Hans Brügelmann

PISA & Co: Nutzen und Grenzen von Leistungsvergleichen[1] – auf internationaler, institutioneller und individueller Ebene[2]

Zusammenfassung: Mit den internationalen Leistungsvergleichen wie PISA hat im Bildungswesen die sog. „Output-Orientierung“ an Einfluss gewonnen: Standards für Fachleistungen und Kompetenztests zu ihrer Überprüfungen werden zum zentralen Maßstab für die Qualität von Schule, von Unterricht und individuellem Lernerfolg. Als Ergänzung der in Deutschland traditionell „input“-orientierten Evaluation kann diese Sicht das Repertoire pädagogischer Interventionen bereichern. Gegenwärtig droht allerdings eine Monokultur, in der die Leistungsfähigkeit von Tests erheblich überschätzt wird. Im Kontroll-Kontext von Schulaufsicht und Zeugnissen kann die Output-Orientierung sogar die pädagogische Praxis gefährden, deren Entwicklung sie unterstützen will. Nach einer Diskussion der Potenziale und Risiken von Leistungstests werden deshalb als Alternativen Konzepte der „Peer-Review“ und einer dialogischen Leistungsbeurteilung vorgestellt.

Schlüsselbegriffe: Blick über den Zaun, Leistungsvergleiche, Pädagogische Leistungskultur, Output-Orientierung, Kompetenzmessung, PISA, Tests

Hans Brügelmann, Dr. rer. soc., Jg. 1946, ist Professor für Erziehungswissenschaft mit den Schwerpunkten Grundschulpädagogik und -didaktik an der Universität Siegen
Seine Arbeitsschwerpunkte sind: Evaluation; offener Unterricht; Schriftspracherwerb

PISA und weitere internationale wie nationale Leistungsstudien, z. B. PIRLS/IGLU, TIMSS, LAU, KESS usw., haben unser Repertoire zur Eva-

[1] Vorbereitungstext für einen Beitrag zur Enzyklopädie Erziehungswissenschaft Online. Teile dieses Textes habe ich in überarbeiteter und aktualisierter Form übernommen aus: Fieber genau zu messen ist noch keine Diagnose, Fieber erfolgreich zu senken keine Therapie. Wie Leistungstests in ihren Leistungsmöglichkeiten durch PISA & Co überfordert werden. Beitrag zum Forum „Schule ist mehr als PISA - Zur Bedeutung reformpädagogischer Ansprüche an die schulische Bildung von heute“ der ZEIT-Stiftung in Hamburg am 6./7. März 2008.
Vgl. ausführlicher zu einzelnen Teilen des Beitrags: Brügelmann (1980; 2005a, Kap. 46-50; 2006; 2007).

[2] Weil die internationale Leistungsstudie PISA inzwischen zum Modell für Bildungsforschung allgemein und für die Evaluation nicht nur des Bildungssystems, sondern auch einzelner Schulen, Lehrer/innen und Schüler/innen geworden ist, muss man sich auch mit der paradigmatischen Wirkung dieses Projekts auseinander setzen (vgl. zur Kritik etwa Bracey 2009). Wenn dieser Modell-Effekt gemeint ist, spreche ich im Folgenden von „PISA & Co.“.

luation von Schule und Unterricht forschungsmethodisch bereichert – und ihre bildungspolitische Wirksamkeit nachhaltig gestärkt. Übersehen wird allerdings oft, dass ihre Stärken, die wir in der Tat nutzen sollten, unvermeidlich mit spezifischen Schwächen verbunden sind.

Steckbriefe wichtiger zitierter Studien

Die „Trends in International Mathematics and Science Study" (**TIMSS**[3]) wurde von der International Association for the Evaluation of Educational Achievement (IEA) 1995 am Ende der Grundschulzeit (ohne Deutschland) und am Ende der Sekundarstufe I bzw. II durchgeführt, in anderen Ländern inzwischen auch wiederholt. In der Grundschule hat Deutschland erst 2007 teilgenommen[4].

Mit *PISA* erhebt die OECD mit wechselnden fachlichen Schwerpunkten die Leistungen von 15-Jährigen (und damit in verschiedenen Teilnehmerländern unterschiedliche Klassenstufen!) im Lesen, in Mathematik und in den Naturwissenschaften seit 2000 im dreijährigen Turnus, in Deutschland (als PISA-E) ergänzt um Stichproben einzelner Bundesländer[5].

PIRLS/ IGLU[6] untersuchte 2001 und 2006 und 2011 Fachleistungen am Ende der 4. Klassen und ist für den Bereich Lesen eine Nachfolgestudie zur IEA-Lesestudie von 1991 (damals mit 9- und 14-Jährigen[7]).

Die „Lernausgangsuntersuchung" (*LAU*) in Hamburg ist eine Längsschnittstudie von der 5. bis zur 11. Klasse, in der die Leistungsentwicklung – anders als bei den Gruppenvergleichen in den o. g. Querschnitterhebungen – auf der Ebene einzelner Schüler/innen erhoben und (anonymisiert) analysiert wurde[8]. Analog angelegt weitere Studien wie KESS (Hamburg) und Element (Berlin).

Als *VerA* werden jährlich seit 2004 in 3. und 8. Klassen (inzwischen aller Bundesländer) flächendeckende Leistungsvergleiche in zentralen Fächern

[3] Vgl. zu den Ergebnissen deutscher Schüler/innen am Ende der Pflichtschulzeit Baumert u. a. (2000a) und am Ende der gymnasialen Oberstufe Baumert u. a. (2000b).
[4] Vgl. Bos u. a (2008).
[5] Vgl. Anm. 10 und zum innerdeutschen Vergleich: Baumert u. a. (2003); Prenzel u. a. (2005).
[6] Vgl. zu den Ergebnissen der deutschen Viertklässler/innen im internationalen Vergleich und im Vergleich der Bundesländern untereinander: Bos u. a. (2003a; 2005; 2007a).
[7] Vgl. Lehmann u. a. (1995).
[8] Vgl. Lehmann u. a. (1997; 2004).

durchgeführt, deren Ergebnisse an die beteiligten Schulen - mit Daten zu den einzelnen Kindern - zurückgemeldet werden[9].

1. Warum wir PISA brauchen – aber die Ergebnisse mit Vorsicht nutzen sollten

Die als repräsentativ ausgewählten Stichproben verringern die Gefahr zufälliger Eindrücke. Das ist ein großer Vorzug gegenüber persönlicher Erfahrung und gegenüber Fallstudien. Je mehr Personen einbezogen werden, desto höher werden andererseits die logistische Anfälligkeit für Fehler und die Gefahr von Ausfällen 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.

Eine Formalisierung der Erhebungen und ihrer Auswertung wiederum bedeutet eine stärkere Kontrolle von Subjektivität bei der Wahrnehmung von Sachverhalten und ihrer Deutung. Gleichzeitig verhindert sie ein Aushandeln von Bedeutungen und privilegiert den Zugang der Autor/inn/en sowie ihre Interpretationen der Antworten/ Lösungen (z. B. Wahl bzw. Gestaltung der Testaufgaben, Entscheidung für bestimmte statistische Rechenverfahren). Andererseits bietet Standardisierung – zumindest auf der Oberfläche von Fragebogenantworten und Aufgabenlösungen – einfachere Möglichkeiten für Vergleiche. Bezahlt wird die (scheinbare) Eindeutigkeit jedoch mit einem Verlust an kontextueller Bedeutung („Wie hat der Schüler die Aufgabe verstanden?“) und an inhaltlicher Tiefe („Wie ist er zu seiner Lösung gekommen?“).

Bildungspolitisch bedeutet PISA ebenfalls einen Gewinn. Die Öffentlichkeit ist in eindringlicher Schärfe mit Problemen unseres Schulwesens konfrontiert worden, die lange verdrängt wurden. So gehören die in Deutschland lebenden 15-Jährigen - gemessen an den eingesetzten Tests – nicht zu den besonders erfolgreichen Schüler/inne/n. Das gilt nicht nur insgesamt, sondern auch für die Vorzeigeschule Gymnasium, wenn man diese mit den Spitzengruppen anderer Länder vergleicht[10]. Im internationalen Vergleich lagen die deutschen 15-Jährigen 2000 unter dem Durchschnitt und auch 2009 reicht es trotz relativer Positionsgewinne nur zu einem Platz im – je

[9] Vgl. Helmke/ Hosenfeld (2003); Zimmer-Mueller u. a. (2008).
[10] Vgl. im Einzelnen: Baumert u. a. (2001); Prenzel u. a. (2004; 2007); Klieme u. a. (2010a).

nach Fach – (oberen) Durchschnitt. Besondere Probleme bestehen nach wie vor im unteren Leistungsbereich mit – je nach Schwellenwert – 15-25% Schüler/inne/n, denen durch die Tests elementare Schwierigkeiten mit der Schriftsprache, Mathematik und den Naturwissenschaften aufgezeigt wurden. Die Zugehörigkeit zu dieser sog. „Risikogruppe" steht zudem einen engen Zusammenhang mit der Herkunft aus der sozialen Unterschicht und mit dem Migrationshintergrund. Auch die Streuung zwischen Leistungsspitze und –ende ist, trotz einer vergleichsweise gering besetzten Spitzengruppe im internationalen Vergleich besonders groß.

Viele, die als Lehrer/in, in der Schulaufsicht oder durch Forschung Erfahrungen mit Schule in größerer Breite sammeln konnten, hatten diese Probleme schon vorher wahrgenommen und diskutiert. Aber Expertenwissen allein zählt nicht. Zu Recht, wenn man die Fehleranfälligkeit persönlicher Erfahrung in Rechnung stellt; zu Unrecht, wenn solche Expertise systematisch erhoben wird – und wenn man gleichzeitig die Begrenztheit von PISA & Co. bedenkt. Erst die Aufmachung als Leistungsolympiade mit Ranglisten hat ihren Einschätzungen Durchschlagskraft verliehen. Diese mediale Inszenierung war wirkungsvoll, in ihr stecken aber auch Probleme.

Schon diese einleitenden Hinweise machen deutlich: Evaluation von bildungspolitischen Programmen und unterrichtspraktischen Aktivitäten stellt komplexe Anforderungen. Ihnen kann kein methodischer Ansatz allein gerecht werden. Gewinne in einer Dimension sind nicht ohne Verluste in einer anderen zu haben. Meine Kritik an PISA (& Co.) bedeutet deshalb keine grundsätzliche Absage an diesen Ansatz. Es geht mir vielmehr darum, die Autorität, die diesem Paradigma teils von den Forscher/inne/n 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 von und Unterricht geworden; der sog. „Output"-Steuerung, die testbare Fachleistungen zum Maßstab für Schulerfolg macht[11].

PISA fair zu beurteilen ist also nicht leicht. Das fängt mit der Frage an, wer oder was eigentlich „PISA" ist So äußern sich verschiedene Beteiligte – z. B. aus dem internationalen Konsortium und aus der deutschen Forschergruppe - durchaus unterschiedlich. Zudem muss man unterscheiden zwischen Methodenartikeln der Autor/inn/en und grundsätzlichen Selbstdarstellungen[12] einerseits sowie konkrete Auswertungen andererseits. Letztere wiederum finden sich in Fachbeiträgen oder in Artikeln und Interviews in Alltagsmedien. Schließlich gehört zu PISA auch die Rezeption und Nutzung durch Fachkolleg/inn/en einerseits, durch Journalist/inn/en andererseits

[11] Ausführlicher meine Darstellung in: Brügelmann (2005, 46-50).
[12] Z. B. Klieme u. a. (2010); Klieme (2011).

bis hin zu Bildungspolitik, Schulverwaltung, Pädagog/inn/en in der Praxis und interessierten Laien. Denn Testprogramme können je nach gesellschaftlichem und politischem Kontext – und ihrer Funktion in diesen Kontexten - ganz unterschiedliche (Neben-)Wirkungen entfalten.

Vor allem die Art der Präsentation hat/te Nebenwirkungen, die bedacht werden müssen, wenn man Funktion und Anspruch solche Art von Bildungsforschung diskutiert. Dabei sind verschiedene Ebenen zu unterscheiden[13]. Ein besonderes Problem ergibt sich daraus, dass PISA mal als Forschungsprogramm, mal als Evaluationsstudie präsentiert wird – obwohl jeweils andere Anforderungen gelten. Auch von PISA-Forscher/inne/n selbst sind diese Probleme der Selbstdarstellung bzw. ihrer Rezeption zunehmend wahrgenommen worden, so dass sie sich in neueren Veröffentlichungen um eine klarere Positionierung bemühen[14].

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. 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 Kap. 3). In Form von VERA und anderen flächendeckenden Lernstandserhebungen haben die Bundesländer diesen Ansatz allerdings zu einem *Programm* des System-Monitoring gemacht. Dieses erhebt darüber hinaus den Anspruch, die Arbeit von Schulen und Lehrer/innen, ja sogar den Leistungsstand einzelner Schüler/innen bewerten zu können. Insofern ist diese Testmentalität für viele sogar zum generellen *Paradigma* der Evaluation von Unterricht und von individuellen Lernprozessen geworden. Wenn von „PISA &Co." gesprochen ist diese Modellwirkung gemeint (s. dazu Kap. 4).

