PISA: Nachträge zu einer nicht geführten Debatte

Joachim Wuttke

1 *Ein Thema für die Mathematikdidaktik*

PISA, TIMSS, IGLU & Co, obwohl allenfalls marginal von fachdidaktischem Interesse getragen, müssen die Mathematikdidaktik aus einer ganzen Reihe von Gründen interessieren:

(1) Die Studien beanspruchen, den Erfolg von Mathematikunterricht zu *vermessen*.
(2) Sie wirken massiv zurück auf den Mathematik-Unterricht.
(3) Sie haben Bewegung in die Bildungspolitik gebracht und beeinflussen somit indirekt Bedingungen, unter denen in Zukunft Mathematik-Unterricht stattfinden wird.
(4) Sie unterminieren akademische Standards wie die vollständige Offenlegung von Instrumenten, Daten und Auswertemethoden und die Debattierbarkeit in Fachzeitschriften.
(5) Als akademischer Machtfaktor und Karrieretreibstoff werden sie auf Jahrzehnte hinaus beeinflussen, wie in Deutschland universitäre Pädagogik betrieben wird.
(6) Stoffdidaktisch gewendet, zeigen sie beispielhaft, wie wirkungsmächtig und wie interpretationsbedürftig Statistik sein kann.

In den bald acht Jahren, die seit dem initialen PISA-Schock vergangen sind, ist unüberschaubar viel zur Exegese der Testergebnisse gesagt worden, weitaus weniger aber über deren Zustandekommen. Manches Bedenkenswerte ist weithin unbeachtet geblieben, da der öffentliche Diskurs, bis in die Sprachregelungen hinein, von der Selbstdarstellung der Testveranstalter dominiert wird. Daher erscheint es nicht unangemessen, eine von PISA & Co besonders betroffene Fachgemeinschaft noch einmal eindringlich auf problematische Seiten testgetriebener Schulgestaltung hinzuweisen.

2 *Wissenschaft ohne Debatte?*

Trocken wie eine amtliche Statistik, inszeniert wie eine politische Kampagne, tritt PISA zugleich mit dem Anspruch auf, innovative Wissenschaft zu sein – ohne aber formalen Mindestanforderungen für wissenschaftliches Arbeiten zu genügen. Indem die Studie auf Pressekonferenzen statt in Fachzeitschriften veröffentlicht wurde, wurde sie als politisches Faktum etabliert, bevor außenstehende Wissenschaftler die Möglichkeit bekamen, Methoden und Ergebnisse zu prüfen. Viele eingesetzte Testaufgaben wurden erst Jahre später, manche bis heute nicht veröffentlicht. Die Technischen Berichte erscheinen erst viele Monate nach den Auswertungen und beschreiben die Datenreduktion nur unvollständig.

Der Medienerfolg von PISA hat die meistgenannten Ergebnisse längst zu zeitgeschichtlichen Fakten eigenen Rechts gemacht, die völlig unabhängig von ihrem Realitätsgehalt den Verlauf der politischen Debatte prägen. Die inhaltliche Auseinandersetzung mit der Studie wird dadurch erschwert. Während mathematisch-naturwissenschaftlich Geschulten die Grenzen einer solchen Statistik unmittelbar plausibel sind, neigen wissenschaftsferne Beobachter leicht dazu, die Studie für durch ihre Wirkungen legitimiert zu halten. Da sich die Politik festgelegt hat, PISA fortzuführen und durch weitere standardisierte Tests zu ergänzen, gibt es keinen Zweifel mehr, auf welcher Seite ein Bildungsforscher Karriere machen kann. Erziehungswissenschaftliche Zeitschriften nehmen keine PISA-kritischen Manuskripte an. Pädagogische Verlage lehnen PISA-kritische Titel ab, weil sie sich das Geschäft mit Testvorbereitungsbüchern nicht verderben wollen. Das PISA-Konsortium verweigert den Diskurs mit Kritikern und erklärt offen, diesen kein Forum bieten zu wollen.

Für Außenstehende ist es in dieser Lage nicht leicht, sich ein eigenes Urteil über die konzeptionellen und praktischen Mängel von PISA zu bilden. Viele Einwände sind nur an entlegenen Stellen publiziert worden. Während die Test-Szene international vernetzt ist, haben viele Kritiker lange Zeit nichts voneinander gewusst. Erst zwei Sammelbände (Jahnke/Meyerhöfer 2007; Hopmann/Brinek/Retzl 2007) haben deutlich gemacht,

aus wieviel verschiedenen Richtungen der Ansatz, die Durchführung und die gängige Interpretation von PISA in Frage gestellt werden.

