Hans Brügelmann (2008a, 2.12.2007)

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

Wie Leistungstests in ihren Leistungsmöglichkeiten durch PISA&Co überfordert werden[1]

0. Was kann pädagogische Lernbeobachtung von medizinischer Diagnose lernen?

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 genau 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. 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 die Ursache für die erhöhte Temperatur zu finden. Gute Ärzte verzichten auch darauf, Fieber senkende Medikamente zu verschreiben, nur um die störenden Symptome möglichst rasch verschwinden zu lassen. Oft könnte dies sogar riskant sein, weil eine nur oberflächliche Anpassung an die Standardtemperatur verhindert, die eigentliche Krankheit wahrzunehmen.

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 Bil-

[1] Vorläufiges Ms. für: Brügelmann, H. (2008a): 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)

dungswesen 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"…

1. PISA als Projekt, Programm und Paradigma

Man kann das Potenzial und die Schwächen von PISA aus verschiedenen Blickwinkeln diskutieren. Als konkretes **Projekt** untersucht PISA alle paar 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. In Form von VERA und anderen flächendeckenden Lernstandserhebungen haben die Bundesländer PISA&Co allerdings zu einem **Programm** des System-Monitoring gemacht. Dieses erhebt darüber hinaus den Anspruch, die Arbeit von Schulen und Lehrerinnen, ja sogar den Leistungsstand einzelner SchülerInnen bewerten zu können. Inzwischen ist PISAnismus mit seiner Testmentalität für viele sogar zum generellen **Paradigma** der Evaluation von Unterricht und von individuellen Lernprozessen geworden.

1.1 PISA als Projekt

Zunächst ist als PISA als konkrete, periodisch stattfindende internationale Vergleichsstudie von Fachleistungen am Ende der Pflichtschulzeit zu diskutieren. Solche regelmäßigen Bestandsaufnahmen sind durchaus sinnvoll, wenn man sie als *einen* Baustein in einem umfassender konzipierten System-Monitoring sieht[2] – und wenn man die Grenzen von Design und Instrumenten nicht nur kennt, sondern in der Bewertung der Befunde und in den anschließenden Entscheidungen berücksichtigt.

Denn PISA allein
- sagt uns nicht, wie gut oder schlecht unser Schulsystem ist - dazu wird ein zu kleiner Ausschnitt der Schulqualität mit zu begrenzten Mitteln erfasst;
- kann nicht erklären, wo die Ursachen für Schwächen liegen, denn Korrelationen, vor allem in einem Querschnitt, erlauben keine Kausalschlüsse;

[2] Vgl. das schon breiter angelegte „Bildung auf einen Blick" der OECD (z. B. 2007).

- bietet keine Anweisungen für deren Überwindung, denn es ist nicht als experimenteller Vergleich von alternativen Handlungsmöglichkeiten angelegt.

PISA zeigt nicht, wie gut – oder auch nur: wie erfolgreich - ein Schulsystem ist, sondern was 15-Jährige in PISA-Tests leisten. Diese sind bewusst nicht auf das jeweilige Curriculum abgestimmt, untersuchen also nicht, wie gut ein Schulsystem *seine* Ziele erreicht. Aber selbst bezogen auf die PISA-Kriterien dürfen Unterschiede nicht ohne Weiteres der Qualität von Unterricht zugerechnet werden, lassen sie sich doch auch auf andere Faktoren wie Familie und Medien zurückführen. Soweit Schätzungen zu schulbedingten Unterschieden innerhalb von Ländern veröffentlicht sind, liegen sie bei 5-15%[3]. Dieser niedrige Anteil ist plausibel, wenn man Erkenntnissen zur Bedeutung informellen Lernens folgt[4].

