipulationen und Interventionen beruht. Handlungs- oder Veränderungswissen sind danach die Wissensformen, die die größte Nähe zum politischen Entscheiden und zum praktischen Handeln aufweisen. Die Autoren – das muss man betonen – wissen natürlich, dass es keinen direkten Weg weder vom Erklären zum Intervenieren noch von der experimentellen Intervention zum politischen Entscheiden oder praktischen Handeln gibt. So stellen Bromme und Kienhues (2014, S. 60) klar, dass es nur „einen indirekten Zusammenhang zwischen Theorien und Daten, die Sachverhalte beschreiben und erklären (Beschreibungs- und Erklärungswissen), und solchen, die gezielte Veränderungen im Sinne von Interventionen begründen (Veränderungswissen; …)", gebe. Auch Wissenschaftler, die sich in der Problembeschreibung und Problemerklärung einig seien, könnten durchaus zu unterschiedlichen und widersprüchlichen Interventionsempfehlungen kommen. Dies gelte auch für den Schritt vom experimentell erzeugten Veränderungswissen zum politischen oder praktischen Handeln. Es gebe „viele praktische Probleme, die nach wissenschaftsbasierten Lösungen verlangen, [für die] mit *rein* wissenschaftlichen Methoden und Ergebnissen gar keine Lösung gefunden werden kann", weil sie immer auch normative Entscheidungen implizieren (Bromme und Kienhues 2014, S. 61). Dennoch legt die Differenzierung von Wissensformen eine Abstufung der politischen und praktischen Handlungsrelevanz nahe und vernachlässigt dabei, dass bei der Lösung politischer und praktischer Probleme nicht nur normative Entscheidungen zu treffen sind, sondern diese auch unterschiedlichen Handlungslogiken folgen.

Im Hintergrund der Unterscheidung zwischen Erklärungs- und Veränderungswissen stehen zwei unterschiedliche Vorstellungen von Verursachung, die Goldthorpe (2001) in einem grundlegenden Artikel zu Kausalitätsvorstellungen in den Sozialwissenschaften diskutiert. Er spricht von „Verursachung als robuster Abhängigkeit" und von „Verursachung als zu Konsequenzen führender Manipulation". Im ersten Fall werden Erklärungen für regelmäßig nachweisbare Zusammenhänge und im zweiten

Fall der Nachweis von Effekten einer Manipulation gesucht. Goldthorpe analysiert die Leistungsfähigkeit und die Leistungsgrenzen beider Konzeptionen für das Verstehen und Erklären sozialer Phänomene und Zusammenhänge. Die erste Konzeption beruht auf der kovarianzanalytischen Vorstellung, Ursachen für zu erklärende Tatbestände durch schrittweises Auspartialisieren von konfundierten Einflussgrößen eingrenzen zu können. Das in den Sozialwissenschaften weitverbreitete schrittweise Modellfitting ist hier der Prototyp. Goldthorpe weist darauf hin, dass mit einer rein technischen Anwendung von *conditioning* weder das Problem der unbeobachteten Heterogenität gelöst noch ein Beitrag zur Aufklärung der Vermittlungsschritte zwischen vermeintlicher Ursache und dem zu erklärenden Phänomen geleistet werde. Die zweite Konzeption folgt der Maxime „keine Kausalität ohne Manipulation", die Goldthorpe dem kontrafaktischen Modell der Kausalität (Rubin 1974; Holland 1986) unterlegt. Goldthorpe analysiert die Brauchbarkeit dieses Konzepts in sozialen Kontexten, indem er zunächst noch einmal auf die begrenzte Manipulierbarkeit sozialer Tatbestände hinweist. Für die meisten soziologisch interessanten Fragestellungen verbiete sich das Experiment. Tiefer trifft die Kritik, wenn Goldthorpe die Probleme herausarbeitet, die mit dieser Kausalitätsvorstellung verbunden sind, wenn stabile Personenmerkmale und vor allem zielgerichtetes Handeln vernunftbegabter Subjekte als Ursachen betrachtet werden. Der Kern der Kritik an beiden Konzeptionen ist der Einwand, dass die Rolle eines handlungs- und entscheidungsfähigen Subjekts in sozialen Kontexten und die historische Bedingtheit ihres Handelns unberücksichtigt blieben.

Goldthorpe entwickelt als Antwort auf diese Problemlage ein für sozialwissenschaftliche Fragestellungen angemesseneres Modell der Verursachung als eines generativen Prozesses. In diesem Modell beginnt die wissenschaftliche Arbeit mit der theoriegeleiteten dichten Beschreibung des zu erklärenden Phänomens: Der Gegenstand wird theoretisch rekonstruiert. Damit erhält die deskriptiv-analytische Funktion der Sozialwissenschaften, die häufig und zu Unrecht gering geschätzt wird, eine angemessene Bedeutung im Prozess des Verstehens und Erklärens sozialer Phänomene. Es folgt dann der handlungstheoretisch angeleitete Versuch, die individuellen und/oder kollektiven Vermittlungs- und Verarbeitungsschritte zwischen Manipulation und Folge bzw. zwischen Ursache und Effekt zu rekonstruieren, um den generativen Prozess, der zum Auftreten des zu erklärenden Phänomens führt, zu klären und zu verstehen. Auch theoretisch alternative Rekonstruktionen sind möglich und wünschenswert. Daran schließt sich die empirische Prüfung der handlungstheoretischen Narrative an, die je nach Fragestellung experimentell, quasi-experimentell oder durch Anpassung und Vergleich theoretisch konkurrierender Modelle erfolgen kann (zur Verträglichkeit mit dem *counterfactual model* vgl. Winship und Morgan 2007, S. 230 ff). In dieser Konzeption kommen sowohl die prinzipielle Revidierbarkeit und historische Kontingenz sozialwissenschaftlicher Erklärungen als auch das Prinzip der Falsifikation zu ihrem Recht. Fasst man in dieser Weise Verursachung unter einer handlungstheoretischen Perspektive als generativen Prozess auf, gibt es keine Abstufung der Bedeutung von Wissensformen – weder innerwissenschaftlich noch im Hinblick auf die Anwendung sozialwissenschaftlicher Erkenntnis in anderen Sozialsystemen. Mit der Einführung von *agency* – also des handlungs- und entscheidungsfähigen Subjekts als zentraler theoretischer Komponente für die

Erklärung sozialer Sachverhalte ist auch die Vorstellung der Nutzung sozialwissenschaftlicher Erkenntnisse im Modell der technischen Anwendung keine überzeugende Option mehr. Im Folgenden soll Goldthorpe's Modell der Verursachung als eines generativen Prozesses als Rahmen genutzt werden, um Leistungen von *Large-Scale Assessments* in den vergangenen beiden Jahrzehnten an prototypischen Beispielen darzustellen.

5 Deskriptive Zustandsdiagnose, Trendbeschreibungen und konzeptuelle Ordnung des Feldes

LargeScale Assessments haben Ergebnisse von Bildungsprozessen in zentralen Bereichen des Bildungsprogramms operativ beschreibbar und ihre Verteilung in der Population ausgewählter Altersjahrgänge bzw. Jahrgangsstufen des Schulsystems sichtbar gemacht. Damit wurde es möglich, „Bildung" in Deutschland zum ersten Mal auf der Ebene definierter kognitiver Leistungsdispositionen zu thematisieren. Im Rückblick kann man dies durchaus als einen historisch unwahrscheinlichen Schritt zur Transparenz in einem bedeutenden gesellschaftlichen Teilsystem verstehen, in dem sich demokratische Prinzipien nur langsam durchgesetzt haben. Bis vor wenigen Jahren konnte noch der öffentliche Glaube an die Versprechen der Bildungsprogramme Realitätsprüfungen ersetzen. Allein der Gedanke einer systematischen Prüfung war anstößig und mit einer Misstrauenserklärung an Politik und Verwaltung verbunden. Vor diesem Hintergrund kann man die Leistung der LSA mit ihrem Beitrag zur Deskription des Feldes, mit dem sie auf die Frage „Was geschieht?" antworten, oder – etwas anspruchsvoller formuliert – mit der theoretischen und empirischen Rekonstruktion des zu erklärenden Phänomens kaum überschätzen. Dabei handelt es sich um die Darstellung von Zuständen, längerfristigen Entwicklungen und stabilen Zusammenhängen.

Auch Deskription ist nicht voraussetzungslos. Allein die Auswahl des Gegenstandes verlangt eine vorgängige konzeptuelle Vorstellung des zu beschreibenden Phänomens. Mit der theoriegeleiteten, auch normative Optionen einschließende Auswahl von Problemen und Fragestellungen wird aber nicht nur Deskription vorbereitet, sondern implizit auch eine konzeptuelle Ordnung des Feldes vorgenommen, die Aufmerksamkeit vorstrukturiert, und zwar in Abhängigkeit von der Problemstellung wissenschaftlich, öffentlich und politisch. Zu den wichtigen Themen, die durch die LSA angeschlagen wurden, öffentliche und damit auch politische Aufmerksamkeit neu justiert haben und bis heute virulent sind, gehören wahrscheinlich die folgenden:

- die Relativierung des in Deutschland bis zum Ende der Vollzeitschulpflicht erreichten Kompetenzniveaus in basalen Bereichen des Bildungsprogramms im internationalen Vergleich, aber auch im Vergleich mit den Ansprüchen des eigenen Programms und dessen bildungstheoretischer Überhöhung durch ein theoretisch begründetes und empirisch prüfbares Konzept von Basiskompetenzen;
- die Identifikation einer Gruppe von jungen Menschen, die aufgrund unzureichender Basiskompetenzen besonders vulnerabel von gesellschaftlicher Exklusion be-

droht sein könnten, und die Neudefinition von Schulversagen als Versagen der Schule;

- der bis heute unveränderte Befund, dass in Deutschland die Gruppe der hochleistenden Schülerinnen und Schüler trotz eines früh selegierenden Schulsystems relativ schmal besetzt ist, und damit die Verschiebung der Aufmerksamkeit von Hochbegabung auf Hochleistung;
- die Wiederentdeckung sozialer Ungleichheit im Bildungssystem als gesellschaftspolitisches und öffentliches Problem und die Konkretisierung dieser Ungleichheiten nicht nur in Beteiligungs-, sondern auch in Kompetenzmaßen;
- die Legitimierung der Diskussion über Zuwanderung als Tatbestand und zukunftsbedeutsames Disparitätsproblem;
- Leistungsunterschiede zwischen politischen Gebietseinheiten und ihre Bedeutung für das deutsche Berechtigungssystem;
- die Diagnose positiver Entwicklungstrends und ihre möglichen Ursachen.

Fünf dieser Themen sollen im Folgenden teils eingehender, teils kursorisch behandelt werden. Dabei soll auch im Blick bleiben, welcher Typ von Befunden (s. Abschn. 4) besondere öffentliche Aufmerksamkeit gefunden hat und politisch rezipiert wurde.

6 Die Konstitution des Explanandums: Kompetenzen als latente domänenspezifische Leistungsdispositionen

Kompetenzen im engeren Sinne (Weinert 2001) theoretisch als latente domänenspezifische kognitive Leistungsdispositionen aufzufassen (Klieme und Leutner 2006; Klieme et al. 2008) und die latente Fähigkeit der Person und die Schwierigkeitsparameter der Aufgaben, die diese auf manifester Ebene indizieren, in einem mathematischen Modell auf einer gemeinsamen Metrik mit Intervallskalenqualität abzubilden, ist heute üblich. Damit werden Testwerte kriterial verankert und erhalten mit der Beschreibung durch die bedingten Lösungswahrscheinlichkeiten von Aufgaben inhaltliche Bedeutung: Tests haben sprechen gelernt. In der Regel wird in LSA im Rahmen der *Item-Response*-Modelle (IRT-Modelle) ein *Multi-Matrix Sampling* verwendet, bei dem Testteilnehmer systematisch rotierte Untermengen der verfügbaren Testaufgaben bearbeiten. Dieses Design erlaubt es, das zu messende Konstrukt mit einer großen Anzahl von Aufgaben dicht zu beschreiben und gleichzeitig die individuelle Bearbeitungszeit des Tests relativ kurz zu halten. Dies ist heute so selbstverständlich, dass man daran erinnern muss**,** dass bis noch vor gut 15 Jahren IRT-Modelle und das damit verbundene methodische Wissen nur an ganz wenigen Standorten der Erziehungswissenschaft und Psychologie in Deutschland präsent waren. In der Soziologie und in der Ökonomie war dies überhaupt kein Thema[1]. Mit

[1] In den 1980er-Jahren waren dies nur die Arbeitsgruppe um Ingenkamp und Schreiber in Landau und die Arbeitsgruppe um Roeder, Baumert und Schnabel am Max-Planck-Institut für Bildungsforschung in Berlin. In den 1990er-Jahren kamen die Arbeitsgruppe um Lehmann an der Universität Hamburg und die Arbeitsgruppen um Rost und Baumert am Institut für die Pädagogik der Naturwissenschaften (IPN) in Kiel hinzu.

diesen IRT-Modellen war die Voraussetzung geschaffen, bildungstheoretisch, fachdidaktisch oder curricular begründete Konzeptionen domänenspezifischer Fähigkeiten operativ zu beschreiben und empirisch zu überprüfen. Durch die inhaltliche Verankerung der Tests wurde es auch möglich, die stofflichen und kognitiven Ansprüche des in Lehrplänen, Curricula und zugelassenen Lehrbüchern kodifizierten Bildungsprogramms mit den Bildungsergebnissen in den jeweils untersuchten Dimensionen zu vergleichen. Dies war in mancher Hinsicht ein bildungstheoretisches Desillusionierungsprogramm, auch wenn in der Allgemeinen Erziehungswissenschaft gelegentlich noch die enttäuschungsfeste Überzeugung anzutreffen ist, man könne „die Frage nach den wesentlichen ‚Kräften' des Wissens, Handelns, des ästhetischen Sinnes und welttranszendierender Religiosität" reflexiv bearbeiten, auch ohne in der Lage zu sein, „banale Inhalte" im Alter von 15 Jahren richtig zu lesen (Koch 2004, S. 189).