3 *Evidenzbasierte Politik — politikverseuchte Evidenz*

Im März 2007 veranstaltete die neue deutsche PISA-Zentrale, das Deutsche Institut für Internationale Pädagogische Forschung (DIPF) in Frankfurt, eine Fachtagung unter dem Titel „Wissen für Handeln – Forschungsstrategien für eine evidenzbasierte Bildungspolitik". Die Eröffnungsrede des Stellvertreters der Bundesbildungsministerin machte unmissverständlich klar: PISA war nur der Anfang. Es wird dauerhaft ein „Bildungsmonitoring" installiert, das automatisch auf Erfolge und Fehlentwicklungen aufmerksam machen und einen von politischer Opportunität gelösten „Zwang zum Lernen" verankern soll.[1]
Die Forderung, dass professionelle Praxis auf der besten verfügbaren Evidenz basieren solle, wurde in den 1990er Jahren in der Medizin zu einem Formalismus zur Bewertung von wissenschaftlichen Studien verdichtet. Von dort hat sich die Idee der „evidenzbasierten Systemsteuerung" in andere Handlungsfelder ausgebreitet und die Politik erreicht. Heftigen Widerspruch hat dieser Ansatz auf dem Gebiet der Suchtprävention ausgelöst: Mit einseitiger Evidenz konfrontiert, weisen Fachleute darauf hin, dass weder Wissenschaft noch Politik werturteilsfrei sein können, dass die Berufung auf Evidenz ethische Entscheidungen verschleiert, und dass Evidenzbasierung von Politik unvermeidlich zur Politisierung von Evidenz führt.[2]
Aus genau diesen Gründen können auch PISA & Co keine politischen Handlungsanweisungen liefern. Das zeigt sich auf mindestens zwei Ebenen. Erstens auf der Ebene der Lernziele. Wenn im Gefolge von PISA beschlossen wurde, die Unterrichtsanstrengungen in Lesen und Mathematik zu verstärken, ist das keine Konsequenz aus einem empirischen Ergebnis, sondern aus dem zuvor gefassten Beschluss, vor allem Lesen und Mathematik zu testen.[3] Das Testergebnis hat keine andere Funktion, als öffentliche Unterstützung für den zuvor im Expertenzirkel eingefädelten Richtungswechsel zu besorgen. Die Unterstützung fällt umso stärker aus, je besser man das Testergebnisses als eine nationale Katastrophe inszeniert.
Zweitens zeigt sich die Unmöglichkeit evidenzbasierter Politik, wenn suggeriert wird, PISA könne Fragen entscheiden, die zuvor nur „ideologisch" diskutiert worden seien – wie zum Beispiel die umstrittenste Besonderheit der deutschen Schulstruktur, die frühe Aufteilung in verschiedene Schularten. Hier verdeckt die Berufung auf Evidenz den ethischen Standpunkt, Politik solle das Schulwesen so steuern, dass möglichst gute Leistungsmittelwerte und ein möglichst geringes soziales Leistungsgefälle herauskommen. Wo dieser Standpunkt absolut gesetzt wird, erweist er sich selbst als ideologisch. In der Politik geht es kaum je nur darum, einen fürs Gemeinwesen optimalen Zustand herzustellen; es geht immer auch darum, zwischen widerstreitenden Partikularinteressen auszugleichen und Lösungen zu finden, die auch dann noch funktionieren, wenn sie von einer Minderheit abgelehnt werden.
Die deutsche Schulstrukturdebatte ist latent unehrlich, weil oft mitgedacht und selten ausgesprochen wird, dass sich die Interessen einzelner Eltern und einzelner Schichten mit denen des Gemeinwesens nicht decken. Darunter leidet auch die PISA-Rezeption. Je nach unausgesprochenem ethischen Standpunkt und uneingestandenem Interesse werden die empirischen Daten so selektiv rezipiert, dass sie als Argumente pro oder contra Gesamtschule gebraucht werden können. Auch einige Projektverantwortliche, namentlich der OECD-Koordinator Andreas Schleicher, exponieren sich in dieser Weise. In einem solchen Umfeld ist es kaum mehr möglich, PISA-Ergebnisse zu zitieren, ohne dass die eine oder die andere Seite einen politischen Spin heraushört.

4 *PISA als Lebensunterhalt*

Unausgesprochene Eigeninteressen haben auch die Produzenten von PISA. PISA schafft Umsatz, Arbeitsplätze, Karrieremöglichkeiten, Ansehen und

[1] A. Storm, http://www.bmbf.de/pub/psts_20070328.pdf.
[2] A. Uhl: How to camouflage ethical questions in addiction research. In: J. Fountain, D. Korf, eds: The Social Meaning of Drugs. Research from Europe. Oxford: Radcliffe 2007. – B. Baumberg: Against evidence-based policy: over-claiming social research and undermining effective policy. Presented at the Social Policy Association conference, Edinburgh June 2008.
[3] Nähme man die Texte der sechzehn deutschen Landesverfassungen ernst, müsste man die Qualität von Schulunterricht primär an moralischen und sozialen und eher noch an ästhetischen als an kognitiven Erziehungszielen messen.

Macht.
An PISA sind über zweihundert Wissenschaftler beteiligt. Ein solcher Apparat entfaltet eine Eigendynamik. Warum, zum Beispiel, wird PISA alle drei Jahre wiederholt? In Anbetracht der Trägheit von Bildungssystemen ist es illusorisch, in so kurzer Zeit substantielle Änderungen zu beobachten. Zur Feststellung langfristiger Trends würde es genügen, alle fünf oder sieben Jahre einen Test durchzuführen. Der dreijährige Zyklus dient allein dazu, die Wissenschaftler beschäftigt zu halten. Auswertung einer Testrunde und Vorbereitung der nächsten dauern jeweils ungefähr eineinhalb Jahre. Wäre dazwischen eine längere Pause, würden die Teams auseinanderfallen.
Nach außen tritt die OECD als Autor von PISA auf. Tatsächlich leistet sie jedoch nur politische, administrative und redaktionelle Abstimmungsarbeiten. Die Ausgestaltung, Durchführung und Auswertung der Tests wurde ausgeschrieben und an ein Konsortium aus überwiegend privatwirtschaftlichen Instituten unter Führung des Australian Council of Educational Research (ACER) vergeben. Das Geschäftsmodell dieser weltweit tätigen Unternehmen besteht darin, Regierungen Tests zu verkaufen und dann den Markt für Testvorbereitungsmaterialien und -kurse zu erschließen. Von diesen Firmen darf man keine Informationen über Schwächen und Grenzen standardisierter Leistungstests erwarten.
Die empirische Bildungsforschung profitiert ebenso unmäßig wie einseitig von PISA.

