4. PISA & Co.: Nutzen und Grenzen von Leistungsvergleichen auf Systemebene

PISA* und weitere internationale wie nationale Leistungsstudien, z. B. IGLU/PIRLS*, TIMSS*, LAU* oder KESS, haben unser Repertoire zur Evaluation von Schule und Unterricht forschungsmethodisch bereichert – und die bildungspolitische Wirksamkeit empirischer Bildungsforschung nachhaltig gestärkt. Man darf aber nicht übersehen, dass ihre Stärken, die wir in der Tat nutzen sollten, unvermeidlich mit spezifischen Schwächen verbunden sind.

Zudem ist der durch PISA geförderte Forschungsstil inzwischen zum Modell für Bildungsforschung allgemein und für die Evaluation nicht nur des Bildungssystems, sondern auch einzelner Schulen, Lehrer/innen und Schüler/innen geworden, sodass man sich auch mit der paradigmatischen Wirkung dieses Projekts auseinandersetzen muss (vgl. zur Kritik etwa Bracey 2009). Wenn dieser Modell-Effekt gemeint ist, spreche ich im Folgenden von PISA & Co. Sonst konzentrieren sich meine Anmerkungen auf PISA selbst.

4.1 Warum wir PISA & Co. brauchen, die Ergebnisse aber nur mit Vorsicht nutzen sollten

Methodisch kontrollierte Untersuchungen an repräsentativ ausgewählten Stichproben* verringern die Gefahr zufälliger Eindrücke. Das ist ein großer Vorzug von PISA & Co. gegenüber der persönlichen Erfahrung und gegenüber der Untersuchung von Einzelfällen. Je mehr Personen einbezogen werden, desto anfälliger werden Studien aber für logistische Probleme bis hin zu systematischer Verzerrung durch Nicht-Teilnahme bestimmter Gruppen. Außerdem gilt aus ökonomischen Gründen: je größer die Zahl der Teilnehmer/innen, desto weniger Aspekte können (genau) erfasst werden – und desto stärker muss ihre Erhebung fokussiert und formalisiert werden.

Die Formalisierung der Erhebungen und ihrer Auswertung wiederum verspricht weniger Subjektivität bei der Wahrnehmung von Sachverhalten und bei ihrer Deutung. Auch bietet Standardisierung – zumindest auf der Oberfläche von Fragebogenantworten und Aufgabenlösungen – einfachere Möglichkeiten für Vergleiche. Bezahlt wird die (scheinbare) Eindeutigkeit aber mit einem Verlust an kontextueller Bedeutung (»Wie hat der Schüler die Aufgabe verstanden?«) und inhaltlicher Tiefe (»Wie ist er zu seiner Lösung gekommen?«). Standardisierung verhindert ein Aushandeln von Bedeutungen und privilegiert den Zugang der Forscher/innen sowie ihre Interpretation der Antworten oder Lösungen (z. B. Wahl bzw. Gestaltung der Testaufgaben, Entscheidung für bestimmte statistische Rechenverfahren).

Bildungspolitisch bedeuten PISA & Co. einen Gewinn. Die Öffentlichkeit ist in eindringlicher Schärfe mit Problemen unseres Schulwesens konfrontiert worden, die lange verdrängt wurden, auch wenn sie unter Experten durchaus bekannt waren. So gehören die in Deutschland lebenden 15-Jährigen – gemessen an den eingesetzten Tests – nicht zu den besonders erfolgreichen Schüler/inne/n. Das gilt – trotz Verbesserung einiger Vergleichswerte über die Jahre – nicht nur insgesamt, sondern auch für die Vorzeigeschule Gymnasium, wenn man diese mit den entsprechenden Teilgruppen in anderen Ländern vergleicht (vgl. im Einzelnen: Baumert u. a 2001; Prenzel u. a. 2004; 2007a; Klieme u. a. 2010a). Im internationalen Vergleich lagen die deutschen 15-Jährigen 2000 unter dem Durchschnitt, und auch 2009 und 2012 reichte es trotz relativer Positionsgewinne nur zu einem Platz im – je nach Fach – (oberen) Durchschnitt. Besondere Probleme bestehen nach wie vor im unteren Leistungsbereich bei – je nach Schwellenwert – 15 bis 25 Prozent der Schüler/innen, denen durch die Tests elementare Schwierigkeiten mit der Schriftsprache, Mathematik und den Naturwissenschaften attestiert wurden. Die Zugehörigkeit zu dieser sogenannten Risikogruppe steht zudem in einem engen Zusammenhang mit der Herkunft aus der sozialen Unterschicht und mit dem Migrationshintergrund. Insofern hat PISA lange tabuisierte Gründe für die Benachteiligung verschiedener Gruppen wieder in die Diskussion gebracht, etwa die Selektionsmechanismen des zwei- bis fünfgliedrigen Schulsystems (s.

dazu aber auch schon frühere Befunde, etwa aus LAU* in Hamburg, vgl. Lehmann u. a. 1997, S. 98-102).Auch die Streuung zwischen Leistungsspitze und Leistungsende ist – trotz einer vergleichsweise gering besetzten Spitzengruppe im internationalen Vergleich – besonders groß.

Obwohl all diese Probleme hier und da auch vorher schon wahrgenommen und diskutiert wurden: Erst die Aufmachung von PISA als Leistungsolympiade mit Ranglisten hat solchen Einschätzungen Durchschlagskraft verliehen. Diese mediale Inszenierung (Merkens 2003, 184) war wirkungsvoll, in ihr stecken aber auch Probleme. Vor allem wegen der Weiterungen auf andere Bereiche. Denn PISA ist ein, wenn nicht *das* bekannteste Projekt beim Paradigmenwechsel von der Input- zur Output-Steuerung (vgl. etwa Dowling 2008 und ➔ »Output statt Input« 2004).

Schon diese einleitenden Hinweise machen deutlich: Die Evaluation von bildungspolitischen Programmen und unterrichtspraktischen Aktivitäten stellt komplexe Anforderungen, denen kein methodischer Ansatz allein gerecht werden kann. Gewinne in einer Dimension sind nicht ohne Verluste in einer anderen zu haben. Meine Vorbehalte bedeuten deshalb keine grundsätzliche Absage an diesen Ansatz. Es geht mir vielmehr darum, die Autorität, die diesem Paradigma teils von den Forschern selbst, teils in der Rezeption der Befunde von außen zugesprochen wird, zu relativieren. Denn PISA ist inzwischen zum Symbol für einen Paradigmenwechsel in der Evaluation von Schule und Unterricht geworden, der sogenannten Output-Steuerung, die testbare Fachleistungen zum Maßstab für die Qualität von Schule und Unterricht macht (vgl. ausführlicher Brügelmann 2005a, Kap. 46-50).

PISA fair zu beurteilen ist also nicht leicht. Das fängt mit der Frage an, wer oder was eigentlich PISA ist. Verschiedene Beteiligte – z. B. aus dem internationalen Konsortium und aus der deutschen Forschergruppe – äußern sich durchaus unterschiedlich. Zudem muss man unterscheiden zwischen Methodenartikeln der Autoren und grundsätzlichen Selbstdarstellungen (z. B. Klieme u. a. 2010a, b; Klieme 2011) sowie konkreten Auswertungen und deren Kommentierung durch Dritte. Letztere wiederum finden sich in Fachbei-

trägen oder in Artikeln und Interviews aus Alltagsmedien. Schließlich gehört zu PISA auch die Rezeption und Nutzung durch Fachkollegen einerseits, durch Journalisten andererseits bis hin zu Bildungspolitik, Schulverwaltung, Pädagogen in der Praxis und interessierten Laien. Denn Testprogramme können je nach gesellschaftlichem und politischem Kontext ganz unterschiedliche (Neben-)Wirkungen entfalten. So haben Carvalho/ Costa (2014) über den Vergleich von sechs europäischen Ländern gezeigt, dass sowohl das „Instrument PISA" als auch die Ergebnisse der Erhebungen in sechs europäischen Ländern kulturspezifisch und zusätzlich von einzelnen Akteuren interessensgeleitet genutzt wurden.

Vor allem die Art der Präsentation hatte und hat Nebenwirkungen, die bedacht werden müssen, wenn man Funktion und Anspruch einer solchen Bildungsforschung diskutiert. Dabei sind verschiedene Ebenen zu unterscheiden.[1] Ein besonderes Problem ergibt sich daraus, dass PISA mal als Forschungsprogramm, mal als Evaluationsstudie präsentiert wird – obwohl jeweils andere Anforderungen gelten. Auch die PISA-Forscher selbst haben diese Probleme der Selbstdarstellung bzw. ihrer Rezeption zunehmend wahrgenommen, weshalb sie den Geltungsbereich konkreter Aussagen und die Wirkung daran anschließender Maßnahmen in neueren Veröffentlichungen deutlich vorsichtiger formulieren (vgl. etwa Klieme u. a. 2010b, insbesondere S. 287 ff.). Andererseits spielt auch die OECD in der öffentlichen Darstellung mit unterschiedlichen Geltungsansprüchen (vgl. Forster 2014, 896).

[1] Vgl. zur Kritik aus verschiedenen Perspektiven: Collani (2001), Ladenthin (2003), Fertig (2004), Lind (2004; 2009; 2011b), Sjøberg (2004), Huisken (2005), Karg (2005), Meyerhöfer (2005; 2007), Radtke (2005), Rindermann (2006), Wuttke (2006a; 2007a, b; 2009), Klemm (2008), Gaeth (2009; 2010), Mortimore (2009), Münch (2009; 2011), weitere Beiträge in Jahnke/Meyerhöfer (2007) und in Hopmann u. a. (2007) sowie den zusammenfassenden Überblick bei Bank/Heidecke (2009); Repliken, die jeweils auf ausgewählte Punkte der Kritiken eingehen, finden sich bei Blum/Neubrand (2004), Prenzel (o. J.; 2005), Köller (2006); Baumert u. a. (2007; 2008); dazu wiederum Wuttke (2006b; 2007b, S. 99 f.) und Bank/Heidecke (2009, S. 368 f.).

Man kann das Potenzial und die Grenzen des Ansatzes hinter PISA insofern aus verschiedenen Blickwinkeln diskutieren. Als konkretes Projekt untersucht PISA alle drei Jahre die Fachleistungen am Ende der Pflichtschulzeit in sorgfältig ausgewählten Stichproben*. Analog zu TIMSS*, IGLU* und anderen internationalen Leistungsstudien hat PISA damit den begrenzten Anspruch, einen kleinen Ausschnitt der Wirkungen von Schule über verschiedene Länder hinweg zu vergleichen (s. dazu 4.2). Im Verbund von VerA* und anderen flächendeckenden Lernstandserhebungen haben die Bundesländer diesen Ansatz allerdings zu einem Programm des System-Monitoring bis hin zur Schul- und Unterrichtsevaluation gemacht – zum Teil sogar mit dem Anspruch, den Leistungsstand einzelner Schüler/innen bewerten zu können (s. dazu Kap. 5 und 6). Insofern ist Standardisierung für viele zum generellen Paradigma der Evaluation von Unterricht und individuellen Lernprozessen geworden. Wenn von PISA & Co. gesprochen wird, ist diese Modellwirkung gemeint (s. dazu 4.3).

Die Reichweite der Methoden und damit der Geltungsbereich der Befunde ist aber begrenzt: Statt der vielfach behaupteten Evidenz-Basierung handelt es sich um (Ein-)Schätzungen, die in hilfreicher, aber begrenzter Weise empirisch fundiert sind. Insofern plädiere ich am Ende dieses Kapitels für eine gleichgewichtige Ergänzung des Repertoires durch komplementäre Formen der Evaluation. Dabei sehe ich durchaus, dass manche Kollegen mit derselben Intention für eine Stärkung des Standardisierungsparadigmas argumentieren. Da diese in den letzten Jahren in großer Breite und mit starker Wirkung (siehe etwa die Besetzung von Lehrstühlen in der Erziehungswissenschaft) stattgefunden hat, sehe ich aber eher die Notwendigkeit, die Balance in umgekehrter Richtung auszutarieren.

4.2 Stärken, Reichweite und Schwächen von PISA

Reichen die Fachleistungen zum Ende der Pflichtschulzeit aus, um die Qualität des Schulwesens zu beurteilen?

Den Blick der Schule auf das zu wenden, was Schüler/innen aus ihr mitnehmen, ist wichtig. Der kontinentaleuropäischen Tradition entsprechend haben sich im deutschen Bildungssystem Steuerung und Kontrolle lange Zeit auf Input und Rahmenbedingungen wie Lehrpläne, Stundentafeln und Qualifikation der Lehrerschaft beschränkt. Pragmatisch verständlich ist auch, dass sich PISA auf sogenannte Basiskompetenzen konzentriert, und das in den Schlüsselbereichen wie Lesen und Mathematik (vgl. zur Begründung OECD 2004; 2007a). Zugleich sind aber die Einschränkungen zu beachten, die mit diesem Fokus einhergehen.

So zeigt PISA nicht, wie gut – oder auch nur wie erfolgreich – ein Schulsystem ist, sondern was 15-Jährige eines Landes in PISA-Tests leisten. Dies ist ein wichtiger Unterschied. Denn die PISA-Aufgaben sind bewusst nicht auf das jeweilige Curriculum abgestimmt, untersuchen also nicht, wie gut ein Schulsystem seine selbst gesetzten Ziele erreicht.

Die Frage, ob Ländercurricula die PISA-Kompetenzen zureichend abbilden, ist zunächst einmal normativ zu diskutieren: Setzt sich ein System die richtigen Ziele? Nur scheinbar selbstverständlich ist es beispielsweise zu prüfen, ob Schüler/innen lesen können. Denn PISA unterstellt ein bestimmtes Verständnis von Lesen: die sogenannte funktionale »Literacy«, sprich den Umgang mit Schrift im Alltag, vor allem die Nutzung von Sachinformationen (vgl. Artelt u. a. 2001, S. 78; kritisch zu dieser Einengung u. a. Hurrelmann 2003).

Wesentliche Erfahrungen mit Inhalten des Unterrichts machen junge Menschen außerhalb der Schule. Sie betreffen nicht nur die Entwicklung der Lesefähigkeit, sondern auch Kompetenzen in den Sachfächern und in Sport und Musik. PISA selbst verweist auf die hohe Abhängigkeit des Schulerfolgs von der sozialen Herkunft, also auf die Bedeutung von Normen und Lernmöglichkeiten im Alltag. Insofern dürfen beobachtete Leistungsunterschiede nicht allein der Qualität von Schule zugerechnet werden, lassen sie sich doch auch auf externe Faktoren wie Familie und Medien zurückführen. Soweit empirisch gestützte Schätzungen dazu vorliegen, wie weit sich Leistungsunterschiede durch schulische Einflüsse erklären lassen, schwanken sie sehr stark: zwischen weniger als 5 Prozent (bei

Jencks u. a. 1972) und 35 bis 40 Prozent (Largo 2014, S. 22, und Köller 2012, S. 74; vgl. zu den forschungsmethodischen Schwierigkeiten Siewert 2013, S. 132). Der zum Teil überraschend niedrige Anteil wird plausibel, wenn man Schätzungen zur Bedeutung des informellen Lernens hinzunimmt. So finden einer Auswertung der einschlägigen Forschung zufolge (Dohmen 2001, S. 7) zwei Drittel bis drei Viertel aller Lernprozesse außerhalb der Schule statt.

Neben dem Unterricht spielen Systembedingungen wie Stundentafel oder Halbtags- bzw. Ganztagsschule eine große Rolle für die Leistungsentwicklung der Schüler/innen. Hinzu kommen außerschulische Faktoren, beispielsweise die Bildungseinstellungen der Eltern und die Arbeitsmarktchancen der Absolventen. Unterschiede allein der öffentlichen Schule zuzurechnen verkennt auch die unterschiedliche Rolle von zusätzlichen, privat bezahlten Fördereinrichtungen wie etwa der Jukus in Japan oder des individuellen Nachhilfeunterrichts in Deutschland.

Die Qualität von Schule kann nicht allein an den messbaren Wirkungen von Tests festgemacht werden. *Wie* etwas gelernt wird, hat Bedeutung für die Dauerhaftigkeit und den Transfer von Wissen bzw. Können und bestimmt zudem die Erfahrung von Schule als Lernraum. Prozessqualität und Rahmenbedingungen stellen deshalb eigene Merkmale einer guten Schule dar (vgl. etwa die Kriterien des Deutschen Schulpreises – s. Kap. 5.1.3 – und die pädagogischen Standards des Schulverbundes »Blick über den Zaun« bei Groeben u. a. 2005 – s. Kap. 5.3). Aus problematischen Fachleistungen in Tests darf nicht vorschnell auf Schwächen des Unterrichts, auf Schwächen von Schulen oder Schulsystemen geschlossen werden. Mit der Erhebung der Leistungen bei 15-Jährigen erfasst PISA erworbene Kompetenzen nur punktuell. Wie aussagekräftig diese als Indikatoren für längerfristige Leistungsmöglichkeiten sind, muss gesondert nachgewiesen werden – was bisher fehlt.

Zwischenfazit: Durch die Prominenz der Leistungsstudien gerät Schule als Ort der Persönlichkeitsentwicklung, des sozialen und politischen Lernens, als Raum, in dem eine Gesellschaft durch die Begegnung ihrer Teilkulturen und der Generationen zusammenwächst (➔ »Pädagogik von gestern – in der Schule heute – für ein Leben

morgen?« 2010), in der öffentlichen Diskussion ebenso aus dem Blick wie das Kind »hinter PISA« mit seinen lebensweltlichen Erfahrungen (s. Kap. 3.2). Zudem lenkt der Fokus auf Unterricht bzw. auf das Bildungssystem als Ursache von Problemen allzu leicht ab von der Bedeutung der Wirtschaftspolitik und den Anforderungen an die Sozialpolitik, die die Rahmenbedingungen des Aufwachsens (und Lernens) von jungen Menschen in unserer Gesellschaft bestimmen.

Erfassen die PISA-Tests die Fachleistungen in angemessener Form?

Studien zur Notengebung zeigen seit vielen Jahren, wie fehleranfällig das Lehrerurteil ist (s. Kap. 6.1). Insofern liegt es nahe, nach Alternativen zu suchen. Standardisierte Tests haben mehrere Vorteile: Fokussierung der Datenerhebung, Transparenz der Anforderungen, Kalibrierung* der Maßstäbe (durch Bezug auf Normstichproben), Unabhängigkeit von persönlichen Besonderheiten. Insofern sind standardisierte Tests in heuristischer Funktion (eben als »Thermometer«, s. Kap. 2.2) stärker als bisher in das Repertoire pädagogischer Leistungsbeurteilung einzubeziehen.

Aber standardisierte Tests bedeuten auch eine Einschränkung der Lernbeobachtung und Leistungsbeurteilung durch die Auswahl bestimmter Fächer, durch die Beschränkung auf bestimmte Bereiche in diesen Fächern und durch die Begrenzung auf bestimmte Aufgabenformate. Zudem führt die bewusst begrenzte Zeit für das Lösen der Aufgaben dazu, dass geringe Prüfungsangst, Vertrautheit mit den Testformaten, testökonomisches Arbeiten und Verzicht auf Reflexion »belohnt« werden und damit die fachliche Kompetenz überlagern (vgl. Lind 2009b; Wuttke 2009, S. 26). Mädchen und Jugendliche aus schulfernen Milieus schneiden bei solchen Leistungssituationen tendenziell schlechter ab als in Situationen ohne starken Zeitdruck (vgl. einerseits Wieczerkowski/Jansen 1990; Waelbrock 1992; Ratzka 2003, S. 148 f.; andererseits Sacks 1999; Beilock u. a. 2004).

Die hohen Korrelationen zwischen den Lösungshäufigkeiten in Mathematik und Naturwissenschaften einerseits und im Lesen andererseits (vgl. Baumert u. a. 2003, S. 221) verweisen auf eine mangelnde fachliche Trennschärfe der Aufgabentypen – schon allein wegen der sprachlichen Einkleidung der Aufgaben in allen Bereichen (Lind 2004). Schließlich darf nicht übersehen werden, dass der forschungsmethodische Zugang das inhaltliche Bild des untersuchten Gegenstands bestimmt (vgl. Rindermann 2006, S. 71; dagegen: Baumert u. a. 2007; aber erneut: Lind 2009b). Wählt man beispielsweise in einer Untersuchung Test(typ) A, erhält man ein anderes Ergebnis, als wenn man Test(typ) B einsetzt.

Es muss nachdenklich stimmen, dass die USA bei internationalen Vergleichen der Leseleistung in den letzten 15 Jahren auf vorderen, auf mittleren und auf hinteren Plätzen gelandet sind:

Kohorte Schulanfang	Rangplatz	von xx Nationen	Studie
1934-1982	7	8	IALS
1988	2	16	IEA-GS
1991	15	31	PISA-1
1994	15	29	PISA-2
1998	4-12[2]	35	PIRLS

Diese Studien wurden mit unterschiedlichen Verfahren an jeweils anderen Stichproben* durchgeführt. Dagegen hat sich im US-internen *National Assessment of Educational Progress* das Niveau der Leseleistung über 30 Jahre hinweg weder in der Grundschule noch auf der Sekundarstufe bedeutsam verändert (vgl. Rampey u. a. 2009, NCES 2010). Diese Diskrepanz ist (neben dem Einfluss unterschiedlich gezogener Stichproben) nur erklärbar durch den Einsatz verschiedener Tests.