Die Reichweite der Methoden und damit der Geltungsbereich der Befunde sind aber begrenzt: Statt der vielfach behaupteten „Evidenz"-Basierung[15]

---

[13] Vgl. zur Kritik aus verschiedenen Perspektiven: Collani (2001), Ladenthin (2003), Fertig (2004), Lind (2004; 2009; 2011), Sjøberg (2004), Gaeth (2005), Huisken (2005), Karg (2005), Meyerhöfer (2005), Radtke (2005), Rindermann (2006), Wuttke (2006; 2007; 2009), Klemm (2008), Mortimore (2009), Münch (2011), die Beiträge in Jahnke/ Meyerhöfer (2007) und in Hopmann u. a. (2007) sowie den zusammenfassenden Überblick bei Bank/ Heidecke (2009); Repliken, die allerdings nur 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, 99-100) und Bank/ Heidecke (2009, 368-369).

[14] Vgl. etwa Klieme u. a. (2010b).

[15] Vgl. zur Diskussion dieses Anspruchs einer empirischen Fundierung von Aussagen u. a. die Beiträge zu Böttcher u. a. (2009) und Bellmann (2011). Zur Kritik einseitiger Einforderung bestimmter Forschungsdesigns hat sich seit 1972 die „Cambridge-Gruppe" mehrfach

handelt es sich um (Ein-)Schätzungen, die in hilfreicher, aber begrenzter Weise empirisch fundiert sind. Insofern plädiere ich am Ende meines Beitrags für eine gleichgewichtige Ergänzung des Repertoires[16] durch komplementäre Formen der Evaluation (s. dazu Kap. 5).

Das Grundproblem kann ein Vergleich mit der Arbeit von Ärzten veranschaulichen:

2. Fieber genau zu messen ist noch keine Diagnose, Fieber erfolgreich zu senken noch keine Therapie

Ärzte und Schulreformer teilen ein Problem: Sie müssen aus beobachtbaren Symptomen erschließen, ob ihr „Patient" wirklich krank ist und welche tiefer liegenden Störungen es sind, die dessen Unwohlsein verursachen. Erst aus einer umfassenden Diagnose können sie wirksame Maßnahmen zur Überwindung der Krankheit entwickeln.

Die meisten Patienten fühlen selbst, ob sie Fieber haben oder nicht. Thermometer sind aber ein wichtiges Hilfsmittel, um präziser festzustellen, ob die Körpertemperatur dem Standard entspricht bzw. wie weit sie tatsächlich erhöht ist. Für Ärzte ist sie ein bedeutsamer Krankheitsindikator – interpretierbar allerdings nur im Kontext weiterer Symptome. Denn die „Grundtemperatur" verschiedener Menschen ist unterschiedlich. Manche bekommen sogar, wenn sie sehr krank sind, kein Fieber. Das heißt, Fieber ist ein *Symptom* und es ist nur *ein* Symptom.

Ärzte messen nicht nur die Temperatur, sie erheben je nach Konstellation eine Fülle weiterer Daten, um auch die Ursache für die erhöhte Temperatur festzustellen. Gute Ärzte verzichten im Regelfall zudem darauf, Fieber senkende Medikamente zu verschreiben, da diese nur die störenden Symptome möglichst rasch verschwinden lassen. Oft könnte dies sogar riskant sein, weil eine nur oberflächliche Anpassung an die Standardtemperatur verhindert, die eigentliche Krankheit wahrzunehmen.

---

geäußert, vgl. zuletzt: Elliott/ Kushner (2007); speziell in den USA gibt es entschiedene Stellungnahmen der American Evaluation Association als zuständiger Fachgesellschaft gegen eine Bevorzugung sog. „quantitativer" Ansätze als „Gold-", Silber-" und „Bronze"-Standard, vgl. Mabry (2010, 23-28).

[16] Ich sehe durchaus, dass manche Kolleg/inn/en umgekehrt - mit derselben Intention - für eine Stärkung des Standardisierungs-Paradigmas argumentieren. Da diese in den letzten Jahren in großer Breite und mit starker Wirkung (s. etwa die Besetzung von Lehrstühlen in der Erziehungswissenschaft) stattgefunden hat, sehe ich aber eher die Notwendigkeit eines Austarierens der Balance in umgekehrter Richtung.

Auch in der Pädagogik nutzen wir Thermometer: Leistungstests. Sie können ein wichtiges Hilfsmittel sein, um Problemen des Bildungssystems auf die Spur zu kommen. Aber nicht jede erhöhte Temperatur ist Anzeichen für ernste Probleme. Für die Frage, ob das Bildungswesen insgesamt oder ob eine Schule „gesund" ist, ist eine Fülle weiterer Indikatoren zu beachten. Umgekehrt muss eine Annäherung der Testergebnisse an vorgegebene Standards nicht bedeuten, dass das untersuchte System wirklich besser geworden ist, gibt es doch auch in der Pädagogik fiebersenkende Medikamente. Das bekannteste wurde dank PISA&Co aus den USA auch nach Deutschland importiert: „teaching to the test"…

Dass man mit verschiedenen Thermometern unterschiedliche Werte bekommt, sollte zwar nicht sein, ist aber im Rahmen enger Bandbreiten akzeptiert, eine Eichung fällt aufgrund der technischen Normierbarkeit vergleichsweise leicht. In der Bildungsforschung gibt es erhebliche Messunterschiede - je nachdem, welchen Test man einsetzt. Bei PISA-2006 hat sich sogar die für Mediziner erstaunliche Situation ergeben, dass die deutschen und die internationalen Autoren an demselben Thermometer unterschiedliche Temperaturen ablasen. Spätestens hier wird deutlich, wie naiv diejenigen sind, die glauben, man könne psychische Merkmale mit vergleichbarer Genauigkeit messen wie physische. PISA ist ein heikles Instrument und deshalb noch vorsichtiger zu handhaben als selbst ein Quecksilber-Thermometer…

3. Stärken, Reichweite und Schwächen von PISA

Fragen an PISA und zur Bedeutung bzw. Aussagekraft seiner Ergebnisse sind auf verschiedenen Ebenen zu stellen:

(1) Wie bedeutsam ist die Qualität von Schule für die *Lebenschancen des einzelnen* und für die wirtschaftliche *Entwicklung einer Gesellschaft*?

Eine wichtige Begründung für den Einsatz von PISA ist die behauptete Bedeutung seiner Ergebnisse, also des untersuchten „Outputs" des Bildungssystems, für die wirtschaftliche Entwicklung des Landes und für die Berufschancen seiner Bürger. Statistiken sowohl zu den Arbeitslosigkeitsquoten als auch zum Einkommensniveau bestätigen, dass Ausbildung und Bildung in der Tat wichtige Ressourcen für den individuellen Lebens- und Berufserfolg sind[17]. 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 findet sich bei PISA die behauptete

[17] Vgl. Klemm (1998) und Geißler (2006, 280ff.).

Abhängigkeit des Wirtschaftswachstums und der Arbeitslosigkeit von Leistungspunkten nur eingeschränkt[18]. Der Einfluss von - mit dem (Aus-)Bildungsniveau eines Landes korrelierenden - Drittfaktoren darf insofern nicht übersehen werden:

Der Zusammenhang zwischen PISA-Ranking und *Wirtschaftskraft eines Landes* sind nicht sehr hoch. Er wird durch weitere Bedingungen stark relativiert bzw. kann seinerseits durch weitere Faktoren bedingt sein. 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 der 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.
Der Einfluss von (Aus-)Bildung wird also in hohem Maße durch andere Bedingungen moderiert. Zu diesen gehören ebenfalls die Konjunktur einer Branche oder demografische Veränderungen der Einstellungsbedingungen. 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. 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.

. Die *ökologische Validität* standardisierter Tests erweist sich generell als problematisch. Schon die Prognosekraft für den weiteren Schulerfolg[19], und

---

[18] Vgl. die Kritik von Münch (2011, 281-287) und seine Hinweise auf Wolf (2002) bzw. Ramirez u. a. (2006) sowie die Relativierung durch die OECD (2010b) selbst.

[19] Vgl. die 20% erfolgreichen Sekundarabschlüsse unter den 17% nach PISA angeblich funktional leseunfähigen „Illiterates“ in Dänemark (Dolin 2007, 114) und die sogar 62% auf Lesestufe 1 bzw. darunter, die in Canada einen High-School-Abschluss erwerben (Knighton/ Bussière 2006, 21); vgl. auch die abnehmenden Unterschiede DDR vs BRD in freien Texten gegenüber Diktaten (Brügelmann u. a. 1994) und die niedrigen Testleistungen vieler berufsfähiger Handwerker und LehrerInnen vs. GrundschülerInnen in der Studie LUST (Brügelmann 2004). 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 den publizierten Studien (Knighton/ Bussière 2006, 21, und Shaienks/ Gluszynski 2007, 33) nicht entneh-

erst recht die Alltagsrelevanz von Testleistungen für den Berufs- und Lebenserfolg außerhalb der Schule sind für PISA nicht überzeugend belegt[20].

➔ PISA-Werte sind als Indikatoren für individuelle Leistungsmöglichkeiten außerhalb der Testsituation und 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 hat einen Eigenwert unabhängig vom Nutzen in Leistungssituationen.

(2) Reichen die *Fachleistungen* am Ende der Pflichtschulzeit – als Indikator für den Output" des Schulwesens – aus, um dessen Qualität zu beurteilen?

Den Blick auf das zu wenden, was Schüler/innen aus der Schule mitnehmen, ist wichtig. Der kontinentaleuropäischen Tradition entsprechend haben sich im deutschen Bildungssystem Steuerung und Kontrolle lange Zeit auf „Input" und Rahmenbedingungen wie Lehrpläne, Stundentafeln, Qualifikation der Lehrerschaft beschränkt. Einleuchtend ist auch, dass sich PISA auf sog. Basiskompetenzen konzentriert, und das in den Schlüsselbereichen Lesen und Mathematik[21].
Gleichzeitig sind aber die Einschränkungen dieses Fokus im Blick zu behalten:
- PISA zeigt 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* Ziele erreicht.
- Die Frage, ob Länder-Curricula die PISA-Literacy zureichend abbilden, ist erst 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 sog. funktionale „Literacy"[22].

- Wesentliche Erfahrungen mit Inhalten des Unterrichts machen junge Menschen außerhalb der Schule, z. B. für die Entwicklung der Lesefähigkeit oder in den Sachfächern, aber auch in Sport und Musik. PISA selbst ver-

---

men, vgl. Arbeitsgruppe Primarstufe (2006, 1.1.3.3) zu ähnlich schwachen Korrelationen in anderen Studien.
[20] Vgl. zu den Schwierigkeiten repräsentativer Vergleiche mit Erwachsenen die PISA-Nachfolgestudie bei Eltern von Schüler/inne/n: Ehmke/ Siegle (2006). Auch die Ergebnisse unsere eigenen Versuche mit einem Lesetest bei Berufsschüler/inne/n und Berufstätigen mahnen zur Vorsicht, vgl. Brügelmann (2004).
[21] Vgl. zur Begründung OECD (2004; 2007a).
[22] Vgl. Artelt u. a. (2001, 78); kritisch zu dieser Einengung u. a. Hurrelmann (2003).

weist auf die hohe Abhängigkeit des Schulerfolgs von der sozialen Herkunft, also auf die Bedeutung von Normen und Lernmöglichkeiten im Alltag. Insofern dürfen Unterschiede 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 zu schulbedingten Unterschieden innerhalb von Ländern veröffentlicht sind, liegen sie bei 5-15%. Dieser niedrige Anteil ist plausibel, wenn man Erkenntnissen zur Bedeutung informellen Lernens folgt.

- Schulintern spielen neben dem Unterricht zudem Systembedingungen wie Stundentafel oder Halbtags- vs. Ganztagsschule eine große Rolle für die Leistungsentwicklung der Schüler/innen. Hinzu kommen außerschulische Faktoren (beispielsweise die Bildungseinstellungen von Eltern und die Arbeitsmarktchancen der Absolvent/inn/en). Unterschiede der öffentlichen Schule zuzurechnen verkennt auch die unterschiedliche Rolle von zusätzlichen, privat bezahlten Fördereinrichtungen wie etwa den Jukus in Japan oder dem Nachhilfeunterricht in Deutschland.

- Qualität von Schule kann nicht allein an den Wirkungen festgemacht werden. *Wie* etwas gelernt wird hat Bedeutung für die Dauerhaftigkeit und Transfermöglichkeit und bestimmt zudem die Erfahrung von Schule. Prozessqualität und Rahmenbedingungen stellen deshalb eigene Merkmale einer guten Schule dar[23]. Aus Problemen in 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 kurzfristig erworbene Kompetenzen. Wie aussagekräftig diese als Indikatoren für längerfristige Leistungsmöglichkeiten sind, muss gesondert nachgewiesen werden (s.o. Anm. 24 und 26).

➔ 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 Teil-Kulturen und der Generationen zusammenwächst (Demokratiefunktion der Schule), in der öffentlichen Diskussion ebenso aus dem Blick wie das Kind „hinter PISA“ mit seinen lebensweltlichen Erfahrungen[24]. 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 Sozialpolitik.

---

[23] Vgl. etwa die Kriterien des Deutschen Schulpreis (Fauser u. a. 2007; STERN 2010).

[24] Vgl. zu einem anderen Blick auf die Bedeutung von Schule für Kinder und Jugendliche die Beiträge in Hellermann u. a. (2008).

(3) Erfassen die *PISA-Tests* die Fachleistungen in angemessener Form?

Studien zur Notengebung[25] zeigen seit vielen Jahren, wie fehleranfällig das Lehrerurteil ist. 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 Zufälligkeiten. Insofern sind standardisierte Tests in heuristischer Funktion (eben als „Thermometer") 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 ihnen und als Folge der Begrenzung durch mögliche Aufgabenformate.