Punktdifferenzen bei PISA werden überschätzt. 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 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 einen Unterschied von zwei richtigen Aufgaben[5]. Ist das eine pädagogisch oder politisch relevante Differenz? Und wie schnell kann selbst die verschwinden, wenn sich allein durch die Übersetzung, etwa vom Englischen ins Französische, die Textmenge um 10-20% erhöht?[6]

Anders als in der Medizin kann im Bildungsbereich ein anderes „Thermometer" auch zu anderen Ergebnissen führen[7]. Schon PISA-intern wird dies sichtbar in unterschiedlichen Leistungsprofilen verschiedener Länder oder Teilgruppen (z. B. Mädchen vs. Jungen). Je nach der Gewichtung von Teilbereichen und Aufgabentypen können sich Rangfolgen von Ländern und Anteile von Leistungsgruppen innerhalb von Ländern verändern. Auch wer keine grundlegenden Verschiebungen erwartet, sollte die Präzision von Punktwerten nicht überschätzen. Konkret sind folgende Einschränkungen zu bedenken, wenn man die Ergebnisse von PISA sinnvoll nutzen will[8]:

[3] Vgl. Hopmann (2007, 390) mit Bezug auf die PISA-Auswertung von Watermann u. a. (2003)..
[4] Vgl. Dohmen (2001); Otto/ Rauschenbach (2004).
[5] Vgl. Wuttke (2007a, 244-245).
[6] Vgl. Wuttke (2007a, 257).
[7] S. ausführlicher dazu unten 1.3 .
[8] Vgl. zu einzelnen Punkten auch eine Reihe von Beiträgen in Jahnke/ Meyerhöfer (2007) und Hopmann u. a. (2007)

- Die Itemanalyse nach Kompetenzstufen mit einem unterstellten **eindimensionalen** Aufbau wird der Entwicklung von Wissen und Können nicht gerecht und schränkt das Spektrum möglicher Aufgaben noch einmal ein[9].

- **Ohne** Kenntnis der **Lösungswege** sind die Aufgabenlösungen oft mehrdeutig[10].

- Die ökologische **Validität** solcher Tests erweist sich immer wieder als problematisch[11] bzw. es kann nicht geklärt werden, ob Variablen *Ursache* oder nur *Indikator* für Drittfaktoren sind.

- Die **Stichproben** in den verglichenen Ländern sind nicht immer vergleichbar[12].

- **Interkulturelle** Vergleiche sind auch inhaltlich schwierig und die Herkunft von Testitems kann deren Schwierigkeit für SchülerInnen aus verschiedenen Ländern unterschiedlich beeinflussen (Vertrautheit des Testformats; Bekanntheit der Inhalte; Verständlichkeit der Übersetzungen; Länge der Sätze; unterschiedliche Reaktionen auf Zeitdruck und Multiple Choice)[13].

Zusätzlich ist im Blick zu behalten:

---

[9] Vgl. zusammenfassend Wuttke (2006, 144).

[10] 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 Kap. 1.2 und Anm. 29).

[11] 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 vs. 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 entnehmen.

[12] Vgl. die Hinweise von Wuttke (2006, 143) auf die unterschiedlichen Schulbesuchquoten mit 15; auf die unterschiedliche Einbeziehung von SchülerInnen der untersten Leistungsgruppen; auf die unterschiedliche Teilnahmequoten; auf nicht plausible Verteilungen in einzelnen Ländern.

[13] Vgl. Artelt/ Baumert (2003); Wuttke (2007a, 254-257)

- Mit der Beschränkung auf **Wirkungen** („output") erfasst PISA nur einen Aspekt der Qualität von Schule – diese drückt sich aber auch in Prozessmerkmalen und Rahmenbedingungen von Schule aus[14].

- Innerhalb der Wirkungen stellen die untersuchten Fächer bzw. der aus ihnen jeweils einbezogenen **Fachanteile** nur einen kleinen Ausschnitt wesentlicher Effekte von Schule dar.

- Die notwendige **Standardisierung** begrenzt die möglichen Aufgabenformate, Rahmenbedingungen wie Zeitdruck verändern die Ergebnisse zusätzlich.

- **Kategorien** wie „Migrationshintergrund" oder „Gesamtschule" lassen sich zwar als Variablen technisch operationalisieren, haben aber in verschiedenen Kontexten unterschiedliche Bedeutung[15].

- Die Korrelationen aus **Querschnitt**-Untersuchungen erlauben keine Erklärungen, suggerieren aber Kausalitäten (Lesefreude ➔ Leseleistung) oder werden zumindest so gelesen[16].

- Durch die große Zahl der getesteten SchülerInnen werden auch inhaltlich irrelevante Unterschiede statistisch „**signifikant**", ohne dass sie deshalb auch inhaltlich bedeutsam wären.