Ähnlich verwirrend ist das Bild innerhalb Deutschlands, wenn man sich anschaut, welchen Teilgruppen mangelnde Lesefähigkeit attestiert wird. Selbst wenn man nur die internationalen Lesestudien

[2] Die Abstände zwischen den Rangplätzen im Bereich 4 bis 12 waren so gering, dass eine Abstufung nicht möglich war.

von 1991 bis 2001 betrachtet, werden in den verschiedenen Auswertungen als »Risikogruppe« zwischen 2 Prozent und 25 Prozent der einbezogenen Jahrgänge benannt:

Studie	Alter	Schulanfang	Jahr	Anteil »Risiko«
IALS	16+	1934-1982	1993	14 %
IEA-Sek	14	1983	1991	2 %
PISA-I	15	1991	2000	25 %
IGLU/PIRLS	9	1998	2001	10 %

Kann sich die Quote innerhalb von zehn Jahren von 2 Prozent auf 25 Prozent mehr als verzehnfacht haben – um dann wieder in sieben Jahren auf 10 Prozent abzusinken? Plausibler ist folgende Erklärung: Je nach Aufgabentyp und je nach Definition der Schwellenwerte ergeben sich ganz unterschiedliche Anteile (vgl. zu deutlich unterschiedlichen Quoten je nach inhaltlicher Bestimmung der Schwellenwerte Grotlüschen/Riekmann 2012; Venn-Brinkmann 2012). Beispielsweise zeigt unsere LUST-Studie in verschiedenen Schulbezirken in Nordrhein-Westfalen, dass am Ende der 4. Klasse selbst in der Gruppe der unteren 5 Prozent viele Kinder mehrere Sätze pro Minute lesen können (Brügelmann 2003, S. 13, 16).

Entsprechend ist auch die Rangverbesserung deutscher 15-Jähriger von PISA-2003 auf PISA-2006 (s. Prenzel 2007a) diskutierbar: »Danach erreichen die 15-jährigen Deutschen in der neuen Pisa-Untersuchung Rang 13 – von diesmal 57 Teilnehmerländern. 2003 hatte Deutschland noch auf Platz 18 unter 40 Staaten gelegen. Laut OECD sind beide Tests wegen ihrer geänderten Aufgabenstruktur allerdings nicht vergleichbar« (www.spiegel.de/schulspiegel/wissen/0,1518,520291,00.html). Nun kann man sagen: Die Deutschen haben sich nicht verbessert, der neue Test bevorteile sie (so urteilt

der OECD-Experte Andreas Schleicher, das Testverfahren 2006 habe bestimmte Stärken von deutschen Schülern begünstigt); oder aber: Die Deutschen waren schon 2003 besser (sprich: Das damalige Testverfahren habe bestimmte Stärken deutscher Schüler/innen ausgeblendet und sie deshalb benachteiligt).

Ratzka (2003) hat Aufgaben aus der 1995 in deutschen Grundschulen nicht durchgeführten TIMS-Studie nachträglich in einer eigenen Stichprobe eingesetzt – parallel mit zwei anderen Tests. Sie fand, dass selbst bei einer groben Aufteilung nach Leistungsniveaus nur 41 Prozent der Schüler/innen in allen drei Tests in derselben Gruppe landeten und dass sogar beim gleichem Aufgabentyp (Textaufgaben) viele Schüler/innen in verschiedenen Tests unterschiedliche Ergebnisse erzielten – ja, sogar in demselben Test (den TIMSS-Aufgaben für die Grundschule), je nachdem, ob die Aufgaben mit oder ohne Zeitdruck zu bearbeiten waren (vgl. für den Abgleich von Kompetenztests in Lesen und Mathematik ähnlich Böhme u. a. 2014, 362).

Zwischenfazit: Die aufgaben- und kontextabhängige Ausschnitthaftigkeit von Instrumenten und ihren Ergebnissen muss in der öffentlichen Diskussion bewusst gemacht und gehalten werden. PISA & Co. dürfen nicht mehr über *die* Lese- und Mathematikleistung oder gar *die* Lesekompetenz der deutschen Schüler/innen urteilen – geschweige denn so tun, als ob sie mit ihrer Sondierung ausgewählter Wirkungsausschnitte die Qualität des vorhergegangenen Unterrichts oder gar den individuellen Lernstand einzelner Schüler/innen erfassen könnten.

Sind die Auswertungs- und Darstellungsformen der erhobenen Daten aussagekräftig und verständlich?

PISA hat Probleme von Teilgruppen ins öffentliche Bewusstsein gerückt, die vorher zwar in Fachkreisen, die aber in der Bildungspolitik nicht ernst genommen wurden. Das betrifft die (schrift-)sprachlichen Schwächen von Jungen (vgl. schon Kemmler 1967; die Bei-

träge von Aufenanger und Robine zu Brinkmann u. a. 1990; Wagemaker u. a. 1993 und den Sammelband von Richter/Brügelmann 1994), gilt aber auch für die Leistungsrückstände von Migranten (vgl. schon die Arbeitsgruppe am Max-Planck-Institut für Bildungsforschung 1979; Glumpler 1985; Bommes/Radtke 1993) und generell von Kindern und Jugendlichen aus Familien mit niedrigem sozioökonomischen Status (vgl. schon Rolff 1967; Geißler 1992 ff.; Müller 1998).

Gleichzeitig sind folgende Einschränkungen der Ergebnisse zu bedenken, wenn man PISA sinnvoll nutzen will (vgl. ergänzend die Beiträge in Jahnke/Meyerhöfer 2007 und Hopmann u. a. 2007):

Die Itemanalyse nach Kompetenzstufen mit einem unterstellten eindimensionalen Aufbau wird der komplex verschachtelten Entwicklung von Wissen und Können nicht gerecht und schränkt zudem das Spektrum möglicher Aufgaben ein (vgl. zusammenfassend Wuttke 2006, S. 144, und Lind 2009a, b).

Falsch vs. richtig auf der Oberfläche ist interpretationsbedürftig im Blick auf die Tiefenstruktur. Ohne Kenntnis der Lösungswege sind die Aufgabenlösungen mehrdeutig. Ob jemand z. B. ein bestimmtes Wort oder einen Text schnell und mit Verständnis lesen kann, hängt von seinen inhaltlichen Interessen, von seinem persönlichen Weltwissen und seinem Wortschatz ab, sodass richtige wie falsche Aufgabenlösungen sich nicht eindeutig bestimmten Formalstufen zuordnen lassen.

Durch die große Zahl der getesteten Schüler/innen werden auch inhaltlich irrelevante Unterschiede statistisch signifikant (s. zur statistischen Signifikanz* vs. praktischen Bedeutung von Unterschieden Kap. 2.3).

Um Ergebnisse aus verschiedenen Untersuchungen besser vergleichen zu können, werden die Punktwerte in einem Test oft auf eine Standardskala umgerechnet, auf der das arithmetische Mittel bei 500 angesetzt wird und zwei Drittel der Fälle innerhalb der sogenannten Standardabweichung* von +/- 100 liegen. Durch diese Spreizung können Punktdifferenzen (z. B. bei PISA) von Laien leicht überschätzt werden. Eine Steigerung der durchschnittlichen Fehlerquote um 50 Prozent bedeutet in einem Diktat mit 50 Wörtern

bei einer Zunahme von zwei auf vier Fehler etwas völlig anderes als bei einer Zunahme von zehn auf 20 Fehler.

Trotz der methodischen Sorgfalt der Forscher/innen bleiben so viele potenzielle Fehlerquellen unkontrolliert, dass deren Einfluss auf die Länder- und Untergruppen-Ergebnisse die Rangfolgen kräftig durcheinander schütteln kann. Konkret: Wenn 9 Punkte Unterschied im Ländervergleich als statistisch signifikant gelten, dann braucht diese Differenz bei den richtigen Antworten nur 2 Prozent zu betragen. Das bedeutet bei 26 Aufgaben (in PISA-Mathematik-2003) eine halbe Aufgabe oder auf 100 Aufgaben hochgerechnet einen Unterschied von zwei richtig gelösten Aufgaben (vgl. Wuttke 2007a, S. 244 f.). Ist das eine pädagogisch oder politisch relevante Differenz? Gaeth (2009) hat an einem fiktiven Beispiel illustriert, wie die Übertragung von einstelligen Rohwerten* in die PISA-Skala mit einem Mittelwert* von 500 zu einer Verzerrung der Wahrnehmung führen kann, wenn die Qualität der Werte und der Bezugsrahmen (Höchst- und Mindestwerte) nicht explizit gemacht werden. Die Übersetzung der Punktwerte in Alltagsbegriffe (z. B. »45 Punkte entsprechen etwa einem Schuljahr«) macht Kennwerte zwar anschaulicher, kann aber auch in die Irre führen (s. das Beispiel in Kap. 2.3).

Auch die Wahl von Schwellenwerten ist diskussionsbedürftig, so z. B. die Bildung des OECD-Mittelwerts aus den Ergebnissen der teilnehmenden Staaten und nicht der Schüler/innen oder die Schnitte auf der kontinuierlichen Leistungsskala bei der Bildung von Kompetenzstufen (vgl. Wuttke 2009, S. 24, der für PISA bei einer Berechnung nach Schüler/inne/n statt Staaten auf einen Durchschnittswert von 490 statt 500 kommt, womit die »schwachen« deutschen Ergebnisse in einem anderen Licht erscheinen).

Zwischenfazit: Tests sind wie Präzisionsgeräte, die eine Messung von minimalen Differenzen erlauben. Diese führen zu klaren Ranglisten, deren Unterschiede damit aber noch nicht im Alltag bedeutsam sind. Entsprechend vorsichtig sollte man Unterschiede zwischen Bildungssystemen bei PISA interpretieren. Und: *Zahlen sprechen nicht für sich*. Schon die Wahl der statistischen Verfahren, die Darstellung von Testergebnissen, ihre Übersetzung in Schaubilder

und Texte sind immer Akte der Interpretation und können bei entsprechender Absicht zur Manipulation der Wahrnehmung genutzt werden. Dass PISA an vielen Stellen stark ist, liegt auch in der hohen *theoretischen* Kompetenz der beteiligten Forscher/innen – und an dem Reichtum anderer Studien, die sie zur Interpretation ihrer Ergebnisse heranziehen.

Bilden Veränderungen in PISA-Tests tatsächliche Zuwächse und Verluste verlässlich ab?

Die Wiederholung von Erhebungen im Drei-Jahres-Rhythmus ermöglicht Reihenanalysen, in denen Veränderungen erkannt werden, jedenfalls wenn sie deutlich aus dem Rahmen fallen. Ein Wechsel in den Rangplätzen kann allerdings unterschiedliche Gründe haben:

Da die Rohwerte in den Tests durch Bezug auf die Ergebnisse anderer Länder in PISA-Punkte umgerechnet werden, können positive Entwicklungen trotzdem zu einer (relativen) Verschlechterung führen (und umgekehrt) – je nachdem, wie sich die gesamte Stichprobe* verändert.

Die zunehmende Vertrautheit mit den Testformaten verändert die Erfolgschancen – vor allem unter dem Zeitdruck der Tests. Das ist für Länder wie Deutschland bedeutsam, in denen standardisierte Tests traditionell eine geringe Rolle im Schulalltag gespielt haben. Bessere Testergebnisse müssen deshalb nicht zwingend eine Folge besserer fachlicher Kompetenzen sein.

Werden die Ergebnisse in den Tests von Beteiligten als bedeutsam für ihre eigene Situation oder Position wahrgenommen, wächst die Versuchung einer gezielten Vorbereitung. Verlage werben mit Etiketten wie »PISA-fit« für ihre Materialien, Lehrer/innen bauen gezielt Übungen in den Unterricht ein. Mit Veränderungen in den inhaltlichen Schwerpunkten können fachliche Schwächen ausgeglichen, mit der Gewöhnung an bestimmte Aufgabenformate vorhandene Kompetenzen stärker zur Geltung gebracht werden, ohne dass sie sich tatsächlich verbessern.

Zwischenfazit: Eine Wiederholung der Erhebungen alle drei Jahre ist zu kurzfristig angelegt (vgl. die bereits erwähnte Konstanz der Testleistungen im *National Assessment of Educational Progress* in den USA über mehr als 30 Jahre hinweg, obwohl sich in dieser Zeit das Schulsystem erheblich verändert hat). Diese Erhebungen binden zudem Aufmerksamkeit sowie fachliche und finanzielle Ressourcen, die an anderer Stelle für die Entwicklung von Schule und Unterricht oft bitter vermisst werden.

Wie bedeutsam ist die Qualität von Schule für die Lebenschancen des Einzelnen und für die wirtschaftliche Entwicklung einer Gesellschaft?

Eine wichtige Begründung für PISA ist die behauptete Bedeutung seiner Ergebnisse, also des untersuchten Outputs des Bildungssystems, für die wirtschaftliche Entwicklung eines Landes und für die Berufschancen seiner Bürger (kritisch dazu Meyer/Benavot 2013). Statistiken sowohl zu den Arbeitslosenquoten als auch zum Einkommensniveau bestätigen, dass Ausbildung und Bildung in der Tat wichtige Ressourcen für den individuellen Lebens- und Berufserfolg sind (vgl. Klemm 1998; Geißler 2006, S. 280 ff.). Zumindest als Korrelation lässt sich auch ein Zusammenhang zwischen allgemeinem Bildungsniveau und Wohlstand einer Gesellschaft feststellen. Aber die Frage, was Ursache und was Wirkung ist, lässt sich nicht so leicht beantworten. Konkret bestätigt sich in den PISA-Ergebnissen die behauptete Abhängigkeit des Wirtschaftswachstums und der Arbeitslosigkeit von Leistungspunkten nur eingeschränkt (vgl. die Kritik von Münch 2011, 281ff, und seine Hinweise auf Wolf 2002 bzw. Ramirez u. a. 2006 sowie die Relativierung durch die OECD 2010b). Der Einfluss von Drittfaktoren, die mit dem (Aus-)Bildungsniveau eines Landes korrelieren, darf insofern nicht übersehen werden.

Die Zusammenhänge zwischen dem PISA-Ranking und der Wirtschaftskraft von Ländern sind nicht sehr hoch. Sie werden durch weitere Bedingungen und Faktoren stark relativiert. Dazu zählen

etwa politische Veränderungen wie der Zusammenbruch der Sowjetunion und sein Einfluss auf die volkswirtschaftliche Entwicklung im Ostseeraum. Beispielsweise sind die Korrelationen zwischen Indikatoren für die Qualität des Bildungssystems und Indikatoren für die Wirtschaftskraft eines Landes höher, wenn man Industrieländer mit Entwicklungsländern vergleicht, als wenn man nur von den Randbedingungen her ähnliche Länder einbezieht.

Auch auf der Individualebene belegen Untersuchungen, dass Zusammenhänge zwischen Schulerfolg bzw. bestimmten Fachleistungen und späterem Berufserfolg bestehen. Aber die soziale Herkunft, Migrationshintergrund und Geschlecht spielen nicht nur für den Schulerfolg, sondern auch für die berufliche Laufbahn nach wie vor eine große Rolle – von der Bedeutung der sozialen Beziehungen ganz zu schweigen (vgl. zum »Mythos von den Leistungseliten« Hartmann 2002).

Der Einfluss von (Aus-)Bildung wird also in hohem Maße durch andere Bedingungen moderiert (vgl. Seeber/Maaz 2014; Cortina 2015, 225f., 238f.). Zu diesen gehören ebenfalls Konjunkturen einer Branche oder demografische Veränderungen. Zu bedenken ist zudem die ökonomische Entwertung von Qualifikationen durch deren Zunahme in der Breite: Dasselbe Zertifikat, aber auch dieselbe Kompetenz verliert auf einem Markt mit knappen Angeboten an Bedeutung, wenn mehr Menschen über sie verfügen. Aktuelle Wertigkeiten bestimmter Qualifikationen können deshalb nicht linear fortgeschrieben werden, auch wegen der Veränderung der Anforderungsprofile (vgl. Heuser 2014). Vor allem sind die Zusammenhänge nicht linear: So nimmt das Risiko von Arbeitslosigkeit bei Abschlüssen an Fachschulen über Fachhochschulen zu Universitäten nicht ab, sondern sogar geringfügig zu. Ein aktuelles Beispiel für die hohe Bedeutung von ökonomischen Kontextbedingungen: In China hat 2013 ein Drittel der Hochschulabsolventen keine Stelle bekommen – aber von den Abgängern der Berufsschulen wurden nur 2,5 Prozent arbeitslos (Schmidt 2014).

Die ökologische Validität standardisierter Tests, also ihre Gültigkeit für den Erfolg im Alltag, erweist sich generell als problematisch. Schon die Prognosekraft für den weiteren Schulerfolg und erst

recht die Alltagsrelevanz von Testleistungen für den Berufs- und Lebenserfolg außerhalb der Schule sind für PISA nicht überzeugend belegt (vgl. zu den Schwierigkeiten repräsentativer Vergleiche mit Erwachsenen die PISA-Nachfolgestudie bei Eltern von Schüler/inne/n: Ehmke/Siegle 2006). Dass die Vorhersagekraft der PISA-Lesetests besser sei als das Lehrerurteil oder die Schulnoten – wie etwa Schleicher (OECD) mit Verweis auf entsprechende Befunde der YITS auf mehreren Tagungen betont hat –, lässt sich diesen Studien (vgl. auch Shaienks/Gluszynski 2007, S. 33) nicht entnehmen. In Dänemark finden sich z. B. unter den 17 Prozent der nach PISA geblich funktional leseunfähigen „Illiterates“ immer noch 20 Prozent erfolgreiche Sekundarabsolventen (Dolin 2007, S. 114); und in Kanada erwerben sogar 62 Prozent der Schüler/innen auf Lesestufe 0 und 1 einen Highschool-Abschluss (Knighton/Bussière 2006, S. 21). Andere Beispiele für die Schwierigkeit, aus Testleistungen auf die Alltagskompetenz zu schließen, sind die Unterschiede der Rechtschreibleistungen in freien Texten gegenüber Diktaten (Brügelmann u. a. 1994) und die niedrigen Testleistungen vieler berufsfähiger Handwerker und sogar Lehrer/innen gegenüber Grundschüler/inne/n in der Lese-Studie LUST (Brügelmann 2004).

Zwischenfazit: PISA-Daten sind als Indikatoren für individuelle Leistungsmöglichkeiten außerhalb der Testsituation und als Indikatoren für gesellschaftliche Entwicklungen nur schwer einzuschätzen. Unabhängig davon ist zu bedenken: Die Vermittlung von *Qualifikationen* kann man instrumentell für ihre wirtschaftliche Verwertbarkeit untersuchen, betrachten und planen. *Bildung* als eine bestimmte Form der Aneignung von Welt(wissen) hat einen Eigenwert unabhängig vom Nutzen in Leistungssituationen.

Erlauben Unterschiede zwischen PISA-Kennwerten Vergleiche zwischen verschiedenen Bildungssystemen oder zumindest zwischen Teilgruppen innerhalb eines Systems?

Interkulturelle Vergleiche sind wegen der unterschiedlichen Bedeutung der Aufgabeninhalte generell problematisch (s. Kap. 2.3). Die

Herkunft von Testitems kann deren Schwierigkeit für Schüler/innen aus verschiedenen Ländern unterschiedlich beeinflussen (Vertrautheit des Testformats; Bekanntheit der Inhalte; Verständlichkeit der Übersetzungen; Länge der Sätze; Wortfrequenz »derselben« Wörter in verschiedenen Sprachen; unterschiedliche Reaktionen auf Zeitdruck und den Eindeutigkeitszwang bei Multiple-Choice-Aufgaben; vgl. Artelt/Baumert 2003; Puchhammer 2007; Wuttke 2007a, S. 254-257). Selbst statistisch signifikante Unterschiede können verschwinden, wenn sich allein durch die Übersetzung, etwa vom Englischen ins Französische, die Textmenge und damit der Zeitbedarf für das Lesen der Aufgabe um 10 bis 20 Prozent erhöht (vgl. Wuttke 2007a, S. 257).

Einen Einfluss auf die Ergebnisse kann auch haben, wie die Testsituation und ihre Bedeutung von den Teilnehmer/inne/n in den einzelnen Ländern wahrgenommen und bewertet wird. Zum Einfluss der Motivation auf die Testleistungen stellen allerdings Jude/Klieme (2010, S. 18) fest: »Es gibt keine Hinweise darauf, dass von motivationsbedingten Verzerrungen der Ergebnisse auszugehen wäre.« Nicht zu vernachlässigen ist zudem, dass die deutschen 15-Jährigen in der Regel erst die 9. Klasse besucht haben, während sie sich in erfolgreichen Bildungssystemen wie Finnland, Korea oder Japan bereits im 10. oder 11. Schuljahrgang befinden. Damit schwindet die Bedeutung der Punktdifferenzen weiter – vor allem, wenn man an ihnen die Leistungsfähigkeit unterschiedlicher Bildungssysteme festmachen will. Vor diesem Hintergrund kann auch die Veränderung der Altersstruktur durch den Systemwechsel von G9 auf G8 Einfluss auf die Verbesserung der deutschen Ergebnisse seit 2000 gehabt haben (vgl. Ehmke u. a. 2013), ohne dass der Unterricht besser geworden sein muss.