- Die *begrenzte Zeit* führt dazu, dass geringe Prüfungsangst, Vertrautheit mit den Testformaten, test-ökonomisches Arbeiten, Verzicht auf Reflexion „belohnt" werden und damit die fachliche Kompetenz überlagern[26]. Mädchen[27] und Schüler/innen aus schulfernen Milieus[28] schneiden bei solchen Leistungssituationen tendenziell schlechter ab als in Situationen ohne so starken Zeitdruck.

- Die hohen Korrelationen zwischen den Lösungshäufigkeiten in Mathematik und Naturwissenschaften einerseits und im Lesen andererseits verweisen auf eine mangelnde fachliche *Trennschärfe* der Aufgabentypen – entweder wegen des Gewichts der allgemeinen Intelligenz oder wegen der sprachlichen Einkleidung der Aufgaben in allen Bereichen[29].
Schließlich darf nicht übersehen werden: Der forschungsmethodische Zugang bestimmt das inhaltliche Bild des untersuchten Gegenstands. Wählt man beispielsweise in einer Untersuchung Test(typ) A, erhält man ein anderes Ergebnis, als wenn man Test(typ) B einsetzt[30].

Die USA sind bei internationalen Vergleichen der Leseleistung in den letzten 15 Jahren auf vorderen, auf mittleren und auf hinteren Plätzen gelandet (Tab. 1). Diese wurden mit unterschiedlichen Verfahren an jeweils anderen Stichproben durchgeführt. Im US-internen National Assessment of Educati-

[25] Vgl. zusammenfassend: Arbeitsgruppe Primarstufe (2006).
[26] Vgl. Lind (2009) und Wuttke (2009, 26).
[27] Vgl. Waelbroeck (1992) und Ratzka (2003, 148-149), Sacks mit ergänzenden Hinweisen auf Goldstein u. a. (1990) und Wieczerkowski/ Jansen (1990).
[28] Vgl. Sacks (1999); Beilock u. a. (2004).
[29] Vgl. Rindermann (2006, 71), dagegen Baumert u. a. (2007), aber erneut Lind (2009).
[30] S. dazu auch unten Ratzka (2004).

onal Progress dagegen hat sich das Niveau der Leseleistung über 30 Jahre hinweg weder in der Grundschule noch auf der Sekundarstufe bedeutsam verändert:

Kohorte Schulanfang	Rang-platz	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**	**35**	**PIRLS**

Tab. 1: Rangplätze der USA in internationalen Vergleichen der Leseleistung

Diese Differenzen sind (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 von 1991 bis 2001 betrachtet, so werden in den verschiedenen Auswertungen als „Risikogruppe" zwischen 2% und 25% der einbezogenen Jahrgänge benannt (Tab. 2).

Studie	Alter	Schulanfang	Jahr	Anteil „Risiko"
IALS	**16+**[31]	**1934-1982**	**1993**	**14 %**
IEA-Sek	**14**	**1983**	**1991**	**2 %**
PISA-I	**15**	**1991**	**2000**	**25 %**
IGLU/PIRLS	**9**	**1998**	**2001**	**10 %**

Tab. 2: Größenordnung sog. „Risikogruppen" leseschwacher Schüler/innen in Deutschland in verschiedenen Leistungsstudien

Kann sich die Quote innerhalb von 10 Jahren im Verhältnis von 12.5 zu 1 verändert haben – dazu noch mit derart kurzfristig wechselnder Tendenz? Plausibler ist folgende Erklärung: Je nach Aufgabentyp und je nach Defini-

[31] Beim Textlesen kommen die 16- bis 25-Jährigen gemeinsam mit Canada auf Platz 3 von 8 (OECD 1995, 172).

tion der Schwellenwerte ergeben sich ganz unterschiedliche Anteile[32]. Beispielsweise zeigt unsere eigene LUST-Studie in verschiedenen Schulbezirken in Nordrhein-Westfalen, dass am Ende der vierten Klasse selbst in der Gruppe der unteren 5% viele Kinder mehrere Sätze pro Minute lesen können (Brügelmann 2003, 13, 16).

Ratzka (2004) setzte bei einer Replikation der Mathematikuntersuchung TIMSS mit den deutschsprachigen Tests der österreichischen Studie in deutschen Grundschulen drei verschiedene Tests ein. Sie fand, dass selbst bei einer groben Aufteilung nach Leistung nur 41% der Schüler/innen in allen drei Tests in derselben Gruppe landeten und dass sogar beim gleichem Aufgabentyp (Textaufgaben) von vielen Schüler/innen in verschiedenen Tests unterschiedliche Ergebnisse erzielt wurden – ja, sogar in demselben Test (den TIMSS-Aufgaben für die Grundschule), je nachdem, ob die Aufgaben mit oder ohne Zeitdruck zu bearbeiten waren.

➔ Diese aufgaben- und kontextabhängige Ausschnitthaftigkeit von Instrumenten und ihren Ergebnisse muss in der öffentlichen Diskussion bewusst gemacht und gehalten werden. PISA & Co dürfen nicht mehr über „die" Mathematikleistung oder gar „die" Lesekompetenz der deutschen Schüler/innen urteilen[33] – 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.

(4) 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, aber nicht in der Öffentlichkeit bekannt waren und vor allem in der Bildungspolitik nicht ernst genommen wurden: die

[32] Vgl. zu einer auch qualitativen Differenzierung der Schwellenwerte durch eine Auswertung nach Teilkompetenzen *vor* bzw. *neben* dem Textlesen: Venn-Brinkmann (2011).
[33] Vgl. auch die Relativierung der Rangverbesserung deutscher 15-Jähriger von PISA-2003 auf PISA-2006 (so Prenzel 2007): „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 [Abruf: 29.11.2007]). Nun kann man sagen: Die Deutschen haben sich nicht verbessert, der neue Test bevorteile sie (so sagt Schleicher, das spätere Testverfahren habe bestimmte Stärken von deutschen Schülern begünstigt) oder aber: Die Deutschen waren schon 2003 besser (sprich: das damalige Testverfahren habe bestimmte Stärken deutscher SchülerInnen ausgeblendet und sie deshalb benachteiligt)… (s. dazu auch unten Anm. 58).

Schwächen von Jungen[34], von Migrant/inn/en[35] und generell von Kindern und Jugendlichen aus Familien mit niedrigem sozio-ökonomischer Hintergrund[36].
Gleichzeitig sind folgende Einschränkungen zu bedenken, wenn man die Ergebnisse von PISA sinnvoll nutzen will[37]:

- Die Itemanalyse nach Kompetenzstufen mit einem unterstellten *eindimensionalen* Aufbau wird der komplex verschachtelten Entwicklung von Wissen und Können nicht gerecht und schränkt das Spektrum möglicher Aufgaben noch einmal ein[38].

- falsch vs. richtig der Oberfläche ist mehrdeutig im Blick auf Tiefenstruktur. Ohne Kenntnis der *Lösungswege* sind die Aufgabenlösungen mehrdeutig[39].

- Durch die große Zahl der getesteten Schüler/innen werden auch inhaltlich irrelevante Unterschiede statistisch „signifikant", ohne dass sie deshalb auch *inhaltlich bedeutsam* wären (statistische vs. praktische Bedeutung von Unterschieden).

- Durch *Spreizung der Rohwerte* über eine Standardskala mit dem arithmetischen Mittel von 500 und einer Standardabweichung von 100 können Punktdifferenzen bei PISA von Laien leicht überschätzt werden. Trotz anerkennenswerter methodischer Sorgfalt bleiben so viele Fehlerquellen, dass deren Beitrag zu den Länder- und Untergruppen-Ergebnissen die Rangfolgen kräftig durcheinander schütteln kann. Konkret: Wenn 9 Punkte Unterschied im Ländervergleich als statistisch signifikant gelten, dann muss die Differenz der richtigen Antworten 2% betragen. Das bedeutet bei 26 Aufgaben in Mathematik-2003 eine halbe Aufgabe oder auf 100 Aufgaben hochgerechnet einen Unterschied von zwei richtigen Aufgaben[40]. Ist das eine

---

[34] Vgl. schon Kemmler (1967), die Beiträge von Aufenanger und Robine zu Brinkmann u. a. (1990); Wagemaker u. a. (1992) und die Beiträge zu Richter/ Brügelmann (1994).
[35] Vgl. schon Arbeitsgruppe am Max-Planck-Institut für Bildungsforschung (1979); Glumpler (1985); Bommes/ Radtke (1993).
[36] Vgl. schon Rolff (1967); Geißler (1992 ff.); Müller (1998).
[37] Vgl. zu einzelnen Punkten auch eine Reihe von Beiträgen in Jahnke/ Meyerhöfer (2007) und Hopmann u. a. (2007)
[38] Vgl. zusammenfassend Wuttke (2006, 144) und Lind (2009).
[39] Insofern kann eine Lösung auf verfügbares Wissen, auf aktuelle Problemlösung oder auch auf Raten zurückgeführt werden. Deshalb war auch der Grund für die besseren Testleistungen der deutschen SchülerInnen bei PISA-2003 umstritten, den z. B. Klemm (2004) in der größere Vertrautheit der deutschen SchülerInnen mit dem 2000 noch kaum bekannten Aufgabenformat der PISA-Tests vermutet hat. (s. dazu auch unten Anm. 69).
[40] Vgl. Wuttke (2007a, 244-245).

pädagogisch oder politisch relevante Differenz? [41] Die Übersetzung der Punktwerte in Alltagsbegriffe z. B. („45 Punkte entsprechen etwa einem Schuljahr“) macht Kennwerte anschaulicher, kann aber auch in die Irre führen. So bedeutet ein Unterschied von „einem Schuljahr“ für das Lesen am Anfang der Grundschule eine große Leistungsdifferenz, während sie bei 15-Jährigen nur einen geringen Unterschied signalisiert.

- 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[42] oder die Bildung von Kompetenzstufen[43].

➔ Tests sind wie Präzisionsgeräte, die eine Messung von minimalen Differenzen, z. B. Hundertstel Sekunden bei olympischen Spielen, erlauben; sie 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 werdem. Die Stärke und zugleich die Schwäche von PISA &Co liegt in der hohen fachlichen Kompetenz der beteiligten Forscher/innen – und dem Reichtum anderer Studien, die sie zur Interpretation ihrer Ergebnisse heranziehen.

---

[41] Vgl. auch das folgende Beispiel in Gaeth (2009): „Noch ein Argument zur PISA-Skala an sich. Nehmen wir zum einfachen Verständnis ein stark reduziertes Zahlenbeispiel, dann wird deutlich, wie eine Transformation auf eine PISA-Skala aus einer Mücke einen Elefanten machen kann und umgekehrt. Nehmen wir einmal die Messwerte 1 ; 2 ; 3, die bei einem Test gemessen worden sein könnten und transformieren sie auf eine PISA-Skala mit μ=500 und s=100, dann erhalten wir die transformierten Werte 400, 500 und 600. Das klingt doch schon ganz anders. Zwischen den Werten liegt jetzt in den Worten der Autoren „eine ganze Standardabweichung“. Das ist zwar richtig, aber in den absoluten Werten umfasst die Differenz eben nur einen einzigen Punkt. Und was bedeuten 3 Punkte: Handelt es sich um 3 Punkte von 3 insgesamt, von 30 von 300? Ist jemand mit 3 von 300 Punkten ein „Sieger“ nur weil er 600 PISA-Punkte erreicht hat? Erreichen unsere Teilnehmer in einem weiteren fiktiven Zahlenbeispiel von insgesamt 1.000 möglichen absoluten Punkten 997, 998 und 999 Punkte, wären wohl sämtliche Teilnehmer „Sieger“. Nicht jedoch nach PISA-Maßstäben. Nach der Transformation auf die PISA-Skala sieht es nämlich ganz anders aus. Hier erreichen sie ebenfalls 400, 500 und 600 Punkte. Plötzlich haben wir „Sieger“ und „Verlierer“. Nur weshalb? Wäre ein Schüler, der 99,7% aller Aufgaben korrekt gelöst hat, ein „Verlierer“ gar ein Problemfall? Gehört ein Wert von absolut 997 Punkten einer anderen „Kompetenzstufe“ an als ein absoluter Wert von 998 Punkten? Wäre eine Steigerung um 0,1 Prozent-Punkte auf der absoluten Skala, also etwa von 997 auf 998 ein „Kompetenzzuwachs“ von einem Schuljahr?“

[42] Vgl. Wuttke (2009, 24), der bei einer Orientierung am zweiten Kriterium auf einen Durchschnitt von 490 statt 500 kommt, womit die („schwachen“) deutschen Ergebnisse in einem anderen Licht erschienen.

[43] Ebda.

(5) Bilden *Veränderungen* in PISA-Tests tatsächliche Zuwächse/ Verluste verlässlich ab?

Eine Wiederholung von Erhebungen ermöglicht Reihenanalysen, in denen Sonderergebnisse als Ausreißer erkennbar werden. Veränderungen in den Rangplätzen können aber unterschiedliche Gründe haben:

- Da die Rohwerte in den Tests durch Bezug auf die Ergebnisse anderer Länder in PISA-Punkten umgerechnet werden, können positive Entwicklungen trotzdem zu einer (relativen) Verschlechterung führen (und umgekehrt).