> Man kann in diesem Bereich geradezu von einer Überhitzung der Konjunktur sprechen. Seitenweise werden Professuren in dieser Disziplin ausgeschrieben. Man fragt sich, was diese Armada in den nächsten Jahrzehnten ihrer Berufstätigkeit, so sie nicht über diesen Horizont hinauswachsen, alles wird messen.[4]

5 *Das Prinzip der schlechten Nachricht*

In der Berichterstattung über PISA und ähnliche Studien gibt es eine klare Tendenz, schlechte Nachrichten hervorzuheben und Ergebnisse in einem negativen Licht darzustellen.
Der initiale PISA-Schock von 2001 beruhte wesentlich darauf, dass sich Deutschland in allen drei Teiltests unter dem Mittelwert von 500 Punkten fand. Dieser Mittelwert wird allerdings in sehr eigenwilliger Weise unter Gleichgewichtung aller OECD-Staaten berechnet: isländische Schüler werden 800-mal stärker berücksichtigt als US-amerikanische. Wenn man das korrigiert, sinkt der Mittelwert auf rund 490. Diese Korrektur, die, verglichen mit anderen Ungenauigkeiten des Messverfahrens nicht einmal besonders schwerwiegend ist, genügt bereits, damit sich die öffentliche Wahrnehmung, Deutschland habe in PISA schlecht abgeschnitten, als durch die Daten nicht gedeckt erweist: Deutschland hat, bezogen auf die Gesamtpopulation der OECD, niemals signifikant unterdurchschnittlich abgeschnitten, und erzielt seit 2003 konsistent überdurchschnittliche Ergebnisse.
Tief festgesetzt hat sich das Gerücht, PISA habe gezeigt, dass ein knappes Viertel aller 15-Jährigen nicht richtig lesen und rechnen könne. Hintergrund ist ein Versuch des Konsortiums, PISA-Punkte inhaltlich zu interpretieren. Dazu wird die Punkteskala in sechs sogenannte Kompetenzstufen und eine darunter liegende Inkompetenzstufe eingeteilt. Da Lesen und Rechnen als selbstverständlich vorausgesetzt wird, müssen Schüler schon, um Stufe 1 zu erreichen, Aufgaben lösen, die mehr als nur diese Grundfertigkeiten erfordern. Dementsprechend redet der internationale Bericht von »Risiko« nur mit Bezug auf die Inkompetenzstufe. Hingegen wird im deutschen Bericht die Stufe 1 zur „Risikogruppe“ hinzugezählt. Die Öffentlichkeit wurde nicht darüber informiert, dass sich der hohe Anteil an Risikoschülern kein empirisches Ergebnis ist, sondern sich primär aus der mathematischen Konstruktion der Kompetenzstufen ergibt. Die erbrachten Leistungen werden nämlich gerade so in Punkte umgerechnet, dass sich OECD-weit knapp 8% in der Inkompetenzstufe und weitere 12–13 % in Stufe 1 befinden – völlig unabhängig davon, wie gut oder schlecht die Schüler tatsächlich mit dem Test zurechtgekommen sind.
Die schlechte Nachricht besteht also alleine darin, dass sich in Deutschland eher 22 % statt der OECD-weiten 21 % in Stufe 0 oder 1 befinden — und das liegt nicht am unterdurchschnittlichen Können der deutschen Schüler, sondern an der

[4] T. Jahnke, Die PISA-Unternehmer. Forschung & Lehre, 15, 26, 2008.
[5] Kap. 2 in Wuttke: Die Insignifikanz signifikanter Unterschiede: Der Genauigkeitsanspruch von PISA ist illusorisch. In Jahnke/Meyerhöfer 2007. Englische Kurzfassung (Uncertainties and Bias in PISA) in Hopmann/Brinek/Retzl 2007. Beide Aufsätze auch unter http://www.messen-und-deuten.de/pisa.

Schlafende Robbe
Eine Robbe muss atmen, auch wenn sie schläft. Martin hat eine Robbe eine Stunde lang beobachtet. Zu Beginn seiner Beobachtung befand sich die Robbe an der Wasseroberfläche und holte Atem. Anschließend tauchte sie zum Meeresboden und begann zu schlafen. Innerhalb von 8 Minuten trieb sie langsam zurück an die Oberfläche und holte Atem. Drei Minuten später war sie wieder auf dem Meeresboden, und der ganze Prozess fing von vorne an.