Die Stichproben* in den einzelnen Ländern sind nicht immer vergleichbar, vor allem was die Einbeziehung schwächerer Schüler/innen betrifft. So verweist Wuttke (2006a, S. 143) auf die – je nach Land – unterschiedlichen Schulbesuchsquoten mit 15 Jahren, auf die unterschiedliche Einbeziehung von Schüler/inne/n der untersten Leistungsgruppen, auf die unterschiedlichen Teilnahmequoten und auf nicht plausible Verteilungen in einzelnen Ländern.

Selbst innerhalb eines nationalen Bildungssystems sind Gruppenvergleiche nicht einfach. Kategorien wie »Migrationshintergrund« oder »Gesamtschule« lassen sich zwar für Deutschland als Variablen operationalisieren, haben aber je nach Bundesland eine ganz unterschiedliche Bedeutung. Besonders deutlich wird das beim Vergleich der Migrantenanteile und ihrer unterschiedlichen Zusammensetzung (damit auch ihrer Bedeutung für die Leistungsverteilung) in den verschiedenen Bundesländern.

Das lässt sich am Bundesländer-Vergleich »PISA-E« von 2000 verdeutlichen (vgl. Baumert u. a. 2002; 2003). Presse und Bildungspolitik handelten damals fast ausschließlich mit den Punktwerten für den Durchschnitt der 15-Jährigen insgesamt: 510 Punkte in Bayern, 482 in Nordrhein-Westfalen und 448 in Bremen – eine Differenz von fast einer ganzen Kompetenzstufe oder rund zwei Schuljahren. Bezieht man sich auf relevante Untergruppen, so verändert sich das Bild aber beträchtlich – und zwar je nach Kategorie unterschiedlich stark (vgl. Carle 2002; Klemm 2002): Vergleicht man nur die Kinder deutscher Staatsangehöriger, schrumpft der Rückstand von Nordrhein-Westfalen auf Bayern von 28 auf 12 Punkte, der Bremer Rückstand von 62 auf 41 Punkte. Bezieht man sich nur auf die 9. Klasse in Gymnasien und rechnet man zusätzlich die höhere Besuchsquote in Nordrhein-Westfalen und Bremen heraus, vergleicht man also gleiche Anteile des Jahrgangs in der gleichen Schulform, dann liegt Bremen nur noch 18 Punkte zurück und Nordrhein-Westfalen sogar rund 20 Punkte vor Bayern. Auch die oft gerühmte Förderung von Migranten in Bayern erscheint nicht mehr wesentlich besser als in Nordrhein-Westfalen, wenn man deren unterschiedliche soziale Zusammensetzung in beiden Bundesländern berücksichtigt (nur 14 statt 41 Punkte Differenz). Zu bedenken sind zusätzlich die unterschiedlichen Anteile verschiedener Nationalitäten. In Nordrhein-Westfalen und Bremen ist zum Beispiel der Anteil von türkischen Schüler/inne/n besonders hoch, die im Durchschnitt schlechtere Ergebnisse erzielen als etwa Spanier/innen oder gar Iraner/innen. Damit verändert sich die Situation noch einmal (vgl. auch Holtappels/Herdeegen 2005).

Sind Erklärungen möglich – und Beschreibungen ohne Erklärungen hilfreich?

Korrelationen aus Querschnittuntersuchungen* erlauben per se keine Erklärungen. Sie suggerieren jedoch oft Kausalitäten und werden auch so gelesen. Allerdings kann im Querschnitt* nicht geklärt werden, ob erhobene Variablen Ursache oder nur Indikator für Drittfaktoren sind. Verwirrung stiften immer wieder Korrelationsstudien zu Klassengröße und Schülerleistung. In ihnen wird beispielsweise oft nicht berücksichtigt, dass Schulleitungen Klassen verkleinern, wenn in der Lerngruppe besonders viele schwache Schüler/innen sind. (vgl. etwa Prais 2007, S. 147-148; kritisch zu diesen Folgerungen aus Querschnittstudien Hagemeister 2007 mit Verweis auf den US-amerikanischen STAR-Längsschnitt). Leider haben forschungsmethodisch anspruchsvollere Längsschnittstudien wie BIJU (»Bildungsprozesse und psychosoziale Entwicklung im Jugendalter«, Max-Planck-Institut 1994; 1996) und LAU (»Aspekte der Lernausgangslage und der Lernentwicklung«, Lehmann u. a. 1997; 2004) weniger Aufmerksamkeit gefunden als Querschnittstudien wie PISA und IGLU.

Selbst wenn plausible Erklärungen angeboten werden, die zu den Daten passen, gibt es oft weitere bedeutsame Randbedingungen, die genauso erklärungskräftig wären. Bezogen auf die Differenzen bei PISA-E 2000 hätte man etwa auf die Zahl der Deutschstunden in Klasse 1 bis 9 verweisen können (z. B. über 2 000 in Sachsen vs. weniger als 1 450 in Nordrhein-Westfalen) oder auf die unterschiedlichen Quoten der Sozialhilfeempfänger unter den Eltern (z. B. 2 Prozent in Bayern vs. 14 Prozent in Bremen). Je nachdem, wie man diesen und weitere Faktoren gewichtet, purzeln die Rangplätze und Abstände der Länder heftig durcheinander. Zwischen den entsprechenden Alternativmodellen fällt eine Entscheidung oft schwer. Zwar werden bei PISA über Fragebögen Daten zu einer Fülle von Bedingungen erhoben, aber forschungsökonomisch bedingt in einer zu abstrakten Form, um die Qualität der Lebens- und Lernbedingungen zureichend zu erfassen.

Zwischenfazit: Sowohl für die internationalen wie für die nationalen Vergleiche werden viele Ressourcen absorbiert. Oft wären Sekundär- und Meta-Analysen vorhandener Studien aussagekräftiger als neue Querschnittstudien. Für eine methodisch abgesicherte und inhaltlich ergiebige Interpretation der PISA-Korrelationen sind sie ohnehin zusätzlich notwendig, wie die vielfältigen Verweise auf solche Studien in den PISA-Bänden zeigen (z. B. zum Sitzenbleiben, zur Bedeutung kleiner Klassen oder zum Einfluss der Vorschulerziehung auf den späteren Schulerfolg). Die dort vorgelegten Interpretationen und Folgerungen beruhen insofern nur zum Teil auf den PISA-Befunden selbst. Durch inhaltlich überzeugende Folgerungen entsteht aber der Eindruck, diese Erkenntnisse seien dem spezifischen forschungsmethodischen PISA-Ansatz zu verdanken, sodass dadurch dessen Ausweitung auf andere Ebenen der Evaluation legitimiert, ja geboten erscheint.

4.3 Eignet sich PISA als Paradigma für die Evaluation von Schule und Unterricht generell?

Die Bedeutung standardisierter Verfahren nimmt in der Schullandschaft zu: verpflichtende Sprachstandserhebungen bei Vier- oder Fünfjährigen vor Schulbeginn, die landesweiten Vergleichsarbeiten »VerA« in den Jahrgängen 3 und 8, zentrale Prüfungen am Ende der verschiedenen Schulzweige, mit Computerprogrammen auswertbare Beobachtungsraster der »Qualitätsanalyse« bei Inspektionen. Zu beobachten ist ein Multiplikationseffekt des Evaluationsstils von Projekten wie PISA unter dem Etikett »evidence-based policy«. Deshalb lohnt ein zweiter Blick und eine grundsätzlichere Diskussion dieses Ansatzes (vgl. zu unterschiedlichen Einschätzungen die Beiträge zu Böttcher u. a. 2009; 2010, und Bellmann/Müller 2011).

Mein Fazit aus dem letzten Kapitel: Die angeblich »objektive« Bildungsforschung ist nicht schon deshalb wertvoll, weil sie mit standardisierten Instrumenten eine große Zahl von Fällen erfasst. Sie ist es erst, wenn die Forscher/innen über sinnstiftende Theorien verfügen. Dabei können sie dann auch zu unterschiedlichen Einschätzungen kommen, was für Laien nicht immer so leicht sichtbar

wird wie in der Kontroverse über die Bewertung der besseren Ergebnisse der deutschen Schüler/innen in den Naturwissenschaften bei PISA-2006 gegenüber 2003 und 2000 (s. Kap. 4.2.3 und Füller 2007a, b).

Schon House (1980, S. 81; meine Übers.) stellte fest: »Aus quantitativen Studien kann man oft wertvolle Einsichten gewinnen, aber sie sind gewöhnlich die Folge davon, dass der Forscher intuitiv richtig urteilt, und weniger daraus, dass er richtig gerechnet hat. Diese Erfolge werden oft der quantitativen Methodologie selbst zugerechnet statt dem Urteil. Kritik bezieht sich gewöhnlich auf die technische Qualität der mathematischen Analyse, nicht auf die Qualität des Urteils, die mit ihrer Formulierung und Interpretation verbunden ist.«

Umso dringlicher ist zu fordern, dass schon in den Forschungsberichten selbst die Interpretationsbedürftigkeit von Daten sichtbar gemacht wird: »Aus empirisch gewonnenen Daten unmittelbar praktisches Handeln oder Handlungsoptionen abzuleiten, ist aus meiner Sicht ein Problem missverstandener Evidenzbasierung. Forschungsbefunde geben Anlässe zu Kommunikation in Interpretations-, Selektions- und Übertragungsprozessen bei durchaus unterschiedlichen Logiken politischen, wissenschaftlichen und praktischen Handelns« (Baumert 2014, S. 22).

Wichtiger als eine weitere technische Perfektionierung der Erhebungen wird damit die soziale Kontrolle ihrer Interpretation (s. Kap. 2.6). In dieser Hinsicht könnte die empirische Bildungsforschung viel aus Erfahrungen im Rechtswesen lernen. Juristen haben über Jahrhunderte hinweg eine Tradition der Evaluation von menschlichem Verhalten und sozialen Situationen entwickelt, und zwar in zweierlei Hinsicht: als empirische Feststellung von Sachverhalten, z. B. Würdigung von Zeugenaussagen und von Sachverständigengutachten, sowie als ihre normative Bewertung, z. B. Auswahl und Gewichtung von einschlägigen Normen.

Das Fazit der rechtswissenschaftlichen Diskussion: Diese beiden Aktivitäten lassen sich methodisch zwar strukturieren, aber sie bleiben an die beteiligten Personen gebunden und lassen sich von der Ausschnitthaftigkeit und Perspektivität jeder Erkenntnis und jedes

Urteils nicht befreien. Um die Gefahr zu minimieren, dass deren unvermeidliche Subjektivität einseitig durchschlägt, schaffen die Prozessordnungen ein System von institutionellen Kontrollen – ein Vorbild für die Evaluation des Schulwesens und einzelner Bildungseinrichtungen (➔ »Towards checks and balances« 1974 Brügelmann 2007). So müssten von Anfang an alternative Deutungen und Folgerungen explizit gegeneinander gestellt werden, um den Anschein von Expertenautorität zu relativieren und damit auch der Öffentlichkeit die Interpretationsbedürftigkeit der nur scheinbar eindeutigen Daten bewusst zu machen.

Standardisierte Tests als Kontrollelement im Programm der Output-Steuerung

PISA hat sich inzwischen von einem spezifischen Projekt zu einem Modell der Leistungsmessung über das System-Monitoring hinaus gemausert. Alle Bundesländer führen flächendeckende Lernstandserhebungen in der Grundschule und in der Sekundarstufe durch. Damit einher geht ein Wechsel zur sogenannten Output-Steuerung, der mit einem Versagen der in Europa traditionell üblichen input-orientierten Maßnahmen begründet wird. Diese Input-Steuerung erfolgte über Lehrpläne, über die Qualifikationsanforderungen an Lehrer/innen und über die Prozesskontrolle durch die Schulaufsicht. Demgegenüber definiert die Output-Steuerung erwartete Leistungen im Voraus und überprüft deren Erreichen mit standardisierten Kompetenztests. Die Forderung nach einem Wechsel von der Input- zur Output-Steuerung folgt einer Argumentationsfigur, die sich kurz so zusammenfassen lässt:

- Deutschlands 15-Jährige haben bei PISA schlecht abgeschnitten.
- Deutschland hat ein Input-System.
- Viele der erfolgreicheren Länder haben eine Output-Steuerung.
- Also ist das Input-System schuld an der »PISA-Katastrophe«.
- Damit die Schülerleistungen besser werden, muss Deutschland auch ein Output-System einführen.

- Dessen Effektivität hat sich darin erwiesen, dass die Leistungen der deutschen Schüler/innen im Vergleich zu 2000 besser geworden sind.

Damit wird der Vorwurf erhoben, bisherige Formen der Steuerung und Evaluation hätten keine Veränderungen bewirkt. Dem ist entgegenzuhalten, dass sich im Grundschulbereich schon zwischen 1980 und 2000 vergleichsweise viel verändert hat. Genau diese evolutionären Reformen von unten drohen nun durch PISA & Co. verloren zu gehen. Dennoch behaupten Bos u. a. (2007b, S. 1): »Wie die Erfahrungen aus den Ländern, die in internationalen Schulleistungsstudien regelmäßig sehr gut abschneiden (z. B. Kanada, England, Finnland, die Niederlande und Schweden), zeigen, wirkt sich die regelmäßige ›Outputkontrolle‹ lang- und mittelfristig ausgesprochen förderlich auf die Kompetenzentwicklung der Schülerschaft aus.«

Für Deutschland schätzen andere Forscher/innen die Folgen des Systemwechsels zu einer Output-Steuerung sehr viel zurückhaltender ein (vgl. etwa die Beiträge zu dem Sammelband von Altrichter/Maag-Merki 2010). Denn wie verträgt sich die Einschätzung von Bos u. a. mit den doch recht unterschiedlichen Befunden zum Erfolg des deutschen Bildungssystems, wie wir sie etwa in den verschiedenen internationalen Leseuntersuchungen finden (vgl. Baumert u. a 2001; Bos u. a. 2003; 2007a; Prenzel u. a. 2004; 2007a)? So hat Deutschland mit Input-Steuerung 2001 und 2006 in der Grundschulstudie IGLU* deutlich besser abgeschnitten als 2000 und 2006 bei PISA am Ende der Pflichtschulzeit – und parallel dazu die USA mit Output-System ebenfalls in der Grundschule besser als in der Sekundarstufe. Zweitens: Die USA haben zwar mit ihrem schon lange etablierten Output-System bei PISA 2000 besser abgeschnitten als Deutschland mit seinem Input-System, doch 2001 bei IGLU liegen beide gleichauf. Zudem haben sich die Schülerleistungen in den USA trotz jährlicher Schulleistungstests in den vergangenen 30 Jahren kaum verbessert (vgl. NCES 2010).

Weiter: Wie ist zu erklären, dass in Kanada mit Output-System verschiedene Provinzen bei IGLU-2006 von 533 bis 560 Punkte streuen und damit teilweise vor und teilweise hinter Deutschland (548 Punkte) liegen? Und innerhalb Deutschlands gibt es größere

Unterschiede zwischen den Bundesländern, die alle input-orientiert gesteuert werden, als zwischen Deutschland und Vergleichsländern mit Output-Steuerung. Nachdenklich stimmt auch, dass Schweden, England und die Niederlande mit Output-System bei IGLU 2001 zur Spitzengruppe zählen, 2006 dagegen trotz der nun schon länger etablierten Output-Steuerung zurückfallen, dass sich dagegen die deutsche Grundschule unter dem Input-System von der IEA*-Studie 1991 bis zu IGLU 2001 rechnerisch pro Jahr doppelt so stark verbessert wie von 2001 bis 2006 unter der allmählichen Umstellung auf ein Output-System (Bos u. a. 2007a, S. 152) – während sich der Punktedurchschnitt der Sekundarstufe von PISA 2000 bis 2006 im gleichen Umstellungsprozess fast gar nicht verändert.

Das Problem mit der eingangs skizzierten Argumentation ist insgesamt: Nicht alle Länder, die besser als Deutschland abgeschnitten haben, sind output-gesteuert. Länder, die besser abgeschnitten haben, unterscheiden sich auch in weiteren Merkmalen vom deutschen System (Dauer der gemeinsamen Schulzeit; Zeitpunkt, ab dem benotet wird, usw.). Zudem zeigen sich in Ländern, die bei PISA oder IGLU besser als Deutschland abgeschnitten haben, in anderen Bereichen Probleme, die mit der Output-Orientierung verbunden sind, sodass dem behaupteten Nutzen auf der einen Seite Kosten auf der anderen gegenüberstehen, die man sich beim Transfer des Systems mit »einkauft«.

Insofern scheint die Leistungsfähigkeit eines Systems nur sehr begrenzt mit dem Steuerungsmodell zusammenzuhängen bzw. in starkem Maße durch Drittfaktoren beeinflusst (vgl. zu kritischen Einschätzungen der KMK-Strategie in Deutschland: Kuhn 2014; Maritzen 2014). Hinzu kommt, dass der Begriff der »Steuerung« ein technisches Verständnis gesellschaftlicher Interventionen unterstellt, das der nur sehr begrenzten Planbarkeit und Kontrollierbarkeit solch komplexer Systeme wie des Bildungswesens nicht gerecht wird (s. Kap. 2.5 und die ausführliche Kritik an diesem Paradigma bei Herzog 2013, Kap. 3 und 5). Wir haben es mit einem Paradox zu tun (Taleb 2014, S. 26): Je mehr wir wissen, desto komplexere – technische wie soziale – Systeme entwickeln wir; je komplexer ein System, desto schwieriger wird es, sein Verhalten vorauszusagen.

Das heißt: je höher unser Wissensstand, desto geringer unsere Fähigkeit zur Vorhersage und Steuerung der Umwelt.

Selbst Hermann Lange, als Hamburger Staatsrat einer der Protagonisten der »empirischen Wende« im Verhältnis von Erziehungswissenschaft und Bildungspolitik, hat darauf verwiesen, »dass die Steuerung eines komplexen Systems nach dem einfachen Muster von Regelung und Ausführung nicht funktioniert. In der Organisationsliteratur wird deshalb auch gelegentlich davon gesprochen, dass man komplexe soziale Systeme nicht ›steuern‹, sondern allenfalls ›segeln‹ könne« (2000, S. 20 f., zit. nach Tillmann 2008, S. 1).

Alternative Perspektiven

Wie schon in Kapitel 2.7 erläutert, ist ein nach Stufen differenziertes System der Rechenschaftspflichten und Evaluationsformen nötig (vgl. zu kulturell bedingt unterschiedlichen Modellen in den OECD-Ländern ➔ »Experimental decision making and responsive accountability« 1980). Ideen für ein demokratisches Verfahren, in dem die Sicht der Betroffenen vor Ort stärker berücksichtigt wird, skizziert Retzl (2010, S. 360-363). Außerdem ist es wichtig, zentrale System-Monitorings (über punktuelle Leistungstests) durch andere Indikatoren für die Qualität des Bildungswesens zu ergänzen, wie es »Bildung auf einen Blick« der OECD (2007b und folgende Jahre) und innerhalb Deutschlands das Konsortium Bildungsberichterstattung (2006 ff.) versuchen. Das aber reicht nicht aus.

Für die Entwicklung einzelner Schulen ist eine umfassendere und kontextsensible Evaluation notwendig, und die individuelle Leistungsrückmeldung ist zu einer »dialogischen Lernbeobachtung« anhand förderorientierter Aufgaben zu entwickeln, die einen Blick unter die Oberfläche des Messbaren erlauben (s. zum Konzept einer »Pädagogischen Leistungskultur« Kap. 6.3). Standardisierte Tests und flächendeckende Erhebungen können – bei entsprechender Einordnung – einen bedeutsamen Beitrag zur Einschätzung der Fachleistungen von Schulen und Schüler/inne/n leisten. Insofern ist es wichtig, einerseits die latente Test-Aversion vieler Lehrer/innen zu überwinden und ihnen andererseits Pools von (fachdidaktisch

kommentierten!) Testaufgaben mit repräsentativen Vergleichswerten zur Verfügung zu stellen. PISA & Co. müssen inhaltlich und methodisch durch komplementäre Formen informeller Evaluation ergänzt werden. Vor allem ist es notwendig, die Evaluation vor Ort als eigenständige Aufgabe mit besonderen Anforderungen und Möglichkeiten zu fördern.

Damit dies keine abstrakte Forderung bleibt, stelle ich in Kapitel 5.3 den reformpädagogischen Schulverbund »Blick über den Zaun« als Beispiel für eine Evaluation durch Peer-Review auf Schul- und Unterrichtsebene vor.

4.4 Fazit

Politisch bedeutet die Zuständigkeit für Evaluation Macht – und damit für die Betroffenen auch Abhängigkeit. Man darf sie nicht allein Experten überlassen, denn die mit ihr verknüpften Probleme sind technisch nicht zu lösen.

Schule ist zwar ein öffentlicher Raum, und dieser Status schließt ein, dass alle Beteiligten rechenschaftspflichtig sind. Dafür brauchen sie aber je nach Aufgabe unterschiedliche Verfahren und Unterstützung. Ein demokratisches Verständnis von Evaluation fordert, die Betroffenen durch externe Beurteilungen nicht zu entmündigen, sondern sie in ihrer persönlichen Evaluations- und Problemlösekompetenz zu stärken.