- Zunehmende Vertrautheit mit den Testformaten verändert – vor allem unter Zeitdruck – die Erfolgschancen, vor allem in Ländern wie Deutschland, in denen standardisierte Tests traditionell eine geringe Rolle im Schulalltag gespielt hatten.

- Werden die Ergebnisse in den Tests von Beteiligten als bedeutsam für ihre eigene 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 Formate vorhandene Kompetenzen stärker zur Geltung gebracht werden, ohne dass sie sich tatsächlich verbessern.

➔ Eine Wiederholung der Erhebungen alle drei Jahre ist zu kurzfristig angelegt, jedenfalls bei der Schwerfälligkeit von Unterrichtstraditionen nicht nötig[44]. Sie bindet zudem Aufmerksamkeit - und an anderer Stelle für die Entwicklung von Schule und Unterricht bitter vermisste Ressourcen.

(6) Erlauben Unterschiede zwischen PISA-Kennwerten Vergleiche zwischen verschiedenen *Bildungssystemen* oder auch nur zwischen *Teilgruppen innerhalb eines Systems*?

- *Interkulturelle* Vergleiche sind auch inhaltlich schwierig und 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; unter-

[44] Vgl. z. B. die Konstanz der Testleistungen im National Assessment of Educational Progress in den USA über 30 Jahre hinweg (s. dazu auch unten 2.1).

schiedliche Reaktionen auf Zeitdruck und den Eindeutigkeitszwang bei Multiple-Choice-Aufgaben)[45].
- Und wie schnell können selbst statistisch signifikante Unterschiede verschwinden, wenn sich allein durch die Übersetzung, etwa vom Englischen ins Französische, die *Textmenge* der Aufgabenstellung um 10-20% erhöht?[46]
- Einen Einfluss auf die Ergebnisse kann auch haben, wie die Testsituation und ihre Bedeutung von den Teilnehmer/innen in den einzelnen Ländern *wahrgenommen* und bewertet werden [47].
- Nimmt man hinzu, 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, schwindet die Bedeutung der Punktdifferenzen weiter[48], will man an ihnen die Leistungsfähigkeit unterschiedlicher Bildungssysteme festmachen.
- Die *Stichproben* in den verglichenen Ländern sind nicht immer vergleichbar, vor allem was die Einbeziehung schwächerer Schüler/innen betrifft[49].

- Auch innerhalb eines Bildungssystems sind *Gruppenvergleiche* nicht einfach. *Kategorien* wie „Migrationshintergrund“ oder „Gesamtschule“ lassen sich zwar untersuchungstechnisch als Variablen operationalisieren, haben aber in verschiedenen Kontexten unterschiedliche Bedeutung[50].

---

[45] Vgl. Artelt/ Baumert (2003); Puchhammer (2007); Wuttke (2007a, 254-257).
[46] Vgl. Wuttke (2007a, 257).
[47] Zum Einfluss der Motivation auf die Testleistungen stellen allerdings Jude/ Klieme (2010, 18) fest: „Es gibt keine Hinweise darauf, dass von motivationsbedingten Verzerrungen der Ergebnisse auszugehen wäre.“
[48] Damit kann auch der *System*-Wechsel von G9 auf G8 Einfluss auf die deutschen Entwicklung der deutschen Ergebnisse seit 2000 gehabt haben, ohne dass der *Unterricht* besser geworden sein muss.
[49] Vgl. die Hinweise von Wuttke (2006, 143) auf die unterschiedlichen Schulbesuchsquoten mit 15; auf die unterschiedliche Einbeziehung von Schüler/inne/n der untersten Leistungsgruppen; auf die unterschiedliche Teilnahmequoten; auf nicht plausible Verteilungen in einzelnen Ländern. Vgl. konkret zur deutschen Stichprobe Jude/ Klieme (2010, 17): „Die Daten für Schülerinnen und Schüler aus beruflichen Schulen sowie Sonder- und Förderschulen werden nicht einzeln berichtet, gehen aber jeweils in den dargestellten Gesamtwert ein.“
Naumann u. a. (2010, 50): „Von den in der PISA-Erhebung 2009 getesteten 179 Schülerinnen und Schülern an Sonder- und Förderschulen müssen etwa drei Viertel (74%) als schwache Leserinnen und Leser klassifiziert werden, die die Kompetenzstufe II nicht erreichen. Diese kleine Stichprobe lässt nur recht ungenaue
Verallgemeinerungen zu. Immerhin kann man sagen: Von den fünfzehnjährigen Schülerinnen und Schülern an Sonder- und Förderschulen in Deutschland erreicht eine Minderheit – zwischen einem Sechstel und einem Drittel – Kompetenzniveau II, während eine deutliche Mehrheit zu den schwachen Leserinnen und Lesern gerechnet werden muss. Die Zusammensetzung der Sonder- und Förderschülerinnen und -schüler nach Migrationsstatus unterscheidet sich dabei nicht von der entsprechenden Zusammensetzung der Gesamtstichprobe. In die Stichprobenziehung gingen Schulen mit den Förderschwerpunkten Lernen, Sprache sowie soziale und emotionale Entwicklung ein.“
[50] Vgl. etwa den Vergleich der Migrantenanteile und ihrer Zusammensetzung in verschiedenen Bundesländern bei Klemm (2002) und die unterschiedlichen Erfolge von Mig-

(7) Sind Erklärungen möglich – und Beschreibungen *ohne Erklärungen* hilfreich?

PISA hat lange tabuisierte Gründe für die Benachteiligung verschiedener Gruppen wieder in die Diskussion gebracht wie die Selektionsmechanismen des zwei- bis fünfgliedrigen Schulsystems[51].

- Die Korrelationen aus *Querschnitt*-Untersuchungen erlauben keine Erklärungen, suggerieren aber leicht Kausalitäten (Lesefreude ➔ Leseleistung) oder werden zumindest so gelesen[52].

- Zudem kann nicht geklärt werden, ob erhobene Variablen *Ursache* oder nur *Indikator* für k sind.

➔ Für die internationalen Vergleiche werden viele Ressourcen absorbiert. Oft wären Sekundär- und Metaanalysen vorhandener Studien zureichend – und sogar aussagekräftiger[53]. 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. 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, so dass dadurch dessen Ausweitung auf andere Ebenen der Evaluation geboten und legitimiert erscheint.

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

---

rant/inn/en verschiedener Nationalität im deutschen Schulsystem (z. B. bei Holtappels/ Herdeegen 2005).

[51] S. aber auch dazu schon frühere Befunde, etwa aus LAU (vgl. Lehmann u. a 1997, 98-102)

[52] Vgl. die Beispiele, z. B. zu Klassengröße und Schülerleistung, bei Prais (2007, 147-148). Kritisch zu diesen Folgerungen aus Querschnittstudien: Hagemeister (2007) mit Verweis auf den US-amerikanische STAR-Längsschnitt. Erstaunlicherweise 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.

[53] Vgl. etwa die Längsschnittstudien zum Sitzenbleiben, zur Bedeutung kleiner Klassen oder zum Einfluss der Vorschulerziehung auf den späteren Schulerfolg

Die Bedeutung standardisierter Verfahren in der Schullandschaft nimmt zu: verpflichtende Sprachstandserhebungen bei Vier- oder Fünfjährigen vor Schulbeginn, die landesweiten Vergleichsarbeiten „VerA“ in Klasse 3 und 8, zentrale Prüfungen am Ende der verschiedenen Schulzweige, mit Computerprogramm en auswertbare Beobachtungsraster der „Qualitätsanalyse“ bei Inspektionen. Zu beobachten ist ein Multiplikationseffekt von Projekten wie PISA unter dem Etikett „ evidence-based policy“[54]. Deshalb lohnt ein zweiter Blick und eine grundsätzlichere Diskussion dieses Ansatzes.

Mein Fazit aus dem zweiten Kapitel: Die angeblich „objektive“ Bildungsforschung ist nicht deshalb wertvoll, weil sie mit standardisierten Instrumenten eine große Zahl von Fällen erfasst, sondern nur, sofern sich in den Forschungsteams kluge Bildungswissenschaftler mit einem fundierten Hintergrundwissen zusammengefunden haben. Dank ihrer Erfahrung wissen sie die Zahlen sinnvoll zu deuten[55].

Umso dringlicher ist zu fordern, dass schon in den Forschungsberichten selbst die Mehrdeutigkeit von Daten sichtbar gemacht wird. Wichtiger als eine weitere technische Perfektionierung der Erhebungen wird damit die soziale Kontrolle ihrer Interpretation. In dieser Hinsicht könnte die empirische Bildungsforschung viel aus Erfahrungen im Rechtwesen 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 Rechtswissenschaft: 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 *checks and balances.* In der Evaluation des Schulwesens und einzelnen Bildungseinrichtungen könnten alternative Deutungen und Folgerungen explizit gegeneinander gestellt werden, um den Anschein von Expertenautorität zu relati-

[54] Vgl. zu diesem Ansatz und seiner Kritik die Beiträge zu: Böttcher u. a. (2009; 2010); Bellmann (2011).

[55] … kommen dabei aber auch zu unterschiedlichen Einschätzungen, 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 (vgl. oben Anm. 38 und Füller 2007a+b).

vieren und damit auch der Öffentlichkeit die Interpretationsbedürftigkeit der nur scheinbar eindeutigen Daten bewusst zu machen[56].

4.1 Standardisierte Tests als Kontrollelement im Programm der „Output"-Steuerung

PISA hat sich inzwischen von einem spezifischen Projekt zu einem Modell des System-Monitoring gemausert. Alle Bundesländer führen Lernstandserhebungen zum Ende der Grund- und der Pflichtschulzeit durch. Begründet wird dieser Wechsel zur sog. „Output"-Steuerung mit einem Versagen der in Kontinentaleuropa traditionell input-orientierten Maßnahmen.

Die bisherige „Input"-Steuerung erfolgte über Lehrpläne, über die Qualifikationsanforderungen an Lehrer/innen und über die Prozesskontrolle durch die Schulaufsicht. Demgegenüber definiert die sog. „Output-Steuerung" erwartete Leistungen im Voraus und überprüft deren Erreichen mit standardisierten Kompetenztests. Die Forderung nach einem Wechsel von der Input- zur Output-Steuerung wird wie folgt begründet:

- Deutschlands 15-Jährige haben bei PISA schlecht abgeschnitten.
- Deutschland hat ein Input-System.
- Viele der bei PISA 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 erweist sich darin, dass im Vergleich zu 2000 die Leistungen der deutschen Schüler/innen besser geworden sind.

Von manchen[57] wird nämlich 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 Reformen „von unten" drohen nun durch PISA & Co. verloren zu gehen. Dennoch behaupten auch Bos u. a. (2007b, 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 ausge-

[56] Vgl. zur vertiefenden Diskussion dieses Problems und zu konkreten Vorschlägen Brügelmann (2007, 23 ff., 35 ff.)

[57] Z. B. von Andreas Schleicher auf dem GEW-PISA-Kolloquium am 21.11.2007 in Berlin.

sprochen förderlich auf die Kompetenzentwicklung der Schülerschaft aus."

Aber wie kommt es dann in den verschiedenen internationalen Leseuntersuchungen dazu[58],…

- dass Deutschland „mit Input-System" 2001 und 2006 in der Grundschulstudie IGLU deutlich besser abschneidet 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?

- dass die USA mit „Output-System" zwar bei PISA 2000 besser abschneiden als Deutschland „mit Input-System", aber 2001 bei IGLU nur gleichauf liegen – und dass sich die Schülerleistungen in den USA trotz jährlicher Schulleistungstests in den vergangenen 20 Jahren kaum verbessert haben[59]?

- dass in Kanada „mit Output-System" verschiedene Provinzen bei IGLU-2006 von 533 bis 560 Punkten streuen und damit teilweise vor und teilweise hinter Deutschland „mit Input-System" mit 548 Punkten liegen?[60]

- dass Schweden, England und die Niederlande „mit Output-System" bei IGLU 2001 zur Spitzengruppe zählen, 2006 dagegen mit der nun schon länger etablierten Output-Steuerung zurückfallen?

- dass sich dagegen die deutsche Grundschule „mit Input-System" von der IEA-Studie 1991 bis zu IGLU 2001 rechnerisch pro Jahr doppelt so stark verbessert wie von 2001 bis 2006 mit allmählicher Umstellung „auf Output-System" (Bos u. a. 2007a, 152) – und dass sich der Punkte-Durchschnitt der Sekundarstufe von PISA 2000 bis 2006 im gleichen Umstellungsprozess sogar 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[61].

---

[58] Vgl. die Ländervergleiche in Baumert u. a. (2001); Bos u. a. (2003; 2007a); Prenzel u. a. (2004; 2007).
[59] Vgl. NCES (2010).
[60] … und auch innerhalb Deutschlands gibt es größere Unterschiede zwischen den Bundesländern, die alle input-orientiert gesteuert werden, als zwischen Deutschland und Vergleichsländern mit Output-Steuerung.
[61] Am Rande sei angemerkt, dass der Begriff der „Steuerung" ein technisches Verständnis gesellschaftlicher Interventionen unterstellt, dass der nur sehr begrenzten Planbarkeit und Kontrollierbarkeit so komplexer Systeme wie des Bildungswesens nicht gerecht wird.

- Länder, die besser abgeschnitten haben, unterscheiden sich auch in weiteren Merkmalen vom deutschen System (Dauer der gemeinsamen Schulzeit; Zeitpunkt, ab dem benotet wird, usw.)