- Eine Wiederholung der Erhebungen alle drei Jahre ist zu **kurzfristig** angelegt, jedenfalls bei der Schwerfälligkeit von Unterrichtstraditionen nicht nötig[17] - und sie bindet Aufmerksamkeit und an anderer Stelle bitter vermisste Ressourcen.

Manche dieser Probleme gelten für standardisierte Testprogramme generell und lassen sich testtechnisch nur begrenzt reparieren. Aber andere Erkenntnisformen haben auch ihre Schwächen. Insofern sollte man das Instrument nicht grundsätzlich verwerfen, sondern

[14] Vgl. etwa die Standards des „Blick über den Zaun" (von der Groeben u. a. 2005) oder des Grundschulverbands (2003). Die Verfasser des Klieme-Gutachtens loben die Prozess-Standards der NCTM in den USA ebenfalls ausdrücklich, lassen diese Alternative bei ihren weiteren Überlegungen aber ohne Begründung links liegen (vgl. Klieme u. a. 2003).

[15] Vgl. etwa den Vergleich der Migrantenanteile und ihrer Zusammensetzung in verschiedenen Bundesländern bei Klemm (2002).

[16] 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.

[17] 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).

in Kenntnis seiner spezifischen Stärken und Schwächen als *eine* Informationsquelle neben anderen nutzen[18].

Für die Einschätzung von PISA als Projekt sind dabei seine Nebenwirkungen mit zu bedenken. Ich beginne mit den positiven Effekten:

- PISA hat Probleme von Teilgruppen ins öffentliche Bewusstsein gerückt, die vorher zwar in Fachkreisen schon bekannt waren, aber in der Bildungspolitik nicht ernst genommen wurden: die Schwächen von Jungen[19], von MigrantInnen[20] und generell von Kindern und Jugendlichen aus Familien mit niedrigem SÖS[21].

- PISA hat lange tabuisierte Gründe für die Benachteiligung dieser Gruppen wieder in die Diskussion gebracht wie die Selektionsmechanismen des zwei- bis fünfgliedrigen Schulsystems[22].

Dem stehen negative Nebenwirkungen gegenüber:

- Es entsteht durch inhaltlich überzeugende Folgerungen 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.

- Für die internationalen Vergleiche werden viele Ressourcen absorbiert. Oft wären Sekundär- und Metaanalysen vorhandener Studien zureichend – und sogar aussagekräftiger[23]. 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.

- Schule als Ort der Persönlichkeitsentwicklung, des sozialen und politischen Lernens, als Raum, in dem eine Gesellschaft durch die Begegnung ihrer Teil-Kulturen und der Generati-

[18] Vgl. zur Notwendigkeit komplementärer Forschungsansätze in den Humanwissenschaften u. a. aus der Medizin Lurija (1993, 177 ff.)
[19] 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).
[20] Vgl. schon Arbeitsgruppe am Max-Planck-Institut für Bildungsforschung (1979); Glumpler (1985); Bommes/ Radtke (1993).
[21] Vgl. schon Rolff (1967); Geißler (1992 ff.); Müller (1998).
[22] S. aber auch dazu schon frühere Befunde, etwa aus LAU (vgl. Lehmann u. a 1997, 98-102)
[23] Vgl. etwa die Längsschnittstudien zur Bedeutung kleiner Klassen oder zum Einfluss der Vorschulerziehung auf den späteren Schulerfolg

onen zusammenwächst, gerät in der öffentlichen Diskussion ebenso aus dem Blick wie das Kind „hinter PISA" mit seinen lebensweltlichen Erfahrungen.

Man kann nun einwenden: Nichts ist perfekt und lieber das bessere statt des schlechteren Instruments nehmen. Aber um wie viel besser ist PISA als andere Instrumente und was ist der Preis für die Akzeptanz des aktuell besten? Wenn man bei einer Vorsorgeuntersuchung für Hautkrebs zwischen zwei Ärzten wählen kann, deren Brillen um jeweils 6 Dioptrien korrigiert werden müssten, und der eine hat eine Brille mit 2 Dioptrien, der andere mit 4 – soll man dann einfach den mit der besseren Brille nehmen? Oder nimmt man lieber einen Laien mit gesunden Augen? Oder verzichtet man lieber erst einmal auf die Untersuchung und wartet auf einen Arzt mit besser passender Brille? Solche Kosten-Nutzen-Rechnungen setzen eine hohe Transparenz auf Seiten der AutorInnen empirischer Studien und eine hohe Kompetenz auf der Seite ihrer LeserInnen voraus.