Nach einer Stunde war die Robbe:
(a) auf dem Meeresboden
(b) auf dem Weg nach oben
(c) beim Atemholen
(d) auf dem Weg nach unten

Abbildung 1. *Beispielaufgabe aus dem Feldtest zu PISA 2000 – Bereich: Mathematische Grundbildung*

überdurchschnittlichen Gründlichkeit der deutschen Stichprobenziehung.[5]
Nirgends sonst entscheidet die soziale Herkunft so stark über den Testerfolg wie in Deutschland, lautet ein weiteres PISA-Ergebnis, das tief ins öffentliche Bewusstsein eingesunken ist. Es stammt aus PISA 2000 – und hat sich in den Folgerunden nicht bestätigt. In PISA 2003 musste die deutsche Projektleitung sogar den Indikator wechseln (Korrelationskoeffizient statt Gradient), um wenigstens einen signifikant überdurchschnittlichen Zusammenhang vorweisen zu können. Zugleich bestreiten deutsche PISA-Autoren, dass der von der OECD verwendete Herkunftsindex sachgerecht ist.
Der Medienerfolg von PISA beruht übrigens auch in anderen Staaten auf dem Prinzip der schlechten Nachricht. In Finnland war man über die große Leistungsdifferenz zwischen Mädchen und Jungen schockiert. Erst als die Pilgerfahrten deutscher Schulpolitiker einsetzten, gewöhnte man sich an die Rolle des Testsiegers.

6 Was PISA testet

PISA orientiert sich nicht an der Schnittmenge nationaler Curricula, sondern postuliert einen eigenen Bildungsbegriff, der auf Englisch als literacy bezeichnet wird: „das Wissen, die Fähigkeiten, die Kompetenzen, … die relevant sind für persönliches, soziales und ökonomisches Wohlergehen."[6] „Hinter diesem Konzept verbirgt sich der Anspruch, über die Messung von Schulwissen hinauszugehen und die Fähigkeit zu erfassen, bereichsspezifisches Wissen und bereichsspezifische Fertigkeiten zur Bewältigung von authentischen Problemen einzusetzen."[7]
Ob dieser Anspruch erfüllt wird, kann sich einzig und allein an den real gestellten Testaufgaben erweisen. Aus Platzgründen kann hier nur ein einziges, kurzes Beispiel diskutiert werden: die Aufgabe „Schlafende Robbe" (Abb. 1). Ich empfehle, sich zunächst selbst mit dieser Aufgabe auseinanderzusetzen und dann erst hier weiterzulesen
…
Der letzte Halbsatz des Einleitungstextes ist schlicht falsch: Zu Beginn von Martins Beobachtung befindet sich die Robbe an der Wasseroberfläche. Somit kann „der ganze Prozess" nicht zu einem Zeitpunkt von vorne anfangen, zu dem sich die Robbe am Meeresboden befindet. Interessanterweise steht dieser falsche Halbsatz weder in der englischen noch in der französischen Originalfassung, wo es stattdessen heißt: „Martin bemerkte, dass der ganze Prozess sehr regelmäßig war." Außer in Deutschland ist die falsche Aufgabenfassung jedoch in mindestens einem weiteren Staat, Argentinien, eingesetzt worden. Demnach ist der falsche Halbsatz nicht erst durch den deutschen Übersetzer hinzugefügt worden; es ist vielmehr zu vermuten, dass er sich in der ursprünglichen internationalen Vorlage befand, dass diese Vorlage später (vor oder nach Einsatz der Aufgabe?) korrigiert wurde, und dass diese Korrektur in Deutschland und Argentinien übersehen wurde. Man kann nur spekulieren, ob diese Unstimmigkeit ursächlich dafür war, dass die Aufgabe

[6] Measuring Student Knowledge and Skills. A New Framework for Assessment. Paris: OECD 1999. Insbes. S. 11.
[7] http://www.mpib-berlin.mpg.de/pisa/intgrundkonzeption.htm.

„Schlafende Robbe“ es nicht aus dem Feldtest in den Haupttest geschafft hat. Übernahmekriterium war allein das „psychometrische“, d. h. rein statistische „Funktionieren“ der Aufgabe. Eine inhaltliche Kontrolle fand an dieser Stelle nicht mehr statt. Es ist keineswegs ausgeschlossen, dass auch die im Haupttest eingesetzten Aufgaben Übermittlungsfehler enthalten.
Auch ohne den falschen Halbsatz vermag die Aufgabe nicht zu überzeugen: Ist dies ein „authentisches“ Problem? Um die Aufgabe eindeutig lösen zu können, muss man, gestützt auf Einleitungstext und physiologische Selbsterfahrung, die Annahme treffen, dass die Robbe zum Atemholen weniger als zwei Minuten braucht. Schülern eine solche außermathematische Eigenleistung abzuverlangen, mag man grundsätzlich begrüßen – nur sollte man dann auch strenge Maßstäbe an die Stimmigkeit des außermathematischen Kontextes anlegen. Es ist unwahrscheinlich, dass eine Robbe den beschriebenen Schlafrhythmus mit der zur Aufgabenlösung erforderlichen Präzision befolgt, und es ist vollkommen unersichtlich, wie Martin eine Robbe, die drei Minuten braucht, um bis zum Meeresgrund zu tauchen, kontinuierlich im Blick behalten kann.
Viele andere Aufgaben sind im gleichen Stil verfasst und leiden unter denselben Mängeln: Der Versuch, schülernahe, authentische Kontexte zu konstruieren, scheitert fast immer; den Schülern wird nicht mathematische Modellierung abverlangt, sondern die Decodierung von Texten, die keine realen Situationen, sondern unglaubwürdige mathematische Modelle beschreiben.[8] Wenn solche Aufgaben als vorbildlich hingestellt werden, kann man sich ausmalen, wie schlecht erst die epigonalen Aufgaben sind, die im Gefolge von PISA und zur Vorbereitung auf die nächsten Vergleichstests en masse produziert werden.
Während sich viele Mathematikaufgaben als Leseaufgaben mit leichten Rechenanforderungen erweisen, besteht die „Text“grundlage mancher Leseaufgaben aus Tabellen und Graphiken. Kein Wunder, dass die Ergebnisse aus den Teiltests Lesen, Mathematik, Naturwissenschaften miteinander hoch korreliert sind. Von daher könnte man die drei Teilwertungen eigentlich in eine einzige Gesamtwertung zusammenziehen. Das aber wird in den offiziellen PISA-Auswertungen strikt vermieden, denn einen solchen Generalfaktor kognitiver Fähigkeiten könnte man allzu leicht als *Intelligenz* interpretieren,[9] was ein Tabu der Bildungsforschung verletzen würde. Auch dürfte es die politische Unterstützung gefährden, wenn sich herumspricht, dass PISA im Kern ein sprachlastiger Intelligenztest ist.
Bei der Interpretation der Testergebnisse ist außerdem zu berücksichtigen, dass PISA unter starkem Zeitdruck stattfindet: es wird nicht nur Textverständnis, sondern auch Lesetempo getestet. Insbesondere die Lesefertigkeit mancher Immigranten dürfte deshalb massiv unterschätzt werden.