- In Ländern, die bei PISA und/ oder IGLU besser als Deutschland abgeschnitten haben, zeigen sich in anderen Bereichen Probleme, die mit der Output-Orientierung verbunden sind, so dass dem behaupteten Nutzen auf der einen Seite Kosten auf der anderen gegenüber stehen, die man sich beim Transfer des Systems mit „einkauft".

Insofern scheint die Leistungsfähigkeit eines Systems nur sehr begrenzt mit dem „Steuerungs".-Modell zusammenzuhängen bzw. in starkem Maße durch Drittfaktoren beeinflusst.
Denn das grundsätzliche Problem teilen die Lernstandserhebungen mit PISA: Die Fokussierung auf den Output lässt andere wichtige Merkmale der Qualität von Schule und Unterricht außer Acht[62]. Evaluation wird auf „measurement" verkürzt[63] (dazu noch auf *ein* Messmodell). Damit gewinnen sehr spezifische teststatistische Anforderungen Vorrang vor inhaltlicher Relevanz: nur wenige Fächer, passende Inhalte und Aufgabenformate in diesen Fächern, Auswertung nach einseitig bestimmter „Richtigkeit" der Lösung statt nach Produktivität des Lösungswegs (s. 1.1)[64].

Es gibt aber auch wesentliche Unterschiede zu PISA: flächendeckende Durchführung statt bloß Stichproben, jährliche Wiederholung und vor allem Identifizierbarkeit der beteiligten Schulen, Lehrer/innen und Schüler/innen. Diese Merkmale erzeugen zusätzliche Probleme:

- die flächendeckende Durchführung ist sehr viel aufwändiger und zugleich fehleranfälliger;

- die jährliche Durchführung ist angesichts der Trägheit von Schulsystemen nicht nötig, zieht finanzielle und personelle Ressourcen von anderen Aufgaben ab und hat sich für alle Beteiligten als sehr belastend erwiesen, zumal die Ergebnisse im Schulalltag nicht als Hilfe erlebt werden;

- die fehlende Anonymisierung führt zu all' den unerwünschten Nebenwirkungen, die aus angelsächsischen Studien zum High-Stakes-Testing bekannt

---

[62] Umfassender z. B. „Bildung auf einen Blick", vgl. OECD (2007b und Folgejahre).
[63] Vgl., dazu schon vor mehr als 30 Jahren die Kritik und das Angebot von Alternativen in dem Sammelband „Beyond the numbers game" (Hamilton u. a. 1977).
[64] Belegt werden diese Probleme durch detaillierte Analysen einzelner Testaufgaben, vgl. z. B. die Diskussion zwischen Bartnitzky (2005; 2007), Selter (2005) und Bremerich-Vos u. a. (2005) sowie aktuell zwischen Wittmann und Vertreter/inne/n von VerA. ➔ http://www.mathematik.uni-dortmund.de/didaktik/mathe2000/vera3.html [Abruf: 22.1.22]

sind[65]: fachliche Verengung des Curriculum[66]; Ausrichtung des Unterrichts an Testformaten[67]; unterschiedliche Auslegung und Durchführung der Aufgaben bis hin zu bewussten Täuschungsversuchen[68].

Insbesondere Studien in den USA belegen, dass die Schulleistungen in Bundesstaaten mit hohen Sanktionen für Institutionen und/ oder Personengruppen in der Regel schlechter sind als in Bundesstaaten mit niedrigen Sanktionen: Über dem nationalen Durchschnitt liegen 60-70% der Bundesstaaten mit niedrigen Sanktionen, aber nur 10-20% der Staaten mit hohen Sanktionen[69]. Und wo die Leistungen in standard-bezogenen Tests zunehmen, fallen sie mehrheitlich in unabhängigen Tests ab. So zeigt eine Studie von Amrein/ Berliner (2002), dass der berichtete Leistungsanstieg in der Regel nur für den engen Bereich der im jeweiligen Bundesstaat etablierten Tests galt. In unabhängigen Tests, z. B. für die Zulassung zu den Hochschulen, wurde für 2/3 der 28 Staaten eine *Abnahme* der Testleistungen festgestellt. Zusätzlich wurden wachsende Dropout-Raten berichtet, d. h. dass leistungsschwächere Schüler/innen aus dem System ganz herausfielen, was die gemessenen Testleistungen zusätzlich in die Höhe trieb, ohne dass sich der „Output" tatsächlich verbesserte[70]. Kirkland (1971) und Linn (2000) konnten sogar zeigen, dass die Leistungen nur so lange „stiegen", wie sich der Tests nicht änderte: Bei Einführung einer neuen Version desselben Tests sanken die Leistungen wieder auf das Ursprungsniveau. Danach „stiegen" sie erneut – bis die erste Testform wieder eingesetzt wurde und die Leistungen ein zwei-

[65] Generell kritisch zu den (Neben-)Wirkungen des High-Stakes Testing: House (1975); Brügelmann (1980); Haladyna u. a. (1991); Smith/ Rottenburg (1991); Sacks (1999); Bauer (2000); Amrein/ Berliner (2002); Yeh (2005); Au (2007); Nichols/ Berliner (2007); Amrein-Beardsley (2009) – und wissenschaftspolitisch übergreifend die American Educational Research Association (s. AERA 2003).

[66] Vgl. die Lehrerbefragung von Pedulla u. a. (2003, 5-9). Eine grundsätzlichere Kritik am Fokus auf „measurement" hat die international besetzte „Cambridge group" schon nach der ersten Evaluations-Welle in den 1970er Jahren geübt, s. u. a. MacDonald/ Parlett (1973), die Beiträge zu Hamilton u. a. (1977) und aktuell Elliott/ Kushner (2007).

[67] Die zur Widerlegung einer Steigerung von Leistungen durch Testgewöhnung zitierte Studie von Klieme/ Maichle (1989; 1990) beschränkte sich auf ein sechsstündiges Testtraining – kein Vergleich mit den Wirkungen eines insgesamt auf „high stakes testing" abgestimmten Schulsystems (vgl. Meyerhöfer 2007, 66-67)

[68] Vgl. die Zusammenstellung bei Brügelmann (2005b, 8) und ergänzend zu VERA in der Grundschule Helmke/ Hosenfeld (2003) und Pant (2011) sowie die kontroversen Beiträge in „Grundschule aktuell" Nrn. 89, 90, 99, 103 und Grundschulzeitschrift, 22. Jg., H. 217. Bereits anlässlich der ersten Welle test-basierter Evaluation Anfang der 1970er Jahre hat Campbell (1976) als „Gesetz" formuliert: „Je häufiger ein quantitativer sozialer Indikator für gesellschaftliche Entscheidungen genutzt wird, desto stärker verführt er zum Täuschen und desto nachhaltiger wird er die sozialen Prozesse stören, die er messen soll." (eig. Übers.)

[69] Vgl. Sacks (1999, 89-90).

[70] Vgl. dazu die Zusammenfassung und den Kommentar von Winter in der New York Times v. 28.12.2002 .

tes Mal auf ihr Ursprungsniveau „sanken“[71]. Diese Befunde machen deutlich, wie sehr Sanktionen zu einem „teaching to the test“ verführen, das die angestrebte inhaltliche Verbesserung des Unterrichts überlagern oder sogar gefährden kann.

Besondere Probleme ergeben sich aber aus dem Anspruch, Aussagen über die Qualität des Unterrichts einzelner Lehrer/innen oder gar über den Lernstand von Schüler/innen machen zu können. Damit wird der Kredit von Tests endgültig überzogen, wie die folgenden Hinweise zeigen.

4.2 Standardisierte Tests als Instrument für die Evaluation von Schule und Unterricht

Dass punktuelle Erhebungen mit standardisierten Instrumenten wie PISA & Co. inzwischen für viele im Bildungswesen[72] zum generellen *Paradigma* der Evaluation von Leistungen geworden ist, stellt aus meiner Sicht das eigentliche Problem dar. Der grundsätzliche forschungsmethodische Fehler: Für die Untersuchung von Populationen angemessene Methoden eignen sich nur sehr eingeschränkt für die Evaluation von Einzelfällen wie Schulen, Lehrer/innen oder Schüler[73]. Sie können als „Warnlampe“ *heuristisch* hilfreich sein, haben aber immer den Status von Hypothesen, die mit Hilfe weiterer Daten überprüft werden müssen.

Wer Schule und Unterricht verbessern will, hat zwei Optionen: Er kann auf der zentralen Ebene ansetzen und versuchen, die Rahmenbedingungen zu verändern: andere Schulstruktur, neue Lehrpläne, mehr Ressourcen. Wenn man die Klagen von Lehrer/innen hört „Wir können nicht, weil…“, kann man diesem Weg viel abgewinnen. Andererseits fällt auf, wie unterschiedlich der Alltag vor Ort innerhalb derselben Strukturen aussieht. Entsprechend lauten die Klagen von Politik und Verwaltung: „Aber die Lehrer/innen…“.

Interventionen scheinen also auf beiden Ebenen nötig. Erfolg versprechen sie aber nur, wenn sie gut informiert sind. Hier setzt die Aufgabe von Evaluation an, und entsprechend umfassend sind neue Versuche wie etwa die flächendeckenden Lernstandserhebungen (VerA usw.) der Bundesländer auch angelegt. Es gibt nur ein Problem: Methoden, die für statistische Aus-

[71] Vgl. Linn (2000).
[72] Vgl. zur Diskussion über das GATS-Abkommen: Bulmahn (2002) und zu analogen Entwicklungen in anderen Bereichen öffentlicher Daseinsvorsorge die Beiträge in GEW (2000).
[73] S. dazu auch die kritischen Anmerkungen von Strietholt/ Bos (2010, 173-175) zur eingeschränkten Aussagekraft und zum begrenzten Nutzen von Tests wie VERA für die Lerndiagnostik.

sagen über Zusammenhänge in großen Populationen taugen, sind unangemessen, wenn es um Einzelfallentscheidungen geht. Auch hier kann die Medizin als Beispiel dienen[74]. Ergebnisse aus der Erprobung von Medikamenten werden als Durchschnittsaussagen berichtet und führen zu allgemeinen Dosierungsanweisungen. Meist wird nur nach Kindern und Erwachsenen unterschieden – ohne Rücksicht auf Differenzen innerhalb dieser Untergruppen. Dabei kann dieselbe Arznei unterschiedlich wirken bei Dicken und Dünnen, bei mehr oder weniger reizempfindlichen Personen, bei Frauen während oder außerhalb der Menstruation, je nachdem, wie sich jemand ernährt oder welche Medikamente sonst noch genommen werden. Es bedarf der langjährigen Praxiserfahrung eines Arztes und zusätzlich seiner Vertrautheit mit dem Patienten, um die Therapie optimal anzupassen. Eine solche „Einstellung" der Medikamentierung ist im Praxisalltag aber oft nicht üblich, wenn es nicht um so heikle Medikamenten wie bei der Behandlung von Parkinson geht.

In großen Stichproben gleichen sich die auf der Individualebene unvermeidlichen Messfehler (bedingt z. B. durch die persönliche Tagesform) weitgehend aus. Auf der Fallebene aber sind sie erheblich. Für Politiker mag es interessant sein, dass Buchbesitz in der Familie stärker mit der Leseleistung von Schüler/innen korreliert als andere Merkmale. Lassen wir beiseite, dass „Buchbesitz" vielleicht nur Indikator für das umfassendere „kulturelle Kapital" oder dass das damit korrelierende „häufige Vorlesen" die eigentliche Ursache ist. Für den Einsatz öffentlicher Mittel, z. B. in einem Programm, Eltern bei den Vorsorgeuntersuchungen U4 bis U9 jeweils ein Bilder- oder Kinderbuch zu schenken, kann diese Korrelation ein wichtiges Argument sein. Bei der Erklärung von Lernschwierigkeiten auf der Individualebene kann das Merkmal Buchbesitz dagegen völlig in die Irre führen. Hier gewinnen Faktoren an Bedeutung, die man mit kurzen Fragen in standardisierter Form nicht erfassen kann, z. B. *wie* mit diesen Büchern in der Familie umgegangen wird.

Ebenso kann die Bildung von Untergruppen mit besonderem Risiko wie Jungen, wie Migranten- oder Unterschichtkinder durchaus helfen, den bildungspolitischen Blick zu differenzieren. Aber auf der Schul- und Unterrichtsebene verschließen solche Kategorien den Blick auf den Einzelfall und seine Besonderheit: das Mädchen mit den Leseproblemen, das unmotivierte Oberschichtkind, den gut lesende Migranten. Es kommt nicht nur zu normativen Etikettierungen, sondern auch zu so aberwitzigen Vorschlägen wie der Aufhebung der Koedukation von Mädchen und Jungen. Aberwitzig, denn die Streuung der Leistungen, der Interessen und der Lernstile innerhalb solcher Gruppen ist um ein Vielfaches größer als die Differenzen zwischen ihnen.

---

[74] Vgl. Albrecht (2005).

Mit PISA eingeführte Begriffe wie „Migrationshintergund“ erfassen den relevanten Sachverhalt zwar besser als „Staatsangehörigkeit“. Doch in standardisierter Form führt auch diese Klassifikation in die Irre. Nach Deutschland migrierte Familien unterscheiden sich beispielsweise von kanadischen Einwanderern in ihrer sozialen Schichtzugehörigkeit deutlich, so dass in Deutschland dieser Faktor zusätzlich ins Spiel kommt: nicht der Migrationshintergrund oder die Zweisprachigkeit an sich sind das Problem, sondern die Lebensbedingungen, die familiäre Unterstützung und andere durch das sozio-ökonomische Milieu bedingte Einflüsse. Die Differenzierung lässt sich fortsetzen. So gibt es unter den deutschen Immigranten große Unterschiede je nach nationaler Herkunft. Migranten-Jungen wiederum haben – nicht anders als unter den „eingeborenen“ Deutschen – mehr Probleme als Mädchen. Schließlich landet man beim Individuum.