Das gilt auch inhaltlich. Denn: „Zahlen sprechen nicht für sich". Schon die Darstellung von Testergebnissen, ihre Übersetzung in Schaubilder und Texte ist immer ein Akt der Interpretation. Die Stärke und zugleich die Schwäche von PISA&Co liegt in der hohen fachlichen Kompetenz der beteiligten ForscherInnen – und dem Reichtum anderer Studien, die sie zur Interpretation ihrer Ergebnisse heranziehen. Die angeblich so „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.

Umso mehr wünschte man sich, dass schon in den Forschungsberichten selbst die Mehrdeutigkeit von Daten sichtbar gemacht würde. Wichtiger als eine 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 die scheinbare Expertenautorität zu relativieren .und damit auch der Öffentlichkeit die Interpretationsbedürftigkeit der nur scheinbar eindeutigen Daten bewusst zu machen[24]

1.2 PISA & Co als Programm

PISA hat sich inzwischen von einem spezifischen Projekt zu einem Modell des System-Monitoring gemausert, das darüber hinaus den Anspruch erhebt, die Arbeit von Schulen und Lehrerinnen, ja sogar den Leistungsstand einzelner SchülerInnen bewerten zu können. 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 LehrerInnen 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[25].

Das Problem mit dieser Argumentation ist allerdings:

- Nicht alle Länder, die besser abgeschnitten haben, sind output-gesteuert.

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

[25] Von manchen, z. B. von Andreas Schleicher auf dem GEW-PISA-Kolloquium am 21.11.2007 in Berlin, wird der Vorwurf erhoben, bisherige Formen der Steuerung und Evaluation hätten keine Veränderungen bewirkt. Dem ist entgegenzuhalten, dass sich im Grundschulbereich zwischen 1980 und 2000 vergleichsweise viel verändert hat. Genau diese Reformen „von unten" drohen nun durch PISA&Co. verloren zu gehen. Auch die IGLU-Studien zeigen, dass der Zuwachs von 1991 bis 2001 vergleichsweise größer ist als der von 2001 bis 2006, also nach der breiten Einführung von Kompetenztests und Outputsteuerung, vgl. Bos u. a. (2007, 20).

- Länder, die besser abgeschnitten haben, unterscheiden sich auch in weiteren Merkmalen vom deutschen System (Dauer der gemeinsamen Schulzeit; Zeitpunkt, ab dem benotet wird, usw.)
- Im deutschen input-orientierten System hat die Grundschule bei IGLU wesentlich besser abgeschnitten als die Sekundarstufe bei PISA – und selbst innerhalb Deutschlands gibt es größere Unterschiede zwischen den Bundesländern, die alle input-orientiert gesteuert werden, als zwischen Deutschland und Vergleichsländern mit Output-Steuerung.
- 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".

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[26]. Evaluation wird auf „measurement" verkürzt[27] (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 Richtigkeit der Lösung statt nach Produktivität des Lösungswegs (s. 1.1)[28].

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

- die flächendeckende Durchführung ist sehr viel aufwändiger und zugleich fehleranfälliger, wäre für ein System-Monitoring auf bildungspolitischer Ebene auch gar nicht nötig;

- 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;

[26] Umfassender z. B. „Bildung auf einen Blick", vgl. OECD (2007).
[27] Vgl., dazu schon 1997 die Kritik und Alternativen in dem Sammelband „Beyond the numbers game" (Hamilton u. a. 1977).
[28] 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).

- die fehlende Anonymisierung führt zu all' den unerwünschten Nebenwirkungen, die aus angelsächsischen Studien zum High-Stakes-Testing bekannt sind: fachliche Verengung des Curriculum[29]; Ausrichtung des Unterrichts an Testformaten[30]; unterschiedliche Auslegung und Durchführung der Aufgaben bis hin zu Täuschungsversuchen[31].