7 Sprachliche und kulturelle Verzerrungen

Lehrer werden immer wieder überrascht, wie Schüler vermeintlich eindeutig formulierte Aufgaben missverstehen. Änderungen im Wortlaut, die man für völlig nebensächlich gehalten hätte, können Schülern unüberwindliche Schwierigkeiten bereiten. Schon von daher erscheint die Annahme, man könne in sprachlich und kulturell neutraler Weise einen weltweiten Lesetest erstellen, als unbegründet optimistisch.
In PISA wird nicht einmal der Versuch unternommen, diese Annahme a posteriori zu validieren; empirische Untersuchungen werden sogar gezielt verhindert: Die Angabe der Testsprache fehlt im internationalen Datensatz. Ein Grund für diese Auslassung wird nicht genannt: Ist es der Druck von Regierungen, die Vergleiche zwischen Sprachgruppen verhindern wollen? Oder der zum Dogma erhobene Glaube, die Testsprache habe keinen Einfluss auf die Testergebnisse?
Die Übersetzung und Endredaktion des seit PISA 2000 eingesetzten Grundbestandes an Testaufgaben ist unter starkem Zeitdruck erfolgt; in etlichen Staaten wurde nur die englische, nicht aber die gleichberechtigte französische Ausgangsversion berücksichtigt. So ist es nicht überraschend, dass es zu Fehlern und Ungereimtheiten gekommen ist.
Schwerwiegender ist jedoch das grundsätzliche Problem, dass es keine Methode gibt, zu messen, ob und wie stark sich die Schwierigkeit einer Aufgabe durch Übersetzung verändert. Eine Kalibrierung auf Populationsmittelwerte verbietet sich, da man ja gerade Unterschiede zwischen Popula-

[8] Besonders gründlich hat das W. Meyerhöfer in seiner Dissertation untersucht: Tests im Test: Das Beispiel PISA. Leverkusen: Budrich 2005.
[9] Heiner Rindermann, Psychol. Rundsch. 57, 69, 2006, Eur. J. Personality 21, 667, 2007.

tionen feststellen möchte. Eine Kalibrierung auf sprachunabhängige Intelligenztests verbietet sich, da man behauptet, von Intelligenz verschiedene Fähigkeiten zu messen. Man könnte immerhin zu einer Abschätzung der durch das Übersetzungsproblem verursachten Unsicherheit gelangen, wenn man verschiedene, unabhängig voneinander erstellte Übersetzungen einsetzen würde, aber eine solche Kontrolle wurde in PISA nicht unternommen.

Kaum trennbar vom Übersetzungsproblem, aber noch fundamentaler ist die Frage, ob nicht unterschiedliche Sprachen an sich unterschiedliche Anforderungen an das Textverständnis, zumal unter Zeitdruck, stellen. Statistisch nachweisbar sind Unterschiede in der Textlänge: die Einleitungstexte der PISA-Aufgaben umfassen in der französischen Version 12 % mehr Wörter und 19 % mehr Buchstaben als in der englischen. Das PISA-Konsortium hat lediglich den direkten Effekt der unterschiedlichen Textlänge auf die Häufigkeit richtiger Lösungen untersucht und für gering befunden; tatsächlich dürfte es eine wesentlich größere Verzerrung dadurch geben, dass für längere Texte mehr Zeit benötigt wird, die dann am Ende des Tests, bei anderen Aufgaben, fehlt.

Kulturell bedingte Verzerrungen sind schon deshalb wahrscheinlich, weil die Mehrheit der Testaufgaben aus einigen wenigen Staaten stammt und weil sämtliche Aufgabentexte in Australien von hauptberuflichen Testaufgabenredakteuren homogenisiert wurden. Die PISA-Aufgaben sind in einem Stil verfasst, an den Schüler in Australien und Neuseeland, den USA und Kanada von der Grundschule an gewöhnt sind. In vielen europäischen Staaten ist dieser Stil ungewohnt.