Dezentrale Evaluation kann bildungspolitisch orientierte Studien wie PISA nicht ersetzen. Beide Ansätze haben ihre spezifische Funktion in einem umfassenderen Rechenschaftssystem. Dabei sind mir abschließend drei Punkte wichtig:

- Der PISA-Stil muss auf Systemevaluation begrenzt werden und darf nicht zum Paradigma für Evaluation und Bildungsforschung generell werden. Insbesondere sind der Sinn und die Form flächendeckender jährlicher Lernstandserhebungen zu überdenken (s. Kap. 5.2).
- Auch auf der Systemebene muss die Scheinpräzision von Zahlen aus Erhebungen mit Leistungstests immer wieder bewusst ge-

macht und ihre mehrperspektivische Deutung schon bei der Veröffentlichung gesichert werden (vgl. unseren Vorschlag für die Einrichtung einer ➔ »Evaluationsstelle für nutzerorientierte Bildungsforschung« 2006).

- Die Ressourcen für Evaluation dürfen nicht auf die zentrale Evaluation konzentriert werden. Wir brauchen eine vergleichbare politische, wissenschaftliche und finanzielle Unterstützung für die Entwicklung der Evaluationskompetenz vor Ort – und dafür, Folgerungen aus den Ergebnissen ziehen zu können, z. B. für die Entwicklung schwacher Schulen und die Förderung von Kindern mit besonderen Schwierigkeiten.

PISA & Co. können unser Evaluationsrepertoire nur bereichern, wenn wir das Instrumentarium nicht überfordern.

7. Zehn Thesen zur Diskussion

Um den Nutzen empirischer Studien zu begründen, darf man nicht gute Forschung mit schlechter Praxis vergleichen. In konkreten Handlungssituationen sind erfahrene Praktiker/innen klugen Forscher/inne/n meist überlegen. Weder ist Bildungsforschung zu Vorschriften berechtigt noch ist sie in der Lage, Handlungsentscheidungen aus Daten abzuleiten. Ihre Aufgabe ist Anregung, Herausforderung und kritische Befragung.

Zur Begründung:

1 **Fieber zu messen, ist noch keine Diagnose – und Fieber zu senken, keine Therapie**. Denn gleiche Testwerte bedeuten nicht gleiche Kompetenzen und ein Anstieg richtiger Lösungen nicht immer eine Verbesserung der Leistungen.

2 **Menschliche Verhaltensweisen – auch in standardisierten Tests – sind immer mehrdeutig:** Dieselben Aufgaben, aber auch deren Lösungen, werden von anderen Beteiligten je nach ihren Erfahrungen und Theorien unterschiedlich gelesen.

3 **Die Wirkungen pädagogischer Maßnahmen sind konzept- und kontextabhängig**. Menschen und ihr Lernen lassen sich nicht vermessen wie Maschinen und technische Prozesse – und noch vermessener wäre es zu glauben, sie ließen sich vergleichbar steuern und planen.

4 **Erfolge und Misserfolge von Interventionen streuen innerhalb eines jeden Ansatzes je nach Kontext breit**. Deshalb kann die Forschung zwar jeweils besondere Potenziale und Risiken benennen, aber nur in dieser vorsichtigen Form generelle Qualitätsurteile über konkurrierende Konzepte fällen.

5 **Zahlen sprechen nicht für sich.** Je komplexer die statistischen Verfahren, desto begründungsbedürftiger ist ihre Wahl und

desto interpretationsbedürftiger sind ihre Ergebnisse. Sobald es zwei oder mehr Studien gibt, haben wir – wie bei der Alltagserfahrung – oft widersprüchliche Ergebnisse – und damit denselben Interpretationsbedarf, der ebenfalls personabhängig befriedigt wird.

6 **Die Stärken von Großstudien mit standardisierten Instrumenten sind zugleich ihre Schwächen**: Sie können nur wenige Faktoren erfassen, die Kategorien der Untersuchungen sind notwendig abstrakt und individuelle Bedeutungen gehen verloren.

7 **Wahrscheinlichkeitsaussagen, die aus der die Untersuchung großer Populationen gewonnen werden, eignen sich nur sehr eingeschränkt für die Evaluation einzelner Schulen, Lehrer/innen oder Schüler/innen**. Forschung muss sich auf unterschiedliche Zielgruppen einstellen: Politik und Verwaltung sind primär an Häufigkeiten und Durchschnitten interessiert, Lehrer/innen und Eltern an Besonderheiten und Streuungen.

8 **Anders als im Verhältnis von Physik und Bautechnik lassen sich Allgemeinaussagen in der Pädagogik nicht einfach »anwenden«.** Sie müssen kontextbezogen interpretiert werden. Insofern kann „Evidenzbasierung“ Subjektivität nicht ausschalten.

9 **Die sozialen Kontexte und die politischen Implikationen von Evaluation werden nicht zureichend reflektiert.** Technisch-methodisch wird in der heutigen Bildungsforschung auf einem sehr viel höheren Niveau gearbeitet und diskutiert.

10 **Evaluation als Blick von außen ist wichtig, um scheinbare Selbstverständlichkeiten in der Praxis infrage zu stellen - „von außen“ heißt aber nicht zwingend »von oben«.** Dialog und Unterstützung statt Autoritätsanspruch und Vorschrift – nur so wird Evaluation im Bildungswesen produktiv. Die besondere Kompetenz von Praktiker/inne/n anzuerkennen bedeutet: Man

kann nicht für sie Fortbildung zum Verständnis empirischer Methoden fordern, wenn man nicht auch von die Wissenschaftler/innen verlangt zu lernen, was es heißt, unter besonderen Bedingungen und unter Zeitdruck Fallentscheidungen zu fällen.

Literatur

Ackeren, I. van, u. a. (2011): Evidenzbasierte Schulentwicklung. Ein Forschungsüberblick aus interdisziplinärer Perspektive. In: Die Deutsche Schule, 103. Jg., H. 2, 170-184.

Ackeren, I. van, u. a. (2013): Evidenzbasierte Steuerung im Bildungssystem? Befunde aus dem SteBis-Verbund. Die Deutsche Schule, 12. Beiheft. Waxmann: Münster.

Adelman, C./Elliott, J. (1975): Lehrer als Evaluatoren. In: Thema Curriculum, H. 5, 57-64.

AERA (2003): Standards and tests: Keeping them aligned. In: Research Points (American Educational Research Association), Vol. 1, No. 1, 1-4.

Albrecht, H. (2005): Kritik der reinen Norm. Klinische Forschung hilft vor allem Standardpatienten. In: DIE ZEIT, Nr. 2, v. 5.1.2005, S. 25.

Albrecht, H. (2008): Jeder Patient zählt. Was klinische Großstudien übersehen. in: Die Zeit. Nr. 22 v. 21.5.2008, 33.

Alkin, M. C. (ed.) (2004): Evaluation roots: A wider perspective of theorists' views and influences. Sage Publication: Thousand Oaks, CA (2nd. ed. 2013).

Allport, G. W. (1949): Persönlichkeit. Struktur, Entwicklung und Erfassung der menschlichen Eigenart. Klett: Stuttgart (engl. 1937).

Altrichter, H. (2010). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In: Altrichter/Maag (2010, 219-254).

Altrichter, H./Maag Merki, K. (Hrsg.) (2010): Handbuch neue Steuerung im Schulsystem. VS Verlag für Sozialwissenschaften: Wiesbaden.

Altrichter, H./Posch, P. (2007): Lehrerinnen und Lehrer erforschen ihren Unterricht. Klinkhardt: Bad Heilbrunn (4. erw. Aufl.; 1. Aufl. 1998).

Amrein, A. L./Berliner, C. D. (2002): High-stakes testing, uncertainty, and student learning. Download: www.epaa.aus.edu/epaa/v10n18/

Amrein-Beardsley, A. (2009): The unintended, pernicious consequences of »staying the course« on the United States' No Child Left Behind policy. In: International Journal of Education Policy and Leadership, Vol. 4, No. 6, 1-13.

Antes, G. (Hrsg.) (2013): Wo ist der Beweis? Plädoyer für eine evidenzbasierte Medizin. Huber: Bern.

Arbeitsgruppe am Max-Planck-Institut für Bildungsforschung (Hrsg.) (1979): Das Bildungswesen in der Bundesrepublik Deutschland – Ein Überblick für Eltern, Lehrer, Schüler. Rowohlt: Taschenbuch: Reinbek.

Arbeitsgruppe Primarstufe (2006): Sind Noten nützlich und nötig? Zifferzensuren und ihre Alternativen im empirischen Vergleich. Eine wissenschaftliche

Expertise des Grundschulverbandes, erstellt von der Arbeitsgruppe Primarstufe an der Universität Siegen (Hans Brügelmann mit Axel Backhaus u. a.). Grundschulverband e.V.: Frankfurt (2. erg. Aufl. 2014).

Argyris, C./Schön, D. (1978): Organization learning: A theory of Action perspective. Addison Wesley: Reading (MA).

Arnold, R. (2014): Begriffe sind Fenster. Systemische Pädagogik von A bis Z. Antworten, Algorithmen und Akronyme. Schneider Hohengehren: Baltmannsweiler.

Artelt, C., u. a. (2001): Lesekompetenz: Testkonzeption und Ergebnisse. In: Baumert u. a. (2001, 69-137).

Artelt, C./Baumert, J. (2004): Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs. In: Zeitschrift für Pädagogische Psychologie, 18. Jg., H. 3-4, 171-185.

Asendorpf, J. (1995): Persönlichkeitspsychologie: Das empirische Studium der individuellen Besonderheit aus spezieller und differentieller Perspektive. In: Psychologische Rundschau, 46. Jg., 235-247.

Aspland, R. (1997): Managing quality in the UK education system. StudienVerlag: Innsbruck.

Au, W. (2007): High-stakes testing and curricular control: A qualitative metasynthesis. In: Educational Researcher, Vol. 36, No. 5, 258-267.

Backhaus, A. (2011): LESEN&SCHREIBEN. Ein Aufgabenset für die dialogische Förderdiagnostik in der Alphabetisierung. In: Alfa-Forum, Nr. 76, 21-23.

Backhaus, A., u. a. (2009): »Blick über den Zaun«: Schulen lernen von Schulen. Vorschläge zur Planung und organisatorischen Ausgestaltung von Peer-Reviews durch kritische Freunde. Reformpädagogische Arbeitsstelle »Blick über den Zaun« an der Universität: Siegen. www.blickueberdenzaun.de/index.php/textarchiv/167-schulen-lernen-von-schulen-handreichung-fuer-schulbesuche

Backhaus, A., u. a. (2010): »Was ist Sache?« Didaktisches Verfahren und Dialogisches Instrument für die Arbeit mit Gebrauchstexten. Handreichung und Arbeitsheft. Bundesverband Alphabetisierung: Münster.

Backhaus, A., u. a. (2011): »Lesen & Schreiben«. Lese- und Schreibaufgaben für die Lernbeobachtung in der Erwachsenenalphabetisierung. Bundesverband Alphabetisierung: Münster.

Backhaus, A./Brügelmann, H. (Hrsg.) (2010): Was ist eine gute Schule? Unsere Standards für Kinder (Erprobungsfassung). Reformpädagogische Arbeitsstelle des Verbunds »Blick über den Zaun« an der Universität: Siegen (Download: www.blickueberdenzaun.de/images/downloads/kinderstandardssynopse.pdf)

Backhaus, A./Knorre, S. (Hrsg.) (2008): Demokratische Grundschule - Mitbestimmung von Kindern über ihr Leben und Lernen. Arbeitsgruppe Primarstufe/FB2 der Universität. Universi Verlag: Siegen.

Backhaus, A./Knorre, S. (2010): Dialogische Diagnostik. Deutungen und Einsichten von Kindern ernst nehmen. In: Die Grundschulzeitschrift, 24. Jg., H. 234, 10-12.

Balhorn, H./Brügelmann, H. (Hrsg.) (1995): Rätsel des Schriftspracherwerbs. Neue Sichtweisen der Forschung. Auswahlband Theorie der DGLS-Jahrbücher 1-5. Libelle: CH-Lengwil.

Bank, V. (Hrsg.) (2005): Vom Wert der Bildung. Bildungsökonomie in wirtschaftspädagogischer Perspektive neu gedacht. Haupt: Bern.

Bank, V./Heidecke, B. (2009): Gegenwind für PISA. Ein systematisierender Überblick über kritische Schriften zur internationalen Vergleichsmessung. In: Vierteljahresschrift für wissenschaftliche Pädagogik, 85. Jg., H. 3, 361-372.

Bartnitzky, H. (2005): VerA Deutsch 2004: Ungeeignet und bildungsfern. In: Grundschule aktuell, H. 89, 10-16.

Bartnitzky, H. (2007): VerA Deutsch 2007: »Alles Geschmackssache«? – Nein, auch eine Sache der Qualität! In: Grundschule aktuell, H. 99, 5-10.

Bartnitzky, H. (2008): VerA Deutsch 2008: Wie Kindern durch Vergleichsarbeiten Unrecht geschieht. In: Grundschule aktuell, H. 103 4-8.

Bartnitzky, H., u. a. (1999): Zur Qualität der Leistung – 5 Thesen zu Evaluation und Rechenschaft der Grundschularbeit. Grundschulverband – Arbeitskreis Grundschule e. V.: Frankfurt. Auch in: Schmitt (1999, 165-196). Download: www.grundschulverband.de/bildungspolitik/zur-qualitaet-der-leistung/

Bartnitzky, H., u. a. (Hrsg.) (2005&2006&2007): Pädagogische Leistungskultur: Materialien für Klasse 1/2 und Klasse 3/4. Beiträge zur Reform der Grundschule, Bd. 119 & 121 & 123. Grundschulverband: Frankfurt.

Bartnitzky, H., u. a. (Hrsg.) (2012): Fördern – warum, wer, wie, wann? Individuell fördern – Kompetenzen stärken in der Eingangsstufe (Kl. 1 und 2). Grundschulverband: Frankfurt.

Bartnitzky, H./Speck-Hamdan, A. (Hrsg.) (2004): Pädagogische Leistungskultur: Leistungen der Kinder wahrnehmen – würdigen – fördern. Beiträge zur Reform der Grundschule, Bd. 118. Grundschulverband: Frankfurt.

Bauer, S. C. (2000): Should achievement tests be used to judge school quality? In: Education Policy Analysis Archives, Vol. 8, No. 46, 1-19.

Bauersfeld, H. (1972): Einige Bemerkungen zum Frankfurter Projekt und zum »alef«-Programm. In: Schwartz u. a. (1972, 237-246).

Baumert, J. (2014): Im Fokus der Lehrerfortbildung: Optimierung des Lernens. Jürgen Baumert im Gespräch mit Botho Priebe. In: Lernende Schule, H. 68, 22-24.

Baumert, J., u. a. (1980): Bildung in der Bundesrepublik Deutschland. Bd. 2: Gegenwärtige Probleme. Hrsg. von der Projektgruppe Bildungsbericht am Max-Planck-Institut für Bildungsforschung. Frankfurt.

Baumert, J., u. a. (2007): Was messen internationale Schulleistungsstudien? Resultate kumulativer Wissenserwerbsprozesse. Eine Antwort auf Heiner Rindermann. In: Psychologische Rundschau, 58. Jg., H. 2, 118–145.

Baumert, J., u. a. (2008): Steuerungswissen, Erkenntnisse und Wahlkampfmunition: Was liefert die empirische Bildungsforschung? – Eine Antwort auf Klaus Klemm. Institut für Pädagogik der Naturwissenschaften: Kiel. Download: www.ipn.uni-kiel.de/pisa/ReplikKlemm.pdf

Baumert, J., u. a. (Hrsg.) (2000a): TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Leske+Budrich: Opladen.

Baumert, J., u. a. (Hrsg.) (2000b): TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 2: Mathematische und naturwissenschaftliche Grundbildung am Ende der gymnasialen Oberstufe. Leske+Budrich: Opladen.

Baumert, J., u. a. (Hrsg.) (2001): PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske+Budrich: Opladen.

Baumert, J., u. a. (Hrsg.) (2003): PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Leske+Budrich: Opladen.

Baumert, Jürgen u. a. (Hrsg.) 2002: PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich. Opladen: Leske+Budrich

Bayrhuber, H., u. a. (Hrsg.) (2004): Konsequenzen aus PISA. Perspektiven der Fachdidaktiken. StudienVerlag: Innsbruck.

Beck-Bornholdt, H.-P./Dubben, H.-H. (2001): Der Hund, der Eier legt. Erkennen von Fehlinformationen durch Querdenken. Rororo: Reinbek (überarb. Neuausgabe; 1. Aufl. 1997).

Behnken, I., u. a. (2005): Lernen, Bildung, Partizipation. Die Perspektive der Kinder und Jugendlichen. Befragung zum 8. Kinder und Jugendbericht des Landes NRW (im Auftrag des Ministeriums für Schule, Jugend und Kinder). Siegener Zentrum für Kindheits-, Jugend- und Biographieforschung der Universität (SIZE): Siegen/ProKids: Herten.

Behnken, I., u. a. (Hrsg.) (1999): Leistung. Schüler ’99. Erhard Friedrich Verlag: Seelze.

Behnken, I./Jaumann, O. (Hrsg.) (1995): Kindheit und Schule. Kinderleben im Blick von Grundschulpädagogik und Kindheitsforschung. Juventa: Weinheim/München.

Beilock, S. L., et al. (2004): More on the fragility of performance: Choking under pressure in mathematical problem solving. In: Journal of Experimental Psychology: General, Vol 133, No. 4, 584-600.

Beywl, W./Zierer, K. (2013): Lernen sichtbar machen. Zur deutschsprachigen Ausgabe von »Visible Learning«. In: Hattie (2013, VI-XXVI).

Bellmann, J. (2013): Projekt »Nicht-intendierte Effekte Neuer Steuerung im Schulsystem«. Kurzzusammenfassung erster Projektergebnisse (Arbeitspapier). Abteilung Bildungstheorie und Bildungsforschung/Westfälische Wilhelms-Universität: Münster (ref. in Böttcher 2013).

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

Berliner, D. C. (2002): Educational research: The hardest science of all. In: Educational Researcher, Vol. 31, No. 8, 18-20.

Biesta, G. (2011): Warum »What works« nicht funktioniert: Evidenzbasierte pädagogische Praxis und das Demokratiedefizit der Bildungsforschung. In: Bellmann/Müller (2011, 95-121).

Biesta, G. (2014): The beautiful risk of education. Paradigm Publishers: London.

Biesta, G. J. J. (2010a): Good education in an age of measurement: Ethics, politics, democracy. Paradigm Publ.: Boulder, Col.

Biesta, G. J. J. (2010b): Valuing what we measure or measuring what we value? On the need to engage with the question of purpose in educational evaluation, assessment, and measurement. In: Böttcher u. a. (2010, 35-46).

BLLV u. a. (2014): 10 Jahre VerA – das Ziel ist verfehlt. Schulen brauchen Unterstützung statt Testeritis. In: Grundschule aktuell, H. 126, 25.

Blossfeld, H.-P., u. a. (2011): Grundidee, Konzeption und Design des Nationalen Bildungspanels für Deutschland (NEPS Working Paper No. 1). Otto-Friedrich-Universität: Bamberg. Download: www.uni-bamberg.de/fileadmin/inbil/Publikationen/Working-Papers/WP_I.pdf

Blum, W./Neubrand, M. (2004): Der schiefe Blick auf Pisa. In: Süddeutsche Zeitung v. 11.12.2004, 13.

BMBF (Hrsg.) (2008): Rahmenprogramm zur Förderung der empirischen Bildungsforschung. Bildungsforschung Band 22. Bundesministerium für Bildung und Forschung: Berlin.

BMBF (Hrsg.) (2014): Bildungsforschung 2020 – Herausforderungen und Perspektiven. Bundesministerium für Bildung und Forschung: Berlin.

BMFSFG (Hrsg.) (1993): Übereinkommen über die Rechte des Kindes. UN-Konvention im Wortlaut mit Materialien. Texte in amtlicher Übersetzung. Bundesministerium für Familie, Senioren, Frauen und Gesundheit: Bonn.

Bohl, T./ Kiper, H. (2009) (Hrsg.): Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren. Julius Klinkhardt: Bad Heilbrunn.

Bommes, M./Radtke, F.-O. (1993): Institutionalisierte Diskriminierung von Migrantenkindern. In: Zeitschrift für Pädagogik, 39. Jg., 483-497.

Bos, W., u. a. (Hrsg.) (2003): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Waxmann: Münster.

Bos, W., u. a. (Hrsg.) (2005): IGLU. Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien. Waxmann: Münster.

Bos, W., u. a. (Hrsg.) (2007a): IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Waxmann: Münster.

Bos, W., u. a. (2007b): Zusammenfassung wichtiger Ergebnisse zu Kompetenzen und Einstellungen von Hamburger Schülerinnen und Schülern. In: Bos/Pietsch (2007, 1-8)

Bos, W., u. a. (Hrsg.). (2008). TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Waxmann: Münster.

Bos, W., u. a. (2012): IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Waxmann: Münster.

Bos, W./Pietsch, M. (Hrsg.) (2007): KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen. Waxmann: Münster.

Bosse, U., u. a. (2015): Noten, Raster und Berichte. Nach der Bielefelder Herbsttagung: Ulrich, Hans Brügelmann und Ulrich Hecker im Gespräch. In: Grundschule aktuell, Nr. 129, 3-5.

Böhme, K., u. a. (2014): Messen wir dasselbe? Zur Vergleichbarkeit des IQB-Ländervergleichs 2011 mit den internationalen Studien IGLU und TIMSS 2011. In: Unterrichtswissenschaft, 42. Jg., H. 4, 342-365.

Böttcher, W. (2013): Das Monitoring-Paradigma – Eine Kritik der deutschen Schulreform. In: Empirische Pädagogik, 27. Jg., H. 4, 5-21.