Diese Auffächerung zeigt ein grundsätzliches Dilemma des Forschungsparadigmas auf, dem PISA & Co. verpflichtet sind[75]: Die Kategorien, mit denen sich Daten einer derart großen Stichprobe statistisch verarbeiten lassen, sind zu grob, um die Vielfalt relevanter Bedingungen zu erfassen. Dies schränkt die Möglichkeit bildungspolitischer Folgerungen ein: Für welche (Teil-)Gruppen genau ist ein Förderprogramm zu entwickeln, wenn man die knappen Ressourcen konzentrieren muss?

Wenig ergiebig sind die Befunde insofern für diejenigen, die vor Ort mit einzelnen Schülern zu tun haben. Sie helfen nicht beim Umgang mit einem muslimischen Mädchen aus einer türkischen Mittelschichtfamilie. Solche Schubladen mögen nützlich sein, um die Fülle der erhobenen Variablen etwas zu ordnen. Auf der Unterrichtsebene verstellen sie die Wahrnehmung für die je besondere Lernbiographie einzelner Schüler. Das Sprachproblem von Migranten ist eben nicht nur ein sprach-„technisches“, das sich durch entsprechende Kurse ausgleichen ließe, sondern ein sprachkulturelles, wie es auch deutsche Unterschichtkinder in die Schule mitbringen. Dieses aber ist nur mit genauen qualitativen Beobachtungen zu erfassen. Der in den letzten zwanzig, dreißig Jahren in der pädagogischen Praxis mühsam errungene „Blick auf das Individuum“ droht im Nach-PISA-Diskurs, der von Grobkategorien geprägt ist, wieder verloren zu gehen – übrigens entgegen der erklärten Absicht der Autoren: eine der ungewollten, aber absehbaren Nebenwirkungen. Von PISA & Co. profitiert die Politik, nicht die pädagogische Praxis.

In der Pädagogik muss man deshalb in einem ähnlichen Sinn „experimentell“ vorgehen wie in der Medizin: mit einer erfahrungsoffenen Haltung, nicht als bloß technische Anwendung von „Erkenntnissen“. Testergebnisse

[75] Ausführlichere Begründungen und Beispiele in: Brügelmann (2011a).

können als Ausgangshypothese dienen, um geeignete Aufgaben zu suchen. Aber dann werden diese – wie ein verordnetes Medikament – selbst zum Diagnoseinstrument. Erst die Reaktion der Schülerin, des Schülers zeigt, ob die Aufgabe passt oder nicht. Im gemeinsam reflektierten Probieren wird versucht, die Passung zu verbessern[76]. Das ist damit gemeint, dass Tests (nur) *heuristisch* sinnvoll sein können.

5. Alternative Perspektiven

Notwendig ist ein nach Stufen differenziertes System der Rechenschaftspflichten und Evaluationsformen[77]. Insofern ist es wichtig, aber nicht ausreichend, zentrale System-Monitorings (über punktuelle Leistungstests) zu ergänzen durch andere Indikatoren für die Qualität des Bildungswesens, wie es „Bildung auf einen Blick" der OECD (2007) und innerhalb von Deutschland das Konsortium Bildungsberichterstattung (2006) versuchen.

Für die Entwicklung einzelner Schulen ist eine umfassendere und kontextsensible Evaluation notwendig, und die individuelle Leistungsbewertung ist zu einer „dialogischen Lernbeobachtung" anhand förderorientierter Aufgaben zu entwickeln, die einen Blick unter die Oberfläche des Messbaren erlauben („Pädagogische Leistungskultur" im Sinne des Grundschulverbands, s.u.). 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 andererseits ihnen Pools von (fachdidaktisch kommentierten!) Testaufgaben mit repräsentativen Vergleichswerten zur Verfügung zu stellen. PISA & CO müssen aber inhaltlich und methodisch durch komplementäre Formen der Evaluation ergänzt werden. Vor allem ist es notwendig, die Evaluation vor Ort als eigenständige Aufgaben mit besonderen Anforderungen und Möglichkeiten zu fördern.

Damit dies keine abstrakte Forderung bleibt, stelle ich im Folgenden - in aller Kürze - zwei Beispiele vor:
- den reformpädagogischen Schulverbund „Blick über den Zaun" als Beispiel für eine Evaluation durch Peer-Review auf Schulebene und
- das Konzept „Pädagogische Leistungskultur" des Grundschulverbands als Beispiel für eine dialogische Form der Lernbeobachtung und Leistungsbeurteilung.

---

[76] S. dazu unten 3.2

[77] Vgl. den konkreten Vorschlag in Bartnitzky u. a. (1999) und die Übersicht über alternative Modelle in den OECD-Ländern bei Brügelmann (1980). Ideen für ein „demokratisches" Verfahren, in dem die Sicht der Betroffenen vor Ort stärker berücksichtigt werden, skizziert Retzl (2010, 360-363).

5.1 Schulevaluation durch „kritische Freunde“

Interne Evaluation hat den Vorteil intimer Situationskenntnis und fehlender Bedrohung. Andererseits: Der Fremdblick von außen ist wichtig, um scheinbare Selbstverständlichkeiten in Frage zu stellen. Zudem braucht jede/r Distanz zum Alltag, um sich den erkannten Schwächen zu stellen. Diese Einsichten stehen hinter den Lernstandserhebungen und den Schulinspektionen. Sie machen die Forderung nach externen Evaluationen stark. Aber die Frage ist, in welcher Rolle die Externen kommen: als ExpertInnen, gar als Autoritäten – oder als PartnerInnen?

Externe Evaluation ist notwendig. Aber für die Schulentwicklung wird sie produktiver, wenn sie von Kontrolle abgekoppelt wird. Die Schulaufsicht sollte deshalb durch ein kollegiales Peer-Review ergänzt werden, das den Unterrichtsbetrieb nicht rechtlich kontrolliert, sondern fachliche Rückmeldung gibt (dies geht nicht durch dieselben Personen, wie auch beispielsweise die Bewertungen und Beratungen von Lehramtsanwärterinnen durch ihre sie bewertenden Fachleiter zeigen). Die Schule kann den „Fremdblick“ durch eine interne Bestandsaufnahme vorbereiten: Was sind unsere Ziele, wo liegen unsere Stärken, welche Probleme haben wir? Und sie sollte die Berichte der externen BeobachterInnen öffentlich kommentieren: Welche Einschätzungen teilen wir, was wollen wir unternehmen, um unsere Arbeit weiter zu verbessern? Auch ohne formelle Sanktionsbefugnisse kann eine solche Inspektion durch ihre Kompetenz und durch den sozialen Druck, den die Berichterstattung bedeutet, wirken.

Einen Versuch, externe Sicht und notwendige Vertrautheit auszutarieren, ist die Initiative "Blick über den Zaun". Dieser Verbund von inzwischen über 120 reformpädagogischen Schulen besteht seit 1989. In ihm schließen sich jeweils 7-10 Schulen (bewusst aus verschiedenen Schulformen, Regionen und Traditionen) zu einem Arbeitskreis zusammen. Gemeinsam sind den Schulen die Standards des "Blick über den Zaun", die sich vor allem auf die Qualität der pädagogischen *Prozesse* beziehen[78]. Zweimal im Jahr wird eine Schule von jeweils zwei VertreterInnen der anderen Schulen des gleichen Arbeitskreises besucht[79], die zwei bis drei Tage in der Schule mitleben. Die gastgebende Schule kann einen spezifischen Beobachtungsauftrag formulieren, aber daneben sind die Gäste frei, das zu beobachten, was ihnen wichtig erscheint. Sie nehmen am Unterricht teil, sie unterhalten sich mit Kolleg/inn/en, mit Schüler/inne/n und Vertreter/inne/n der Eltern. In einer

[78] Vgl. von der Groeben u. a. (2005).

[79] Aus den Erfahrungen der vergangenen 20 Jahre ist ein Leitfaden für die Schulbesuche entstanden, der auch in anderen Netzwerken oder – in Abwandlung – von einzelnen Schulen genutzt werden kann (Backhaus u. a. 2009).

Schlussrunde spiegeln die BesucherInnen einzeln, d. h. aus ihrer je individuellen und damit unterschiedlichen Perspektive ihre Eindrücke dem Kollegium zurück.

Die Erträge dieses kollegialen Austauschs sind vielfältig (vgl. Seydel 2007, 5-7):

> „Auf die Zaungäste haben diese Besuche in der Regel drei wichtige Wirkungen:
> 1. Die oft geradezu verwirrende Konfrontation mit einer anderen, z. T. sehr fremden, Schulkultur klärt den Blick auf die eigene Schule.
> 2. Die Übernahme von neu in der besuchten Schule Gesehenem geschieht selten direkt, sondern zeitverzögert und mit einer Reihe von Transformationen, manchmal mit einem ‚sleeper effect', wenn erst im Nachhinein klar wird: ‚Das hatte ich ja dort und dort gesehen.' Nach der Rückkehr des Grenzgängers ist die Neugier der daheimgebliebenen Kollegen auf das, was er gesehen hat, nur von kurzer Dauer. Mit einiger zeitlicher Verzögerung kommt dann aber bei passender Gelegenheit die Frage: ‚Du warst doch in der Bodenseeschule - wie haben die denn die Organisationsprobleme des Epochenunterricht gelöst?' etc. Der Bericht über eine andere Schule bekommt eine ganz andere Qualität, wenn es im Kollegium jemanden gibt, der ihre Schwelle überschritten hat. Er hat nicht nur über deren Zaun geblickt, sondern mit dem Nachbarn selbst gesprochen.
> 3. Mindestens genauso wichtig – wenn nicht sogar wichtiger – im Vergleich zum ‚sachlichen' Transfereffekt ist der motivationale Aspekt der Ermutigung und Rückenwärme: ‚Meine Schule ist im Vergleich zu der anderen gar nicht so schlecht.' ‚Meine Arbeit wird von den anderen wahrgenommen und wertgeschätzt.' ‚Andere haben auch ungelöste pädagogische Probleme und sind trotzdem nicht verzagt.' …
> Das Gastgeschenk, das die Besucher als Dank zurücklassen, ist von ganz besonderer Art.
> Das Kollegium bekommt am Ende des Besuches einen ungewöhnlichen Spiegel über das Gesamtbild der Schule, über ihre Stärken, Schwächen und Entwicklungspotentiale. Die unterschiedliche Herkunft der Besucher – die Differenz z.B. zwischen einem antiautoritär geprägten Glockseelehrer und einem durch gemeinsame Formen und Werte geleiteten Montessorilehrer - ergibt differenzierte Blickwinkel und Einfärbungen der zurückgemeldeten Bilder. Die Fragen, die diese ‚kritischen Freunde' stellen, die Bebachtungen, die sie mitteilen, die Anregungen, die sie vorsichtig formulieren, tragen die Chance einer ganz anderen Wirkung in sich als der Besuch des Schulrates oder gar Inspektors. Sie sind Angebote auf Augenhöhe. Und weil es – auf Grund der Unterschiedlichkeit der Herkünfte der

> Besucher – nie ein ‚konsistentes' Bild ergibt, bleibt die Deutungshoheit bei der besuchten Schule. Die Differenz der Bilder fordert heraus. Oft waren noch Jahre nach einem Besuch die Provokationen der Zaungäste im Schulentwicklungsprozess präsent."

Als Rahmen für die Gespräche mit einzelnen Kolleg/inn/en können Fragen wie die folgenden hilfreich sein, wenn man den Unterricht in einer Klasse in den Blick nimmt:

- Was sind Ihre wichtigsten Ziele und Prinzipien?

Nachfrage: „Wie stehen Sie zu folgenden Vorgaben der Richtlinien/ Lehrpläne, zu folgenden Standards oder Positionen der schulpädagogischen und (fach-)didaktischen Diskussion?"

- Wo steht Ihre Lerngruppe, wo stehen einzelne Kinder in den zentralen Entwicklungsdimensionen?

Nachfrage: „Ist Ihnen aufgefallen, dass Marc...?"

- Wo sehen Sie die Stärken und die Schwächen Ihrer Arbeit, also Ihrer Versuche, die eigenen Ansprüche und die vorgegebenen Anforderungen umzusetzen?

Nachfrage: „Mir ist bei der Beobachtung Ihres Unterrichts aufgefallen, dass..."

- Welche Umstände erschweren es Ihnen, Ihre Ansprüche im Alltag umzusetzen?

Nachfrage: „Könnte es auch daran liegen, dass ...?"

- Was haben Sie sich als nächste Schritte zur Entwicklung Ihrer Arbeit vorgenommen?

Nachfrage: „Haben Sie auch an folgende Möglichkeiten gedacht:...?"

- Welche Unterstützung/ welche Rahmenbedingungen wären nötig?

Nachfrage: „Würde es Ihnen helfen, wenn...?"