Dass es durch die unterschiedliche Vertrautheit mit dem Aufgabenformat zu quantitativ bedeutsamen Verzerrungen kommt, lässt sich eindeutig bei den Multiple-Choice-Aufgaben nachweisen. Dieser Aufgabentyp wird bei gut einem Viertel aller PISA-Aufgaben eingesetzt. Von vier oder fünf Antwortalternativen wird genau eine als richtig gewertet. In Australien ist diese Regel eine so selbstverständliche Gewohnheit, dass es die PISA-Leitung nicht einmal für nötig gehalten hat, die Teilnehmer nachhaltig darauf hinzuweisen. Im deutschen Sprachraum aber haben bei manchen Aufgaben bis zu 10% aller Probanden mehr als eine Antwort angekreuzt – was als falsch gewertet wurde, obwohl sich diese Schüler unter Umständen wesentlich tiefere Gedanken gemacht haben als die, die wussten, dass immer nur eine Antwort richtig sein kann.

Aber auch die Aufgaben mit offenem Antwortformat stehen auf dem Boden einer ganz bestimmten Prüfungstradition. Einerseits wird erhebliche Reflexionsfähigkeit gefordert, wenn ein Lesetext auf verschiedenen Ebenen beleuchtet wird. Andererseits zeigen die Korrekturhinweise, dass die Anforderungen eher formaler als inhaltlicher Art sind und schematisch geübt werden können: die Schüler sollen Vorgaben aus dem Textmaterial aufgreifen, ohne dabei die Originalformulierungen zu verwenden.

8 Illusorische Genauigkeit

Ein weiteres fundamentales Problem von PISA besteht darin, dass die Unterschiede zwischen einzelnen Schülern um ein Vielfaches größer sind als die Unterschiede zwischen ganzen Staaten. In jedem einzelnen Staat schneiden rund 30 % aller Teilnehmer um mehr als 100 Punkte schlechter oder besser als der Durchschnitt ab. Unterschiede zwischen einigermaßen vergleichbaren Staaten liegen hingegen im einstelligen oder niedrigen zweistelligen Bereich. Deshalb sind Staatenvergleiche fehleranfällig: Schon kleine Uneinheitlichkeiten zum Beispiel bei der Stichprobenziehung können die nationalen Mittelwerte so weit verzerren, dass sich Ranglisten ändern.

PISA ist eine statistische Untersuchung an einer zufällig gezogenen Stichprobe und gehorcht insofern denselben mathematischen Gesetzen wie eine Meinungsumfrage. Jede solche Untersuchung ist mit zwei Arten von Ungenauigkeit behaftet: stochastisch und systematisch. Stochastische Ungenauigkeit hat ihre Ursache in der Zufälligkeit der Stichprobenziehung. Systematische Ungenauigkeit kann verschiedenste Ursachen haben, von nichtzufälligen Verzerrungen bei der Stichprobenziehung bis hin zu Teilnehmern, die nicht kooperieren. Die stochastische Ungenauigkeit kann man reduzieren, indem man die Stichprobe vergrößert. Ab einem gewissen Punkt überwiegt jedoch die systematische Ungenauigkeit, und eine weitere Vergrößerung der Stichprobe bringt keinen nennenswerten Vorteil mehr. Aus diesem Grund beschränken sich Meinungsumfragen auf 1000 bis 2000 Teilnehmer.

Zu PISA werden hingegen pro Staat mindestens 5000 Schüler herangezogen. Nur dadurch gelingt es, die stochastischen Fehler so klein zu machen, dass Staaten-Unterschiede von 9 oder 10 Punkten für statistisch „signifikant“ erklärt werden können. Die Möglichkeit systematischer Ungenauigkeiten wird dabei vollständig ausgeblendet. Tatsächlich aber gibt es in PISA eine ganze Reihe

von Regelverstößen, Verzerrungen und Unsicherheiten;[10] gar nicht zu reden, von der unterschiedlichen Motivation der Testteilnehmer.[11] Deshalb sind die Signifikanz-Urteile in den offiziellen Berichten illusorisch, die übertrieben großen Stichproben sind ökonomisch nicht zu rechtfertigen, und viele statistische Ergebnisse sind kaum mehr als Zufallszahlen.