Böttcher, W., u. a. (Hrsg.) (2009): Evidenzbasierte Bildung: Wirkungsevaluation in Bildungspolitik und pädagogischer Praxis. Waxmann: Münster.

Böttcher, W., u. a. (Hrsg.) (2010): Evaluation, Bildung und Gesellschaft. Steuerungsinstrumente zwischen Anspruch und Wirklichkeit. Waxmann: Münster.

Böttcher, W./Keune, M. (2010): Funktionen und Effekte der Schulinspektion. Ausgewählte nationale und internationale Forschungsbefunde. In: Böttcher u. a. (2010, 151-164).

Bracey, G. W. (2005): Education's groundhog day. In: Education Week, Vol. 24, No 21, 38-39. Auch Download: www.edweek.org/ew/articles/2005/02/02/21bracey.h24.html

Bracey, G. W. (2009): Education Hell: Rhetoric vs. Reality. Education Research Service: Alexandria/Va.

Bremerich-Vos, A., u. a. (2005): Stellungnahme zur Kritik an VerA in »Grundschule aktuell«, Heft 89. In: Grundschule aktuell, H. 90, 3-6.

Brinkmann, A., u. a. (Red.) (1990): Lesen im internationalen Vergleich. Materialien zur Leseförderung und Leseforschung, Teil I. Materialien zur Leseförderung und Leseforschung, Bd. 2. Stiftung Lesen: Mainz.

Brinkmann, E., u. a. (2008ff.): ABC-Lernlandschaft. vpm/Klett: Stuttgart.

Brinkmann, E., u. a. (2013): ABC-Lernlandschaft. Lehrerhandreichung. vpm/Klett: Stuttgart (2. Aufl.)

Brinkmann, E./Brügelmann, H. (2005): Pädagogische Leistungskultur – Materialien für Klasse 1 und 2: Deutsch. Heft 3 in: Bartnitzky u. a. (2005).

Bromme, R./Prenzel, M. (Hrsg.) (2014): Von der Forschung zur evidenzbasierten Entscheidung. Die Darstellung und das öffentliche Verständnis der empirischen Bildungsforschung. Zeitschrift für Erziehungswissenschaft, Sonderheft, Bd. 17. VS Verlag für Sozialwissenschaften: Wiesbaden.

Bromme, R., u. a. (2014): Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. Eine Analyse von Anforderungen an die Darstellung, Interpretation und Rezeption empirischer Befunde. In: Bromme/Prenzel (2014, 3-54).

Brügelmann, H. (1974): Towards checks and balances in educational evaluation – On the use of social control in research design. In: MacDonald/Walker (1974, 25-66; dt. Kurzfassung in: betrifft:erziehung H. 7, 1975, 37 – 42).

Brügelmann, H. (1980): Experimental decision making and responsive accountability. Expert report for »Basic Education Policies Project«. OECD/CERI: Paris. Download der Kurzfassung http://www2.agprim.uni-siegen.de/print-brue.htm.

Brügelmann, H. (1995): Kultur und Kontext. Gesellschaftliche und situative Bedingungen der Medienwirkung. In: Behnken/Jaumann (1995, 95-98).

Brügelmann, H. (1998): Öffnung des Unterrichts Befunde und Probleme der empirischen Forschung. In: Brügelmann u. a. (1998, 8-42, rev. in Brügelmann/Brinkmann 2008).

Brügelmann, H. (Hrsg.) (1999): Was leisten unsere Schulen? Erhard-Friedrich-Verlag: Seelze.

Brügelmann, H. (2003): Lese-Untersuchung mit dem Stolperwörter-Test. Abschlussbericht des Projekts LUST-1. Download: http://www2.agprim.uni-siegen.de/lust

Brügelmann, H. (2004): Leseleistungen von Lehrer/innen und Lehramtsstudierenden im Stolperwörter-Lesetest. Erste Befunde und ihre Deutung. Projekt LUST/FB 2 der Universität: Siegen. Download: http://www2.agprim.uni-siegen.de/lust/stolper_auswertung_lehramt.pdf

Brügelmann, H. (2005a): Schule verstehen und gestalten – Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil (bis 2008 fortlaufend aktualisiert unter: http://www2.agprim.uni-siegen.de/schuleverstehen).

Brügelmann, H. (2005b): Wahrheit durch Vera? Anmerkungen zum ersten Durchgang der landesweiten Leistungstests in sieben Bundesländern. In: Grundschule aktuell, Nr. 89, S. 7-9.

Brügelmann, H. (2006): International tests and comparisons in education performance: A pedagogical perspective on standards, core curricula, and the measurement of the quality of schooling. In: Rotte (2006, 21-44).

Brügelmann, H. (2007): Scharfe Brillen, wache Augen und ein einfühlsamer Blick. Wie Reformschulen den fremden Blick kritischer Freunde am besten nutzen können: Zur Bedeutung von technischer Präzision und sozialer Kontrolle bei der Evaluation pädagogischer Standards. In: Schulverbund »Blick über den Zaun« (2007, 8-52).

Brügelmann, H. u. a. (1999): Was leisten unsere Schulen? Qualität und Evaluation von Unterricht in der Diskussion. Kallmeyersche Verlagsbuchhandlung: Seelze.

Brügelmann, H., u. a. (1994): »Schreibvergleich BRDDR« 1990/91. In: Brügelmann/Richter (1994, 129 134).

Brügelmann, H., u. a. (Hrsg.) (1998): Jahrbuch Grundschule. Fragen der Praxis Befunde der Forschung. Erhard Friedrich Verlag: Seelze.

Brügelmann, H.,/Backhaus, A. (2011): Kinder als Kinderschützer. In: Buchholz u. a. (2011, 309-318).

Brügelmann, H./Brinkmann, E. (1998): Die Schrift erfinden – Beobachtungshilfen und methodische Ideen für einen offenen Anfangsunterricht im Lesen und Schreiben. Libelle: CH-Lengwil (2. Aufl. 2005; Vorfassung »Die Schrift entdecken« 1984).

Brügelmann, H./Brinkmann, E. (2009): Öffnung des Anfangsunterrichts. Theoretische Prinzipien, unterrichtspraktische Ideen und empirische Befunde. Arbeitsgruppe Primarstufe/Universität. Univers Verlag: Siegen (1.Aufl. 2008).

Brügelmann, H./Brinkmann, E. (2013): Freies Schreiben im Anfangsunterricht? Eine kritische Übersicht über Befunde der Forschung. Grundschulverband: Frankfurt www.grundschulverband.de/fileadmin/grundschulverband/Download/Forschung/Zur_Leseforschung_Feb._07.pdf [Erstveröffentlichung 2006]

Brügelmann, H./Richter, S. (Hrsg.) (1994a): Wie wir recht schreiben lernen. Zehn Jahre Kinder auf dem Weg zur Schrift. Libelle Verlag: CH-Lengwil (2. Aufl. 1996).

Brügelmann, H./Richter, S. (1994b): Zur Bedeutung der »Bedeutung«: Unterschiede in den Rechtschreibleistungen von Jungen und Mädchen bei Wörtern mit geschlechtsspezifischer Bedeutung. In: Richter/Brügelmann (1994, 121-131).

Buchholz, T. , u. a. (Hrsg.) (2011): Kinderschutz in gemeinsamer Verantwortung von Jugendhilfe und Schule. VS Verlag für Sozialwissenschaften: Wiesbaden.

Buhren, C. (2011): Kollegiale Hospitation. Verfahren, Methoden und Beispiele aus der Praxis. Carl Link. Köln.

Buhren, C. G., u. a. (1998): Wege und Methoden der Selbstevaluation. Ein praktischer Leitfaden für Schulen. Beiträge zur Bildungsforschung Nr. 6. Institut für Schulentwicklungsforschung: Dortmund.

Bulmahn, E. u. a. (1999): Was müssen Schüler heute wissen? Eine Psychologie Heute – Umfrage unter deutschen Kultusministern. In: Psychologie Heute, August (1999), 60–64.

Burk, K., u. a. (Hrsg.) (2003): Kinder beteiligen – Demokratie lernen? Beiträge zur Reform der Grundschule, Bd. 116. Grundschulverband: Frankfurt.

Carle, U. (2002): Blick hinter den PISA-Spiegel: Bayern nicht mehr deutscher Meister, Bremen. Download: www.grundschulpaedagogik.uni-bremen.de/archiv/aggpaed_archiv.html

Carvalho, L. M./ Costa, E. (2014): Seeing education with one's own eyes and through PISA lenses: considerations of the reception of PISA in European countries. In: Discourse – Studies in the Cultural Politics of Education. Download: http://dx.doi.org/10.1080/01596306.2013.871449.

Clausen, J./Albrecht, H. (2013): Specialised databases of the Karl and Veronica Carstens Stiftung. Research Updates – Homeopathy, Vol. 2, 48-55.

Collani, E. v. (2001): OECD PISA - An Example of Stochastic Illiteracy? In: Economic Quality Control, Vol. 16, No. 2, 227-253.

Cortina, K. S. (2015): PIAAC und PISA: Pädagogisch paradoxe Parallelen. In: Zeitschrift für Pädagogik, 61. Jg., H. 2, 223-242.

Cronbach, L. (1975): Beyond the two disciplines of scientific psychology. In: American Psychologist, Vol. 30, 671–684.

Dammer, K.-H. (2015): Vermessene Bildungsforschung. Wissenschaftsgeschichtliche Hintergründe zu einem neoliberalen Herrschaftsinstrument. Schneider Hohengehren: Baltmannsweiler.

de Boer, H./Reh, S. (Hrsg.) (2011): Beobachtung in der Schule – Beobachten lernen. VS Verlag für Sozialwissenschaften: Wiesbaden (in Vorbereitung).

DeGEval (Hrsg.) (2002): Standards für Evaluation. Deutsche Gesellschaft für Evaluation e.V.: Köln.

Dehn, M. (1994): Schlüsselszenen zum Schriftspracherwerb. Arbeitsbuch zum Lese- und Schreibunterricht in der Grundschule. Beltz: Weinheim.

Dehn, M. (2013): Zeit für die Schrift – Lesen und Schreiben im Anfangsunterricht. Mit Beiträgen von Petra Hüttis-Graff. Cornelsen Scriptor: Berlin.

Demmer, M. (Hrsg.) (2000): Was leisten Leistungsvergleiche (nicht)? Hrsg. vom Bildungs- und Förderungswerk der Gewerkschaft Erziehung und Wissenschaft im DGB e. V.: 60489 Frankfurt (Reifenberger Str. 21).

Demmer, M., u. a. (2007): Mit Qualitätsanalyse Schule entwickeln – Konzepte mit und ohne externe Evaluation. PISA-Info 19/2007 (Nachdruck aus: Dokumentation zum »forum bildung« didacta – die Bildungsmesse 2007 Köln). Gewerkschaft Erziehung und Wissenschaft: Frankfurt.

Demmer, M. (2014): VerA (Vergleichsarbeiten): Praktiken, Einstellungen und Erfahrungen der GEW-Mitglieder in Grundschulen. Pilotstudie auf der Basis einer GEW - Mitgliederbefragung. Download: www.gew.de/Binaries/Binary112804/MD_VerA-Befragung_GEW_2014-Auswertung_und_KommentierungDECKBLATT.pdf

Devine, D. (2003): Children, power and schooling: How childhood is structured in the primary classroom. Trentham Books: Stoke-On-Trent.

Dilthey, W. (1957): Gesammelte Schriften. Bd. 5. Teubner: Stuttgart.

DiSalvo, D. (2014):Was Ihr Gehirn glücklich macht ... und warum Sie genau das Gegenteil tun sollten. Springer Spektrum: Berlin/ Heidelberg.

Ditton, H. (2010): Wie viel Ungleichheit durch Bildung verträgt eine Demokratie? In: Zeitschrift für Pädagogik, 56. Jg., H. 1, 53-68.

Dohmen, G. (2001): Das informelle Lernen. Die internationale Erschließung einer bisher vernachlässigten Grundform menschlichen Lernens für das lebenslange Lernen aller. Bundesministerium für Bildung und Forschung: Bonn. Download: www.bmbf.de/pub/das_informelle_lernen.pdf .

Dolin, J. (2007): PISA – an example of the use and misuse of large-scale comparative tests. In: Hopmann u. a. (2007, 93-126).

Dowling, A. (2008): Output measurement in education. Download: http://research.acer.edu.au/policy_analysis_misc/2

Ehmke, T./Siegle, T. (2006): Wie gut schneiden Erwachsene bei PISA ab? Befunde zur mathematischen Kompetenz von Eltern und Kindern in der deutschen PISA-Stichprobe. Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN). Universität: Kiel.

Ehmke, T., u. a. (2013): Veränderungen der Lesekompetenz von PISA 2000 nach PISA 2009. Die Rolle von Unterschieden in den Bildungswegen und der Zusammensetzung der Schülerschaft. In: Jude/Klieme (2013, 132-150).

Elley, W.B. (1992): How in the world do students read? International Association for the Evaluation of Educational Achievement: s'Gravenhage.

Elliott, J./Adelman, C. (1973): Supporting teachers' research in the classroom. In: New Era, Vol. 54, 210-213.

Elliott, J./Kushner, S. (2007): The need for a manifesto for educational programme evaluation. In: Cambridge Journal of Education, Vol. 37, No. 3, 321-336.

Faber, G./Billmann-Mahecha, E. (2010): Notengebung im Spiegel wissenschaftlicher Untersuchungen. Probleme, Erfordernisse und Möglichkeiten aus pädagogisch-psychologischer Sicht. In: Lernchancen, 13. Jg., H. 74, 30-33.

Fauser, P., u. a. (Hrsg.) (2007): Was für Schulen! Portraits der Preisträgerschulen und der nominierten Schulen des Wettbewerbs 2006. Klett Kallmeyer: Stuttgart.

Fauser, P., u. a. (Hrsg.) (2008): Was für Schulen! Profile, Konzepte und Dynamik guter Schulen in Deutschland. Portraits der Preisträgerschulen und der nominierten Schulen des Wettbewerbs 2007. Klett Kallmeyer: Stuttgart.

Fauser, P., u. a. (Hrsg.) (2009): Was für Schulen! Wie gute Schule gemacht wird - Werkzeuge exzellenter Praxis. Der Deutsche Schulpreis 2008. Robert-Bosch-Stiftung: Stuttgart. Piper: München.

Fay, J. (2010): Die Entwicklung der Rechtschreibkompetenz beim Textschreiben. Eine empirische Untersuchung von Klasse 1 bis 4. Peter Lang: Frankfurt.

Fend, H. (1998): Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung. Juventa: Weinheim.

Fend, H. (2008): Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität. VS Verlag für Sozialwissenschaften: Wiesbaden.

Fertig, M. (2004): What can we learn from international student Performance studies? Some methodological remarks. In: RWI Discussion Papers 23. Rheinisch-Westfälisches Institut für Wirtschaftsforschung: Essen.

Fickermann, D./ Maritzen, N. (Hrsg.) (2014). Grundlagen für eine daten- und theoriegestützte Schulentwicklung. Waxmann: Münster.

Fischer, D. (Hrsg.) (1982): Fallstudien in der Pädagogik. Ekkehard Faude: Konstanz.

Fischer, D. (Hrsg.) (1983): Lernen am Fall. Zur Interpretation und Verwendung von Fallstudien in der Pädagogik. Ekkehard Faude: Konstanz.

Flitner, E. (2006): Pädagogische Wertschöpfung. Zur Rationalisierung von Schulsystemen durch public-private-partnerships am Beispiel von PISA. In: Oelkers u. a. (2006, 245-266).

Forster, E. (2014): Kritik der Evidenz. Das Beispiel evidence-informed policy research der OECD. In: Zeitschrift für Pädagogik, 60. Jg. H. 6, 890-907.

Freud, S. (1954): Zur Psychopathologie des Alltagslebens. Fischer-Taschenbuch 68: Frankfurt (1. Aufl. 1941)..

Fritz-Vannahme, J. (1996): Waren damals alle besser? In Frankreich traten Schüler von 1995 gegen die von 1925 an. In: Die Zeit, Nr.17 v. 19.4.1996, 41.

Füller, C. (2007a): Streit um die Pisa-Interpretation. Faktisch nur im Mittelmaß. in: TAZ vom 2.12.2007. Download: www.taz.de/1/zukunft/wissen/artikel/1/faktisch-nur-im-mittelmass/?src=SE&cHash=40e8c46e85

Füller, C. (2007b): Widersprüche bei Pisa. Pisa hat einen kleinen, fröhlichen Bruder. In: TAZ vom 5.12.2007. Download: www.taz.de/1/zukunft/wissen/artikel/1/pisa-hat-einen-kleinen-froehlichen-bruder/?src=SE&cHash=4fe51fe346

Funke, R. (2014): Erstunterricht nach der Methode Lesen durch Schreiben und Ergebnisse schriftsprachlichen Lernens – Eine metaanalytische Bestandsaufnahme. In: Didaktik Deutsch, 19. Jg., H. 36, 20-41.

Gaeth, F. (2009): PISA: Zukunft der Bildung oder Joulupukki? In: Politik-Poker v. 1.10.2009. Download: www.politik-poker.de/pisa-zukunft-der-bildung-oder-joulupukki.php

Gaeth, F. (2010): Sieger im Mittelmaß - die neue PISA Studie. In: Politik-Poker v. 29.6.2010. Download: www.politik-poker.de/sieger-im-mittelmass-die-neue-pisa-studie.php

Gallin, P./Ruf, U. (1990): Sprache und Mathematik in der Schule. Auf eigenen Wegen zur Fachkompetenz. Illustriert mit sechzehn Szenen aus der Biographie von Lernenden. Verlag Lehrerinnen und Lehrer Schweiz: Zürich (Neuauflage Friedrich/Kallmeyer: Seelze 1998).

Geißler, R. (1992): Die Sozialstruktur Deutschlands. Zur gesellschaftlichen Entwicklung mit einer Bilanz zur Vereinigung, Hrsg. von der Bundeszentrale für politische Bildung. VS Verlag für Sozialwissenschaften: Wiesbaden (4. überarbeite und aktualisierte Aufl. 2006).

Gigerenzer, G. (2007): Bauchentscheidungen. Die Intelligenz des Unbewussten und die Macht der Intuition. Bertelsmann: München.

Glumpler, E. (1985): Schullaufbahn und Schulerfolg türkischer Kinder. ebv Rissen: Hamburg.

Gräsel, C. (2114): Die Praxisrelevanz der Empirischen Bildungsforschung. Vortrag auf dem ZfE-Form »Kritik empirischer Bildungsforschung« am 5.12.2014 an der Universität Hamburg (voraussichtlich Veröffentlichung in Zeitschrift für Erziehungswissenschaft 2015).

Groeben, A. v. d., u. a. (2005): Unsere Standards. Ein Diskussionsentwurf, vorgelegt von »Blick über den Zaun » – Bündnis reformpädagogisch engagierter Schulen. In: Neue Sammlung, 45. Jg., H. 2, 253-297 (Download über www.BlickUeberDenZaun.de).

Grotlüschen, A./Riekmann, W. (Hrsg.) (2012): Funktionaler Analphabetismus in Deutschland. Ergebnisse der ersten leo. – Level-One Studie. Waxmann: Münster.

Grundschulverband (2003): Bildungsansprüche von Grundschulkindern - Standards zeitgemäßer Grundschularbeit. In: Grundschulverband aktuell, Nr. 81, 1-24.

Hagemeister, V. (2007): Langfristige Wirkung geringer Klassenfrequenzen. Download: www.pisa-kritik.de/files/Langfristige-Wirkung-geringer-Klassenfrequenzen.pdf

Haladyna, T. M. u. a. (1991): Raising standardized achievement test scores and the origins of test score pollution. In: Educational Researcher, Vol. 20, No. 5, 2-7.

Hamilton, D. (1976): Curriculum evaluation. Open Books: London.

Hamilton, D., et al. (eds.) (1977): Beyond the numbers game. Macmillan: London

Hartmann, M. (2002): Der Mythos von den Leistungseliten. Spitzenkarrieren und soziale Herkunft in Wirtschaft, Politik, Justiz und Wissenschaft. Campus: Frankfurt/New York.

Hattie, J.A.C. (2013): Lernen sichtbar machen. Überarbeitete deutschsprachige Ausgabe von »Visible Learning« besorgt von Wolfgang Beywl und Klaus Zierer. Schneider Hohengehren: Baltmannsweiler (engl. 2009).

Hecker, U. (1998): Anforderungen der Betriebe an die Auszubildenden. Ministerium für Berufsbildung: Berlin.

Hellermann, M., u. a. (2008): Das Kind hinter PISA. Wie die junge Generation fühlt, was sie denkt und wie sie lernt. Frage (Auszüge aus der Wissenschaft), Nr. 4. Presse- und Informationsstelle der Universität: Siegen. Download: www2.agprim.uni-siegen.de/inprint/Frage0408.pdf

Helmke, A. (2008): Unterrichtsqualität. Erfassen, Bewerten, Verbessern. Kallmeyer: Seelze (6. völlig neubearb. Auflage 2008; 1. Aufl. 2003).