Insbesondere Studien in den USA belegen, dass die Schulleistungen in Bundesstaaten mit hohen Sanktionen für Institutionen und/ oder Personengruppen schlechter sind als in Bundesstaaten mit niedrigen Sanktionen[32]: Über dem nationalen Durchschnitt liegen 60-70% der Bundesstaaten mit niedrigen Sanktionen, aber nur 10-20% der Staaten mit hohen Sanktionen[33]. 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 SchülerInnen aus dem System ganz herausfielen, was die gemessenen Testleistungen zusätzlich in die Höhe trieb, ohne dass sich der „Output" tatsächlich verbesserte[34]. Linn (2000) konnte 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 zweites Mal auf ihr Ursprungsniveau „sanken"[35]. 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 LehrerInnen oder gar über den Lernstand von SchülerInnen machen zu können. Damit wird der Kredit von Tests endgültig überzogen, wie die folgenden Hinweise zeigen.

---

[29] Vgl. die Lehrerbefragung von Pedulla u. a. (2003, 5-9)
[30] 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)
[31] Vgl. die Zusammenstellung bei Brügelmann (2005b, 8).
[32] Generell kritisch zu den (Neben-)Wirkungen des High-Stakes Testing: AERA (2003).
[33] Vgl. Sacks (1999, 89-90).
[34] Vgl. dazu die Zusammenfassung und den Kommentar von Winter in der New York Times v. 28.12.2002 und ausführlicher unten (8.).
[35] Vgl. Linn (2000).

1.3 PISAnismus als Paradigma

Dass punktuelle Erhebungen mit standardisierten Instrumenten wie PISA&Co. inzwischen für viele im Bildungswesen[36] 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, Lehrerinnen oder Schüler. 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 Lehrerinnen 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 Lehrerinnen…".

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 Aussagen über Zusammenhänge in großen Populationen taugen, sind unangemessen, wenn es um Einzelfallentscheidungen geht. Auch hier kann die Medizin als Beispiel dienen[37]. 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 aber nur selten üblich (z. B. bei Parkinson-Patienten).

---

[36] Vgl. zur Diskussion über das GATS-Abkommen: Bulmahn (2002) und zu analogen Entwicklungen in anderen Bereichen öffentlicher Daseinsvorsorge die Beiträge in GEW (2000).

[37] Vgl. Albrecht (2005).

In großen Stichproben gleichen sich Messfehler in der Regel aus. Auf der Fallebene sind sie erheblich. Für Politiker mag es interessant sein, dass Buchbesitz in der Familie stärker mit der Leseleistung von SchülerInnen 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.

In der Pädagogik muss man deshalb ähnlich „experimentell" vorgehen wie in der Medizin. Testergebnisse 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[38] Das ist damit gemeint, dass Tests (nur) *heuristisch* sinnvoll sein können.

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.

[38] S. dazu unten 3.2

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

Diese Differenzen sind (neben dem Einfluss unterschiedlich gezogener Stichproben) nur erklärbar durch den Einsatz verschiedener Tests. Ähnlich verwirrend ist das Bild innerhalb von Deutschland, 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:

Studie	Alter	Schulanfang	Jahr	Anteil „Risiko"
IALS	16+[39]	1934-1982	1993	14 %
IEA-Sek	14	1983	1991	2 %
PISA-I	15	1991	2000	25 %
IGLU/PIRLS	9	1998	2001	10 %

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

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

Je nach Aufgabentyp und je nach Definition der Schwellenwerte ergeben sich ganz unterschiedliche Anteile. 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 in deutschen Grundschulen drei verschiedene Tests ein. Sie fand, dass selbst bei einer groben Aufteilung nach Leistung nur 41% der SchülerInnen in allen drei Tests in derselben Gruppe landeten und dass sogar beim gleichem Aufgabentyp (Textaufgaben) von vielen SchülerInnen 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 gehalten werden. PISA&Co dürfen nicht mehr über „die" Mathematikleistung oder gar „die" Lesekompetenz der deutschen SchülerInnen räsonnieren[40] – geschweige denn so tun, als ob sie mit ihrer Sondierung ausgewählter Wirkungsausschnitte „die Qualität" des vorhergegangenen Unterrichts oder gar den individuellen Lernstand einzelner SchülerInnen erfassen könnten

2. Alternative Perspektiven

Notwendig ist ein nach Stufen differenziertes System der Rechenschaftspflichten und Evaluationsformen[41]. 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.