Von den Möglichkeiten der IRT-Modelle wurde im *Literacy*-Konzept von TIMSS und elaborierter von PISA systematisch Gebrauch gemacht. In PISA werden im Anschluss an die angelsächsische *Literacy*-Tradition die Beherrschung der jeweiligen Verkehrssprache in Form von Lesekompetenz, mathematische Modellierungsfähigkeit und naturwissenschaftliches Verständnis als Basisqualifikationen, die Voraussetzung für gesellschaftliche Teilhabe darstellen, privilegiert. Es handelt sich nach dieser bildungstheoretischen Konzeption um Kompetenzen, die jeder Angehörige der nachwachsenden Generation ausnahmslos zu erwerben hat, soll er nicht von gesellschaftlicher Exklusion bedroht sein. Mit dieser Annahme ist die *Literacy*-Konzeption an die deutschsprachige Tradition der *allgemeinen* Bildung prinzipiell anschließbar (vgl. Baumert et al. 2001). Historische Grundlage der allgemeinen Bildung ist die Universalisierung des Zugangs zu formaler Bildung in der modernen Pflichtschule. Mit Tenorth (1994) kann man dann unter allgemeiner Bildung das Versprechen auf die Universalisierung der Prämissen für die Teilhabe an gesellschaftlicher Kommunikation durch die Garantie des Bildungsminimums und die Kultivierung der Lernfähigkeit verstehen (vgl. auch Tenorth in diesem Band). Es steht außer Frage, dass die Beherrschung der Verkehrssprache, insbesondere Lesekompetenz, und mathematische Modellierungsfähigkeit Basisqualifikationen darstellen, die den Zugang zu den symbolischen Gegenständen der Kultur überhaupt erst eröffnen und damit auch die Grundlage jedes selbstständigen Weiterlernens bilden. Es gibt weitere basale „Kulturwerkzeuge" (vgl. Baumert 2002; Bildungskommission 2003, S. 75 ff.; Tenorth 2004), aber Lesekompetenz und mathematisches Verständnis sind wahrscheinlich die wichtigsten und insofern Teil einer *Grundbildung*, in deren Rahmen auch das Bildungsminimum – historisch und gesellschaftlich variabel – zu bestimmen ist. Man darf die Basisqualifikationen allerdings nicht als einfache Techniken auffassen, die abstrakt und inhaltsindifferent vermittelt oder erworben werden könnten (Koch 2004), und schon gar nicht als Bildung auf deutschem Grundschulniveau, auf dem dann die reflexive Begegnung mit „Kunden" und/oder Wissenschaften aufbaut (Benner 2002). Diese Piaget nachempfundene Stufenkonzeption ist schon ontogenetisch und entwicklungspsychologisch unzutreffend (Schneider und Hasselhorn 2012; Kray und Schäfer 2012). Die Basiskompetenzen entwickeln sich vielmehr in einem langfristigen und kumulativen Prozess der Begegnung und aktiven Auseinandersetzung mit den Gegenständen der Kultur. In modernen Schulsystemen heißt dies

in der Begegnung mit unterschiedlichen, aber nicht beliebigen und nicht wechselseitig substituierbaren Modi des Weltverstehens, die – mögen die Fächerzuschnitte auch unterschiedlich sein – im Kanon des Bildungsprogramms universell institutionalisiert sind (Baumert 2002, vgl. auch Messner in diesem Band). Dies ist die Grundstruktur moderner Allgemeinbildung, die im Anschluss an Humboldt (1809) von Flitner (1965), Wilhelm (1969) oder Tenorth (1994, 2004) ähnlich konzipiert wird (vgl. auch Bildungskommission 2003). Insofern ist Grundbildung immer auch Allgemeinbildung, und zwar von Schulbeginn an. Für die Naturwissenschaften, die prototypisch für die Rationalität des instrumentellen Zugriffs auf Wirklichkeit stehen, gilt dies allemal. Der besondere Beitrag von PISA (und der deutschen Erweiterungen) zur Deskription des Feldes und zur theoretischen Konstitution des „zu erklärenden Phänomens" – und dieser Beitrag soll hier herausgearbeitet werden – liegt einmal in der Identifikation des Überschneidungsbereichs zweier unterschiedlicher bildungstheoretischer Traditionen und zum anderen in der fachdidaktisch bzw. im Falle der Lesekompetenz kognitionspsychologisch begründeten Konzeption der latenten Dispositionen, ihrer breiten Beschreibung auf operativer Ebene und in der empirischen Prüfung der theoretischen Konstrukte. Mittlerweile liegt eine große Anzahl von publizierten PISA-Aufgaben einschließlich der Item-Parameter vor, sodass ohne Weiteres parallele Testversionen konstruiert und in Konkurrenz zu alternativen Konzepten geprüft werden können – eine Einladung an die Kritiker, ihrer Kritik Forschung folgen zu lassen.

Die öffentliche und politische Rezeption von Befunden der LSA-Studien wird in der Regel durch die Definition von inhaltlich beschriebenen Kompetenzstufen erleichtert. Bei PISA hat die Einteilung des Fähigkeitskontinuums in Kompetenzabschnitte besondere politische Bedeutung erhalten, da im ersten deutschen Bericht das „Bildungsminimum" an das Erreichen einer Kompetenzstufe gekoppelt und beim Unterschreiten dieser Stufe von einem Bildungsrisiko gesprochen wurde (Artelt et al. 2001; Klieme et al. 2001; Baumert und Schümer 2001). Man muss daran erinnern, dass Kompetenzstufen zunächst nichts weiter als kommunikative Hilfsmittel sind, mit denen eine latente kontinuierliche Fähigkeitsdimension in Fähigkeitsabschnitte zerlegt wird, die durch typische kognitive Operationen beschrieben werden können, die für die Lösung von Aufgaben notwendig sind, die bei gegebenem Fähigkeitsniveau mit als hinreichend definierter Wahrscheinlichkeit erfolgreich bearbeitet werden können. Die Festlegung von Kompetenzstufen sind also *arbiträre* Entscheidungen, die in der Regel kommunikativ von Experten unterschiedlicher Verfahren getroffen werden (Cizek und Bunch 2007; Bejar 2008; Pant et al. 2010). Wenn die Schwellenwerte kriteriale Bedeutung erhalten – z. B. Mindestniveau erreicht oder verfehlt –, spricht man von *Standard-Setting*. In PISA wird ein Verfahren angewandt, bei dem die Experten der jeweiligen Domäne technische Vorgaben bezüglich der Breite der Kompetenzstufen und der Lösungswahrscheinlichkeiten am unteren und oberen Ende einer Stufe erhalten. Auf der Basis der empirischen Schwierigkeitsparameter der Testaufgaben und interpretativ erschlossener schwierigkeitsgenerierender Merkmale versuchen sie Schwellen festzulegen, die eine möglichst distinkte Beschreibung der Kompetenzstufen auf operativer Ebene erlauben. (OECD 2009, S. 283 ff.). Kompetenzstufen bündeln Informationen und erleichtern eine nicht technische Kommunikation. Bei der Festlegung der normierenden Bildungsstandards in Deutschland

wird ähnlich nach der sogenannten *Bookmark*-Methode verfahren (Çetin und Gelbal 2013). Das *Standard-Setting* ist hier jedoch ein politisch-administrativer Prozess, der zu einem Verwaltungsakt führt. Im ersten Schritt bereiten Expertengruppen aus Fachdidaktikern, Bildungspraktikern, Lehrplanexperten der Länder und Psychometrikern unter Leitung des IQB und in enger Abstimmung mit den Koordinatoren der Länder auf Amtschefebene einen Vorschlag vor. Die Entscheidung wird im zweiten Schritt auf politischer Ebene im Einvernehmen aller Länder getroffen.

Die Arbeiten der IEA (TIMSS und PIRLS) und der OECD (PISA) waren in verschiedener Hinsicht beispielgebend. Innerwissenschaftlich haben sie als Katalysatoren für eine bildungstheoretische und fachdidaktische Klärung des Bildungsprogramms ganz unterschiedlicher Unterrichtsfächer gewirkt. Die Reihe der mittlerweile entwickelten domänenspezifischen Kompetenzmodelle, die überwiegend auch empirisch geprüft wurden, reicht von den differenzierten Entwürfen des DESI-Konsortiums für Deutsch und Englisch (Klieme et al. 2008), den Bildungsstandards für die Fächer Deutsch, Englisch, Französisch, Mathematik und die Naturwissenschaften (Köller et al. 2010; Stanat et al. 2012; Pant et al. 2013), den Entwürfen für Geschichte und politische Bildung (Schreiber et al. 2006; Körber et al. 2007; Weißeno et al. 2010; Detjen et al. 2012; Trautwein et al. 2016), Wirtschaft, Arbeit & Technik (Leucht et al. 2016), über berufsbildende Kompetenzen (Lehmann et al. 2005) bis hin zum konstitutiv-religiösen Weltverstehen, für das eine Arbeitsgruppe um Benner an der HU-Berlin ein bildungstheoretisch begründetes Kompetenzmodell entwickelt und empirisch geprüft hat (Benner et al. 2007; Nikolova et al. 2007).

Bildungspolitisch waren die IRT-Modelle Voraussetzung der Arbeit der Länder und der KMK an den Bildungsstandards (Klieme et al. 2003), insbesondere für die inhaltliche Definition abschlussbezogener Regel- und Mindeststandards (Pant et al. 2010). Die Arbeit an den Bildungsstandards ist ein gutes Beispiel für die Verschränkung von Wissenschaft und Bildungspolitik in dem in Abschn. 3 skizzierten neu entstandenen Kommunikationssystem. Auch hier ist die wechselseitige Anerkennung differenter Handlungslogiken in Wissenschaft und Politik Voraussetzung der Kommunikation. Kritisch kann es dann werden, wenn bei der Interpretation der Befunde nämlich im *Standard-Setting* beide Perspektiven aufeinandertreffen und die normierende Entscheidung der Funktionslogik der Politik folgt.

7 Bildungsminimum und Hochleistungen

Mit der Definition von Kompetenzstufen und deren inhaltlicher Beschreibung auf der Ebene domänenspezifischer kognitiver Operationen hat PISA den entscheidenden Schritt getan, um Gruppen von Jugendlichen identifizieren zu können, die aufgrund unzureichender Basiskompetenzen potenziell von gesellschaftlicher Exklusion bedroht sind. Allmendinger (1999) spricht in diesen Fällen von Kompetenzarmut. Der Gedanke der notwendigen Universalisierung von Basisqualifikationen wird in der angelsächsischen *Literacy*-Diskussion mit dem Argument neuer und infolge des sich beschleunigenden Wandels von der Industrie- zur postindustriellen Wissensgesellschaft steigender Qualifikationsanforderungen verknüpft. Damit wird die Messlatte für sprachliche, mathematische und naturwissenschaftliche Literalität höher gelegt:

Schlichte Alphabetisierung genügt diesem Anspruch nicht. Wo aber liegt theoretisch und empirisch das Mindestniveau oder mit Tenorth (1994) das Bildungsminimum, unterhalb dessen mit erhöhter Vulnerabilität im Lebenslauf und insbesondere mit einem erhöhten Risiko bei dem Übergang in eine zukunftsfähige berufliche Erstausbildung zu rechnen ist?

Im ersten deutschen PISA-Bericht wurde versucht, auf der Grundlage der Beschreibung von Kompetenzstufen eine kritische Schwelle zu definieren, unterhalb derer man von einem Bildungsrisiko sprechen kann. Die Autoren, die den Begriff der „Risikogruppe" einführten, argumentieren sehr vorsichtig. Im Fall der Lesekompetenz sprechen sie bei 15-Jährigen von der Zugehörigkeit zu einer Risikogruppe, wenn die unterste Kompetenzstufe nicht erreicht wird, und von einem potenziellen Risiko für den Übergang in eine *zukunftsfähige* Berufsausbildung, wenn die erste Kompetenzstufe nicht überschritten wird (Artelt et al. 2001; Baumert und Schümer 2001). Grund für diese Zurückhaltung war der Mangel an Information über die prognostische Validität des PISA-Lesetests. Mittlerweile liegen durch die an PISA gekoppelte kanadische Längsschnittstudie *Youth In Transition Survey* (YITS) und vor allem die schweizerische Langzeitstudie *Transitions from Education to Employment* (TREE) Belege für die prognostische Validität des PISA-Lesetests vor, die es rechtfertigen, bei Personen, die im Lesen die erste Kompetenzstufe nicht überschreiten, von einer Risikogruppe zu sprechen (Bussière et al. 2009; Stalder et al. 2008; OECD 2010; Stalder 2012).

