Manuskript
Stand: 09. April 2003

Jenseits von PISA –
Für eine neue Evaluationskultur[1]

Georg Lind

PISA hat eine breite öffentliche Diskussion angestoßen und viele Fragen aufgeworfen. In Frage steht vor allem, ob PISA und ähnliche Tests (IGLU, BIJU etc.) tatsächlich helfen, unser Bildungssystem zu verbessern, oder ob wir am Ende paradoxerweise das Gegenteil dessen erreichen, was wir beabsichtigen. Die Idee, mit Hilfe von Schulleistungstests eine Qualitätsentwicklung in Gang zu setzen, ist auf den ersten Blick bestechend. Ob sie in der Realität funktioniert, hängt aber offenbar davon ab, *wie* die Tests eingesetzt werden: Zur Evaluierung von Personen oder aber von Methoden. Nur im letzteren Fall ist Optimismus angebracht.

Hermann Lange, Leiter der für die PISA-Studie zuständigen Amtschefkommission der Kultusministerkonferenz (KMK) warnte schon 1999 vor falscher Testeuphorie: “Wachsamkeit ist unerlässlich”. Wie recht Lange und andere mit ihrer Warnung hatten, belegt die in 2002 erschienene Studie von Audrey Amrein und David Berliner, zwei renommierten US-Bildungsforschern, in der Zeitschrift *Education Policy Analysis Archives* über die Auswirkungen der Einführung von Schulleistungstests in vielen US-Bundesstaaten, die zum Teil schon über 20 Jahre zurückliegen. Das Ergebnis ist weitgehend negativ: Nach Einführung solcher Tests kam es überwiegend zu sinkenden Schulleistungen! Zudem stieg die Zahl der Sitzenbleiber und der vorzeitigen Schulabbrecher an und der Lehrplan verengte sich immer mehr auf die durch Tests vorgegebenen Inhalte und Lernformen. Schlimmer noch: Amrein und Berliner und andere Studien belegen, dass vergleichende Schulleistungstests soziale Benachteiligungen im Bildungssystem nicht bloß aufdecken, sondern verstärken oder gar selbst verursachen.

---

[1] Inzwischen erschienen in: Pädagogische Hochschule Schwäbisch Gmünd, Hrsg., (2004), Evaluation, Standards..... Baltmannsweiler: Schneider Verlag Hohengehren.

Die US-Erfahrungen bestätigen eigentlich nur, was Bildungsforscher schon lange wissen: Der bloße Einsatz von Tests und komplexen Auswertungsverfahren garantiert noch keine Qualitätsverbesserung, auch wenn noch so viele Schülerinnen und Schüler noch so oft getestet werden. Jede Evaluation muss selbst strengen Kriterien genügen. Über diese Kriterien herrscht unter Evaluationsexperten eigentlich weitgehend Konsens.

1. Erstens muss Evaluation – am besten als Teil einer Maßnahme zur Qualitätsentwicklung der Schule – so angelegt sein, dass Fragen nach der Verursachung eines Defizits oder zur Effektivität einer Maßnahme eindeutig beantwortet werden können. Zu den Minimalanforderungen an das Evaluationsdesign gehört daher die Durchführung von Vor- und Nachtests sowie – zur Kontrolle – möglichst zeitnahen Vergleichserhebungen und – um die Nachhaltigkeit zu prüfen – *Follow-up*-Studien.

Schon dieses Kriterium trifft auf PISA nicht zu. PISA ist nicht als Evaluationsstudie angelegt, sondern als eine Momentaufnahme und erlaubt daher keine Aussagen über die Ursachen von (vermeintlichen) Defiziten unserer Schulen oder über die Effektivität von bestimmten Aspekten unseres Bildungssystems. Bei reinen Momentaufnahmen wie PISA hängt irgendwie alles mit allem zusammen und welche Kausalitäten angenommen werden ist reine Glaubenssache. So gehört schon viel Glaube dazu, den Einfluss der Beschulungsdauer auf die PISA-Unterschiede zu ignorieren. Während z.B. in Ländern mit hohen PISA-Werten wie Australien und Neuseeland 90 Prozent aller 15-jährigen die 10. Klasse besuchen, sind es in Deutschland nur 24 Prozent. Diese Manko von Momentaufnahmen lässt sich auch nicht im Nachhinein durch komplizierte statistische Analysen ausgleichen. “Internationale Vergleichsuntersuchungen (z.B. TIMSS und PISA)”, so stellt Franz-Emanuel Weinert, der Nestor der deutschen Bildungsforschung, lapidar fest, “schaffen Orientierungswissen, das in der Regel aber nicht geeignet ist, bildungspolitische Entscheidungen [. . .] direkt zu fundieren oder zu steuern”.