---

[40] Vgl. auch die Relativierung der Rangverbesserung deutscher 15-Jähriger von PISA-2003 auf PISA-2006: „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 jüngste 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…)

[41] Vgl. den konkreten Vorschlag in Bartnitzky u. a. (1999) und die Übersicht über alternative Modelle in den OECD-Ländern bei Brügelmann (1980).

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). PISA&CO müssen deshalb inhaltlich und methodisch durch komplementäre Formen der Evaluation ergänzt werden. Damit wird es notwendig, die Evaluation vor Ort als eigenständige Aufgaben mit besonderen Anforderungen und Möglichkeiten zu fördern.

Im Folgenden stelle ich in aller Kürze zwei Beispiele vor:
- den reformpädagogischen Schulverbund "Blick über den Zaun" als Beispiel für eine Evaluation auf Schulebene und
- das Konzept „Pädagogische Leistungskultur" des Grundschulverbands als Beispiel für eine dialogische Form der Lernbeobachtung und Leistungsbeurteilung.

2.1 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, und Distanz ist nötig, um sich den erkannten Schwächen zu stellen. Diese Einsichten stehen hinter den Lernstandserhebungen und den Schulinspektionen. Sie machen externe 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 sie wird produktiver, wenn sie von Kontrolle abgekoppelt wird. Die Schulaufsicht sollte deshalb durch ein kollegiales Peer-Review ergänzt erden, das den Unterrichtsbetrieb nicht rechtlich kontrolliert, sondern fachliche Rückmeldung gibt. Die Schule sollten 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? Ohne Sanktionsbefugnisse kann die Inspektion nur durch ihre Kompetenz und durch den sozialen Druck, den die Berichterstattung bedeutet, wirken.

Ein gelungenes Beispiel für das Austarieren von externer Sicht und notwendiger Vertrautheit ist die Initiative "Blick über den Zaun". Dieser Verbund von inzwischen über 60 reformpädagogischen Schulen besteht seit 1989. In ihm schließen sich jeweils 7-10 Schulen (bewusst aus verschiedenen Schulformen) 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[42]. Zweimal im Jahr wird eine Schule von jeweils zwei VertreterInnen der anderen Schulen des gleichen Arbeitskreises besucht, die zwei bis drei Tage in der Schule mitleben. Die gastgebende Schule kann einen Beobachtungsauftrag formulieren, aber die Gäste sind frei, das zu beobachten, was ihnen wichtig erscheint. Sie nehmen am Unterricht teil, sie unterhalten sich mit KollegInnen, mit SchülerInnen und Vertretern der Eltern. In einer 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 ei-

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

nem 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 Kolleginnen 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 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 bzw. hilfreich?

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 Diskussion über die *Kriterien* für „guten Unterricht" ist zu trennen von der 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 SchülerInnen können eine Infor-

mationsquelle sein, um Schwächen auf die Spur zu kommen. Aber ertragreicher ist das in langjähriger Erfahrung fundierte Urteil der KollegInnen.

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. 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.

Diese Anforderungen gelten auch für die Leistungsbewertung auf der Schülerebene.

2.2 „Dialogische Lernbeobachtung" in einer pädagogischen Leistungskultur

Der Grundschulverband hat bereits vor der Veröffentlichung von PISA ein Konzept für ein umfassendes Evaluationssystem vorgeschlagen. In ihm werden nicht nur die Rechenschaftspflichten von Politik und Verwaltung, sondern auch die Evaluationsaufgaben der einzelnen Schule, der Lehrpersonen und der SchülerInnen 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; standardisierte Tests, die als Alternative angeboten werden, haben wiederum ihre eigenen Probleme[43]. Der Grundschulverband hat deshalb sein Konzept „Pädagogische Leistungskultur" entwickelt und darin folgende Kriterien für Aufgaben zur Lernbeobachtung formuliert. Sie sollten...

- der Lehrperson Informationen erbringen
  - nicht nur über aktuelle Einzelleistungen,
  - sondern auch über die Strategien („Tiefenstrukturen")
  - und über deren Entwicklung („Lerngeschichte");
- für die SchülerInnen auch inhaltlich eine produktive Lernsituation darstellen;

vor allem aber

- dialogisch angelegt sein als wechselseitige Verständigung über Lernziele, Bewertungskriterien und tatsächliche Leistungen und damit

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

- die Fähigkeit der Kinder zur Selbsteinschätzung eigener Arbeiten entwickeln.