Für Mathematik fiel die Entscheidung leichter, da ein Abgleich zwischen PISA und den von den Industrie- und Handelskammern bei der Vergabe von Ausbildungsplätzen benutzten Mathematiktests möglich war. Hier zeigte sich, dass Schulabsolventen, die in Mathematik das unterste Kompetenzniveau bei PISA, das im Wesentlichen durch Aufgaben auf Grundschulniveau beschrieben wird, nicht überschreiten, praktisch keine Chance haben dürften, die Aufgaben der Einstellungstests zu bewältigen. Deshalb wurden diese Personen in PISA 2000 von Anfang an als Risikogruppe klassifiziert (Klieme et al. 2001). Nach dieser Definition gehörten im Jahre 2000 22,5 % der 15-Jährigen im Lesen und im Jahr 2003 21,6 % in Mathematik zu einer Risikogruppe. Bis 2012 verkleinerten sich die Risikogruppen deutlich auf 14,5 bzw. 17,7 % des Altersjahrgangs.

Argumentiert man im bildungstheoretischen Rahmen der allgemeinen Bildung, wird mit der Festlegung der kritischen Schwelle auch eine Gerechtigkeitsfrage normativ entschieden. Mit dem Versprechen auf die Universalisierung der Prämissen für Teilhabe an gesellschaftlicher Kommunikation durch die Garantie des Bildungsminimums und die Kultivierung der Lernfähigkeit gilt im Hinblick auf das Erreichen der Mindeststandards das Gleichheitsprinzip: Alle Schülerinnen und Schüler sollen die Lerngelegenheiten und Unterstützung erhalten, die sie benötigen, um die kognitiven, sozialen und selbstregulativen Basisqualifikationen zu erwerben, die sie befähigen, am wirtschaftlichen, sozialen, politischen und kulturellen Leben in Selbstachtung teilzunehmen. Die Affinität zur *Capability*-Konzeption Amartya Sens (1980, 2011) und Martha Nussbaums (2011) ist deutlich zu erkennen. Damit werden auch Verantwortlichkeiten neu verteilt. Die Sicherung der gesellschaftlichen Teilhabechancen für alle ist auch eine Bringschuld der Schule, und Schulversagen ist nicht mehr allein ein Versagen des Einzelnen, sondern auch ein Versagen

der Schule. Die Einführung und Akzeptanz des Gleichheitsprinzips im Hinblick auf das Bildungsminimum ist innerhalb der Logik formaler Bildungsprozesse nicht selbstverständlich. Denn formale Bildung erzeugt mit der Bereitstellung von Lerngelegenheiten, die im Bildungsprogramm in ihrer Grundstruktur festgelegt sind und der Selbstentwicklung und Selbstwerdung des Individuums dienen, und dem Anspruch der optimalen Förderung jedes Einzelnen notwendigerweise und dauerhaft Differenz. Gleichheit der Ergebnisse kann selbst in einem Einheitsschulsystem kein sinnvolles Regulativ formaler Bildung sein. Umso bemerkenswerter ist das normative Korrektiv der allgemeinen Bildung: Es setzt Bildsamkeit universell voraus und erwartet in der Verfügung über die *Bildungsvoraussetzungen* für ein würdevolles Leben Gleichheit. Dieses Gerechtigkeitsprinzip setzt Vorstellungen meritokratischer Verteilungsgerechtigkeit Grenzen, die in der Öffentlichkeit offensichtlich anerkannt sind. Denn kaum ein PISA-Ergebnis hat für vergleichbare öffentliche und politische Aufmerksamkeit und Kritik gesorgt wie der Befund, dass ein nennenswerter Anteil der nachwachsenden Generation Mindeststandards in Bezug auf Basiskompetenzen nicht erreicht. Er wurde als institutioneller Makel interpretiert. Bildung auf Grundschulniveau zum Ende der Vollzeitschulpflicht unterbietet das erwartete und versprochene Bildungsminimum.

Wie aber hat sich die Definition der Risikogruppe empirisch bewährt? Mit dieser Frage verlassen wir die Ebene der Zustandsbeschreibung. Mit der Untersuchung der prognostischen Validität der Klassifikation wird der erste Schritt zur Erklärung von beruflichen Risiken im Lebenslauf und möglicherweise beginnender gesellschaftlicher Exklusion vollzogen, auch wenn von Rekonstruktion einer Handlungskette noch keine Rede sein kann. Prognostische Entscheidungen sind immer und insbesondere im Bildungsbereich fehlerbehaftet, weil junge Menschen ihren Lebensweg auch selbst gestalten. Bei der Güte von Zuordnungen handelt es sich um probabilistische Zusammenhänge. Bei einer dichotomen Klassifikation wie „erhöhtes“ bzw. „nicht erhöhtes“ Misserfolgsrisiko unterscheidet man zwei voneinander abhängige Fehlerarten. Vom Alphafehler spricht man, wenn Personen fälschlicherweise der Risikogruppe zugeordnet werden („falsche Positive“), und vom Betafehler, wenn Risikopersonen fälschlicherweise in die Gruppe mit nicht erhöhtem Risiko eingeordnet werden („falsche Negative“). Bei konstanter Gesamtfehlerquote hängen beide Fehlerarten direkt voneinander ab: Mit der Verkleinerung des Alphafehlers steigt der Betafehler und umgekehrt. In unserem Fall heißt das: Setzt man die zu erreichenden Mindeststandards sehr niedrig an, minimiert man den Anteil falscher Positiver, während sich gleichzeitig der Anteil falscher Negativer erhöht. Damit vermindert man aber auch die Wahrscheinlichkeit, Personen mit erhöhter Vulnerabilität zu entdecken – die sogenannte Sensitivität der Klassifikation –, während die Entdeckungswahrscheinlichkeit von Nicht-Risikopersonen – die Spezifität der Zuordnung – zunimmt. Bei einer Erhöhung der Mindeststandards wächst die Sensitivität der Klassifikation auf Kosten ihrer Spezifität. Die Wahl der kritischen Schwelle ist eine normative Entscheidung, die von der Bewertung der Folgen des Alpha- und des Betafehlers abhängt. Will man z. B. unter einer Perspektive der Optimierung individueller Entwicklung Personen rechtzeitig fördern, wird man wahrscheinlich einer ausreichend hohen Sensitivität der Zuordnung größeres Gewicht beimessen und gleichzeitig mit einer verminderten Spezifität in Kauf nehmen, auch Personen zu fördern, die der

Förderung weniger bedürfen. Will man öffentliche Kritik in Grenzen halten oder Arbeitgebern – wie Klemm in seinem Beitrag in diesem Band – keinen Vorwand geben, mangelnde Ausbildungsbereitschaft mit mangelnder Ausbildungsreife der Bewerber zu rechtfertigen, wird man konservativ entscheiden und die Spezifität der Klassifikation möglichst hoch setzen.

Bei der Definition der Risikogruppe in PISA war über die prognostische Güte der Zuordnung noch nichts bekannt. Die Klassifikation wurde auf Grundlage der empirischen Beschreibung von Kompetenzstufen interpretativ begründet. Mittlerweile verfügt man für die Güte der Klassifikation aufgrund des an PISA 2000 angekoppelten Schweizer Längsschnitts TREE über empirische Daten. Es ist bekannt, dass das duale System in der Schweiz – stärker noch als in Deutschland – auch schulisch schwach Qualifizierten eine zweite Chance einräumt. Dies bildet sich auch in den Schweizer Längsschnittbefunden ab (Stalder et al. 2008). In der Schweiz gehörten im Jahr 2000 20,6 % der 15-Jährigen im Lesen zur Risikogruppe. Sechs Jahre nach dem Verlassen der Pflichtschule hatten 62 % von ihnen eine Berufsausbildung – etwa die Hälfte davon allerdings auf unterstem Qualifikationsniveau – abgeschlossen oder einen anderen Sekundarstufen-II-Abschluss erworben. 38 % blieben ohne Abschluss. Das ist die Erfolgsgeschichte des Schweizer Berufsbildungssystems. Betrachtet man aber die Karriere der 15-Jährigen, die die unterste Kompetenzstufe in PISA 2000 überschritten hatten, blieben nur 10 % von ihnen ohne Abschluss. Die Wahrscheinlichkeit, keinen Berufsabschluss zu erreichen, war also in der Risikogruppe fast viermal so hoch. Beurteilt man die Güte der Klassifikation an den üblichen Kriterien, ergibt sich folgendes Bild: Mit der dichotomen Zuordnung im Jahr 2000 wurden 79 % der Fälle im Jahr 2006 richtig eingeordnet. Im Jahr 2006 verfügten 16 % der 21-jährigen Schweizer weder über eine abgeschlossene Berufsausbildung noch einen anderen Sekundarstufen-II-Abschluss. Die Hälfte dieser jungen Erwachsenen gehörte im Jahr 2000 im Lesen zur PISA-Risikogruppe. Die Sensitivität der Klassifikation beträgt also 50 %. Die Spezifität dagegen – die richtige Zuordnung der Erfolgreichen – liegt bei 85 %. Die PISA-Definition der Risikogruppe beruht danach auf einer relativ konservativen Entscheidung, bei der – wie man im Nachhinein sieht – nicht beide Fehlerarten gleich gewichtet sind, sondern die Spezifität privilegiert wird. Legt man das vom deutschen PISA-Konsortium vorgeschlagene Erfolgskriterium des Zugangs zu einer *zukunftsfähigen* Berufsausbildung der Beurteilung der Klassifikation zugrunde – dies wären im Schweizer System Ausbildungen oberhalb des untersten Qualifikationsniveaus –, so stiegen die Gesamtrefferquote auf etwa 85 (berechnet nach den Angaben bei Stalder et al. 2008), die Sensitivität auf etwa 64 und die Spezifität auf ungefähr 91% an (berechnet nach den Angaben bei Stalder et al. 2008 und Stalder 2012).

Aufgrund der hohen öffentlichen Aufmerksamkeit, die das Unterschreiten des Bildungsminimums erfährt, ist es nicht verwunderlich, dass die Behandlung von Mindeststandards ein gutes Beispiel ist, an dem man die unterschiedliche Logik von Wissenschaft und Politik darstellen kann. Das PISA-Konsortium in Deutschland ging davon aus, dass die Verfügung über Kompetenzen auf Grundschulniveau – und das heißt in Deutschland nach vierjährigem Schulbesuch – am Ende der Vollzeitschulpflicht keine gute Voraussetzung für den Übergang in die berufliche Erstausbildung darstellt. Dies war eine Hypothese, die prinzipiell auch scheitern konnte.

Im Rahmen der Überprüfung der neuen Bildungsstandards der Länder wurden im Bereich der Mathematik Mindeststandards für die Erreichung des Hauptschulabschlusses definiert (Blum, Roppelt & Müller 2012; Pant et al. 2013b). Die Festlegung der kritischen Schwelle war eine politisch-administrative Entscheidung, auch wenn sie in Kommunikation mit Wissenschaft und Praxis vorbereitet wurde. Nach der Definition der KMK befähigt die Hauptschule Absolventen zur Fortsetzung ihres Bildungswegs vor allem in berufsqualifizierenden Bildungsgängen (KMK 2015). Die Logik politischen Handelns legte danach eine Entscheidung nahe, bei der der Anteil derjenigen, die Mindeststandards verfehlen – also nach der PISA-Definition einer Risikogruppe angehören –, sich nicht zu weit von der Quote derjenigen, die die Pflichtschule ohne Abschluss verlassen, entfernte und in allen Ländern politisch vertretbar war. Die Entscheidung wurde so getroffen, dass im Bereich Mathematik die unterste Kompetenzstufe geteilt und die Kompetenzstufe 1b, auf der Anforderungen bewältigt werden können, die „typischerweise bis etwa zum 7. Schuljahr des Hauptschulbildungsganges“ (Blum et al. 2012, S. 62) beherrscht werden sollten, als Mindeststandard für den Hauptschulabschluss definiert wurde. In der 9. Jahrgangsstufe erreichen dann 5,5 % der Neuntklässler insgesamt dieses Niveau nicht (Pant et al. 2013, S. 166, Tab. 6.3). Dies entspricht ungefähr der Quote der Schulabgänger ohne Schulabschluss. Vergleicht man diese Festlegung mit den schweizerischen Befunden, so wird deutlich, dass hier eine extrem konservative Entscheidung gefällt wurde, bei der die Sensitivität der Klassifikation weitgehend der Maximierung der Spezifität geopfert wurde. Politisch ist dies rational, unter dem Gesichtspunkt der Identifikation von Förderbedarf wahrscheinlich wenig nützlich und unter wissenschaftlichen Gesichtspunkten die normative Lösung eines Optimierungsproblems. Wissenschaftliche Befunde zur prognostischen Qualität der Klassifikation – also Erklärungswissen – spielten dabei keine Rolle.

Das Thema Kompetenzarmut und drohende gesellschaftliche Exklusion traf und trifft bis heute den Nerv der öffentlichen Aufmerksamkeit. Im Vergleich dazu hatte ein zweiter robuster PISA-Befund, den man als Pendant bezeichnen könnte, nicht wirklich das öffentliche Interesse gefunden, obwohl er in allen Berichten immer wieder herausgestellt wurde (vgl. Prenzel et al. 2013). Trotz früher Differenzierung des Schulsystems ist in allen untersuchten Domänen die Leistungsspitze in Deutschland im Vergleich zu führenden OECD-Staaten relativ schwach ausgeprägt. Offensichtlich kommt das Gymnasium seinem Auftrag, auch Hochleistungen herauszufordern und zu fördern. nicht optimal nach. Auch eine Initiative des Bundes, ein entsprechendes wissenschaftlich begleitetes Förderprogramm aufzulegen, konnte sich in Abstimmung mit den Ländern politisch nicht durchsetzen.