Beratung statt Kontrolle bedeutet Konfrontation mit einer anderen Sicht, ohne dass diese sich als Norm versteht. Die Grundidee: Die normative Diskussion über die *Kriterien* für „guten Unterricht" ist zu trennen von der empirischen Frage nach der Qualität seiner *tatsächlichen Umsetzung*. Es macht wenig Sinn, den tatsächlichen Unterricht mit Ansprüchen zu bewerten, die die Bewerteten gar nicht teilen. Diese normativen Fragen sind vorweg zu klären – beispielsweise in einer Diskussion über das Schulprogramm, seine Stärken und Schwächen. Dabei kann zugleich Verständigung über die Kriterien erzielt werden, mit deren Hilfe der beobachtete Unterricht sinnvoll zu beurteilen ist. Testleistungen der Schüler/innen können eine Informationsquelle sein, um Schwächen auf die Spur zu kommen. Aber noch ertragreicher sind die Beobachtungen und das in langjähriger Erfahrung fundierte Urteil der Kolleg/inn/en aus anderen Schulen.

Auch dieses Verfahren hat seine Schwierigkeiten und das Instrumentarium ist entwicklungsfähig. Entscheidend ist der Ansatz: Evaluation „von außen",

aber nicht „von oben“, seien es die wissenschaftlichen ExpertInnen der Lernstandserhebungen, seien es die VertreterInnen der Verwaltung bei der Schulinspektion[80]. Begegnung auf Augenhöhe ist die Grundlage für Offenheit und damit für die Bereitschaft, sich den eigenen Schwächen zu stellen und ernsthaft an ihnen zu arbeiten. Dieser Anspruch schließt ein, Schüler/innen nicht nur als Informationsquelle, sondern als Experten für Kriterien guter Schule und ihre Umsetzung mit in den Prozess der Evaluation einzubeziehen[81].

Diese Anforderungen gelten auch für die Leistungsbewertung auf der Schülerebene.

5.2 „Dialogische Lernbeobachtung“ in einer pädagogischen Leistungskultur

Der Grundschulverband hat bereits vor der Veröffentlichung von PISA ein Konzept für ein umfassendes Evaluationssystem vorgeschlagen[82]. In ihm werden nicht nur die Rechenschaftspflichten von Politik und Verwaltung, sondern auch die Evaluationsaufgaben der einzelnen Schule, der Lehrpersonen und der Schüler/innen konkret entfaltet. Besondere Bedeutung wird der begleitenden Lernbeobachtung beigemessen. Statt nur abzuprüfen, ob Ziele erreicht sind, sollen Lernfortschritte und Schwierigkeiten im Lernprozess erfasst und diagnostisch gedeutet werden.

Dass die üblichen Klassenarbeiten und ihre Benotung diese Funktion nicht erfüllen können, ist seit vielen Jahrzehnten bekannt[83]; standardisierte Tests, die als Alternative angeboten werden, haben – wie oben gezeigt - ihre eigenen Probleme. Der Grundschulverband hat deshalb sein Konzept „Pädagogische Leistungskultur“ entwickelt und darin folgende Kriterien für Aufgaben zur Lernbeobachtung formuliert. Sie sollten nach Möglichkeit...

- der Lehrperson Informationen erbringen
  - nicht nur über aktuelle Einzelleistungen,
  - sondern auch über die zugrunde liegenden Strategien („Tiefenstrukturen“)

[80] Vgl. zum geringen Ertrag der gegenwärtig praktizierten Formen von Schulinspektion den Überblick über die vorliegenden Befunde bei Böttcher/ Keune (2010). S. zu einer kritischen Diskussion des Ansatzes auch Seydel (2005; 2009), der für Bremen eine stärker qualitativ ausgerichtete und weniger stark in die Hierarchie eingebundene Form der Inspektion entwickelt hat.

[81] Vgl. den Versuch, die Standards des „Blick über den Zaun“ für Kinder/ Jugendliche und ihre Eltern in Alltagssprache zu übersetzen, um sie in die Konkretisierung der Ansprüche und ihre Überprüfung mit einzubeziehen: Backhaus/ Brügelmann (2010); Brügelmann/ Backhaus (2011).

[82] Bartnitzky u. a. (1999).

[83] Vgl. die aktuelle Zusammenfassung in: Arbeitsgruppe Primarstufe (2006).

- und über deren Entwicklung („Lerngeschichte“);
- für die Schüler/innen auch inhaltlich eine produktive Lernsituation darstellen;

vor allem aber

- dialogisch angelegt sein als wechselseitige Verständigung über Lernziele,

Bewertungskriterien und tatsächliche Leistungen und damit

- die Fähigkeit der Kinder zur Selbsteinschätzung eigener Arbeiten entwickeln.

Das diagnostische Repertoire von Lehrer/innen kann durch heuristisch eingesetzte Tests, durch Beobachtungsbögen und verschiedene Dokumentationsformen (wie Portfolios) differenziert werden[84]. Das Konzept einer „pädagogischen Leistungskultur“ fordert daneben aber verschiedene „Institutionen“ im Unterrichtsalltag, die den Schüler/innen helfen, ihre eigene Arbeit kritisch-konstruktiv zu überprüfen und an den Arbeiten anderer ihre Maßstäbe zu schärfen. Nur drei Beispiele aus dem Grundschulbereich:

- Schreibkonferenzen, in denen nach bestimmten Regeln Text-Entwürfe vorgestellt, kommentiert und dann mit Hilfe anderer Kinder überarbeitet werden[85];
- Rechendiskussionen in der Klasse, z. B. zum „harten Brocken des Tages“[86], so dass die Kinder Schwierigkeiten, Hypothesen und Lösungsstrategien austauschen und damit voneinander lernen können;
- im Sachunterricht Metagespräche über Stärken und Schwächen von Präsentationen vor der Klasse, über Arbeitsergebnisse von Gruppen oder Einzelnen bis hin zu deren Bewertung durch das Plenum nach vereinbarten Kriterien[87].

Der Lesedidaktiker Schmalohr (1997, 42 f.) hat in seiner Arbeit mit jugendlichen und erwachsenen Analphabeten drei einfache Fragen genutzt, um sie zum Nachdenken über ihre Probleme zu bringen und mit ihnen in ein Gespräch über sinnvolle Lernwege zu kommen[88]:

1. "Wie lese ich, wo habe ich Schwierigkeiten?"
2. "Woran könnte das liegen?"
3. "Was kann ich tun?"

---

[84] Vgl. die vielfältigen Vorschlage für die verschiedenen Lernbereiche in: Bartnitzky u. a., (2005-2007).

[85] Vgl. Spitta (1998) und ergänzende Hilfen für den Sprachunterricht bei Brinkmann/ Brügelmann (2005).

[86] Eingeführt von Erichson (2004) im Rechtschreibunterricht, vgl. für den Mathematikunterricht Küppers (2005); Sundermann/ Selter (2005).

[87] Vgl. die Beispiele für den Sachunterricht bei Schönknecht/ Klenk (2005, 22 ff.)

[88] Einen analogen Zugang für die Grundschule haben Dehn/ Hüttis-Graff (2006) entwickelt.

In dem Ansatz einer „dialogischen Diagnostik“ [89] zeigt sich derselbe Geist wie in der Peer-Review auf Schulebene: Herausforderung und Beratung statt Beurteilung und Kontrolle. Ein solches Verständnis von Evaluation respektiert und nutzt die Kompetenz der Betroffenen, ihre Probleme selbst zu erkennen, Ursachen für diese zu finden und Ideen für ihre Überwindung zu entwickeln - ohne sich ganz auf sie zu verlassen.

6. Resümee

Damit schließt sich der Kreis: Evaluation bedeutet politisch Macht und inhaltlich Abhängigkeit. Man darf sie nicht Expert/inn/en überlassen, denn die mit ihr verknüpften Probleme sind technisch nicht zu lösen[90].

Schule ist zwar ein öffentlicher Raum und dies 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 nicht durch externe Beurteilungen zu entmündigen, sondern in ihrer persönlichen Evaluations- und Problemlösekompetenz zu stärken[91].

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.

- Die Scheinpräzision von Zahlen aus Erhebungen mit Leistungstests muss immer wieder bewusst gemacht und ihre mehrperspektivische Deutung schon bei der Veröffentlichung gesichert werden.

- Die Ressourcen für Evaluation dürfen nicht auf die zentrale Evaluation konzentriert werden. Wir brauchen eine vergleichbare politische, wis-

[89] Vgl. die konkreten Hilfen für eine „dialogische Diagnostik“ im Bereich des Schriftspracherwerbs die von uns im Rahmen des BMBF-Projekts ALPHA-Profess entwickelten Konzepte (Backhaus/ Knorre 2010; Backhaus 2011; Brügelmann 2011b; Rackwitz u. a. 2011) und Materialien: Backhaus u. a. (2010; 2011);
[90] Vgl. schon die Beiträge in MacDonald/ Walker (1974) und aktuell Mabry (2010, 17-23).
[91] Vgl. MacDonald/ Walker (1976); Brügelmann (2007; 2008); Retzl (2010, 360-363); Biesta (2010a).

senschaftliche und finanzielle Unterstützung für die Entwicklung der Evaluationskompetenz vor Ort[92].

Fazit: PISA & Co. können unser Evaluationsrepertoire bereichern – wenn wir das Instrumentarium nicht überfordern.

Quellen

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.

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.

Amrein, A. L./ Berliner, C. D. (2002): High-stakes testing, uncertainty, and student learning. Download: http://www.epaa.aus.edu/epaa/v10n18/ [Abruf: 22.1.11]

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

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

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

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. Weitere Informationen unter http://www.agprim.uni-siegen.de/notengutachten.htm [Abruf: 16.4.11]

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.

[92] Vgl. auch das Plädoyer für die Stärkung einer *methodisch* kontrollierten *Selbst*evaluation bei Lind (2011).

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

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: http://www.blickueberdenzaun.de/images/downloads/kinderstandards.pdf http://www.blickueberdenzaun.de/images/downloads/kinderstandardssynopse.pdf [Abruf: 30.12.2010])

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

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.

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.

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

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., 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: http://www.grundschulverband.de/bildungspolitik/zur-qualitaet-der-leistung/ [Abruf: 17.6.2011]

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.

Bauer, S. C. (2000): Should achievement tests be used to judge school quality? In: Education Policy Analysis Archives, Vol. 8, No. 46, 1-19.
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, 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: http://www.ipn.uni-kiel.de/pisa/ReplikKlemm.pdf [Abruf: 1.1.2011]
Bayrhuber, H., u. a. (Hrsg.) (2004): Konsequenzen aus PISA. Perspektiven der Fachdidaktiken. StudienVerlag: Innsbruck.
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.
Bellmann, J./ Müller, T. (Hrsg.) (2011): Wissen, was wirkt. Kritik evidenzbasierter Pädagogik. VS Verlag für Sozialwissenschaften: Wiesbaden.
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).
Blum, W./Neubrand, M. (2004): Der schiefe Blick auf Pisa. In: Süddeutsche Zeitung v. 11.12.2004, 13.
Böttcher, W./ Keune, M. (2010): Funktionen und Effekte der Schulinspektion. Ausgewählte nationale und internationale Forschungsbefunde. In: Böttcher u. a. (2010, 151-164).

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 u. a.
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./ Pietsch, M. (Hrsg.) (2007): KESS 4 – Kompetenzen und Einstellungen von SchülerInnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen. Waxmann: Münster.
Bos, W., u. a. (Hrsg.) (2003): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Waxmann: Münster u. a.
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.
Bracey, G. W. (2005): Education's groundhog day. In: Education Week, Vol. 24, No 21, 38-39. Auch Download über: http://www.edweek.org/ew/articles/2005/02/02/21bracey.h24.html [Abruf: 4.2.2004]
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./ Brügelmann, H. (2005): Pädagogische Leistungskultur – Materialien für Klasse 1 und 2: Deutsch. Heft 3 in: Bartnitzky u. a. (2005).
Brügelmann, H. (1980): Experimental decision making and responsive accountability. Expert report for "Basic Education Policies Project".