Das diagnostische Repertoire von Lehrerinnen kann durch heuristisch eingesetzte Tests, durch Beobachtungsbögen und verschiedene Dokumentationsformen (wie Portfolios) erweitert und differenziert werden[44]. Das Konzept einer „pädagogischen Leistungskultur" fordert daneben aber verschiedene „Institutionen" im Unterrichtsalltag, die den SchülerInnen 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 Entwürfe vorgestellt, kommentiert und dann mit Hilfe Anderer überarbeitet werden[45];
- Rechendiskussionen in der Klasse, z. B. zum „harten Brocken des Tages"[46], 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[47].

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 bringen und mit ihnen in ein Gespräch über sinnvolle Lernwege zu kommen:

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

Hier zeigt sich derselbe Geist wie in der Peer-Evaluation auf Schulebene: Beratung statt 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.

3. Fazit

Damit schließt sich der Kreis: Evaluation bedeutet Macht und Abhängigkeit - deshalb darf man sie nicht Expert/inn/en überlassen, denn dieses Problem ist nicht technisch zu lösen.

---

[44] Vgl. die vielfältigen Vorschlage für die verschiedenen Lernbereiche in: Bartnitzky u. a., (2005-2007).
[45] Vgl. Spitta (1998) und ergänzende Hilfen für den Sprachunterricht bei Brinkmann/ Brügelmann (2005).
[46] Eingeführt von Erichson (2004) im Rechtschreibunterricht, vgl. für den Mathematikunterricht Küppers (2005); Sundermann/ Selter (2005).
[47] Vgl. die Beispiele für den Sachunterricht bei Schönknecht/ Klenk (2005, 22 ff.)

Aber Schule ist ein öffentlicher Raum und dies schließt ein, dass alle Beteiligten rechenschaftspflichtig sind. Dafür brauchen sie je nach Aufgabe unterschiedliche Verfahren – und Unterstützung. Ein demokratisches Verständnis von Evaluation fordert, die Betroffenen nicht durch externe Beurteilungen zu entmündigen, sondern ihre persönliche Evaluations- und Problemlösekompetenz zu stärken[48].

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

- Der PISA-Stil muss auf Systemevaluation begrenzt werden und darf nicht zum Paradigma für Evaluation 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 massive politische, wissenschaftliche und finanzielle Unterstützung für die Entwicklung der Evaluationskompetenz vor Ort.

[48] Vgl. Brügelmann (2007; 2008).

Quellen

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

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 ➔ http://www.agprim.uni-siegen.de/notengutachten.htm

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.

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).

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. (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.

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

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

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 [14.4.06].

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

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

Brügelmann, H. (2005a): Schule verstehen und gestalten – Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil (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) ➔ http://www.blickueberdenzaun.de/files/Bruegelmann-Evaluation.doc .

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 Export von Hochschul-Dienstleistungen. In: Frankfurter Rundschau v. 8.7.2002.

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 ➔ 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).

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.

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 ➔ www.pisa-kritik.de/files/Langfristige-Wirkung-geringer-Klassenfrequenzen.pdf [Abruf: 20.11.2007]

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

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.

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.

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

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

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

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).

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

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

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

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.

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

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

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

OECD (2007): Bildung auf einen Blick - OECD-Indikatoren 2007. Organisation for Economic Co-Operation and Development: Paris. ➔ http://www.bmbf.de/pub/zusammenfassung_eag.pdf 22.10.2007

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

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 [Zugriff 2.4.03]

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

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.

Rolff, H.-G. (1967): Sozialisation und Auslese durch die Schule. Gesellschaft und Erziehung, Bd. In Verbindung. Quelle und Meyer: Heidelberg.

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

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. Institut für Schulentwicklung, Goldbacherstr. 66: 88662 Überlingen.

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

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

Sundermann, B./ Selter, C. (2005): Mathematikleistungen feststellen, beurteilen und fördern. Beschreibung des Moduls 9 für das Projekt SINUS-Transfer Grundschule ➔ www.sinus-grundschule-de/ [Abruf: 13.1.06]

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

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

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

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).