8 Soziale Disparitäten des Kompetenzerwerbs und der Bildungsbeteiligung

Ungleichheit der Bildungsbeteiligung ist seit Jahrzehnten ein Standardthema der Bildungssoziologie, das allerdings in den 1980er- und 1990er-Jahren aus der Aufmerksamkeit von Öffentlichkeit und Politik praktisch verschwunden war. Man muss nur daran erinnern, dass mit der routinemäßigen Novellierung des Mikrozensusge-

setzes Anfang der 1990er-Jahre der einzige Indikator, an dem sich soziale Disparitäten der Bildungsbeteiligung während der Vollzeitschulpflicht beobachten ließen, aus dem Erhebungsprogramm gestrichen wurde. Bildungsgerechtigkeit war trotz der Umstrukturierung des Schulsystems in den neuen Ländern kein Thema und schon gar kein Problem.

Dies änderte sich erst mit den LSA und mit der Verwendung eines international vergleichbaren Sozialschichtindikators in PISA. Es war einer der überraschenden Befunde, dass in Deutschland die Kopplung von sozialer Herkunft und Kompetenzerwerb so eng wie in keinem anderen OECD-Staat war. Ein Vergleich der sozialen Gradienten der Lesekompetenz zeigte, dass in Deutschland ein nur mittelmäßiges Leistungsniveau mit einem steilen sozialschichtabhängigen Kompetenzgefälle verbunden war. Der internationale Vergleich belegte ferner, dass die sozialen Gradienten im oberen Bereich der Sozialstruktur konvergierten: Die Unterschiede in der Lesekompetenz zwischen den Staaten verringerten sich, wenn man nur Jugendliche mit privilegierter Herkunft verglich, während sich die Schere im unteren sozialen Bereich öffnete. Zwischen 2000 und 2012 hat sich der Zusammenhang zwischen sozialer Herkunft und Kompetenzerwerb in Deutschland aufgrund verbesserter Ergebnisse im unteren Leistungsbereich etwas gelockert. Ein positiver Entwicklungstrend deutet sich an, ist aber noch nicht zufallskritisch abzusichern. Die Kopplung von sozialer Herkunft und Kompetenzerwerb ist bis heute ein Thema, das für öffentliche und politische Aufmerksamkeit sorgt (vgl. Autorengruppe Bildungsberichterstattung 2016).

Mit der Verfügbarkeit von Leistungsdaten und theoretisch begründeten Indikatoren für Merkmale der sozialen Herkunft veränderte sich die Datenlage im Hinblick auf die Möglichkeit, soziale Disparitäten nicht nur zu beschreiben, sondern auch zu erklären. Es war zum ersten Mal möglich, die von Boudon (1974) vorgeschlagene Differenzierung zwischen „primären“, das heißt über Leistung vermittelten, und „sekundären“, direkten Einflüssen der sozialen Herkunft auf Bildungsentscheidungen und Bildungsbeteiligung in Deutschland empirisch darzustellen. Dies war ein wichtiger Schritt zur Rekonstruktion der im Sinne von Goldthorpe (2001) generativen Prozesse, die zur sozialen Ungleichheit der Bildungsbeteiligung führen. Lehmann et al. (1997) waren die ersten, die anhand von LSA-Daten der Hamburger Lernausgangslagen-Untersuchung (LAU) die Entstehung sekundärer sozialer Disparitäten beim Übergang von der Grundschule in die weiterführenden Schulen in Deutschland nachgewiesen haben. Mittlerweile liegt eine Reihe von Grundschulstudien vor, die alle konsistent für den Übergang in die weiterführenden Schulen sekundäre Herkunftseffekte belegen (vgl. Baumert und Maaz 2010; Dumont et al. 2014). Lange Zeit unbefriedigend geklärt war jedoch die quantitative Relation von primären und sekundären Effekten. Methodisch elaborierte Studien aus dem Vereinigten Königreich und Schweden schätzten den Anteil sekundärer Effekte an sozialer Ungleichheit der Bildungsbeteiligung je nach Untersuchungskohorte und Schätzverfahren auf 20 bis maximal 50 % der Variabilität (Erikson et al. 2005; Jackson et al. 2007; Erikson & Rudolphi 2009). Für Deutschland liegen drei Untersuchungen vor, die es erlauben, primäre und sekundäre Disparitätseffekte beim Übergang in die weiterführenden Schulen quantitativ zu bestimmen. Auf der Grundlage des Mannheimer Bildungspanels (MAPS) kommt Stocké (2007) zu einer explorativen Schätzung, dass

primäre und sekundäre Effekte in ähnlicher Stärke an der Übergangsentscheidung beteiligt seien. In Studien für Bayern und Sachsen konnten Ditton und Krüsken (2006) und Ditton (2007) anhand des Grundschullängsschnitts „Kompetenzaufbau und Laufbahn im Schulsystem“ (KOALA) zeigen, dass bis zu 30 % der Disparitäten auf sekundäre und bis zu 70 % oder mehr auf primäre Effekte zurückgehen. Anhand der ÜBERGANG-Studie des Max-Planck-Instituts für Bildungsforschung in Berlin, die sich zwei LSA-Untersuchungen simultan zunutze machte, arbeiteten Maaz und Nagy (2009) heraus, dass in der chronologischen Abfolge von Leistungsbeurteilung (Notenvergabe, Laufbahnbeurteilung, Verteilung der Übergangsempfehlung) und Übergangsentscheidung das Gewicht der Sekundäreffekte im Vergleich zu den primären Einflüssen systematisch zunimmt. Bei der Notenvergabe überwiegen noch primäre Herkunftseffekte, bei der Verteilung der Übergangsempfehlung sind primäre und sekundäre Effekte ausbalanciert, und bei der Übergangsentscheidung dominieren schließlich sekundäre Herkunftseffekte.

An diese Arbeiten schlossen Untersuchungen an, die unter Nutzung derselben Datensätze die Logik der Entscheidungsfindung handlungstheoretisch zu erklären versuchten. Ein in Deutschland verbreitetes theoretisches Erklärungsmodell ist das von Esser (1999) vorgeschlagene Wert-Erwartungsmodell, das an das von Erikson und Jonsson (1996) und Breen und Goldthorpe (1997) vorgeschlagene *Rational-Choice*-Modell anschließt. In dieses Modell gehen Einschätzungen des Nutzens, der Kosten und der Erfolgswahrscheinlichkeit der Wahl einer Bildungslaufbahn und ihre Wechselwirkungen ein. Die drei bereits vorgestellten Übergangsstudien erlauben eine theoretisch und empirisch befriedigende Spezifikation des Modells. Die Befunde zur Bewährung des Wert-Erwartungsmodells für die Erklärung von Disparitäten der Bildungsbeteiligung sind gemischt. Erst wenn das Modell durch die Berücksichtigung sozialer Normen – das Verhalten signifikanter Anderer – und institutioneller Opportunitäten und Restriktionen – Noten und Übergangsempfehlungen – erweitert wurde, konnten soziale Unterschiede der Bildungsbeteiligung zufriedenstellend erklärt werden (Ditton 2007; Stocké 2007; Stubbe 2009; Jonkmann et al. 2010).

Im Hinblick auf Interventionsmöglichkeiten zur Verminderung sozialer Disparitäten der Bildungsbeteiligung sind die Schätzung der primären Effekte und die Bedeutung der Schulleistungen für die Übergangsentscheidung von größter Bedeutung. Diese Befunde sind eine nachträgliche Rechtfertigung der Schwerpunktsetzung auf frühe Förderung insbesondere der Lesekompetenz in der Vor- und Grundschulzeit, die die KMK (2002) in ihren „Handlungsfeldern“ vorgenommen hatte, und eine zusätzliche Begründung für das Forschungs- und Entwicklungsprogramm „Bildung durch Sprache und Schrift“ (BISS) von Bund und Ländern (BISS 2016). Ob man deshalb von evidenzbasierter Steuerung sprechen kann, ist allerdings mehr als fraglich. Die öffentliche Aufmerksamkeit konzentriert sich nach wie vor ausschließlich auf den deskriptiven Befund des Zusammenhangs von Herkunft und Kompetenzerwerb und sorgt damit für politische Dynamik.

9 Zuwanderung als Tatbestand

Von Deutschland als einem Einwanderungsland zu reden, galt noch Ende der 1990er-Jahre als politisch inkorrekt. Bezeichnenderweise verfügte Deutschland in dieser Zeit über eine Wanderungs-, aber über keine Zugewandertenstatistik. Der Mikrozensus erfasste die Staatsangehörigkeit, aber nicht den Migrationshintergrund. Die verfügbaren Angaben zur ausländischen Wohnbevölkerung lieferten bereits im Jahr 2000 ein unzutreffendes Bild der langfristigen Auswirkungen des Zuwanderungsgeschehens auf die Bevölkerungsstruktur. Der Anteil der Personen mit Zuwanderungsgeschichte an der Wohnbevölkerung wurde um mehr als die Hälfte unterschätzt. Im Rückblick offenbart sich hier ein bemerkenswertes Versagen der Migrationsforschung vor politisch beschlossener Unwissenheit.

Dies änderte sich erst mit PISA und dem internationalen Vergleich. Mit der international üblichen Definition des Migrationshintergrunds anhand des Geburtslandes der erfassten Person, ihrer Eltern und Großeltern wurde zum ersten Mal der tatsächliche Umfang der Zuwanderung und die Verteilung auf Zuwanderungsgenerationen sichtbar (Baumert und Schümer 2001). Im Jahre 2000 stammten 27 % der 15-jährigen Schulbevölkerung in den alten Bundesländern aus Familien, in denen mindestens ein Elternteil zugewandert war. Über alle Bundesländer hinweg betrug der Anteil 21 %. In den westdeutschen Großstädten konnte eine Quote von 35 % Jugendlicher mit Migrationshintergrund erreicht werden. 47 % der 15-Jährigen mit Migrationshintergrund war bereits in Deutschland geboren, gehörte also zur zweiten Generation. Allein die Information über den quantitativen Umfang der Zuwanderung stellte die öffentliche Diskussion über Migration in Deutschland auf eine neue Grundlage. Deutschland war ein Einwanderungsland. Erst ab 2005 wird nach der entsprechenden Novellierung des Mikrozensusgesetzes der Migrationsstatus auch regelmäßig im Mikrozensus erfasst. Damit wurde auch die altersabhängige demografische Dynamik der Zuwanderung sichtbar (Rühl 2009; Autorengruppe Bildungsberichterstattung 2016).

Die PISA-Ergebnisse 2000 zur Bildungsbeteiligung und zum Kompetenzerwerb von Jugendlichen mit Migrationshintergrund stellten klar, dass es sich bei Kindern und Jugendlichen aus Zuwandererfamilien um eine doppelt – sozial und ethnisch – benachteiligte Gruppe handelte, die einem erhöhten Ausbildungs- und Beschäftigungsrisiko ausgesetzt war. Die zahlreichen nachfolgenden LSA-Studien klärten, dass die ethnischen Disparitäten nicht erst in der Sekundarstufe I auftraten, sondern ebenso in der Grundschule und schon vor Beginn der Schulzeit nachweisbar waren (Haag et al. 2012; Anders, 2013; Becker et al. 2013; Ebert et al. 2013). Das Thema der ethnischen Ungleichheit hat in den letzten 15 Jahren nicht an öffentlichem und politischem Interesse verloren und im letzten Jahr durch die Flüchtlingszuwanderung neue Dramatik erhalten (Autorengruppe Bildungsberichterstattung 2016).

Hat sich aber auch das Wissen über die Genese der Ungleichheit verbessert? In Bezug auf die Entstehung von Disparitäten der Bildungsbeteiligung wird man diese Frage bejahen. Im Jahre 2007 lag die Übergangswahrscheinlichkeit zum Gymnasium für deutschstämmige Kinder doppelt so hoch (p =0,46) wie für Kinder aus Zuwandererfamilien (p = 0,23) (Jonkmann et al. 2010). Von wissenschaftlicher Seite wurden ganz unterschiedliche Erklärungen und Erklärungskombinationen angeboten, die im

Sinne von Goldthorpe (2001) jeweils für ein unterschiedliches Narrativ der Genese von Benachteiligung stehen. Verantwortlich für das differenzielle Übergangsverhalten könnten sein: eine kulturelle Distanz der zugewanderten Bevölkerung gegenüber dem deutschen Bildungssystem und damit fehlende Bildungsmotivation, mangelnde Vertrautheit mit dem deutschen Berechtigungssystem und der Bedeutung von Schulabschlüssen für den Lebenslauf, institutionelle Diskriminierung in der Grundschule vor allen Dingen bei der Vergabe der Übergangsempfehlungen oder Auswirkungen eines *stereotype threat* (Steele und Aronson 1995), nach dem sich gesellschaftliche Geringschätzung von Zuwanderern und ihren Fähigkeiten subtil ungünstig auf selbstbezogene Kognitionen, Selbstvertrauen, Aspirationen und kognitive Leistungen auswirkt. Gresch und Becker (2010) sowie Kristen und Dollmann (2009) haben diese Annahmen zumindest implizit für die größten Zuwanderungsgruppen überprüft. In ihren Analysen konnten sie zeigen, dass nach Kontrolle von Schulleistungen ethnische Disparitäten der Bildungsbeteiligung am Ende der Grundschulzeit nicht mehr nachweisbar sind und bei zusätzlicher Kontrolle der Sozialschicht die Übergangswahrscheinlichkeit von Kindern mit Migrationshintergrund mehrfach höher ist als die deutschstämmiger Kinder. Lehmann et al. (1997) konnten bereits mit der *Large-Scale-Assessment*-Studie LAU belegen, dass von institutioneller Diskriminierung von Zuwandererkindern an Grundschulen, sofern es sich um die Übergangsempfehlung handelt, keine Rede sein kann – bei gleichen Leistungen erhielten sie vielmehr im Vertrauen auf ihre Leistungsfähigkeit einen Empfehlungsbonus[2]. Bei Zuwanderern handelt es sich in der Regel um hoch bildungsmotivierte Gruppen, die sehr wohl über die Bedeutung von Bildungsabschlüssen in Deutschland Bescheid wissen, nicht nur hohe Aspirationen haben, sondern diese auch bei adäquater Leistung in eine erhöhte Bildungsbeteiligung am Gymnasium umsetzen. Die Barriere, die Zuwandererkinder überwinden müssen, sind ihre Kompetenznachteile vor allem in der Beherrschung der Verkehrssprache bereits zu Beginn der Grundschulzeit. Ob die langfristigen Auswirkungen von *stereotype threat* die Bildungsmotivation dämpfen, lässt sich aufgrund der verfügbaren Daten nicht sagen – wenn ja, können die Auswirkungen auf die Übergangsentscheidungen nicht sehr groß sein.