9 Das Vorbild Finnland

Hätte Polen oder Portugal die Ranglisten angeführt, wäre PISA in Deutschland nicht ernst genommen und schnell vergessen worden. Ein skandinavischer Testsieger hingegen war überaus plausibel und half, die Studie, die ein solches Ergebnis geliefert hatte, zu beglaubigen. Nur zu gerne hörte man, dass Finnland unserer Bildungspolitik als Vorbild dienen könne. Nichtsdestoweniger ist diese Schlussfolgerung fehlerhaft; sie zeigt beispielhaft, wie die öffentliche PISA-Rezeption auf selektiver Wahrnehmung beruht.
Der statistischen Methodenlehre zufolge sollte man vor Beginn einer statistischen Untersuchung Arbeitshypothesen formulieren, um diese am Ende, wenn es die Daten erlauben, zu bestätigen oder zu widerlegen. Hingegen gilt es als problematisch, nach abgeschlossener Untersuchung einzelne auffällige Werte aus einem umfangreichen Datensatz herauszugreifen und weitreichende Interpretationen daran zu knüpfen: groß ist die Gefahr, Artefakten des Messverfahrens aufzusitzen.
Vor PISA hätten viele in Deutschland für eine sinnvolle Arbeitshypothese gehalten: dass die skandinavischen Staaten besonders gut abschneiden würden. Auswertung: Island, Dänemark und Norwegen schneiden mal besser, mal schlechter als Deutschland ab; Schweden liegt konsistent, aber nicht weit über 500; Finnland mit Werten über 535 bildet eine Ausnahme. Ergebnis: die Hypothese hätte verworfen werden müssen.
Die finnischen Daten erwiesen sich jedoch als unwiderstehlich. Auch ohne theoretische Grundlage wurde der Testsieger zum Vorbild erklärt. Unzählige Delegationen deutscher Schulpolitiker pilgerten nach Finnland, ließen sich von den hervorragend ausgestatteten, autonomen und offenkundig menschenfreundlichen finnischen Gesamtschulen begeistern und konnten bei ihrer Rückkehr ein halbes Dutzend Gründe für die Überlegenheit des finnischen Schulsystems aufzählen. Soweit bekannt, reiste keine einzige dieser Delegationen anschließend nach Norwegen, um zu erforschen, weshalb ganz ähnliche materielle, personelle, organisatorische und allgemein zivilisatorische Voraussetzungen dort nur zu mittelmäßigen Testleistungen führen.
In der Fixierung auf das Vorbild Finnland zeigt sich das genaue Gegenteil von „evidenzbasierter Politik“: von Wunschdenken gesteuertes Ausblenden empirischer Evidenz. Nicht nur das mäßige Abschneiden der anderen skandinavischen Staaten wird ignoriert, sondern auch das der schwedischsprachigen Minderheit in Finnland, die je nach Teiltest zwar manchmal der finnischsprachigen Mehrheit, manchmal aber auch Schweden nahekommt. Weiterhin wird ausgeblendet, dass das hervorragende Abschneiden Finnlands zum guten Teil daran liegt, dass es dort extrem wenige Einwanderer gibt. Beschränkt man den internationalen Vergleich auf im jeweiligen Land geborene Schüler, wird Finnland in manchen Teiltests überholt von flämisch Belgien, den Niederlanden und Bayern. Weiterhin müsste man berücksichtigen, dass Finnland Legastheniker von PISA ausgeschlossen hat.
In Anbetracht des halben Dutzends guter Gründe, die als Erklärung für den Testsieg Finnlands plausibel schienen, sollte der Forschungsauftrag für unsere nächsten Delegationen lauten: was dort falsch gemacht wird, wenn die Finnen trotz bester Voraussetzungen keine besseren Ergebnisse erzielen als die Bayern?

10 Folgen

PISA hat Bewegung in die deutsche Schulpolitik gebracht und den politischen Parteien geholfen, von angestammten Positionen abzurücken. Die Linke akzeptiert zentrale Abschlussprüfungen, die Rechte den Ausbau von Krippen, Vorschulen und Ganztagsbetreuung, und sogar der Streit um die Schulstruktur hat mit der Einigung auf ein zweigliedriges System in etlichen Bundesländern eine produktive Wendung genommen. Das sind große Erfolge: die Inszenierung von PISA als nationale Katastrophe hat ein „window of opportunity“ geschaffen und überfällige Entscheidun-

[10] Wuttke, a. a. O.
[11] In Seoul wird vor Beginn des Tests die Nationalhymne gesungen. In Hamburg geben die ersten Schüler nach zwei Minuten ab.

gen legitimiert.[12] Damit ist freilich nichts über den wissenschaftlichen Wert der Studie gesagt, nichts über die Wünschbarkeit weiterer Testdurchgänge, nichts über Risiken und Nebenwirkungen.

Schlussfolgerungen, die nicht schon in der Luft liegen, kann man nicht so leicht aus den PISA-Daten ableiten. Der Volkswirt Ludger Wößmann hat es immerhin versucht – und ist zu dem Ergebnis gekommen, dass es pure Verschwendung sei, mehr Lehrer einzustellen, da die Testleistungen nahezu unabhängig von der Klassengröße sind. Den eklatanten Widerspruch zur Lebenserfahrung, dass bei mehr als 20 bis 25 Schülern ein deutlicher Umschlag der Quantität in verringerte Unterrichtsqualität stattfindet, erkläre ich mir damit, dass PISA primär nicht Unterrichtserfolg, sondern Intelligenz und Schnelligkeit misst, was noch dadurch verstärkt wird, dass PISA Schüler gegen Ende der Pubertät testet, also nach einer zwei- bis dreijährigen Phase, in der Unterricht besonders ineffizient ist.