Helmke, A. (2012) Leistungsorientierung und Leistungsdifferenzierung im Unterricht: unvereinbare Ziele? Folien zum Vortrag: www.ed-bs.ch/bildung/pzbs/veranstaltungen/paedagogische-tagung-05.09.2012/referat-helmke.pdf

Helmke, A./Hosenfeld, I. (2003a+b): Vergleichsarbeiten (VerA): Eine Standortbestimmung zur Sicherung schulischer Kompetenzen – Teil 1 & 2. In: SchulVerwaltung, Ausgabe Nordrhein-Westfalen (2003), H. 4, 107-110, und H. 5, 143-145.

Helmke, A., u. a. (2014): EMU – Unterrichtsdiagnostik. Studienbrief Version 4.0. Universität: Koblenz-Landau

Henrich, J., et al. (2010): The weirdest people in the world. In: Behavioral and Brain Sciences, Vol. 33, No. 2-3, 61-83.

Hentig, H. v. (2003): Die Schule neu denken. Hanser: München (3., erw. Neuausgabe).

Herzog, W. (2013): Bildungsstandards. Eine kritische Einführung. Kohlhammer: Stuttgart.

Heuser, U. J. (2014): Wenn sich Bildung nicht lohnt. Mehr lernen heißt nicht unbedingt mehr Wohlstand, warnen Ökonomen. In: Die Zeit, Nr. 47 v. 13.11.14, 19.

Heymann, H. W. (1997a): Zur Einführung: Allgemeinbildung als Aufgabe der Schule und als Maßstab für Fachunterricht. In. Heymann (1997b, 7-17).

Heymann, H.W. (Hrsg.) (1997b): Allgemeinbildung und Fachunterricht. Bergmann + Helbig: Hamburg.

Heymann, H. W. (1999): Was können wir aus TIMSS lernen –für Leistungsvergleiche deutscher Schulen? Ein Interview. In: Brügelmann (1999, 63-71).

Holtappels, H. G./Herdeegen, M. (2005): Schülerleistungen in unterschiedlichen Lernumwelten im Vergleich zweier Grundschulmodelle in Bremen. In: Bos u. a.(2005, 361-397).

Hopmann, S. T., u. a. (Hrsg.) (2007): PISA zufolge PISA/PISA According to PISA. Lit-Verlag: Wien/Berlin.

House, E. R. (1975): Accountability in the U.S.A. In: Cambridge Journal of Education, Vol. 5, No. 2, 71 78.

House, E. R. (1980): Evaluating with validity. Sage: Beverly Hills/London.

Huber, S. G. (1999): School Effectiveness: Was macht Schulen wirksam? Internationale Schulentwicklungsforschung (I). In: schul-management, 30. Jg., Nr. 1, 10–17.

Huisken, F. (2005): Der »PISA-Schock« und seine Bewältigung. Wieviel Dummheit braucht / verträgt die Republik? VSA-Verlag: Hamburg.

Hurrelmann, B. (2003): Leseleistung – Lesekompetenz. Folgerungen aus PISA. mit einem Plädoyer für ein didaktisches Konzept des Lesens als kulturelle Praxis. In: Beiträge Jugendliteratur und Medien, 55. Jg., H. 4, 243-255. Nachdruck aus: Praxis Deutsch, 29. Jg., H. 176.

Ingenkamp, K. (Hrsg.) (1971): Die Fragwürdigkeit der Zensurengebung. Beltz: Weinheim/Basel (9. Aufl. 1995).

Ingenkamp, K./Schreiber, H. (Hrsg.) (1989): Was wissen unsere Schüler? Überregionale Lernerfolgsmessung in internationaler Sicht. Deutscher Studien Verlag: Weinheim.

Institut der deutschen Wirtschaft (Hrsg.) (1997): Anforderungsprofile von Betrieben – Leistungsprofile von Schulabgängern. Ergebnisse der Betriebsbefragung (Kurzfassung). Institut der deutschen Wirtschaft: Köln.

Institut für Schulentwicklung PH Schwäbisch Gmünd (Hrsg.) (2004): Standards, Evaluation und neue Methoden. Schneider Verlag Hohengehren: Baltmannsweiler.

Jahn, D. (2005): Fälle, Fallstricke und die komparative Methode in der vergleichenden Politikwissenschaft. In: Kropp/Minkenberg (2005, 55-75).

Jahnke, T./Meyerhöfer, W. (Hrsg.) (2007): Pisa & Co. Kritik eines Programms. Franzbecker: Hildesheim (2. überarb. Aufl.; 1. Aufl. 2006).

Jencks, C., et al. (1972): Inequality: A reassessment of the effects of family and schooling in America. Basic Books: New York (dt. 1972: Chancengleichheit. Rowohlt: Hamburg).

Jensen, R. (1969): How much can we boost IQ and scholastic achievement? In: Harvard Educational Review, Vol. 39, Nr. 1, 1–163. dt. in: Skowronek (1973, 63-155).

Jornitz, S. (2009): Evidenzbasierte Bildungsforschung. In: Pädagogische Korrespondenz H. 40, 68-75.

Jude, N./Klieme, E. (2010): Das Programme for International Student Assessment (PISA). In: Klieme u. a. (2010, 11-21).

Jude, N./Klieme, E. (Hrsg.) (2013): PISA 2009 – Impulse für die Schul- und Unterrichtsforschung. 59. Beiheft der Zeitschrift für Pädagogik.

Karg, I. (2005): Mythos PISA. Vermeintliche Vergleichbarkeit und die Wirklichkeit eines Vergleichs. V&R unipress: Göttingen.

Keiner, E., u. a. (Hrsg.) (2011): Metamorphosen der Bildung. Historie – Empirie – Theorie. Festschrift für Heinz-Elmar Tenorth. Klinkhardt: Bad Heilbrunn.

Kelle, U. (2007): Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. VS Verlag für Sozialwissenschaften: Wiesbaden.

Kemmler, L. (1967): Erfolg und Versagen in der Grundschule. Hogrefe: Göttingen.

Kirkland, M. C. (1971). The effects of tests on students and schools. In: Review of Educational Research, Vol. 41, No. 4, 303-350.

Klawans, H. (2005): Die Höhlenfrau, die Sprache und wir. 13 merkwürdige Geschichten über das menschliche Gehirn. Klett-Cotta: Stuttgart (engl. 2000).

Klemm, K. (1998): Vom Nutzen der Bildung. In: Pädagogik, 50. Jg., H. 6, 9-11.

Klemm, K. (2002): Pisa-E zeigt ein aufregend neues Bild unserer Schulen. Die Gegenüberstellung von anspruchsvoller Unions-Erziehung und SPD-Kuschelpädagogik entbehrt jeder Grundlage. Eine differenzierte Analyse. In: Frankfurter Rundschau v. 26.6.2002.

Klemm, K. (2008): Schulforscher, aufgepasst! Manchmal stiften Untersuchungen wie die PISA-Studie mehr Verwirrung, als dass sie Klarheit schaffen. In: Die Zeit, Nr. 21 v. 15.05.2008. Download: www.zeit.de/2008/21/C-Bildungsforschung?page=all&print=true

Klieme, E. (2009): Leitideen der Bildungsreform und der Bildungsforschung. In: Pädagogik, 61. Jg., H. 5, 44-47.

Klieme, E. (2011): Bildung unter undemokratischem Druck? Anmerkungen zur Kritik der PISA-Studie. In: Ludwig u. a. (2010, 289-302).

Klieme, E. u. a. (2003): Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Deutsches Institut für Internationale Pädagogische Forschung: Frankfurt.

Klieme, E., u. a. (2006): Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie »Deutsch Englisch Schülerleistungen International (DESI)«. Deutsches Institut für Internationale Pädagogische Forschung: Frankfurt. Download: www.dipf.de/desi/DESI_Zentrale_Befunde.pdf

Klieme, E., u. a. (Hrsg.) (2008): Unterricht und Kompetenzerwerb in Deutsch und Englisch Ergebnisse der DESI-Studie. Beltz: Weinheim/Basel.

Klieme, E., u. a. (2010b): PISA 2000–2009: Bilanz der Veränderungen im Schulsystem. In: Klieme u. a. (2010a, 277-300).

Klieme, E., u. a. (Hrsg.) (2010a): PISA 2009. Bilanz nach einem Jahrzehnt. Waxmann: Münster.

Klieme, E./Maichle, U. (1989): Zum Training von Techniken des Textverstehens und des Problemlösens in Naturwissenschaften und Medizin. In. Trost (1989, 188-247).

Klieme, E./Maichle, U. (1990): Ergebnisse eines Trainings zum Textverstehen und zum Problemlösen in Naturwissenschaften und Medizin. In: Trost (1990, 258-307).

KMK (1968): Erläuterung der Notenstufen bei Schulzeugnissen und Einzelergebnissen in staatlichen Prüfungszeugnissen. Vereinbarung vom 3.10.1968. Kultusministerkonferenz: Bonn.

Knighton, T./Bussière, P. (2006): Educational outcomes at age 19 assoiciated with reading ability at age 15. Statistics Canada: Ottawa.

Kochan, B. (1995): Von der Untersuchung des »Lernens durch Instruktion« zur Untersuchung des »Lernens durch Gebrauch«. In: Balhorn/Brügelmann (1995, 26-28). Nachdruck aus: Brügelmann/Balhorn 1990, 231-234).

Köller, O. (2006): Kritik an PISA unberechtigt. Interview mit Bildungsklick. Download: http://bildungsklick.de/a/50155/kritik-an-pisa-unberechtigt/

Köller, O. (2012). What works best in school? Hatties Befunde zu Effekten von Schul- und Unterrichtsvariablen auf Schulleistungen. In: Psychologie in Erziehung und Unterricht, 59. Jg., 72-78 (auch in Terhart 2014).

Konsortium Bildungsberichterstattung (2006): Bildung in Deutschland. Bertelsmann: Bielefeld.

Kraul, M./Schlömerkemper, J. (2007): Bildungsforschung und Bildungsreform. Heinrich Roth revisited. 9. Beiheft zur Zeitschrift »Die Deutsche Schule«. München.

Krautz, J. (2013): Bildungsreform und Propaganda. Download: http://forum-kritische-paedagogik.de/start/wp-content/uploads/downloads/2013/01/krautz.pdf

Kropp, S./Minkenberg, M. (Hrsg.) (2005):Vergleichen in der Politikwissenschaft. VS Verlag für Sozialwissenschaften: Wiesbaden.

Krugman, P. (2011): Degrees and Dollars. The New York Times, March 6, 2011. Auch Download unter www.nytimes.com/2011/03/07/opinion/07krugman.html?_r=2&hp

Küppers, H. (2005): Mathematik. In: Bartnitzky u. a. (2005. Heft 4).

Kuhn, H.-J. (2014): Anspruch, Wirklichkeit und Perspektiven der Gesamtstrategie der KMK zum Bildungsmonitoring. In: Die Deutsche Schule, 106. Jg., H. 4, 414-426.

Ladenthin, V. (2003): PISA – Recht und Grenzen einer globalen empirischen Studie. Eine bildungstheoretische Betrachtung. In: Vierteljahrsschrift für wissenschaftliche Pädagogik, 79. Jg. H. 3, 354–375. Download: www.messen-und-deuten.de/pisa/Ladenthin03.pdf

Landwehr, A. (2010): Shanghais Stärke in Pisa-Studie ist seine Schwäche. dpa-Dossier Bildung Forschung 51 v. 20.12.2010. Download: http://bildungsklick.de/a/76502/shanghais-staerke-in-pisa-studie-ist-seine-schwaeche/

Lange, H. (2000): Qualitätssicherung und Leistungsmessung in den Schulen auf internationaler und nationaler Ebene. Ms. Behörde für Schule, Jugend und Berufsbildung: Hamburg.

Lange, H. (2008): Vom Messen zum Handeln: »empirische Wende« der Bildungspolitik? In: Recht der Jugend und des Bildungswesens, 56. Jg., H. 1, 7-15.

Largo, R. (2009): Schülerjahre. Piper: München.

Largo, R. (2014): Wer bestimmt den Lernerfolg: Kind, Schule, Gesellschaft? Archiv der Zukunft - Flugschriften. Beltz Taschenbuch: Weinheim/Basel.

Lassek, M. (2012): VERgleichsArbeiten – und sie bewegen sich doch? In: Grundschule aktuell, H. 119, 29-31.

Lazar, I./Darlington, R. (1982): Lasting effects of early education: A report from the Consortium for Longitudinal Studies. Monographs of the Society for Research in Child Development, Nr. 147, S. 1–151.

Lehman, R. H./Lenkeit, J. (2008): ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin. Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien. Humboldt-Universität: Berlin.

Lehmann, R. H., u. a. (1995): Leseverständnis und Lesegewohnheiten deutscher Schülerinnen und Schüler. Beltz: Weinheim/Basel.

Lehmann, R. H., u. a. (1997): Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen. Behörde für Schule, Jugend und Berufsbildung: Hamburg.

Lehmann, R. H., u. a. (1999): Aspekte der Lernausgangslage von Schülerinnen und Schülern an Hamburger Schulen: Klassenstufe 7. Behörde für Schule, Jugend und Berufsbildung: Hamburg.

Lehmann, R. H., u. a. (2004): Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 11. Ergebnisse einer Längsschnittsuntersuchung in Hamburg. Behörde für Bildung und Sport: Hamburg. Download: www.hamburger-bildungsserver.de/welcome.phtml?unten=/schulentwicklung/lau/

Lenske, G. (2013): Schülerfeedback zur Unterrichtqualität in der Primarstufe: Studien zur Validität. Unveröff. Dissertation an der Universität Koblenz-Landau: Landau. Materialien dazu unter: http://www.unterrichtsdiagnostik.info/downloads/Grundschulversion/

Lind, G. (2004): Jenseits von PISA: Für eine neue Evaluationskultur. In: Institut für Schulentwicklung PH Schwäbisch Gmünd (2004, 1-7). Download des Ms. von 2003: www.uni-konstanz.de/ag-moral/pdf/Lind-2003_evaluationskultur.pdf

Lind, G. (2009a): PISA im Klassenzimmer Wohin führen sanktions-orientierte Schulleistungsvergleiche? Folgen und Alternativen? Eingeladener Vortrag durch das Zentrum für Kindheits- und Jugendforschung, Universität Bielefeld, und die Sektion Jugendsoziologie der Deutschen Gesellschaft für Soziologie,16. - 18. September 2009 (Folien). Download: www.uni-konstanz.de/ag-moral/pdf/Lind-2009_PISA-Folgen-fuer-Jugend_4fach.pdf

Lind, G. (2009b): Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In: Bohl/Kiper (2009, 61-79). Download: www.uni-konstanz.de/ag-moral/b-liste.htm#Lind_2009_Amerika_Vorbild_lang

Lind, G. (2011): Pädagogik oder Struktur? Auch die Hattie-Studie weiß keinen Rat. In: Hessische Lehrerzeitung, Nr. 11/2011, 26-27 (überarb. Fassung unter dem Titel »Meta-Analysen als Wegweiser?« verfügbar über www.uni-konstanz.de/ag-moral/pdf/Lind-2013_meta-analysen-als-wegweiser.pdf).

Lind, G. (2011):Verbesserung des Unterrichts durch Selbstevaluation. Ein Plädoyer für unverzerrte Evidenz. Ms. für: Bellmann/Müller (2011).

Linn, R. L. (2000): Assessments and accountability. In: Educational Researcher, Vol. 29, No. 2, 4-15.

List, J. (Hrsg.) (1998): TIMSS: Mathematisch-naturwissenschaftliche Kenntnisse deutscher Schüler auf dem Prüfstand. Institut der deutschen Wirtschaft. Deutscher Instituts Verlag: Köln.

Ludwig, J. (Hrsg.) (2012): Lernberatung in der Alphabetisierung. Modelle und Handlungsempfehlungen. wbv: Bielefeld.

Ludwig, L., u. a. (Hrsg.) (2011): Bildung in der Demokratie II. Tendenzen – Diskurse – Praktiken. Barbara Budrich: Opladen & Farmington Hills, MI.

Luhmann, N./Schorr, K. E. (1979): Das Technologiedefizit der Erziehung und die Pädagogik. In: Zeitschrift für Pädagogik, 25. Jg., 345 – 365.

Lurija, A. R. (1991): Der Mann, dessen Welt in Scherben ging. [Und: Kleines Porträt eines großen Gedächtnisses]. Rowohlt: Reinbek (russ. 1968 bzw.1971, engl. 1968 und 1972).

Lurija, A. R. (1993): Romantische Wissenschaft. Forschungen im Grenzbezirk von Seele und Gehirn. Rowohlt: Hamburg (russ. 1982).

Maaz, K., u. a. (2011): Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Vodafone Stiftung Deutschland: Düsseldorf.

Mabry, L. (2010): The responsibility of evaluation. In: Böttcher u. a. (2010, 17-33).

Mabry, L. (2010): The responsibility of evaluation. In: Böttcher u. a. (2010, 17-33).

MacDonald, B. (1971): The evaluation of the Humanities Curriculum Project: A holistic Approach. In: Theory into Practice, Vol. 10, 163-167.

MacDonald, B. (1976): Evaluation and the control of education. In: Tawney (1976, 125-136; Nachdruck aus: MacDonald/Walker 1974).

MacDonald, B. (1978): The experience of innovation. Vol. 2 of the revised edition of the publications of the Humanities Curricululum Project Evaluation Unit. Centre for Applied Research in Education. Occasional Publication No. 6. University of East-Anglia: Norwich.

MacDonald, B./Parlett, M. (1973): Rethinking evaluation: Notes from the Cambridge conference. In: Cambridge Journal of Education, Vol. 2, No. 1, 74–81.

MacDonald, B./Walker, R. (eds.) (1974): Innovation, Evaluation, Research and the Problem of Control. Some interim papers. SAFARI project/Centre for Applied Research in Education/U.E.A.: Norwich.

MacDonald, B./Walker, R. (1976): Changing the curriculum. Open Books: London.

Makel, M. C./Plucker, J. A. (2014): Facts are more important than novelty. Replication in the education sciences. In: Educational Researcher, Vol. 43, No. 6, 304–316.

Maritzen, N. (2014): Glanz und Elend der KMK-Strategie zum Bildungsmonitoring. Versuch einer Bilanz und eines Ausblicks. In: Die Deutsche Schule, 106. Jg., H. 4, 398-413.

Maritzen, N./ Tränkmann, J. (2014). Zwischen Empirie und Normativität: Elemente einer Theorie des Bildungsmonitorings. In: Fickermann/ Maritzen (2014, 27-49).

Max-Planck-Institut für Bildungsforschung & Institut für die Pädagogik der Naturwissenschaften (Hrsg.) (1994). Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU), 1. Bericht für die Schulen. Max-Planck-Institut für Bildungsforschung: Berlin/Institut für die Pädagogik der Naturwissenschaften: Kiel.

Max-Planck-Institut für Bildungsforschung (Hrsg.) (1996). Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU), 2. Bericht für die Schulen. Max-Planck-Institut für Bildungsforschung: Berlin.

May, P. (1994): Jungen und Mädchen schreiben »ihre« Wörter. Zur Rolle der persönlichen Bedeutung beim Lernen. In: Richter/Brügelmann (1994, 110-120).

May, P. (2013): HSP 1 – 10. Hamburger Schreib-Probe. Manual / Handbuch: Diagnose orthografischer Kompetenz zur Erfassung der grundlegenden Rechtschreibkompetenzen. Neustandardisierung 2012. Verlag für Pädagogische Medien/Klett: Stuttgart.

Merkens, H. (2003): PISA – Erfolg als Kombination von Vermarktung und wissenschaftlicher Redlichkeit. In: Soziologische Revue, 26. Jg., H. 2, 183-194.

Merkens, H./Zinnecker, J. (Hrsg.) (2004): Jahrbuch Jugendforschung 5. VS Verlag für Sozialwissenschaften: Wiesbaden.

Messner, R. (2004): PISA und Allgemeinbildung. In: Zeitschrift für Pädagogik 49. Jg., H. 3, 400-412.

Metze, W. (2008): Lernwegsorientierter Schriftspracherwerb im Spiegel der Empirie und des Schulalltags. Ms. eines Vortrags in Zürich. Download: www.wilfriedmetze.de/Vortrag_Zurich_31.5.08.pdf

Metzger, K., u. a. (2004): Sprachbezogene Leistungen würdigen. Die aktuelle Leistungsdiskussion und erfolgreicher Deutschunterricht. In: Bartnitzky/Speck-Hamdan (2004, 250-268).

Meyer, H. (2013): Auf die Lehrkraft kommt es an! Ms. des Eröffnungsvortrages auf der Fachtagung des Bundesarbeitskreises der Seminar- und Fachleiter/innen e.V. Niedersachsen (BAK) am 11.3.13 an der Carl von Ossietzky Universität: Oldenburg (auch in Terhart 2014).

Meyer, H.-D./Benavot, A. (eds.) (2013): PISA, power, and policy - the emergence of global educational governance. Oxford Studies in Comparative Education. Symposium Books: Oxford.

Meyerhöfer, W.. (2005): Tests im Test: Das Beispiel PISA. Barbara Budrich Verlag: Leverkusen.

Meyerhöfer. W. (2007): Testfähigkeit – Was ist das? In: Hopmann u. a. (2007, 57-92).

Mortimore, P. (2009):Alternative Modelle zur Analyse und Darstellung der PISA-Ergebnisse einzelner Länder. In: PISA-Info 19/09. Gewerkschaft Erziehung und Wissenschaft: Frankfurt.

Moser, H. (Hrsg. (2011): Aus der Empirie lernen? Forschung in der Lehrerbildung. Lehrerwissen kompakt, Bd. 10. Schneider Verlag Hohengehren: Baltmannsweiler.