Die Folgefrage liegt auf der Hand. Unterscheiden sich die Entwicklungsverläufe von Kindern mit und ohne Zuwanderungshintergrund im Hinblick auf Basisqualifikationen während der Grundschulzeit? Die Befunde sind hier gemischt. In einer von Pfost et al. (2014) vorgelegten Metaanalyse zeigten sich für die Lesekompetenz sowohl kompensatorische als auch Divergenz vergrößernde Entwicklungsverläufe. Besonders während der Vorschul- und zu Beginn der Grundschulzeit scheinen sich Startvorteile zu vergrößern. Für Deutschland ist die Datenlage unbefriedigend. Vergleicht man die Ergebnisse von Querschnittuntersuchungen zum Wortschatz und zur Morphosyntax im Alter von 5 Jahren mit den Untersuchungsergebnissen zum Leseverständnis gegen Ende der 4. Klasse, gehen die Unterschied zwischen Kindern

[2] Ob bei Klassenwiederholung, die in der Grundschule bei Kindern mit MGH mehrfach höher liegt als bei Kindern ohne MGH, institutionelle Diskriminierung auftritt, ist unklar. Um dies zu entscheiden, genügt nicht allein ein Vergleich der Wiederholungswahrscheinlichkeiten unter Kontrolle von Fachleistungen, sondern es sind auch die längerfristigen pädagogischen Auswirkungen auf die Entwicklung von Kindern mit MGH zu prüfen.

mit und ohne Migrationshintergrund von mehr als einer SD auf gut eine halbe SD zurück (Haag et al. 2012; Anders 2013). Baumert et al. (2012) konnten für die späte Grundschulzeit in der 5. und 6. Jahrgangsstufe einen generellen Kompensationseffekt in der Entwicklung der Lesekompetenz nachweisen, von dem Kinder aus Zuwandererfamilien in besonderer Weise profitierten. Pfost et al. (2012) berichten für die 3. und 4. Jahrgangsstufe einen Schereneffekt, bei dem Kinder mit ernsthaften Leistungsdefiziten im Lesen eine langsamere Entwicklung zeigen als unauffällige Klassenkameraden. Eine Erklärung für diese scheinbar widersprüchlichen Befunde könnten nichtlineare Entwicklungsverläufe in Kombination mit Entwicklungsrückständen sein, die zu unterschiedlichen Phasen der Vergrößerung und Verkleinerung von Kompetenzunterschieden führen (vgl. Baumert et al. 2012).

Was bedeuten diese Befunde für politisches Handeln? Die wichtigste Erkenntnis ist wohl, dass differenziertes Erklärungswissen nicht notwendigerweise größere Bedeutung für politisches Handeln hat als die genaue Deskription eines Tatbestandes. Wenn man darüber hinaus zwei weitere Schlüsse aus den Befunden ziehen will, dann die, dass die frühzeitige Sicherung von Basiskompetenzen für alle *die* Maßnahme zur Verringerung herkunftsbedingter Disparitäten ist und man bei demselben Angebot für alle nicht ohne Weiteres mit Kompensationseffekten rechnen kann. Im Grunde bleibt es bei der Schlussfolgerung, die Willms (2002) in seinem Aufsatz *Raising and Leveling the Learning Bar* gezogen hat. Wenn man herkunftsbedingte Disparitäten verringern und gleichzeitig das durchschnittliche Leistungsniveau der Altersgruppe erhöhen will, heißt der Königsweg: Förderung aller Leistungsschwachen unabhängig von ihrer sozialen und ethnischen Herkunft. Angesichts der (theoretisch und empirisch nicht gut begründeten) pädagogischen Hoffnungen, die zurzeit auf Individualisierung und differenzierenden Umgang mit Heterogenität gesetzt werden, ist dies eine erfrischend einfache, aber immer noch hinreichend anspruchsvolle Botschaft.

10 Veränderungswissen – die Lösung des Problems?

Hilft in dieser Situation die Durchführung von LSA, die als quasi-experimentelle oder randomisierte Feldexperimente angelegt sind, weiter? Dieser Typ von Untersuchungen ist nicht nur in Deutschland im Bildungsbereich selten zu finden. Das gilt vor allem für randomisierte Kontrollgruppenstudien (Coalition for Evidence-Based Policy 2016). Zwei Beispiele aus Deutschland lassen sich jedoch heranziehen. Klieme (2014) betrachtet die Ergebnisse einer im Rahmen von PISA 2009 durchgeführten quasi-experimentell angelegten Längsschnittstudie auf institutioneller Ebene als Beispiel für politisch handlungsleitendes Veränderungswissen. Bei der Untersuchung handelt es sich um eine klug geplante Studie mit 54 Gymnasien, die zweimal – im Jahre 2000 und 2009 – im PISA-Sample vertreten waren. Die Anlage der Studie ist quasi-experimentell, da ein Teil der Gymnasien sich in der neunjährigen Karenzzeit zu Ganztagsschulen weiterentwickelt und/oder eine verstärkte interne Evaluationspraxis eingeführt hatte. Ganztagsbetrieb und interne Evaluation wurden als Treatments behandelt und die Stichprobe entsprechend dichotomisiert. Damit ergibt sich ein quasi-experimentelles Zwei-mal-zwei-Design (Bischof et al. 2013). Die

abhängigen Variablen sind die Veränderungen des mittleren Motivations- und Leistungsniveaus über die Zeit. Die Autoren berichten drei Haupteffekte: Die Einrichtung des Ganztagsbetriebs geht mit verbesserter Motivation der Schülerinnen und Schüler einher, und mit der Kultivierung einer internen Evaluationspraxis verbessern sich Leistung und Motivation (die Interaktion beider Treatments wurde nicht geprüft). Ist dieser Befund Veränderungswissen, das unmittelbar politisch handlungsleitend sein kann oder zumindest sein sollte? Wissenschaftlich liegen die Einschränkungen einer kausalen Interpretation der Befunde auf der Hand. Schulleitungen und Lehrkörper, die sich für die Umstellung ihrer Schule auf Ganztagsbetrieb oder eine verstärkte interne Evaluationspraxis verständigen, unterscheiden sich wahrscheinlich nicht nur hinsichtlich dieser Entscheidungen, sondern auch in ihren Vorstellungen von Schulleben und Qualitätssicherung – also relevanten Faktoren der Schulentwicklung. Diese Unterschiede werden aber mit der Kontrolle des mittleren Leistungs- bzw. Motivationsniveaus der Schülerschaft im Jahr 2000 nicht kontrolliert. Damit bleibt das *Assignment*-Problem ungelöst, und es ist mit großer Wahrscheinlichkeit mit unbeobachteter Heterogenität zu rechnen. Aber selbst wenn man von diesen Einwänden einmal absieht, handelt es sich bei diesen Befunden tatsächlich um handlungsleitendes Veränderungswissen? Wenn die Befunde in ein bildungspolitisches Programm passen, in dem die allmähliche Ausweitung von Ganztagsbetrieb und Qualitätssicherung durch interne Evaluation ein Entwicklungsziel darstellt, sind sie sicherlich eine brauchbare argumentative Unterstützung und Vergewisserung politischen Handelns, wie Tillmann et al. (2008) am Beispiel der Rezeption von PISA 2000 gezeigt haben (vgl. Dedering et al. 2007). Die Befunde werden aber keine verantwortliche Ministerin und keinen Minister dazu bringen, Prioritäten ihres oder seines politischen Programms zu ändern – und mit Recht. Denn eine flächendeckende Implementation der Maßnahmen unter veränderten Kontextbedingungen – sei es durch Anordnung oder Inzentivierung – garantiert kein verantwortliches und zielgerichtetes Handeln von Professionellen und damit auch nicht das Auftreten des quasi-experimentell nachgewiesenen Effekts.

Ändert sich dies, wenn man ein tatsächlich randomisiertes Feldexperiment als Beispiel heranzieht? Mit Unterstützung der *Jacobs Foundation* hat das Max-Planck-Institut für Bildungsforschung im Land Bremen eines der wenigen randomisierten Feldexperimente, die in Deutschland den Ansprüchen von LSA annähernd genügen, durchgeführt (Stanat et al. 2005). In einer Sommerschule mit randomisierter Zuweisung zu Treatment- und Kontrollgruppe sollte geprüft werden, ob leseschwache Grundschülerinnen und -schüler durch sprachintensive Erfahrungen bei der Vorbereitung einer Theateraufführung (Treatment 1) bzw. durch eine Kombination von Theaterspiel und zusätzlichem Sprachunterricht (Treatment 2) im Vergleich zu einer unbehandelten Wartegruppe Verbesserungen in der Beherrschung der Verkehrssprache erreichen könnten. In diesem randomisierten Feldexperiment ließ sich ein positiver Sprachförderungseffekt für das kombinierte Treatment, nicht aber für die Immersion in sprachintensive Situationen zeigen, der auch noch mehrere Monate später abklingend nachweisbar war (Stanat et al. 2012). Das Resultat ist mit Ergebnissen anderer Interventionsprogramme vergleichbar. Punktuelle Fördermaßnahmen können sichtbare Erfolge haben, die aber über die Zeit ausklingen, wenn die Förderung nicht systematisch fortgesetzt wird. Handlungsleitende Evidenz? Wohl kaum,

aber eine Orientierung, wenn Bildungspolitik und Bildungsadministration über die intelligente Nutzung von Sommerschulen und die Stabilisierung ihrer positiven Effekte nachdenken.

11 Empirische Bildungsforschung und Politik: Kommunikation trotz unterschiedlicher Handlungslogiken

Politik und Verwaltung sind, um handlungsfähig zu sein, auf kontinuierliche und vor allen Dingen auch auf vorausschauende Informationen über die Funktionsfähigkeit des Bildungssystems angewiesen. Dazu gehören auch wissenschaftliche Informationen insbesondere dann, wenn im Wissenschaftssystem Forschungsergebnisse zur Leistungsfähigkeit des Systems erzeugt werden, die öffentliches Interesse finden und politisierbar sind. Die Bildungsverwaltungen aller Länder haben mittlerweile Instrumente der quantitativen und qualitativen Dauerbeobachtung entwickelt und institutionalisiert, die sie in die Lage versetzen, nicht nur aufgrund amtlicher Statistiken Entwicklungen zu verfolgen, sondern Informationen auch selbst im Modus der Wissenschaft zu erzeugen oder erzeugen zu lassen. Verwaltungsnahe Einrichtungen mit wissenschaftlichem Auftrag können dabei ganz selbstverständlich Teil des Wissenschaftssystems sein, auch wenn die Grundzüge des Arbeitsprogramms und zentrale Aufgabenstellungen im Modus politischen Handelns festgelegt werden. Sie sind Bindeglied eines komplexen Kommunikationsnetzes zwischen Politik, Verwaltung und Wissenschaft (Baumert und Füssel 2012; Tenorth 2014).

Empirische Bildungsforschung definiert sich über den Gegenstandsbereich – auch wenn er sehr breit gefasst ist (vgl. Abschn. 1) – und nicht über eine wissenschaftliche Disziplin. Die empirisch arbeitende Erziehungswissenschaft ist nur ein, wenn auch wichtiger Mitspieler unter anderen. Dies hat Folgen sowohl für die Selektion als auch für die Bearbeitung von Fragestellungen. Fragestellungen sind an das Forschungsfeld gebunden, und ihre Auswahl und Formulierung orientieren sich in der Regel an zwei unterschiedlichen Relevanzkriterien, die in eine Balance zu bringen sind. Sie müssen innerhalb einer Referenzdisziplin wissenschaftlich und d. h. theoretisch und methodisch anschlussfähig sein und sollen gleichzeitig innerhalb des Handlungsfelds gesellschaftliche, politische oder praktische Bedeutung haben.