Da PISA keine spezifischen Erkenntnisse liefert, die Unterricht zu verbessern helfen, bleibt als hauptsächliche Konsequenz, die aus diesem Test gezogen wird: noch mehr zu testen. Durch die Fortführung von PISA & Co, durch zentrale Abschlussprüfungen, durch schulweite, landesweite und bundesweite Vergleicharbeiten. Um Leistungen punktemäßig vergleichen zu können, braucht es „Bildungsstandards“;[13] um deren Einhaltung zu überprüfen, müssen sie durch standardisierte Testaufgaben konkretisiert werden: ein Zirkelschluss, der, einmal politisch gewollt, alsbald institutionell verfestigt wurde. Ohne nennenswerte Diskussion ist so der Übergang von der Systembeobachtung zur Individualdiagnose und von beschreibenden zu sanktionsbewehrten Tests in die Wege geleitet worden. Aus Amerika weiß man, welche zerstörerischen Nebenwirkungen das haben kann: Verengung des Unterrichts auf Testvorbereitung, Verfälschung der Testergebnisse durch Betrug auf allen Ebenen.[14]

Gesellschaftlicher Widerstand gegen eine zunehmende Testorientierung ist nicht zu erwarten, passt sie doch zu einem säkularen Trend, ohne den PISA nicht zum Ereignis hätte werden können: die saturierte Epoche mit den Eckdaten ’68 und ’89 ist zu Ende gegangen; nach leistungsfeindlichen Übertreibungen in Schulgesetzen und Unterrichtspraxis schwingt das Pendel längst in die Gegenrichtung; die ökonomisch verunsicherte Mittelschicht ist empfänglich für die Forderung, dass Schule fit für den globalen Wettbewerb machen soll, und für alle damit verbundenen Kurzschlüsse sowieso.[15]

Unter den Fachdidaktiken ist die Mathematik ganz speziell betroffen, wenn Vorgesetzte, Eltern und Schüler die Lehrer unter Druck setzen, den Unterricht auf die Einübung von Aufgaben im PISA-Stil zu konzentrieren: Aus Schülersicht sind das durchweg *Textaufgaben*, ergo besonders schwere Aufgaben. Bisher gilt als ausgemacht, dass es die Versetzung nicht gefährdet, wenn man mit solchen Aufgaben nicht zurechtkommt, ansonsten aber einigermaßen fleißig ist und die zuletzt eingeübten Rechentechniken in bekanntem Kontext anwenden kann. Wenn nun ein Standard fordert, eine breite Mehrheit der Schülerschaft solle in der Lage sein, Textaufgaben zu lösen, dann erfordert das entweder eine radikale Reduktion mathematischer Inhalte (mit unabsehbarer Schädigung derjenigen, die Mathematik im Studium brauchen werden), oder es werden in stupider Weise ganz bestimmte Aufgabenmuster trainiert werden. Was ich aus einer nordrhein-westfälischen Grundschule höre, deutet in letztere Richtung und erinnert an die alte Polemik, Dreisatz an der Gesamtschule laufe darauf hinaus, das Wort Kartoffel zu unterstreichen: die Grundschüler sollen lernen, Textaufgaben, die sie noch gar nicht zu Ende rechnen können, mit dem Buntstift vorzustrukturieren.

[12] K. J. Tillmann et al., PISA als bildungspolitisches Ereignis. Oder: Wie weit trägt das Konzept der „evaluationsbasierten Steuerung“? In: T. Brüsemeister, K.-D. Eubel (Hrsg.): Evaluation, Wissen und Nichtwissen. Wiesbaden: VS Verlag für Sozialwissenschaften 2008. – B. Payk: Deutsche Schulpolitik nach dem PISA-Schock: Wie die Bundesländer auf die Legitimationskrise des Schulsystems reagieren. Hamburg: Dr. Kovač 2009.

[13] W. Herzog: Bildungsstandards: pädagogische oder politische Notwendigkeit. Vortrag, Bern 2006. http://cmslive1.unibe.ch/lenya/kwb/live/3/33/Erfolge/Diplomfeier/Beitrag_Herzog.pdf. – H.–D. Sill: PISA und die Bildungsstandards. In Jahnke/Meyerhöfer 2007. – T. Jahnke: Deutsche PISA–Folgen. In Hopmann/Brinek/Retzl 2007.

[14] P. Sacks: Standardized Minds.The high price of America’s testing culture and what we can do to change it. Cambridge Mass.: Perseus Publishing 1999. – A. Kohn: The Case Against Standardized Testing. Raising the Scores, Ruining the Schools. Portsmouth NH: Heinemann 2000. – S. L. Nichols, D. Berliner: Collateral Damage. How High-Stakes Testing Corrupts America’s Schools. Cambridge Mass.: Harvard Education Press 2007. – G. Lind: Jenseits von PISA: Für eine neue Evaluationskultur. In Inst. f. Schulentwicklung Schwäb. Gmünd: Standards, Evaluation und neue Methoden. Baltmannsweiler: Schneider Verlag Hohengehren 2004. http://www.uni-konstanz.de/ag-moral/pdf/Lind-2003_evaluationskultur.pdf.

[15] K. P. Liessmann, Theorie der Unbildung. Die Irrtümer der Wissensgesellschaft. München: Piper 2008.

Ob das die mathematische Grundbildung unserer Bevölkerung nachhaltig heben wird? Oder den Fehler von New Math wiederholt, verfrüht Abstraktion erzwingen zu wollen? Mir scheint, da besteht noch erheblicher Diskussionsbedarf.

Literatur

S. T. Hopmann, G. Brinek, M. Retzl (Hrsg.): PISA zufolge PISA – PISA According to PISA. Hält PISA, was es verspricht? Does PISA Keep What It Promises? Reihe Schulpädagogik und Pädagogische Psychologie, Bd.6. Wien: Lit-Verlag 2007.

T. Jahnke, W. Meyerhöfer (Hrsg.): PISA & Co – Kritik eines Programms. Hildesheim: Franzbecker, 2. Auflage 2007.