MSWWF (Hrsg.) (1999): Evaluation eine Handreichung. Schriftenreihe Schule in NRW Nr. 9033. Ministerium für Schule, Weiterbildung, Wissenschaft und Forschung: Düsseldorf.

Mühlhausen, U. (1994): Überraschungen im Unterricht. Situative Unterrichtsplanung. Beltz Praxis: Weinheim/Basel.

Müller, W. (1998): Erwartete und unerwartete Folgen der Bildungsexpansion. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 38/1998, 81-112.

Münch, R. (2009): Globale Eliten, lokale Autoritäten: Bildung und Wissenschaft unter dem Regime von PISA, McKinsey & Co. Suhrkamp: Frankfurt.

Münch, R. (2011): Mit PISA-Punkten zu mehr ökonomischem Wachstum? In: Ludwig u. a. (2011, 277-288).

Naumann, J., u. a. (2009): Lesekompetenz von PISA 2000 bis PISA 2009. In: Kieme u. a (2010, 23-71).

NCES (2010): The Nation's Report Card: Reading 2009. National Center for Education Statistics: Washington.

Negt, O. (2014): Philosophie des aufrechten Ganges. Streitschrift für eine neue Schule. Steidl: Göttingen.

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

Nichols, S.L. & Berliner, D.C. (2007). Collateral damage: How-stakes testing corrupts schools. Cambridge, MA: Harvard Education Press.

Nöh, H.-J. (1997): Nicht Pauken allein, aber auch Pauken muß sein. Neue Studie des Max-Planck-Instituts widerlegt Reformpädagogik. In: Welt am Sonntag v. 29.6.97.

Norris, N. (ed.) (1977): Theory in practice. Papers Two. Occassional. Publications No. 3. Centre for Applied Research in Education/University of East-Anglia: Norwich.

Norris, N. (1990): Key problems in educational evaluation. Kogan Page: London (Paperback 1993).

OECD (2004): Equity in education. Students with disabilities, learning difficulties and disadvantages. Organization of Economic Co-operation and Development: Paris.

OECD (2007a): Understanding the social outcomes of learning. Organization of Economic Co-operation and Development: Paris.

OECD (2007b): Bildung auf einen Blick - OECD-Indikatoren 2007. Organisation for Economic Co-Operation and Development: Paris. Download: www.bmbf.de/pub/zusammenfassung_eag.pdf

OECD (2007c): School accountability, autonomy, choice and the level of student achievement: International evidence from PISA 2003. www.oecd.org/dataoecd/53/59/39839361.pdf

OECD (2007d): School accountability, autonomy, choice and the equity of student achievement: International evidence from PISA 2003. www.oecd.org/dataoecd/53/57/39839422.pdf

OECD (2010a): PISA 2009 - Ergebnisse: Was Schülerinnen und Schüler wissen und können. Bd. I. Organization of Economic Co-operation and Development: Paris.

OECD (2010b): The high cost of low educational performance. The long-run economic impact of improving PISA outcomes. Organisation of Economic Co-operation and Development: Paris.

OECD & Statistics Canada (ed.) (1995): Grundqualifikationen, Wirtschaft und Gesellschaft. Ergebnisse der ersten internationalen Untersuchung von Grundqualifikationen Erwachsener. Paris/Ottawa (engl. 1995).

Oelkers, J., u. a. (Hrsg.) (2006): Rationalisierung und Bildung bei Max Weber. Beiträge zur historischen Bildungsforschung. Julius Klinkhardt: Bad Heilbrunn.

Oelkers, J./Reusser, K. (2008): Qualität entwickeln – Standards sichern – mit Differenz umgehen. Eine Expertise. Bundesministerium für Bildung und Forschung: Berlin.

Olson, D. R. (2004):. The triumph of hope over experience in the search for k »what works«: A response to Slavin. In: Educational Researcher, Vol. 33, 24–26.

Otto, H.-U./Rauschenbach, T. (Hrsg.) (2004): Die andere Seite der Bildung. Zum Verhältnis von formellen und informellen Bildungsprozessen. VS Verlag für Sozialwissenschaften: Wiesbaden.

Panagiotopoulou, A./Rohlfs, C. (Hrsg.) (2001): Lernbiografien im sozialen Kontext. Projekt LISA&KO FB 2 der Universität. Universi Verlag: Siegen.

Pant, H. A. (2011): »So war der Test nicht gemeint«. Die Vergleichsarbeiten (Vera) an Grundschulen sind in die Kritik geraten. Ein Interview von M. Spiewak. In: Die Zeit, Nr. 22 v. 26.5.2011.

Pant, H. A. (2014a): Visible Evidence? Eine methodisch orientierte Auseinandersetzung mit John Hatties Meta-Metaanalysen. In: Terhart (2014, 134-146).

Pant, H. A. (2014b): Aufbereitung von Evidenz für bildungspolitische und pädagogische Entscheidungen: Metaanalysen in der Bildungsforschung. In: Bromme/Prenzel (2014, 79-99).

Pedulla, J. J., et al. (2003): Percieved effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. National Board on Educational Testing and Public Policy. Lynch School of Education: Boston College; s. a. www.bc.edu/research/nbetpp/reports.html

Piaget, J. (1980): Das Weltbild des Kindes. Klett-Cotta im Ullstein-Taschenbuch 39001: Frankfurt u. a. (frz. 1926).

Piel, E. (1999): Schlechte Noten für die Schule. In: GEO-Wissen, Nr. 1/1999, 22–24.

Pirsig, R. (1978): Zen und die Kunst, ein Motorrad zu warten. Fischer-Taschenbuch: Frankfurt (engl. 1974).

Plewnia, A./Witt, A. (Hrsg.) (2014): Sprachverfall? Dynamik – Wandel – Variation. IDS Jahrbuch 2013. de Gruyter: Berlin.

Polanyi, M. (1958): Personal knowledge: Towards a post-critical philosophy. Routledge and Kegan Paul: London.

Polanyi, M. (1966): The tacit dimension. Doubleday: New York.

Pollard, A./Triggs, P. (2000): What pupils say: Changing policy and practice in primary education. Findings from the PACE project. Continuum: London.

Popper, K. R. (1982): Logik der Forschung. Mohr Siebeck: Tübingen (7. erw. Aufl.; 1. Aufl. 1935).

Prais, S. J. (2007): England: Poor survey response and no sampling of teaching groups In: Hopmann u. a. (2007, 139-156).

Prenzel, M. (2005): »Viel gerechnet, aber wenig nachgedacht«. Pisa-Forscher Manfred Prenzel wehrt sich gegen neue Vorwürfe, die Pisa-Ergebnisse gäben ein verzerrtes Bild der deutschen Schulwirklichkeit. Interview von U. Schlicht. In: Der Tagesspiegel v. 1.9.2005.

Prenzel, M. (2007): »Wir haben einen Sprung nach vorn gemacht«. Was Experten jetzt von Schule und Politik erwarten. Pisa-Leiter Manfred Prenzel erklärt die Lernerfolge deutscher Schüler – und verteidigt seine Studie. Interview von T. Warnecke in: Tagesspiegel v. 5.12.2007.

Prenzel, M. (o.J.): Wie solide ist PISA? oder: Ist die Kritik von Joachim Wuttke begründet? Institut für Pädagogik der Naturwissenschaften. Download www.ipn.uni-kiel.de/pisa/Wie_solide_ist_PISA.pdf

Prenzel, M., u. a. (Hrsg.) (2004). PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Waxmann: Münster.

Prenzel, M., u. a. (Hrsg.) (2005): PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche? Waxmann: Münster.

Prenzel, M., u. a. (Hrsg.) (2007a): PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie. Waxmann: Münster.

Prenzel, M., u. a. (Hrsg.) (2007b): Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft, Sonderheft, Bd. 8. VS Verlag für Sozialwissenschaften: Wiesbaden.

Prenzel, M., u. a. (Hrsg.) (2013): PISA 2012. Fortschritte und Herausforderungen. Waxmann: Münster.

Preuss-Lausitz, U. (1999): Demokratische Selbstvergewisserung anstelle von Black-Box-Messungen. In: Brügelmann (1999, 77-83).

Puchhammer, M. (2007): Language-Based Item Analysis. In: Hopmann u. a. (2007, 127-137).

Rackwitz, R.-P. (2010): Tests unter der Lupe: Können sie halten, was sie versprechen? In: Die Grundschulzeitschrift, 24. Jg., H. 234, 4 – 9..

Rackwitz, R.-P. (2011): Dialogische Lernbeobachtung statt standardisierter Tests. Wie finde ich heraus, was meinen Schülern fehlt? In: Pädagogik, 63. Jg., H. 5, 34 – 37.

Rackwitz, R.-P., u. a. (2012): Dialogische Förderdiagnostik in der Alphabetisierung am Beispiel des Schriftspracherwerbs. In: Ludwig (2012, 78-102).

Radtke, F.-O. (2005): Die Schwungkraft internationaler Vergleiche. In: Bank (2005, 355-386).

Ramirez, F. O. u. a. (2006): Student achievement and national economic growth. In: American Journal of Education, Vol. 113, 1-29.

Rampey, B. D., et al. (2009):. NAEP 2008 trends in academic progress. Washington D.C.: U. S. Department of Education: Washington, D. C.

Ratzka, N. (2003): Mathematische Fähigkeiten und Fertigkeiten am Ende der Grundschulzeit – Empirische Studien im Anschluss an TIMSS (Phil. Diss. FB 2 der Universität Siegen). Franzbecker: Hildesheim/Berlin.

Reeves, T. C. (2011): Can educational research be both rigorous and relevant? In: Educational Designer, Vol. 1, No. 4, 1-24.

Reichen, J. (1982): Lesen durch Schreiben. Leselehrgang, Schülermaterial und Lehrerkommentar. Sabe: Zürich (Neufassung 2001 »Hannah hat Kino im Kopf«: Die Reichen Methode Lesen durch Schreiben und ihre Hintergründe für Lehrerinnen und Lehrer, Studierende und Eltern. Heinevetter-Verlag: Hamburg).

Retzl, M. (2010): Schulqualität entwickeln durch nationale Standards? Grundlegungen für einen alternativen Ansatz der Qualitätsentwicklung. In: Böttcher u. a. (2010, 355-365).

Richter, D., u. a. (2014): Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten: Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. In: Zeitschrift für Pädagogik, 60. Jg., H. 2, 225-244.

Richter, S. (1998): Interessenbezogenes Rechtschreiblernen. Methodischer Leitfaden für den Rechtschreibunterricht in der Grundschule. Westermann: Braunschweig.

Richter, S./Brügelmann, H. (Hrsg.) (1994): Mädchen lernen ANDERS lernen Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb. DGLS Reihe »Lesen und Schreiben«. Libelle: CH Lengwil (2. Auf. 1996).

Rindermann, H. (2006): Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? In: Psychologische Rundschau, 57. Jg., H. 2, 69-86.

Rohlfs, C. (2006): Freizeitwelten von Grundschulkindern. Eine qualitative Sekundäranalyse von Fallstudien. Dissertation im FB 2 der Universität Siegen. Weinheim/München: Juventa.

Rolff, H.-G. (1967): Sozialisation und Auslese durch die Schule. Quelle und Meyer: Heidelberg.

Rolff, H.-G. (2013): Die Hattie –Studie: Ein Rorschach-Test. In: Pädagogik, 65. Jg. H. 3, 46-49 (auch in Terhart 2014).

Rolff, H.-G. (2015): Sind schulische Strukturfaktoren wirklich unwichtig? In: b&w, Nr. 3/2015, 26-29.

Rost, H. (Hrsg.) (1998): Handwörterbuch Pädagogische Psychologie. Psychologie Verlags-Union: Weinheim.

Roth, H. (1962): Die realistische Wende in der pädagogischen Forschung. In: Neue Sammlung, 2. Jg., 481-490.

Rotte, R. (ed.) (2006): International perspectives on education policy. Nova Science Publ.: New York.

Rudduck, J./Flutter, J. (2004): How to improve your school: Giving pupils a voice. Continuum: London

Rüppell, M. (2010): Warum der VerA-Lesetest für meine Schüler so problematisch ist. In: Grundschule aktuell, H. 111, 25-26.

Rumpf, H. (1971): Scheinklarheiten. Sondierungen von Schule und Unterrichtsforschung. Westermann: Braunschweig.

Sackett, D. L., et al. (2000): Evidence-based medicine: How to practice and teach EBM (2nd ed.). Churchill Livingstone: Edinburgh (2nd ed.1st ed. 1997).

Sacks, O. (1987): Der Mann, der seine Frau mit einem Hut verwechselte. Rowohlt: Reinbek (engl. 1985, 1987).

Sacks, O. (1993): Lurija und die romantische Wissenschaft. In: Lurija (1993, 7-22; engl. 1990).

Sacks, P. (1999): Standardized minds: The high price of America's testing culture and what we can do to change it. Da Capo Press/Perseus Books: New York.

Sanders, J. R. (Hrsg.) (2006): Handbuch der Evaluationsstandards: Die Standards des »Joint Committee on Standards for Educational Evaluation » (German Edition. VS Verlag für Sozialwissenschaften: Wiesbaden (3. erw. und akt. Aufl.).

Schleicher, A. (2006): The economics of knowledge: Why education is key for Europe's success. Policy Brief. Lisbon Council. Download: www.lisbon-council.net/files/download/Policy_Brief_Economics_of_Knowledge_FINAL.pdf

Schlömerkemper, J. (2010): Konzepte pädagogischer Forschung. Eine Einführung in Hermeneutik und Empirie. Klinkhardt-UTB 3273: Bad Heilbrunn.

Schmalohr, E. (1997): Das Erlebnis des Lesens. Grundlagen einer erzählenden Lesepsychologie. Klett-Cotta: Stuttgart.

Schmidt, M. (2014): Ameisen ohne Job. In China steigt die Zahl der Hochschulabsolventen stetig an. Dabei sind ihre Berufsaussichten schlecht. In: Die Zeit, Nr. 44 v. 23.10.2014, 81.

Schmitt, R. (Hrsg.) (1999): An der Schwelle zum dritten Jahrtausend. BundesGrundschulKongress 1999. Grundschulverband – Arbeitskreis Grundschule: Frankfurt.

Schmoll, H. (1997): Auf den Lehrer kommt es an. Neue Untersuchungen streichen die Bedeutung des professionellen Wissens und pädagogischen Könnens heraus. In: Frankfurter Allgemeine vom 9.6.97.

Schneider, E. (1971): Recht und Gesetz – die Welt der Juristen. Eine Darstellung juristischer Grundbegriffe und eine Einführung in das juristische Denken für Laien. Goldmann Taschenbuch 2728: München (3., völlig neu bearb. Auflage 1998: Herne Verlag für die Rechts- und Anwaltspraxis).

Schön, D. (1983): The reflective practitioner. Temple Smith: London.

Schönknecht, G./Klenk, G. (2005): Sachunterricht. Heft 5 in: Bartnitzky u. a. (2005).

Schratz, M., u. a. (2000): Qualitätsentwicklung. Vorschläge, Methoden, Instrumente. Beltz Verlag: Weinheim/Basel.

Schründer-Lenzen, A. (Hrsg.) (2006): Risikofaktoren kindlicher Entwicklung. Migration, Leistungsangst und Schulübergang, VS Verlag für Sozialwissenschaften: Wiesbaden.

Schründer-Lenzen, A./Merkens, H. (2006): Differenzen schriftsprachlicher Kompetenzentwicklung bei Kindern mit und ohne Migrationshintergrund. In: Schründer-Lenzen (2006, 15-44).

Schuler, H. (1998): Noten und Studien- und Berufserfolg. In: Rost (1998, 370-374).

Schulmeister, R./ Loviscach, J. (2014): Kritische Anmerkungen zur Studie “Lernen sichtbar machen” (Visible Learning) von John Hattie. In: SEMINAR, H. 2/2014, 121-130 (auchin: Lehren&Lernen, H. 9/2014, 33-39).

Schulmeister, R./ Loviscach, J. (2014b): Fehler in John Hatties „sichtbarem Lernen“. Download: www.lehrerfreund.de/schule/1s/hattie-studie-kritik/4572

Schulverbund »Blick über den Zaun« (Hrsg.) (2007): Beobachten, bewerten, beraten. Verfahren und Werkzeuge für eine andere Evaluation. Bezug über: www.blickueberdenzaun.de/index.php/textarchiv/168-evaluation-beobachten-bewerten-beraten

Schwartz, E., u. a. (Hrsg.) (1972): Materialien zum Mathematikunterricht in der Grundschule. Beiträge zur Reform der Grundschule Bd. 13. Arbeitskreis Grundschule: Frankfurt.

Schweinhart, L. J., u. a. (1993): Significant benefits: The High/Scope Perry Preschool Study through age 27. Monographs of the High/Scope Educational Research Foundation, Nr. 10. High/Scope Press: Ypsilanti.

Scriven, M. (1976): Goal-Free Evaluation. In: Hamilton (1976, 134-145).

Seeber, S/Maaz, K. (2014): Bildung nützt der Gesellschaft und Individuen. In: DJI-Bulletin Impulse, Nr.107, H. 3/2014, 30-33.

Segerholm, C. (2012): The quality turn. Political and methodological challenges in contemporary educational evaluation and assessment. In: Education Inquiry, Vol. 3, No. 2, June 2012, 15–122.

Seidel, B. (Hrsg.) (2006): Einstein, Luke Skywalker und all' die anderen. Kinder und ihre Lernbiografien – Beiträge aus dem Projekt LISA&KO. Universi Verlag: Siegen.

Seidel, T./Shavelson, R. J. (2007): Teaching effectiveness research in the last decade: Role of theory and research design in disentangling meta-analysis results. In: Review of Educational Research. Vol. 77, No. 4, 454-499.

Selter, C. (2005): VerA Mathematik 2004: VERbesserungsbedürftige Aufgaben! VERkapptes Ausleseinstrument?. In: Grundschule aktuell, H. 89, 17-20.

Selter, C./Spiegel, H. (1997): Wie Kinder rechnen. Klett: Leipzig u. a.

Seydel, O. (2005): »Hilfe! Der Inspektor kommt« Oder: Sind Schulen Kunstwerke? In: Pädagogik, 57. Jg., H. 9, 10-15.

Seydel, O. (2007): »Blick über den Zaun ». Wie Schulen voneinander lernen können. In: Schulverbund »Blick über den Zaun » (2007, 2-7).

Seydel, O. (2009): »Eine Schule ist kein Auto«. In: Grundschulzeitschrift, 23. Jg., H. 223, 4-7.

Shaienks, D./Gluszynski, T. (2007): Participation in postsecondary education: Graduates, continuers and drop-outs. Results from YITS cycle 4. Statistics Canada: Ottawa.

Shavelson, R.J./Towne, L. (eds.) (2002): Scientific research in education. National Academy Press: Washington, D. C.

Siewert, J. (2013): Herkunftsspezifische Unterschiede in der Kompetenzentwicklung: Weil die Schule versagt? Untersuchungen zum Ferieneffekt in Deutschland. Dissertation Universität Siegen. Waxmann: Münster.

Simons, H. (Hrsg.) (1980): Towards a science of the singular. Centre for Applied Research in Education Occasional Publication No. 10. University of East-Anglia: Norwich.

Sjøberg, S. (2004): Internationale Vergleichsstudien — ihre guten und schlechten Seiten. In: Bayrhuber u. a. (2004, 51-61).

Skowronek, H. (Hrsg.) (1973): Umwelt und Begabung. Klett, Stuttgart.

Smith, M. L./Rottenberg, C. (1991): Unintended consequences of external testing in elementary schools. In: Educational Measurement: Issues and Practice, Vol. 10, No.4, 7-11.

Spiewak, M. (2013): Ich bin superwichtig! In: DIE ZEIT Nr. 2 v. 3.1.2013. Download: www.zeit.de/2013/02/Paedagogik-John-Hattie-Visible-Learning .

Spinner, K. (2005): Der standardisierte Schüler. In: Didaktik Deutsch, H. 18, 4-13.

Spitta, G. (Hrsg.) (1998): Freies Schreiben – eigene Wege gehen. Libelle: CH-Lengwil.

Stake, R. E. (1967): The countenance of educational evaluation. In: Teachers College Record, Vol. 68, 523-540 (dt. 1972).

Stake, R. E. (1972): Verschiedene Aspekte pädagogischer Evaluation. In: Wulf (1972, 92-112).

Stake, R. E. (1974): Responsive evaluation (revised). In: New Trends in Evaluation. Institute of Education No. 35. University of Göteborg (Extract in: Hamilton 1976, 163-164).

Stake, R. E. (2004): Standards-based and responsive Evaluation. Sage Publications: Thousand Oaks, CA.

Steinig, W., u. a. (2009): Schreiben von Kindern im diachronen Vergleich. Texte von Viertklässlern aus den Jahren 1972 und 2002. Waxmann: Münster.

Steinig, W./Betzel, D. (2014.): Schreiben Grundschüler heute wirklich schlechter als vor 40 Jahren? Texte von Viertklässlern aus den Jahren 1972, 2002 und 2012. In: Plewnig/Witt (2014, 353-371).

Stenhouse, L. (1975): An introduction to curriculum research and development. Heinemann Educational Books: London et al. (teilweise dt. Zusammenfassung in: Zeitschrift für Pädagogik, 19. Jg., H. 3, S. 447-452).

Stenhouse, L. (1980): The study of samples and the study of cases. In: British Educational Research Journal, Vol. 6, 1-6..