Die LSA sind ein gutes Beispiel dafür, wie sich innerwissenschaftliche Bedeutung von Fragestellungen und Befunden mit öffentlicher und politischer Relevanz verbinden kann. Die LSA wirkten gerade in dieser Verbindung als Katalysatoren für die Dynamik der empirischen Bildungsforschung insgesamt, die mittlerweile auf einer Reihe von Gebieten jenseits von LSA hoch aktiv ist. Von diesem Prozess haben insbesondere die Erziehungswissenschaft und die Fachdidaktiken profitiert, die – soweit sie empirisch arbeiten – internationale Anschlussfähigkeit gewonnen haben. Die LSA scheinen immer dann besondere wissenschaftliche und öffentliche Aufmerksamkeit gefunden zu haben, wenn sie Sensitivität für gesellschaftliche Problemlagen mit einem methodischen Vorgehen verbinden konnten, in dem sowohl die theoretisch angeleitete, dichte Beschreibung des Gegenstands als auch die Entwicklung eines das Phänomen oder den Zusammenhang erklärenden Narrativs zu ihrem Recht kamen.

Die LSA sind aber auch der Untersuchungstypus, bei dem die institutionalisierte Kommunikation zwischen Politik und Wissenschaft besonders eng und differenziert ist. Umso mehr drängt sich die Frage auf, wie sich dieser Austausch trotz unterschiedlicher Funktionslogik der Systeme relativ konfliktarm auf Dauer stellen ließ. Die Antwort mag im ersten Moment paradox erscheinen. Unter den Bedingungen einer medialen, Transparenz erzeugenden Dauerbeobachtung dürfte die wechselseitige Akzeptanz der unterschiedlichen Handlungsrationalität von Politik und Wissenschaft die Voraussetzung sein, um erfolgreich kommunikative Anschlussstellen zu finden. Solche Anschlussstellen sind am ehesten in der Verständigung über wichtige und wissenschaftlich bearbeitbare Problemlagen im Feld und im Austausch über mögliche Implikationen von empirischen Befunden auszumachen. Die Kommunikation erlaubt den konstruktiven Umgang mit unterschiedlichen Handlungslogiken, setzt ihre Differenz aber nicht außer Kraft. Grenzüberschreitungen gefährden oder beenden die Kommunikation.

Empirische Bildungsforschung ist als Wissenschaft dem Erkenntnisgewinn und seiner diskursiven Validierung verpflichtet. Insofern verblüfft das wissenschaftskritische Aperçu, das Ewald Terhart mit Zustimmung von politischer Seite und Applaus der kritischen Erziehungswissenschaft geprägt hat: „Das Wissen über Leistungsergebnisse von Schulsystemen wächst schneller als das Wissen darüber, was man mit diesem Wissen anfangen kann." (Terhart 2002, S. 108). Man kann darin eine melancholische Beschreibung von Systemdifferenz sehen. Zur Kritik wird die Aussage erst durch einen Kategorienfehler, nämlich wenn man dem Konzept einer praktischen oder politischen Wissenschaft folgt. Wenn man aus der Geschichte lernen kann, müsste die Erziehungswissenschaft klüger geworden sein. Denn mit beiden Konzeptionen hat sie missliche Erfahrungen gemacht. Aber auch hier gilt *vox emissa non revertitur* – oder vielleicht doch? Nämlich als vorwurfsvolles Echo seitens der Politik, die empirische Bildungsforschung möge endlich das tun, was sie nicht tun kann – handlungsleitende Evidenz für eine rationale Politik erzeugen.

Open access funding provided by Max Planck Society.

Abkürzungen

BIJU	*Bildungsverläufe und psychosoziale Entwicklung im Jugend- und Erwachsenenalter*, Max-Planck-Institut für Bildungsforschung, Berlin (ab 1991); https://www.mpib-berlin.mpg.de/de/forschung/beendete-bereiche/erziehungswissenschaft-und-bildungssysteme/forschungsgebiet-i
BIKS	*Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalte*r (BiKS-3-10 und 8–14;

	2005–2014), Universität Bamberg (Artelt et al.); https://www.iqb.hu-berlin.de/fdz/studies/BiKS_3-10
BILWISS	*Bildungswissenschaftliches Wissen als Teil professioneller Kompetenz in der Lehramtsausbildung* (2009–2019), Goethe Universität, Frankfurt/M. (Kunter et al.); http://www.bilwiss.uni-frankfurt.de/index.html
BISTA	*Überprüfen von Bildungsstandards – Ländervergleich*, IQB Berlin; https://www.iqb.hu-berlin.de/bista; https://www.iqb.hu-berlin.de/bista/subject
CIVED	*Civic Education Study,* IEA (1999); http://www.iea.nl/cived.html
COACTIV	*Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz*, Max-Planck-Institut für Bildungsforschung, Berlin (2003–2004); https://www.mpib-berlin.mpg.de/coactiv/index.html
COACTIV-R	*A Study on Teacher Candidates' Acquisition of Professional Competence During Teaching Practice* (2008–2011), Max-Planck-Institut für Bildungsforschung, Berlin; https://www.mpib-berlin.mpg.de/coactivr/englisch/index.php
DESI	*Deutsch-Englisch-Schülerleistungen-International*, Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt/M. (DIPF) (2003–2004); http://www.dipf.de/de/forschung/projekte/deutsch-englisch-schuelerleistungen-international
ELEMENT	*Erhebung zum Lese- und Mathematikverständnis: Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin* (2003–2005); Humboldt-Universität Berlin (Lehmann); https://www.iqb.hu-berlin.de/fdz/studies/Element
IALS	*International Adult Literacy Study* (1994); http://www5.statcan.gc.ca/olc-cel/olc.action?lang=en&ObjId=89M0014X&ObjType=2
Jacobs Sommercamp	*Lernsommer. Jacobs Sommercamp Bremen* (2004, 2005 und 2006), Max-Planck-Institut für Bildungsforschung, Berlin; https://www.mpib-berlin.mpg.de/de/forschung/beendete-bereiche/erziehungswissenschaft-und-bildungssysteme/forschungsgebiet-iii/text-und-bild
KESS	*Kompetenzen und Einstellungen von Schülerinnen und Schülern,* 4.–13. Jahrgang (2003–2011), Schulbehörde Hamburg; http://bildungsserver.hamburg.de/bildungsqualitaet/
KOALA-S	*Kompetenzaufbau und Laufbahnen im Schulsystem. Eine Längsschnittuntersuchung an Grundschulen* in Bayern und Sachsen (2006–2008), LMU (Ditton); https://sofis.gesis.org/sofiswiki/
LAU	*Aspekte der Lernausgangslage und der Lernentwicklung,* 5.–13. Jahrgang (1996–2003), Schulbehörde Hamburg; https://de.wikipedia.org/wiki/LAU-Studie
LEK	Längsschnittliche Erhebung pädagogischer Kompetenzen von Lehramtsstudierenden (2009–2013), Universität Köln (König); https://www.hf.uni-koeln.de/33207

LEK-R	Längsschnittliche Erhebung pädagogischer Kompetenzen von Lehramtsstudierenden und Referendarinnen/Referendaren (2013–2015), Universität Köln (König); https://www.hf.uni-koeln.de/35966
LEO-Level One	*Leo. – Level-One Studie*, Universität Hamburg (Grotlüschen/Riekmann) (2009–2012); http://blogs.epb.uni-hamburg.de/leo/
LISA	*Lesen in der Sekundarstufe* (2004–2007), CAU Kiel (Möller); http://survey.psychpaed.uni-kiel.de/LISA-Lesen-in-der-Sekundarstufe.html
LISA-6	*Eine empirische Studie zu Lernergebnissen an allgemeinbildenden und beruflichen Gymnasien in Schleswig-Holstein* (2013–2014), CAU und IPN Kiel (Köller und Möller); http://www.ipn.uni-kiel.de/de/forschung/projekte/lisa-6
NEPS	*Nationales Bildungspanel Deutschland* (*National Educational Panel Study)* des Leibniz-Instututs für Bildungsverläufe, Bamberg (ab 2009); https://www.neps-data.de/en-us/projectoverview.aspx
PALMA	*Projekt zur Analyse der Leistungsentwicklung in Mathematik* (2002–2007), LMU (Pekrun); http://www.psy.lmu.de/pde/forschung/forsch_projekte/palma/index.html; http://www.uni-regensburg.de/Fakultaeten/nat_Fak_I/BIQUA/
PIAAC	*Programme for the International Assessment of Adult Competencies*, OECD (2011–2012); http://www.oecd.org/skills/piaac/; http://www.gesis.org/piaac/piaac-im-ueberblick/
PIAAC-L	*Programme for the International Assessment of Adult Competencies – Longitudinal* in Deutschland, GESIS und SOEP (2014); https://dbk.gesis.org/dbksearch/SDesc2.asp?no=5989&ll=10&af=&nf=1&db=d&search=piaac&search2=&notabs=1
PIRLS/IGLU	*Progress in International Reading Literacy Study*/Internationale Grundschul-Lese-Untersuchung, IEA; http://timssandpirls.bc.edu/; http://www.iea.nl/pirls_2016.html
PISA	*Program for International Student Assessment,* OECD; https://www.oecd.org/pisa/
PISA-Plus	*Program for International Student Assessment* Längsschnitt Deutschland (2003–2004); http://archiv.ipn.uni-kiel.de/PISA/pisa2003/index.html; https://www.iqb.hu-berlin.de/fdz/studies/PISA-I-Plus
Reading Literacy	*Reading Literacy Study* der IEA (1990/1991); http://www.iea.nl/reading_literacy_study.html
TEDS-M	*Teacher Education and Development Study in Mathematics,* IEA (2007/2008); http://www.iea.nl/teds-m.html
TIMSS	*Trends in International Mathematics and Science Study,* IEA; http://timssandpirls.bc.edu/; http://www.iea.nl/timss_2015.html
TIMSS-II	*Third International Mathematics and Science Study,* IEA, 8. Jahrgangsstufe: Population II (1995/1996); http://www.iea.nl/timss_1995.html

TIMSS-II-L	*Third International Mathematics and Science Study,* Video-Längsschnitt Deutschland Population II (1994/1995), Max-Planck-Institut für Bildungsforschung, Berlin (Kunter & Baumert, 2006)
TIMSS-III	*Third International Mathematics and Science Study,* IEA, Sekundarstufe II: Population III (1995/1996); http://www.iea.nl/timss_1995.html
TOSCA	*Transformation des Sekundarschulsystems und akademische Karr*ieren (2002–2013), Hector-Institut der Universität Tübingen; http://www.uni-tuebingen.de/fakultaeten/wirtschafts-und-sozialwissenschaftliche-fakultaet/faecher/hector-institut-fuer-empirische-bildungsforschung/forschung/laufende-studien/tosca.html
TRAIN	*Tradition und Innovation (TRAIN) – Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen* (2009–2012), Hector-Institut der Universität Tübingen; http://www.wiso.uni-tuebingen.de/faecher/hector-institut-fuer-empirische-bildungsforschung/forschung/laufende-studien/train.html
UEBERGANG	*Übergang von der Grundschule in die weiterführende Schule*, Max-Planck-Institut für Bildungsforschung, Berlin im Rahmen von TIMSS (2007); https://www.mpib-berlin.mpg.de/de/forschung/beendete-bereiche/erziehungswissenschaft-und-bildungssysteme/forschungsgebiet-ii/timss-uebergang

Literatur

Allmendinger, J. (1999). Bildungsarmut: Zur Verschränkung von Bildungs- und Sozialpolitik. *Soziale Welt, 50*, 35–50.

Anders, Y. (2013). Stichwort: Auswirkungen frühkindlicher institutioneller Betreuung und Bildung. *Zeitschrift für Erziehungswissenschaft, 16*(2), 237–275.

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, & W. Schneider (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

Autorengruppe Bildungsberichterstattung (Hrsg.) (2016). *Bildung in Deutschland 2016: Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration.* Bielefeld: Bertelsmann.

Baltes, P. B., Lindenberger, U., & Staudinger, U. M. (1998). Life-span theory in developmental psychology. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 569–664). New York: Wiley.

Baumert, J. (2002). Deutschland im internationalen Bildungsvergleich. In N. Killius, J. Kluge, & L. Reisch (Hrsg.), *Die Zukunft der Bildung* (S. 100–150). Frankfurt a.M.: Suhrkamp.

Baumert, J., & Füssel, H.-P. (2012). Kooperation im föderalen Bildungssystem: Zwischen Wettbewerb und Qualitätssicherung. In I. Härtel (Hrsg.), *Handbuch Föderalismus: Föderalismus als demokratische Rechtsordnung und Rechtskultur in Deutschland, Europa und der Welt, Bd. III: Entfaltungsbereiche des Föderalismu*s (S. 247–273). Heidelberg: Springer.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469–520. doi:10.1007/s11618-006-0165-2

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. doi:10.3102/0002831209345157

Baumert, J., & Maaz, K. (2010). Bildungsungleichheit und Bildungsarmut – Der Beitrag von Large-Scale-Assessments. In K. Hurrelmann & G. Quenzel (Hrsg.), *Bildungsverlierer: Neue Ungleichheiten* (S. 159–179). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J., Nagy, G., & Lehmann, R. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools? *Child Development, 83*(4), 1347–1367. doi:10.1111/j.1467-8624.2012.01779.x

Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb im nationalen Vergleich. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, & W. Schneider (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 159–245). Opladen: Leske + Budrich.

Baumert, J., Stanat, P., & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, & W. Schneider (Eds.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15–68). Opladen: Leske + Budrich.