OECD/ CERI: Paris ➔ Reprint der Kurzfassung http://www.agprim.uni-siegen.de/printbrue.htm [Abruf: 14.4.10].
Brügelmann, H. (2003): Lese-Untersuchung mit dem Stolperwörter-Test. Abschlussbericht des Projekts LUST-1. Download: www.agprim.uni-siegen.de/lust [Abruf: 14.4.10]
Brügelmann, H. (2004): Leseleistungen von LehrerInnen und Lehramtsstudierenden im Stolperwörter-Lesetest. Erste Befunde und ihre Deutung. Projekt LUST/ FB 2 der Universität: Siegen. Download: http://www.agprim.uni-siegen.de/lust/stolper_auswertung_lehramt.pdf [Abruf: 14.4.10]
Brügelmann, H. (2005a): Schule verstehen und gestalten – Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil (bis 2008 fortlaufend aktualisiert unter: www.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 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“ (2007). Download: http://www.blickueberdenzaun.de/files/Bruegelmann-Evaluation.doc [Abruf: 30.12.10]
Brügelmann, H. (2011a): Miss Marple neben PISA & Co. - Plädoyer für eine Bildungsforschung, die der Praxis nützt. In: Moser (2011, 221-234).
Brügelmann, H. (2011b): Individuell beobachten statt (bloß zu) testen. „Können“ und „Wissen“ erschließen sich erst unter der Oberfläche des Verhaltens. Ms. für de Boer/ Reh (2011; in Vorb.).
Brügelmann, H.,/ Backhaus, A. (2011): Kinder als Kinderschützer. Ms. für Buchholz, T., u. a. (Hrsg.) (2011): Kinderschutz in gemeinsamer Verantwortung von Jugendhilfe und Schule. VS Verlag für Sozialwissenschaften: Wiesbaden (in Vorb.).
Brügelmann, H./ Richter, S. (Hrsg.) (1994): Wie wir recht schreiben lernen. Zehn Jahre Kinder auf dem Weg zur Schrift. Libelle Verlag: CH-Lengwil (2. Aufl. 1996).
Brügelmann, H., u. a. (1994): "Schreibvergleich BRDDR" 1990/91. In: Brügelmann/ Richter (1994, 129-134).
Bulmahn, E. (2002): Wir dürfen Bildung nicht als Ware dem Handel überlassen. Die Welthandelsorganisation berät über den Import und Ex-

port von Hochschul-Dienstleistungen. In: Frankfurter Rundschau v. 8.7.2002.
Campbell, D. T. (1976): Assessing the impact of planned social change. In: The Public Affairs Center, Dartmouth College, December, 1976. Reprint: http://www.wmich.edu/evalctr/pubs/ops/ops08.pdf [Abruf: 16.5.07])
Collani, E. v. (2001): OECD PISA - An Example of Stochastic Illiteracy? In: Economic Quality Control, Vol. 16, No. 2, 227-253.
de Boer, H./ Reh, S. (Hrsg.) (2011): Beobachtung in der Schule – Beobachten lernen. VS Verlag für Sozialwissenschaften: Wiesbaden (in Vorb.).
Dehn, M./ Hüttis-Graff, P. (2006): Zeit für die Schrift 2. Beobachtung und Diagnose. Cornelsen Scriptor: Berlin.
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.
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 [Abruf: 17.3.2007]
Dolin, J. (2007): PISA – an example of the use and misuse of large-scale comparative tests. In: Hopmann u. a. (2007, 93-126).
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.
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.
Erichson, C. (2004): Der harte Brocken des Tages. In: Grundschule Deutsch, 1. Jg., H.2, 14-17.
Fauser, P., u. a. (Hrsg.) (2007): Was für Schulen! Portraits der Preisträgerschulen und der nominierten Schulen des Wettbewerbs 2006. Klett Kallmeyer: Stuttgart.
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.
Füller, C. (2007a): Streit um die Pisa-Interpretation. Faktisch nur im Mittelmaß. in: TAZ vom 2.12.2007. Download: http://www.taz.de/1/zukunft/wissen/artikel/1/faktisch-nur-im-mittelmass/?src=SE&cHash=40e8c46e85 [Abruf 4.12.07]

Füller, C. (2007b): Widersprüche bei Pisa. Pisa hat einen kleinen, fröhlichen Bruder. In: TAZ vom 5.12.2007. Download: http://www.taz.de/1/zukunft/wissen/artikel/1/pisa-hat-einen-kleinen-froehlichen-bruder/?src=SE&cHash=4fe51fe346 [Abruf: 5.12.07]

Gaeth, F. (2009): PISA: Zukunft der Bildung oder Joulupukki? In: Politik-Poker v. 1.10.2009. Download: http://www.politik-poker.de/pisa-zukunft-der-bildung-oder-joulupukki.php [Abruf: 1.1.2010]

Gaeth, F. (2010): Sieger im Mittelmaß - die neue PISA Studie. In: Politik-Poker v. 29.6.2010. Download: http://www.politik-poker.de/sieger-im-mittelmass-die-neue-pisa-studie.php [Abruf: 1.1.2011]

Geißler, R. (1992/ 2006): Die Sozialstruktur Deutschlands. Zur gesellschaftlichen Entwicklung mit einer Bilanz zur Vereinigung, Hrsgg. von der Bundeszentrale für politische Bildung. VS Verlag für Sozialwissenschaften: Wiesbaden (4. überarbeite und aktualisierte Aufl. 2006; 1. Aufl. 1992; überarbeitete 2. und 3. Auflage 1996 und 2002).

GEW (2000): Die Privatisierung des Bildungsbereichs. Tagung „Privatisierung des Bildungsbereichs, Eigentum und Wertschöpfung in der Wissensgesellschaft". 15.-17.6.2000 in Hamburg. GEW-Dokumente 10/00.

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

Goldstein, D., u. a. (1990). Sex differences in visual-spatial ability: The role of performance factors. Memory & Cognition, Vol. 8, 546-550.

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).

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 [Abruf: 20.11.2007]

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., 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.

Hellermann, M., u. a. (2008): Das Kind hinter PISA. Wie die junge Generation fühlt, was sie denkt und wie sie lernt. Extrakte (Auszüge aus der Wissenschaft), Nr. 4. Presse- und Informationsstelle der Universität:

Siegen. Download: www.agprim.uni-siegen.de/inprint/extrakte0408.pdf [Abruf 30.12.2010]
Helmke, A. & Hosenfeld, I. (2003). Vergleichsarbeiten (VERA): Eine Standortbestimmung zur Sicherung schulischer Kompetenzen. Teil 1&2 in: SchulVerwaltung (Ausgabe Hessen/Rheinland-Pfalz/Saarland), H. 1/2003, 10-13, und H. 2/2003, 41-43.
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. (2007): Epilogue: No child, no school, no state left behind: Comparative research in the age of accountability. In: Hopmann u. a. (2007, 363-415).
Hopmann, S. T., u. a. (Hrsg.) (2007): PISA zufolge PISA/ PISA According to PISA. Lit-Verlag: Wien/ Berlin.
Hopmann, S. T. (2008): No child, no school, no state left behind: Schooling in the age of accountability. In: Journal of Curriculum Studies, Vol. 40, 417-456.
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.
House, E. R. (1975): Accountability in the U.S.A. In: Cambridge Journal of Education, Vol. 5, No. 2, 71-78.
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.
Institut für Schulentwicklung PH Schwäbisch Gmünd (Hrsg.) (2004): Standards, Evaluation und neue Methoden. Schneider Verlag Hohengehren: Baltmannsweiler.
Jahnke, T./ Meyerhöfer, W. (Hrsg.) (2006): PISA & Co – Kritik eines Programms. Franzbecker: Hildesheim.
Jahnke, T./Meyerhöfer, W. (Hrsg.) (2007): Pisa & Co. Kritik eines Programms. 2. überarb. Aufl. Franzbecker: Hildesheim.
Jude, N./ Klieme, E. (2010): Das Programme for International Student Assessment (PISA). In: Klieme u. a. (2010, 11-21).
Karg, I. (2005): Mythos PISA. Vermeintliche Vergleichbarkeit und die Wirklichkeit eines Vergleichs. V&R unipress: Göttingen.
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.

Klein, H. P. (2010): Die neue Kompetenzorientierung: Exzellenz oder Nivellierung. In: ZfdB, 1. Jg., 15-26. Download: http://www.bildung-wissen.eu/beitraege/artikel_kompetenzorientierung_sicher.pdf [Abruf: 3.2.2011].

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. (2004): Auch die neue Pisa-Studie zeigt, dass die Wahl der Schulform von der sozialen Herkunft der Schüler abhängt. In: Frankfurter Rundschau online v. 08.12.2004

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: http://www.zeit.de/2008/21/C-Bildungsforschung?page=all&print=true [Abruf: 1.1.2011]

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

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).

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

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

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

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

Köller, O. (2006): Kritik an PISA unberechtigt. Interview mit Bildungsklick. Download: http://bildungsklick.de/a/50155/kritik-an-pisa-unberechtigt/ [Abruf: 1.1.2011]

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

Krugman, P. (2011): Degrees and Dollars. The New York Times, March 6, 2011. Auch Download über

http://www.nytimes.com/2011/03/07/opinion/07krugman.html?_r=2&hp [Abruf: 4.6.2011].

Küppers, H. (2005): Mathematik. Heft 4 in: Bartnitzky u. a. (2005).
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: http://www.messen-und-deuten.de/pisa/Ladenthin03.pdf [Abruf: 31.12.2010]
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/ [Abruf: 4.1.11]
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. (2004): Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 11. Ergebnisse einer Längsschnittsuntersuchung in Hamburg. Behörde für Bildung und Sport: Hamburg. Download: http://www.hamburger-bildungsserver.de/welcome.phtml?unten=/schulentwicklung/lau/ [Abruf: 15.1.2008]
Lehmann, R.H., u. a. (1995): Leseverständnis und Lesegewohnheiten deutscher Schülerinnen und Schüler. Beltz: Weinheim/ Basel.
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: http://www.uni-konstanz.de/ag-moral/pdf/Lind-2003_evaluationskultur.pdf [Abruf:1.1.2011]
Lind, G. (2009a): PISA im Klassenzimmer Wohin führen sanktionsorientierte 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: http://www.uni-konstanz.de/ag-moral/pdf/Lind-2009_PISA-Folgen-fuer-Jugend_4fach.pdf [Abruf: 1.1.2011]
Lind, G. (2009b): Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In: Bohl/ Kiper (2009, 61-79). Download: http://www.uni-konstanz.de/ag-moral/b-liste.htm#Lind_2009_Amerika_Vorbild_lang [Abruf: 25.1.2011]
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.
Ludwig, J. (Hrsg.) (2011): Lernberatung in der Alphabetisierung. Modelle und Handlungsempfehlungen. wbv: Bielefeld (im Druck).
Ludwig, L., u. a. (Hrsg.) (2011): Bildung in der Demokratie II. Tendenzen – Diskurse – Praktiken. Barbara Budrich: Opladen & Farmington Hills, MI.

Lurija, A. R. (1993): Romantische Wissenschaft. Forschungen im Grenzbezirk von Seele und Gehirn. Rowohlt: Hamburg (russ. 1982).
Mabry, L. (2010): The responsibility of evaluation. In: Böttcher u. a. (2010, 17-33).
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.
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.
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.
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.
Nichols, S.L. & Berliner, D.C. (2007). Collateral damage: How-stakes testing corrupts schools. Cambridge, MA: Harvard Education Press.

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

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: http://www.bmbf.de/pub/zusammenfassung_eag.pdf [Abruf: 22.10.2007]

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.

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

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.

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 [Abruf: 2.4.03]

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

Prenzel, M. (o.J.): Wie solide ist PISA? oder: Ist die Kritik von Joachim Wuttke begründet? Institut für Pädagogik der Naturwissenschaften. Download http://www.ipn.uni-kiel.de/pisa/Wie_solide_ist_PISA.pdf [Abruf: 1.1.2011]

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., 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.) (2007): PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie. Waxmann: Münster.
Puchhammer, M. (2007): Language-Based Item Analysis. In: Hopmann u. a. (2007, 127-137).
Rackwitz, R.-P., u. a. (2011): Dialogische Förderdiagnostik in der Alphabetisierung am Beispiel des Schriftspracherwerbs. Ms. für Ludwig (2011).
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.
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.
Retzl, M. (2010): Schulqualität entwickeln durch nationale Standards? Grundlegungen für einen alternativen Ansatz der Qualitätsentwicklung. In: Böttcher u. a. (2010, 355-365).
Rindermann, H. (2006): Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? In: Psychologische Rundschau, 57. Jg., H. 2, 69-86.
Rolff, H.-G. (1967): Sozialisation und Auslese durch die Schule. Quelle und Meyer: Heidelberg.
Rotte, R. (ed.) (2006): International perspectives on education policy. Nova Science Publ.: New York.
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.
Schmalohr, E. (1997): Das Erlebnis des Lesens. Grundlagen einer erzählenden Lesepsychologie. Klett-Cotta: Stuttgart.
Schmitt, R. (Hrsg.) (1999): An der Schwelle zum dritten Jahrtausend. BundesGrundschulKongress 1999. Grundschulverband – Arbeitskreis Grundschule: Frankfurt.
Schönknecht, G./ Klenk, G. (2005): Sachunterricht. Heft 5 in: Bartnitzky u. a. (2005).
Schulverbund „Blick über den Zaun“ (Hrsg.) (2007): Beobachten, bewerten, beraten. Verfahren und Werkzeuge für eine andere Evaluation. Re-

formpädagogische Arbeitsstelle 'Blick über den Zaun'. Universität: Siegen.
Selter, C. (2005): VERA Mathematik 2004: VERbesserungsbedürftige Aufgaben! VERkapptes Ausleseinstrument?. In: Grundschule aktuell, H. 89, 17-20.
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).
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.
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.
Sjøberg, S. (2004): Internationale Vergleichsstudien — ihre guten und schlechten Seiten. In: Bayrhuber u. a. (2004, 51-61).
Spitta, G. (Hrsg.) (1998): Freies Schreiben – eigene Wege gehen. Libelle: CH-Lengwil.
STERN (2010): Die beste Schule für mein Kind. Im Kooperation mit dem Deutschen Schulpreis. Ratgeber Bildung 1/2010. Gruner & Jahr: Hamburg.
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).
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/ [Abruf: 13.1.06]
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.
Venn-Brinkmann, U. (2011): Wörter – Sätze – Texte. Eine mehrdimensionale Untersuchung zur Lesekompetenz Jugendlicher am Ende ihrer Regelschulzeit. Noch unveröff. Diss. an der Universität: Oldenburg.
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.

Watermann, R., u. a. (2003): Schulrückmeldungen im Rahmen von Schulleistungsuntersuchungen: Das Disseminationskonzept von PISA-2000. In: Zeitschrift Für Pädagogik, 49. Jg., H. 1, 92-111.
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.
Wolf, A. (2002): Does education matter? Myths about education and economic growth. Penguin: London.
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.
Zimmer-Mueller, M., u. a. (2008): Vergleichsarbeiten in der Grundschule: Wieso, weshalb, warum? In: Grundschulzeitschrift, 22. Jg., H. 215/216, 12-17.