Stenhouse, L. (1982): Pädagogische Fallstudien: Methodische Traditionen und Untersuchungsalltag. In: Fischer (1982, 24-61; engl. 1981).

Strietholt R./Bos, W. (2010): Die Nutzung der Ergebnisse standardisierte Leistungstests und der Zusammenhang zwischen Schülerleistung und Lehrerurteil. In: Böttcher u. a. (2010, 165-177).

Strittmatter, A. (1997): Mythen und Machbares in der Qualitätsevaluation. Erkenntnisse aus Pilotprojekten und Ländern mit längerer Erfahrung. In: Journal für Schulentwicklung, Nr. 3/1997, 22–29.

Stürzer, M. (2003): Unterrichtsformen und die Interaktion der Geschlechter in der Schule. In: Stürzer M. u. a. (Hrsg.) (2003), S. 151-170.

Stürzer, M. u. a. (Hrsg.) (2003): Geschlechterverhältnisse in der Schule. Opladen : Leske + Budrich.

Stufflebeam, D., et al. (1971): Educational evaluation and decision-making. F.E.Peacock: Itasca/Ill.

Sundermann, B./Selter, C. (2005): Mathematikleistungen feststellen, beurteilen und fördern. Beschreibung des Moduls 9 für das Projekt SINUS-Transfer Grundschule. Download: www.sinus-grundschule-de/

Taleb, A. N. (2014): Antifragilität. Anleitung für eine Welt, die wir nicht verstehen. btb Random House: München (1. Aufl. Albrecht Knaus: München; engl. 2012).

Tamke, F., u. a. (2004): Idiografische und personorientierte Analysen zur Validierung bzw. Ergänzung nomothetischer Ergebnisse. In: Merkens/Zinnecker (2004, 179-204).

Tawney, D. (ed.) (1976): Curriculum evaluation today: Trends and implications. Macmillan: London.

Tenorth, H.-E. (2004): Bildungsstandards und Kerncurriculum. Systematischer Kontext, bildungstheoretische Probleme. In: Zeitschrift für Pädagogik, 50. Jg., H. 5, 650-661.

Tenorth, H. E. (2014): „Bildungsforschung“ – Eine Perspektive für das Jahr 2020. In: BMBF (2014, 368-374).

Terhart, E. (1997): Gute Lehrer – schlechte Lehrer. Beltz: Weinheim/Basel.

Terhart, E. (2011): Hat John Hattie tatsächlich den Heiligen Gral der Schul- und Unterrichtsforschung gefunden? Eine Auseinandersetzung mit Visible Learning. In: Keiner (2011, 277-292; auch in Terhart 2014).

Terhart, E. (Hrsg.) (2014): Die Hattie-Studie in der Diskussion. Probleme sichtbar machen. Klett Kallmeyer: Seelze.

Tillmann, K.-J. (1999): »Schulqualität sichern« – erziehungswissenschaftliche Anmerkungen zur aktuellen bildungspolitischen Diskussion. In: Brügelmann (1999, 77-83).

Tillmann, K.-J. (2008): Leistungsvergleichsstudien und Systemsteuerung – oder: Was hat PISA zu bildungspolitischen Entscheidungen beigetragen? Ms. des Vortrags auf dem Symposion zum Gedenken an Hermann Lange am 1.10.2008 in der Universität Hamburg. Download: www.gew.de/Binaries/Binary38502/TillmannVortragHamburg1008.pdf

Tillmann, K.-J./Wischer, B. (Hrsg.) (1998): Schulinterne Evaluation an Reformschulen. Positionen, Konzepte, Praxisbeispiele. Impuls30. Laborschule an der Universität: Bielefeld.

Trost, G. (Hrsg.) (1989): Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 13.Arbeitsbericht. Institut für Test- und Begabungsforschung: Bonn.

Trost, G. (Hrsg.) (1990): Test für medizinische Studiengänge (TMS). 14. Arbeitsbericht. Institut für Test- und Begabungsforschung: Bonn.

VBE (Hrsg.) (2000): Schule und Leistung. Deutscher Lehrertag 1999. Dokumentation einer Veranstaltung am 20.3.1999 in der Universität Potsdam. Verband Bildung und Erziehung: 53175 Bonn (Dreizehnmorgenweg 36).

Venn-Brinkmann, U. (2011): Wörter – Sätze – Texte. Eine mehrdimensionale Untersuchung zur Lesekompetenz Jugendlicher am Ende ihrer Regelschulzeit. Diss. an der Universität: Oldenburg (2012 publ. bei Schneider: Hohengehren).

Vieluf, U., u. a. (2012): Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 12 (KESS 12). Zusammenfassung der zentralen Befunde. Download: http://bildungsverlauf.de/fileadmin/downloads/bsb-kess-12-zusammenfassung.pdf

Vieluf, U., u. a. (2013): Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 13 (KESS 13) – Zusammenfassung der zentralen Befunde. Download: www.hamburg.de/contentblob/4099630/data/studie-kess-13.pdf

Waelbroeck, O. (1993): Geschlechtsunterschiede bei mathematischen Testleistungen: eine meta-analytische Integration von Forschungsarbeiten. Psych. Diplomarbeit. Universität: Bonn

Wagemaker, H., et al. (1993): Gender differences in reading literacy. The International Association for the Evaluation of Educational Achievement: The Hague.

Wagener, A. L. (Hrsg.) (2012): Bill Gates, Kesha und Tom im Siegerland. Kinder und ihre Lernbiografien – Beiträge aus dem Projekt LISA&KO. Arbeitsgruppe Primarstufe in der Fakultät II. Universi Verlag: Siegen.

Weinert, F. E. (Hrsg.) (2001): Leistungsmessungen in Schulen. KMK: Bonn/ Beltz: Weinheim.

Weinert, F.E./Helmke, A. (Hrsg.) (1997): Entwicklung im Grundschulalter. Beltz Psychologie Verlags Union: Weinheim.

Weinert, S. (2007): Kompetenzentwicklung und Kompetenzstruktur im Vorschulalter. In: Prenzel u. a. (2007b, 89-106).

Wells , G. (1986). The meaning makers: Children learning language and using language to learn. Heinemann: Portsmouth, NH., Heinemann.

Wells, G. (1982): Preschool literacy-related activities and success in school. School of Education/University: Bristol.

Wells, G./Raban, E.B. (1978):Children learning to read. Final report to the Social Science Research Council: Bristol.

Weßling, H. (2011): Theorie der klinischen Evidenz. Versuch einer Kritik der Evidenzbasierten Medizin. Lit-Verlag: Münster.

Wieczerkowski, W./Jansen, J. (1990). Mädchen und Mathematik: Geschlechtsunterschiede in Leistung und Wahlverhalten. In: Wieczerkowski/Prado (1990, 134-151).

Wieczerkowski, W./Prado T.M. (Hrsg.) (1990): Hochbegabte Mädchen. Bock: Bad Honnef.

Windelband, W. (1894): Geschichte und Naturwissenschaft. Heitz: Straßburg.

Winter, F. (2004): Leistungsbewertung. Eine neue Lernkultur braucht einen anderen Umgang mit Schülerleistungen. Schneider Hohengehren: Baltmannsweiler (5. überarb. und erw. Neuauflage 2012).

Winter, G. (2002): Make-or-break exams grow, but big study doubts value. In. New York Times v. 28.12.2002. Download: www.nytimes.com/2002/12/28/us/make-or-break-exams-grow-but-big-study-doubts-value.html

Wolf, A. (2002): Does education matter? Myths about education and economic growth. Penguin: London.

Wulf, C. (Hrsg.): (1972). Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München.

Wuttke, J. (2006a): Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung. In: Jahnke/Meyerhöfer (2006, 101-154).

Wuttke, J. (2006b): Antwort auf Pressemitteilungen von Prenzel sowie Prenzel und Walter. Vervielf. Rundbrief v. 17.11.06.

Wuttke, J. (2007a): Uncertainties and bias in PISA In: Hopmann u. a. (2007, 241-264).

Wuttke, J. (2007b): Die Insignifikanz signifikanter Unterschiede: Der Genauigkeitsanspruch von PISA ist illusorisch. In: Jahnke/Meyerhöfer (2007).

Wuttke, J. (2009): PISA: Nachträge zu einer nicht geführten Debatte. In: Mitteilungen der Gesellschaft für Didaktik der Mathematik, 87. Jg., 22-34.

Yeh, S. S. (2005): Limiting the unintended consequences of high-stakes testing. In: Education Policy Analysis Archives, Vol. 13, No. 43, 1-23.

Ziegler, H. (2010): Ist der experimentelle Goldstandard wirklich Gold wert für eine Evidenzbasierung der Praxis Früher Hilfen? In: Bundesgesundheitsblatt, 53. Jg., 1061–1066. Download: www.fruehehilfen.de/fileadmin/user_upload/fruehehilfen.de/pdf/Bundesgesundheitsblatt_Artikel_Ziegler.pdf

Zigler, E./Valentine, J. (eds.) (1979): Project Head Start. A legacy of the War on Poverty. New York.

Zimmer-Mueller, M., u. a. (2008): Vergleichsarbeiten in der Grundschule: Wieso, weshalb, warum? In: Grundschulzeitschrift, 22. Jg., H. 215/216, 12-17.

Zinnecker, J., u. a. (2002): null zoff & voll busy. Die erste Jugendgeneration des neuen Jahrhunderts. Leske+Budrich: Opladen.

Quellennachweise

Für diesen Band wurden Auszüge aus folgenden Texten früherer Veröffentlichungen überarbeitet:

»BABEEK und die Kunst, ein Curriculum zu evaluieren«, Manuskript zu einem Vortrag am 20.2.1978 auf der Tagung »Curriculum-Entwicklung – Erfahrungen und neue Perspektiven« am Institut für die Pädagogik der Naturwissenschaften in Kiel ➔ Kap. 0

»Scharfe Brillen, wache Augen und ein freundlicher Blick. Wie Reformschulen den fremden Blick kritischer Freunde am besten nutzen können: Zur Bedeutung von technischer Präzision und sozialer Kontrolle bei der Evaluation pädagogischer Standards«, in: Schulverbund »Blick über den Zaun« (Hrsg.) (2007): Beobachten, bewerten, beraten. Verfahren und Werkzeuge für eine andere Evaluation ➔ Kap. 0 und 5.2

»Qualität und die Kunst, den Erfolg von Unterricht zu messen – oder: Sieben Mythen der aktuellen Diskussion über Evaluation und Rechenschaft«, in: Burkard, C./Haenisch, H. (Red.) (2000): Was ist guter Fachunterricht? Landesinstitut für Schule und Weiterbildung: Soest/Kettler: Bönen, S. 9-41 ➔ Kap. 0 und 5.1

»Miss Marple neben PISA & Co. – Plädoyer für eine Bildungsforschung, die der Praxis nützt«, in: Moser, H. (Hrsg.) (2011): Aus der Empirie lernen? Forschung in der Lehrerbildung. Professionswissen für Lehrerinnen und Lehrer, Bd. 10. Schneider Verlag Hohengehren: Baltmannsweiler, S. 221-234 ➔ Kap. 1

»Kontroversen um die Schulleistungsmessung in Deutschland. Eine fingierte Diskussion über Positionen und Perspektiven in verteilten Rollen«, in: Weinert, F. E. (Hrsg.) (2001): Leistungsmessungen in Schulen. Beltz: Weinheim, S. 33-44, ➔ Kap. 2.1

»Was leisten unsere Schulen?«, in: Bartnitzky, H., u. a. (1999): Zur Qualität der Leistung – 5 Thesen zu Evaluation und Rechenschaft der Grundschularbeit. Grundschulverband e.V.: Frankfurt a. M. ➔ Kap. 2.7

»Metaanalysen der Unterrichtsforschung. Hat Hatties ›Visible Learning‹ einen Nutzen für die pädagogische Arbeit vor Ort?«, in: Lehren und Lernen, 39. Jg., H. 7, S. 20-22 ➔ Kap. 3.1

»Über die Erkenntnisgrenzen von PISA und notwendige Ergänzungen. Ein Gespräch zwischen Imbke Behnken., Hans Brügelmann und Hans Werner Heymann mit ›extrakte‹«, in: Wagener, A. L. (Hrsg.) (2012): Bill Gates, Kesha und Tom im Siegerland. Kinder und ihre Lernbiografien – Beiträge aus dem Projekt LISA&KO. Universi Verlag: Siegen, S. 241-249 ➔ Kap. 3.2

»PISA & Co: Funktion und Grenzen von Leistungsvergleichen – Wie Tests durch eigene Ansprüche und öffentliche Erwartungen leicht überfordert werden« (2011) http://dl.dropbox.com/u/21116330/Lit-Homepage/brue.pisa.enzyklop%C3%A4die_online.dippelhofer.MS_layout_EEO.110813.pdf ➔ Kap. 4

»Thesen zu den Bildungsstandards«, in: Grundschule aktuell, Nr. 90 (2005), S. 8 ➔ Kap. 5.1

»Wahrheit durch VerA? Anmerkungen zum ersten Durchgang der landesweiten Leistungstests in sieben Bundesländern«, in: Grundschule aktuell, Nr. 89 (2005), S. 7-9 ➔ 5.2

»VerA 2010: Für die Schulen auch dieses Mal wenig Ertrag – bei hohem Aufwand und oft unerfreulichen Nebenwirkungen«, in: Grundschule aktuell, H. 111 (2010), S. 20 ➔ Kap.5.2

»Sind Noten nützlich und nötig? Zifferzensuren und ihre Alternativen im empirischen Vergleich«, Eine wissenschaftliche Expertise des Grundschulverbandes, erstellt von der Arbeitsgruppe Primarstufe an der Universität Siegen. Grundschulverband e.V.: Frankfurt a. M. (2006; 2014) ➔ Kap. 6

Glossar

Die nationale Repräsentativstudie **DESI** konzentrierte sich auf die Schülerleistungen in Deutsch und Englisch zu Beginn und zu Ende des 9. Jahrgangs sowie auf die Leitvorstellungen der Lehrer/innen und ihren Unterricht (Klieme u. a. 2006; 2008).

Längsschnittstudien* sind nur aussagekräftig, wenn über die Zeit hinweg nicht zu viele Teilnehmer/innen verloren gehen. Größere Zahlen an solchen **Dropouts** können die Aussagekraft der Studien inhaltlich beeinträchtigen, weil es oft systematische Gründe sind, die zum Aussteigen von Personen(gruppen) führen.

Die **Effektstärke** ist ein Maß für die Stärke eines Zusammenhangs zwischen zwei Faktoren (z. B. Geschlecht und Leseleistung). Sie ist umso größer, je weiter die Mittelwerte von zwei Gruppen (z. B. Jungen und Mädchen) in einem Test (z. B. zum Lesen) auseinanderliegen und je weniger sich die beiden Gruppen unter diesem Aspekt überlappen. Die Effektstärke wird meist zusätzlich berechnet, da auch praktisch kaum bedeutsame Unterschiede statistisch signifikant* werden, wenn die Stichproben nur groß genug sind.

Die internationale **IEA**-Lesestudie wurde 1991 in dritten und achten Klassen auch in Deutschland durchgeführt – kurz nach der Wende getrennt nach Ost- und Westdeutschland (Lehmann u. a. 1995).

Kalibrierung bezeichnet die Anpassung individueller Maßstäbe an übergreifenden Kriterien. Da die persönlichen Maßstäbe für die Bewertung von Leistungen stark von individuellen Ansprüchen und situativen Erfahrungen beeinflusst werden, ist es wichtig, sie von Zeit zu Zeit durch Vergleiche der eigenen Lerngruppe mit repräsentativen Stichproben zu überprüfen.

Als **Korrelation** bezeichnet man in der Statistik den Zusammenhang zwischen zwei Faktoren, z. B.: Je höher die Intelligenz von Schüler/inne/n ist, desto wahrscheinlicher ist es, dass sie im Schulsystem erfolgreich sind. Diese Parallelität darf aber nicht ohne Weiteres kausal gedeutet werden: Es könnte auch sein, dass erfolgreiches Lernen die Intelligenz steigert oder dass ein dritter Faktor wie das familiäre Milieu auf beide Merkmale wirkt.

In **Längsschnittstudien** nehmen dieselben Personen über einen längeren Zeitraum an Erhebungen zu ihrer Entwicklung teil. Veränderungen ihrer fachlichen Leistungen, z. B. von der ersten bis zur vierten Klasse, im Längsschnitt zu untersuchen, ist überzeugender, als nur zwei Querschnitte* zu vergleichen, stellt aber hohe forschungslogistische Anforderungen, z. B. um den Anteil der Dropouts* gering zu halten.

Die »Lernausgangslagen-Untersuchung« (**LAU**) ist eine Längsschnittstudie als Vollerhebung in Hamburg von der 5. bis zur 11. Klasse, in der die Leistungsentwicklung einzelner Schüler/innen erhoben und (anonymisiert) analysiert wurde (Lehmann u. a. 1997; 1999; 2004). Analog angelegt sind weitere Studien wie KESS in Hamburg (Bos/Pietsch 2007; Vieluf 2012; 2013) und ELEMENT in Berlin (Lehmann/Lenkeit 2008).

Der **Mittelwert** einer Gruppe wird in der Regel als arithmetisches Mittel (= Durchschnitt) berechnet. Bei einer Normalverteilung der Werte (»Glockenkurve«) ist er zugleich der häufigste Wert (z. B. IQ = 100).

PIRLS/IGLU untersuchte 2001, 2006 und 2011 Fachleistungen am Ende der 4. Klassen und ist für den Bereich Lesen eine Nachfolgestudie zur IEA-Lesestudie von 1991 (vgl. Bos u. a. 2003; 2005; 2007; 2012).

In den **PISA**-Studien erhebt die OECD die Leistungen 15-Jähriger (und damit in verschiedenen Teilnehmerländern unterschiedliche

Klassenstufen!) im Lesen, in Mathematik und in den Naturwissenschaften. PISA wird seit dem Jahr 2000 im dreijährigen Turnus durchgeführt, in Deutschland (als PISA-E) ergänzt um größere Stichproben einzelner Bundesländer (Baumert u. a. 2001; 2003; Prenzel u. a. 2004; 2005; 2007a; Klieme u. a. 2010a; Prenzel u. a. 2013).

Population oder **Grundgesamtheit** bezeichnet alle Personen, über die eine Aussage gemacht werden soll (z. B. alle Viertklässler/innen an deutschen Grundschulen). In der Regel werden die Informationen aus ökonomischen Gründen an einer Stichprobe* gewonnen und aus ihr verallgemeinert.

Anders als in Längsschnitten* wird in **Querschnittstudien** aus dem Vergleich von zwei verschiedenen Gruppen zu demselben Zeitpunkt (z. B. Fünft- und Zehntklässlern im Jahre 2015) auf die Entwicklung einer Gruppe geschlossen, z. B. darauf, wie Leseleistungen in einer Gruppe über die Zeit hinweg zunehmen.

Rohwerte sind die erste Stufe der Datenauswertung (z. B. richtige Antworten oder Fehler in einem Test). Um sie über verschiedene Erhebungen hinweg vergleichbar zu machen, werden sie oft in **Standardwerte** umgerechnet, z. B. bei PISA so, dass sie einen Mittelwert* von 500 und eine Standardabweichung* von 100 haben.

Unterschiede zwischen zwei Gruppen können zufällig (z. B. durch eine unbewusst verzerrte Auswahl der Stichprobe) oder durch eine bestimmte Bedingung oder ein bestimmtes Merkmal verursacht sein (z. B. die soziale Herkunft). Mit **Signifikanz**-Tests prüfen Statistiker, wie hoch das Risiko ist, Unterschiede zu Unrecht auf bestimmte Merkmale oder Bedingungen zurückzuführen. Anders formuliert: Wie sicher ist es, dass eine unterstellte Abhängigkeit tatsächlich besteht? Dasselbe gilt für Zusammenhänge zwischen zwei Faktoren (z. B. dem Einsatz einer bestimmten Lehrmethode und den gemessenen Lernerfolgen).

Die **Standardabweichung** ist ein Maß für die Streuung der Einzelwerte um den Mittelwert*. Bei einer Normalverteilung (Glockenkurve) liegen rund zwei Drittel der Werte innerhalb einer Standardabweichung oberhalb und unterhalb des Mittelwerts.

Stichproben sind Teilgruppen aus einer Population* bzw. Grundgesamtheit, die möglichst nach Zufall gezogen werden. Unter dieser Voraussetzung lässt sich statistisch berechnen, mit welcher Wahrscheinlichkeit und mit welchem Fehlerrisiko Befunde aus der Stichprobe auch in der Grundgesamtheit gelten.

Die »Trends in International Mathematics and Science Study« (**TIMSS**) wurde von der *International Association for the Evaluation of Educational Achievement* 1995 am Ende der Grundschulzeit und am Ende der Sekundarstufe I bzw. II durchgeführt, in anderen Ländern inzwischen auch wiederholt (Baumert u. a. 2000a, b). In der Grundschule hat Deutschland erst 2007 im Rahmen von PIRLS/IGLU teilgenommen (Bos u. a. 2008).

Seit 2004 werden in den 3. und 8. Klassen aller Bundesländer jährlich flächendeckende Vergleichsarbeiten (**VerA**) in zentralen Fächern geschrieben, deren Ergebnisse an die beteiligten Schulen – mit Daten zu den einzelnen Kindern – zurückgemeldet werden (Helmke/Hosenfeld 2003a, b).