Becker, B., Klein, O., & Biedinger, N. (2013). The development of cognitive, language and cultural skills from age 3 to 6: A comparison between children of Turkish origin and children of native-born German parents and the role of immigrant parent acculturation to the receiving society. *American Educational Research Journal, 50*(3), 616–649.

Bejar, I. I. (2008). Standard setting: What is it? Why is it important. *R&D Connections, 7*, 1–6.

Bellmann, J. (2006). Bildungsforschung und Bildungspolitik im Zeitalter „Neuer Steuerung". *Zeitschrift für Pädagogik, 52*(4), 487–504. urn:nbn:de:0111-opus-44682

Bellmann, J. (2015). Symptome der gleichzeitigen Politisierung und Entpolitisierung der Erziehungswissenschaft im Kontext datengetriebener Steuerung. *Erziehungswissenschaft, 26*(50), 45–54. urn:nbn:de:0111-pedocs-115010

Bellmann, J., & Müller, T. (Hrsg.) (2011). *Wissen was wirkt: Kritik evidenzbasierter Pädagogik*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Benner, D. (2002). Die Struktur der Allgemeinbildung im Kerncurriculum moderner Bildungssysteme: Ein Vorschlag zur bildungstheoretischen Rahmung von PISA. *Zeitschrift für Pädagogik, 48*(1), 68–90. urn:nbn:de:0111-opus-38217

Benner, D., Krause, S., Nikolova, R., Pilger, T., Schluß, H., Schieder, R., . . . Willems, J. (2007). Ein Modell domänenspezifischer religiöser Kompetenz: Erste Ergebnisse aus dem DFG-Projekt RU-Bi-Qua. In D. Benner (Hrsg.), *Bildungsstandards: Instrumente zur Qualitätssicherung im Bildungswesen. Chancen und Grenzen – Beispiele und Perspektiven* (S. 141–156). München: Schöningh.

Bildungskommission der Länder Berlin und Brandenburg. (2003). *Bildung und Schule in Berlin und Brandenburg: Herausforderungen und gemeinsame Entwicklungsperspektiven.* Berlin: Wissenschaft und Technik Verlag.

Bischof, L. M., Hochweber, J., Hartig, J., & Klieme, E. (2013). Schulentwicklung im Verlauf eines Jahrzehnts: Erste Ergebnisse des PISA-Schulpanels. In N. Jude & E. Klieme (Hrsg.), *PISA 2009 – Impulse für die Schul- und Unterrichtsforschung: Impulse für die Schul- und Unterrichtsforschung* (S. 172–199). Weinheim: Beltz.

BISS. (2016). http://www.biss-sprachbildung.de/biss.html

Blömeke, S., & Delaney, S. (2014). Assessment of teacher knowledge across countries: A review of the state of research. *International perspectives on teacher knowledge, beliefs and opportunities to learn* (pp. 541–585). Dordrecht: Springer.

Blömeke, S., Kaiser, G., & Lehmann, R. (2010). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich.* Münster: Waxmann.

Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016). The relation between content-specific and general teacher knowledge and skills. *Teaching and Teacher Education, 56*, 35–46.

Blum, W., Roppelt, A., & Müller, M. (2012). Kompetenzstufenmodelle für das Fach Mathematik. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I: Ländervergleich* (S. 61–73). Münster: Waxmann.

Botte, A., Sondergeld, U., & Rittberger, M. (2015). *Monitoring Bildungsforschung: Befunde aus dem Forschungsprojekt „Entwicklung und Veränderungsdynamik eines heterogenen sozialwissenschaftlichen Feldes am Beispiel der Bildungsforschung"*. Bad Heilbrunn: Klinkhardt.

Boudon, R. (1974). *Education, opportunity and social inequality*. New York: Wiley.

Breen, R., & Godthorpe, J. (1997). Explaining educational differentials – Towards a formal rational action theory. *Rationality and Society, 9*(3), 275–305.

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

Bromme, R., Prenzel, M., & Jäger, D. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Zeitschrift für Erziehungswissenschaft, 17*(4), 3–54.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design.* Cambridge, MA: Harvard University Press.

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

Bussière, P., Hébert, R., & Knighton, T. (2009). Educational outcomes at age 21 associated with reading ability at age 15. *Education Matters: Insights on Education, Learning and Training in Canada 6(2).* http://www.statcan.gc.ca/pub/81-004-x/2009002/article/10896-eng.htm

Çetin, S., & Gelbal, S. (2013). A Comparison of Bookmark and Angoff Standard Setting Methods. *Educational Sciences: Theory and Practice, 13*(4), 2169–2175.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. London: SAGE.

Coalition for Evidence-Based Policy (2016). http://evidencebasedprograms.org/about/full-list-of-programs.

Dedering, K., Kneuper, D., Kuhlmann, C., Nessel, I., & Tillmann, K.-J. (2007). Bildungspolitische Aktivitäten im Zuge von Pisa: Das Beispiel Bremen. *Die Deutsche Schule, 99*(4), 408–421.

Detjen, J., Massing, P., Richter, D., & Weißeno, G. (2012). *Politikkompetenz: Ein Modell*. Wiesbaden: Springer.

Dicke, T., Parker, P. D., Marsh, H. W., Kunter, M., Schmeck, A., & Leutner, D. (2014). Self-efficacy in classroom management, classroom disturbances, and emotional exhaustion: A moderated mediation analysis of teacher candidates. *Journal of Educational Psychology, 106(2),* 569–583. doi:10.1037/a0035504

Ditton, H. (2007). *Kompetenzaufbau und Laufbahnen im Schulsystem: Ergebnisse einer Längsschnittuntersuchung an Grundschulen*. Münster: Waxmann.

Ditton, H., & Krüsken, J. (2006). Der Übergang von der Grundschule in die Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft, 9*(3), 348–372.

Dumont, H., Maaz, K., Neumann, M., & Becker, M. (2014). Soziale Ungleichheiten beim Übergang von der Grundschule in die Sekundarstufe I: Theorie Forschungsstand, Interventions- und Fördermöglichkeiten. In K. Maaz, M. Neumann, & J. Baumert (Hrsg.), *Herkunft und Bildungserfolg von der frühen Kindheit bis ins Erwachsenenalter* (S. 141–165). Wiesbaden: Springer.

Ebert, S., Lockl, K., Weinert, S., Anders, Y., Kluczniok, K., & Rossbach, H. G. (2013). Internal and external influences on vocabulary development in preschool children. *School Effectiveness and School Improvement, 24*(2), 138–154.

Erikson, R., Goldthorpe, J. H., Jackson, M., Yaish, M., & Cox, D. R. (2005). On class differentials in educational attainment. *PNAS, 102*(27), 9730–9733.

Erikson, R., & Jonsson, J. O. (1996). Explaining class inequality in education: The Swedish test case. In R. Erikson & J. O. Jonsson (Eds.), *Can education be equalized?* (pp. 1–63). Boulder: Westview Press.

Erikson, R., & Rudolphi, F. (2009). Change in social selection to upper secondary school – primary and secondary effects in Sweden. *European Sociological Review, 26*(3), 291–305. doi:10.1093/esr/jcp022

Esser, H. (1999). *Soziologie – Spezielle Grundlagen – Band 1: Situationslogik und Handeln* (Vol. 1). Frankfurt a.M.: Campus.

Flitner, W. (1965). *Grundlegende Geistesbildung*. Heidelberg: Quelle & Meyer.

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociological Review, 17*(1), 1–20.

Greenbaum, H., Meyer, L., Smith, M. C., Barber, A., Henderson, H., Riel, D., & Robinson, D.H. (2016). Individual and institutional productivity in educational psychology journals from 2009 to 2014. *Educational Psychology Review, 28,* 215–223. doi: 10.1007/s10648-016-9360-8

Gresch, C., & Becker, M. (2010). Sozial- und leistungsbedingte Disparitäten im Übergangsverhalten bei türkischstämmigen Kindern und Kindern aus (Spät-)Aussiedlerfamilien. In K. Maaz, J. Baumert, C. Gresch, & N. McElvany (Hrsg.), *Der Übergang von der Grundschule in die weiterführende Schule: Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten* (S. 181–200). Bonn: BMBF.

Haag, N., Böhme, K., & Stanat, P. (2012). Zuwanderungsbezogene Disparitäten. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und MathematiRobinson, D. H., 20k: Ergebnisse des IQB-Ländervergleichs 2011* (S. 209–235). Münster

Waxmann.Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Heidelberg: Springer Medizin Verlag.

Heinrich, M. (2015). Neue „Vergessene Zusammenhänge"? Pädagogisches Unbehagen anlässlich Heinz-Elmar Tenorths Verhältnisbestimmung von Bildungspolitik und Bildungsforschung. *Die Deutsche Schule, 107*(3), 285–298.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (5. Aufl.). Seelze: Klett-Kallmeyer.

Higgins, J. P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions* (Vol. 5). Chichester: Wiley-Blackwell.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Holzberger, D., Philipp, A., & Kunter, M. (2016). „Ein Blick in die Black-Box." Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie *48*(2), 90–105.

Humboldt, W. von. (1809/1969). Der Königsberger und der Litauische Schulplan (1809). In W. von Humboldt, *Werke in fünf Bänden,* Bd. IV., hrsg. von A. Flitner und K. Giel. Darmstadt: Wiss. Buchgesellschaft.

Jackson, M., Erikson, R., Goldthorpe, J. H., & Yaish, M. (2007). Primary and secondary effects in class differentials in educational attainment: The transition to a-level courses in England and Wales. *Acta Sociologica, 50*(3), 211–229. doi:10.1177/0001699307080926

Jones, S. J., Fong, C. J., Torres, L. G., Yoo, J. H., Decker, M. L., & Robinson, D. H. (2010). Productivity in educational psychology journals from 2003 to 2008. *Contemporary Educational Psychology, 35*(1), 11–16.

Jonkmann, K., Maaz, K., McElvany, N., & Baumert, J. (2010). Die Elternentscheidung beim Übergang in die Sekundarstufe I: Eine theoretische Adaption und empirische Überprüfung des Erwartungs-Wert-Modells. In K. Maaz, J. Baumert, C. Gresch, & N. McElvany (Hrsg.), *Der Übergang von der Grundschule in die weiterführende Schule: Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten* (S. 253–282). Bonn: BMBF.

Jonkmann, K., Maaz, K., Neumann, M., & Gresch, C. (2010). Übergangsquoten und Zusammenhänge zu familiärem Hintergrund und schulischen Leistungen: Deskriptive Befunde. In K. Maaz, J. Baumert, C. Gresch, & N. McElvany (Hrsg.), *Der Übergang von der Grundschule in die weiterführende Schule: Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten* (S. 123–149). Bonn: BMBF.

Klieme, E. (2013). *PISA 2009: Nationale Ergänzungsstudien*. Vortrag auf der 68. Sitzung der Amtschefkommission „Qualitätssicherung in Schulen", Berlin, 12. August 2013.

Klieme, E. (2014). *„Steuerungswissen" revisited: Die Bedeutung der Bildungsforschung für Politik und Administration.* Vortrag ZfE-Forum, Hamburg, 5. Dezember 2014.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., . . . Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise*. Berlin: BMBF.

Klieme , E., Eichler, W., Helmke, A., Lehmann, R. H., Nold, G., Rolff, H.-G., Schröder, K., Thomé, G., & Willenberg, H. (Hrsg.) (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie.* Weinheim: Beltz.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 3–22). Cambridge, MA: Hogrefe.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen: Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik, 52*(6), 876–903.

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, & W. Schneider (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 139–190). Opladen: Leske + Budrich.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I. „Aufgabenkultur" un Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS- Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumentation.* (S. 43–67). Bonn: BMBF.

KMK. (2002). *PISA 2000 – Zentrale Handlungsfelder: Zusammenfassende Darstellung der laufenden und geplanten Maßnahmen in den Ländern (Stand: 07.10.2002).* Beschluss der 299. Kultusministerkonferenz vom 17./18.10.2002. Berlin: KMK. (https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2002/2002_10_07-Pisa-2000-Zentrale-Handlungsfelder.pdf)

KMK. (2015). *Das Bildungswesen in der Bundesrepublik Deutschland 2013/14*: *Darstellung der Kompetenzen und Strukturen sowie der bildungspolitischen Entwicklungen für den Informationsaustausch in Europa. In Zusammenarbeit mit der Deutschen EURYDICE-Informationsstelle des Bundes im BMBF.* Bonn. (https://www.kmk.org/dokumentation-und-statistik/informationen-zum-deutschen-bildungssystem.html)

Koch, L. (2004). Allgemeinbildung und Grundbildung, Identität oder Alternative? *Zeitschrift für Erziehungswissenschaft, 7(2)*, 183–191.

Köller, O. (2014). Entwicklung und Erträge der jüngeren empirischen Bildungsforschung. In R. Fatke & J. Oelkers (Hrsg.), *Das Selbstverständnis der Erziehungswissenschaft: Geschichte und Gegenwart* (S. 102–122). Weinheim: Beltz.

Köller, O., Knigge, M., & Tesch, B. (Hrsg.) (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster: Waxmann.

König, J., Blömeke, S., & Kaiser, G. (2015). Early career mathematics teachers' general pedagogical knowledge and skills: Do teacher education, teaching experience, and working conditions make a difference? *International Journal of Science and Mathematics Education, 13*(2), 331–350.

König, J., Lammerding, S., Nold, G., Rohde, A., Strauß, S., & Tachtsoglou, S. (2016). Teachers' professional knowledge for teaching English as a foreign language assessing the outcomes of teacher education. *Journal of Teacher Education*, 1–18. doi:10.1177/0022487116644956

König, J., & Kramer, C. (2016). Teacher professional knowledge and classroom management: On the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM, 48*(1–2), 139–151.

Körber, A., Schreiber, W., & Schöner, A. (Hrsg.) (2007). *Kompetenzen historischen Denkens: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik.* Neuried: ars una.

Kray, J., & Schaefer, S. (2012). Mittlere und späte Kindheit (6–11 Jahre). In W. Schneider & U. Lindenberger (Hrsg.), *Entwicklungspsychologie* (S. 211–233). Weinheim: Beltz.

Kristen, C., & Dollmann, J. (2009). Sekundäre Effekte der ethnischen Herkunft: Kinder aus türkischen Familien am ersten Bildungsübergang. In J. Baumert, K. Maaz, & U. Trautwein (Hrsg.), *Bildungsentscheidungen* (S. 205–229). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuhlmann, C. (2012). *Bildungspolitik und Leistungsvergleichsstudien: PISA 2000 und die Ganztagsschulentwicklung*. Wiesbaden: Springer.

Kunter, M., & Trautwein, U. (2013). *Psychologie des Unterrichts*. Paderborn: Schöningh UTB.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805.

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 97–124). New York: Springer.

Lange, H. (2008). Vom Messen zum Handeln: „empirische Wende" der Bildungspolitik? *Recht der Jugend und des Bildungswesens, 56*(1), 7–15.

Lehmann, R., Gänsfuß, V., & Peek, R. (1997). *LAU 5: Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern, die im Schuljahr 1996/97 eine fünfte Klasse an Hamburger Schulen besuchten; längsschnittliche Studie*. Hamburg: Behörde für Schule, Jugend und Berufsbildung.

Lehmann, R. H., Ivanov, S., Hunger, S., & Gänsfuß, R. (2005). *ULME I: Untersuchung der Leistungen, Motivation und Einstellungen zu Beginn der beruflichen Ausbildung*. Hamburg: Behörde für Bildung und Sport.

Lerner, R. M. (1984). *On the nature of human plasticity*. Cambridge, UK: Cambridge University Press.

Leucht, M., Köller, O., Neumann, M., & Baumert, J. (2016). Berufsbezogene Kompetenzen in der gymnasialen Oberstufe: Vergleich technischer und wirtschaftlicher Gymnasien. *Unterrichtswissenschaft* (in Druck).

Leutner, D., Opfermann, M., & Schmeck, A. (2014). Lernen mit Medien. In T. Seidel & A. Krapp (Hrsg.), *Pädagogische Psychologie* (S. 297–322). Weinheim: Beltz.

Lohse-Bossenz, H., Kunina-Habenicht, O., Dicke, T., Leutner, D., & Kunter, M. (2015). Teachers' knowledge about psychology: Development and validation of a test measuring theoretical foundations for teaching and its relation to instructional behavior. *Studies in Educational Evaluation, 44*, 36–49.

Luhmann, N. (1975). *Macht*. Stuttgart: Enke.

Luhmann, N. (1990). *Die Wissenschaft der Gesellschaft*. Frankfurt a.M.: Suhrkamp.

Luhmann, N. (2000). *Die Politik der Gesellschaft*. Frankfurt a.M.: Suhrkamp.

Maaz, K., & Nagy, G. (2009). Der Übergang von der Grundschule in die weiterführenden Schulen des Sekundarschulsystems: Definition, Spezifikation und Quantifizierung primärer und sekundärer Herkunftseffekte. *Zeitschrift für Erziehungswissenschaft,* 12(2), 153–182.

Mandl, H., & Kopp, B. (Hrsg.) (2005). *Impulse für die Bildungsforschung: Stand und Perspektiven. Dokumentation eines Expertengesprächs.* Berlin: Akademie Verlag.

Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). New Jersey: Pearson.

Meyer-Hesemann, W. (2008). Wissen für Handeln: Forschungsstrategien für eine evidenzbasierte Bildungspolitik. In BMBF (Hrsg.), *Wissen für Handeln: Forschungsstrategien für eine evidenzbasierte Bildungspolitik* (S. 9–15). Bonn/Berlin: BMBF.

Neuweg, G. H. (2015a). *Das Schweigen der Könner: Gesammelte Schriften zum impliziten Wissen*. Münster: Waxmann.

Neuweg, G. H. (2015b). Kontextualisierte Kompetenzmessung: Eine Bilanz zu aktuellen Konzeptionen und forschungsmethodischen Zugängen**.** *Zeitschrift für Pädagogik, 61*(3), 377–383.

Nikolova, R., Schluß, H., Weiß, T., & Willems, J. (2007). Das Berliner Modell religiöser Kompetenz. *Theo-Web. Zeitschrift für Religionspadagogik, 6*(2), 67–87.

Nückles, M., & Wittwer, J. (2014). Lernen und Wissenserwerb. In T. Seidel & A. Krapp (Hrsg.), *Pädagogische Psychologie* (S. 225–252). Weinheim: Beltz.

Nussbaum, M. C. (2011). *Creating capabilities*. Cambridge, MA: Harvard University Press.

OECD. (2009). *PISA 2006: Technical report*. Paris: OECD.

OECD. (2010). *Pathways to success: How knowledge and skills at age 15 shape future lives in Canada.* Paris: OECD.

Oser, F. K., Näpflin, C., Hofer, C., & Aerni, P. (2012). Towards a theory of negative knowledge (NK): Almost-mistakes as drivers of episodic memory amplification. In J. Bauer & C. Harteis (Eds.), *Human fallibility: The ambiguity of errors for work and learning* (pp. 53–70). Dordrecht: Springer Netherlands.

Pant, H. A., Böhme, K., & Köller, O. (2013a). Das Kompetenzkonzept der Bildungsstandards und die Entwicklung von Kompetenzstufenmodellen. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I: Ländervergleich* (S. 53–60). Münster: Waxmann.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Hrsg.) (2013). *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.

Pant, H. A., Tiffin-Richards, S. P., & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment: Projekt Standardsetting. In E. Klieme, D. Leutner, & M. Kenk (Hrsg.), *Kompetenzmodellierung: Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (S. 175–188). Weinheim: Beltz.

Pfost, M., Dörfler, T., & Artelt, C. (2012). Reading competence development of poor readers in a German elementary school sample: An empirical examination of the Matthew effect model. *Journal of Research in Reading, 35*(4), 411–426.

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development: A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research*, 84, 203–244. doi:10.3102/0034654313509492

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

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (Hrsg.) (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Rabe, T. (2013). Ties Rabe und der Bildungsforscher Olaf Köller. *DIE ZEIT*, Nr. 04, 17.01.2013.

Renkl, A. (2008). Lehren und Lernen im Kontext der Schule. In A. Renkl (Hrsg.), *Lehrbuch Pädagogische Psychologie* (S. 109–153). Bern: Huber.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Rühl, S. (2009). *Grunddaten der Zuwandererbevölkerung in Deutschland*. Berlin: Bundesamt für Migration und Flüchtlinge.

Sawyer, R. K. (2006). *The Cambridge handbook of the learning sciences*. Cambridge. UK: Cambridge University Press.

Schneider, W. & Hasselhorn, M. (2012). Frühe Kindheit (3–6 Jahre). In W. Schneider & U. Lindenberger (Hrsg.), *Entwicklungspsychologie* (S. 187–210). Weinheim: Beltz.

Schrader, J. (2014). Analyse und Förderung effektiver Lehr-Lernprozesse unter dem Anspruch evidenzbasierter Bildungsreform. *Zeitschrift für Erziehungswissenschaft, 17*(2), 193–223.
Schreiber, W., Körber, A., von Borries, B., Krammer, R., Leutner-Ramme, S., Mebus, S., . . . Ziegler, B. (2006). *Historisches Denken: Ein Kompetenz-Strukturmodell.* Neuried: ars una.
Schui, G., & Krampen, G. (2015). ZPID-Monitor 2012 zur Internationalität der Psychologie aus dem deutschsprachigen Bereich: Der Kurzbericht. *Psychologische Rundschau, 66*(2),124–130.
Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie. *Zeitschrift für Pädagogik, 60*(6), 850–866.
Seidel, T., & Reiss, K. (2014). Lerngelegenheiten im Unterricht. In T. Seidel & A. Krapp (Hrsg.), *Pädagogische Psychologie* (S. 253–276). Weinheim: Beltz.
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
Sen, A. (1980). Equality of what? *The Tanner Lecture on Human Values, 1*, 197–220.
Sen, A. (2011). *The idea of justice*. Cambridge, MA: Harvard University Press.
Shavelson, R. J., & Towne, L. (Eds.) (2002). *Scientific research in education*. Washington, DC: National Academy Press.
Stalder, B. E., Meyer, T., & Hupka-Brunner, S. (2008). Leistungsschwach – bildungsarm? Ergebnisse der TREE-Studie zu den PISA-Kompetenzen als Prädiktoren für Bildungschancen in der Sekundarstufe II. *Die Deutsche Schule, 100*(4), 436–448.
Stalder, B. E. (2012). School-to-work transitions in apprenticeship-based VET systems: The Swiss approach. In S. Billett, G. Johnson, S. Thomas, C. Sim, S. Hay, & J. Ryan (Eds.), *Experience of school transitions: Policies, practice and participants* (pp. 123–139). Dordrecht: Springer Netherlands.
Stanat, P., Baumert, J., & Müller, A. G. (2005). Förderung von deutschen Sprachkompetenzen bei Kindern aus zugewanderten und sozial benachteiligten Familien: Evaluationskonzeption für das Jacobs-Sommercamp-Projekt. *Zeitschrift für Pädagogik, 51*(6), 856–875.
Stanat, P., Becker, M., Baumert, J., Lüdtke, O., & Eckhardt, A. G. (2012). Improving second language skills of immigrant students: A field trial study evaluating the effects of a summer learning program. *Learning and Instruction, 22*(3), 159–170. doi:10.1016/j.learninstruc.2011.10.002
Stanat, P., Pant, H. A., Böhme, K., & Richter, D. (Hrsg.) (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. doi:10.1037/0022-3514.69.5.797
Stocké, V. (2007). Explaining educational decision and effects of families' social class position: An empirical test of the Breen-Goldthorpe model of educational attainment. *European Sociological Review, 23*(4), 505–519.
Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: The Brookings Institution.
Strasser, J., & Gruber, H. (2015). Learning processes in the professional development of mental health counselors: Knowledge restructuring and illness script formation. *Advances in Health Sciences Education, 20*(2), 515–530. doi:10.1007/s10459-014-9545-1
Stubbe, T. (2009). *Bildungsentscheidungen und sekundäre Herkunftseffekte: Soziale Disparitäten bei Hamburger Schülerinnen und Schülern der Sekundarstufe I*. Münster: Waxmann.
Terhart, E. (2002). Wie können die Ergebnisse von vergleichenden Leistungsstudien systematisch zur Qualitätsverbesserung in Schulen genutzt werden? *Zeitschrift für Pädagogik, 48*(1), 91–110. urn:nbn:de:0111-opus-38229
Tenorth, H.-E. (1994). *„Alle alles zu lehren": Möglichkeiten und Perspektiven allgemeiner Bildung.* Darmstadt: Wiss. Buchgesellschaft.
Tenorth, H.-E. (2004). Stichwort: „Grundbildung" und „Basiskompetenzen". *Zeitschrift für Erziehungswissenschaft, 7*(2), 169–182.
Tenorth, H.-E. (2014). Politikberatung und Wandel der Expertenrolle oder: Die Expertise der Erziehungswissenschaft. *Zeitschrift für Pädagogik, Beiheft 60,* 139–171.
Tenorth, H.-E. (2015). Bildungsforschung und Bildungspolitik im Dialog: Lernprozesse und Irritationen. *Die Deutsche Schule, 107*(3), 264–284.
Tillmann, K.-J. (2015). Empirische Bildungsforschung als Aufklärung? *Die Deutsche Schule, 107*(3), 299–314.

Tillmann, K.-J., Dedering, K., Kneuper, D., Kuhlmann, C., & Nessel, I. (2008). *PISA als bildungspolitisches Ereignis: Fallstudien in vier Bundesländern*. Wiesbaden: Verlag für Sozialwissenschaften.

Voss, T., Kunina-Habenicht, O., & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft, 18*(2), 187–223. doi:10.1007/s11618-015-0626-6

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Seattle: Hofgrefe and Huber Publishers.

Weißeno, G., Detjen, J., Juchler, I., Massing, P., & Richter, D. (2010). *Konzepte der Politik – ein Kompetenzmodell*. Bonn: Bundeszentrale für politische Bildung.

Wilhelm, T. (1969). *Theorie der Schule: Hauptschule und Gymnasium im Zeitalter der Wissenschaften.* Stuttgart: Metzler.

Willms, J. (2002). *Raising and leveling the learning bar: A background report for the HRDC skills and learning task force.* Ottawa: Human Resource Development Canada.

Winship, C., & Morgan, S. L. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.

Trautwein, U., Bertram, C., von Borries, B., Körber, A., Schreiber, W., Schwan, S., Zuckowski, A. et al. (2016). Entwicklung und Validierung eines historischen Kompetenztests zum Einsatz in Large-Scale-Assessments (HiTCH). In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.) (2016), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments. Bildungsforschung Band 44* (S. 97–120). Bielefeld: Bertelsmann.