2. Das zweite wichtige Kriterium für gute Evaluation ist Ziel-Validität. Evaluation muss so angelegt sein, dass nicht nur spezifische Teilaspekte des Bildungssystems

präzise bewertet werden können, sondern auch so, dass sie mit ihren übergeordneten Zielen vereinbar ist. Es wäre wenig sinnvoll, Teilaspekte auf Kosten der Gesamtleistung unseres Bildungssystems zu optimieren.

Die Ziele der Schulbildung werden zum einen durch die Lehrpläne unserer Schulen konkretisiert. Lehrpläne setzen einen Rahmen, der sicher stellt, dass unverzichtbare Dinge in den Schulen behandelt werden, aber sie regulieren heute den Unterrichtsvollzug nicht mehr bis ins Detail, sondern erlauben und erfordern, dass Lehrer auch die spezifischen Bedürfnisse der Schülerinnen und Schüler und ihrer Eltern sowie die sich ständig wandelnden Anforderungen der Berufswelt und der Gesellschaft in den Unterricht mit einbeziehen. Diese Bedürfnisse und Anforderungen müssen bei einer Evaluation ebenfalls berücksichtigt werden, wie Annette Schavan sagte, die als Vorsitzende der Kultusministerkonferenz 1998 den Auftrag an PISA mitformulierte: "*Erfüllt die Schule ihre Aufgabe, junge Menschen zu befähigen, ihren individuellen Ansprüchen, ihren beruflichen Anforderungen und gesellschaftlichen Erwartungen gerecht* zu werden?" Aber weder die "jungen Menschen", noch ihre Eltern oder Lehrer wurden gefragt, welche Leistungen erhoben werden sollten. Die Lehrpläne spielten für PISA nur am Rande eine Rolle. Man gibt sich sogar stolz, dass man sie im Bereich Lesefähigkeit hat großenteils ignorieren können. Statt Gedichte, Romane und andere literarische Texte werden Tabellen und Zeitungsausschnitte verwendet; statt Lesen *und* Schreiben wird nur Lesen getestet.

Viele wichtige Schlüsselfähigkeiten, die für das Leben notwendig sind, bleiben bei PISA links liegen. Statt dessen wird suggeriert ("Lernen für das Leben", "Basiskompetenzen"), dass nach umfassenden Kriterien evaluiert wurde. Die Schule unterricht schon jetzt einen viel zu kleinen Ausschnitt aus den für heutiges Leben notwendigen Kompetenzspektrum; PISA eng es nochmals stark ein. Evaluation ist immer begrenzt; diese Begrenztheit selbst ist nicht problematisch, jedoch der Versuch, sie durch Umwertungen aufheben zu wollen. Lesen, Rechnen und Naturkenntnisse gehören zweifellos zum Unterrichtskanon unserer Schulen. Aber ob sie für die Bewältigung des Lebens immer notwendig oder gar hinreichend sind, muss ernsthaft bezweifelt werden.

Mündige Bürger in einer Demokratie benötigen für ihr Leben heute vor allem Fähigkeiten zum produktiven, kreativen und verantwortungsvollen Umgang mit anderen Menschen und mit der Natur. Viele dieser Fähigkeiten können nicht mit simplen Auswahlantworten und standardisierten Kurzantworten oder vielleicht überhaupt nicht im Schulalter gemessen werden, sondern erst, wenn die Schüler in verantwortungsvollen Positionen zeigen müssen, was sie wirklich können. Was die Schule wert ist, zeigt sich oft erst viel später im Leben. Es wäre, wie das Beispiel USA zeigt, fatal, die Vielfalt der Anforderungen des Lebens auf ein paar "Standards" zu reduzieren, nur um das Testen und die damit verbundenen Sanktionen zu ermöglichen.

Das Etikett "authentisch" täuscht leicht darüber hinweg, dass wirkliche Handlungskompetenzen mit den PISA-Aufgaben kaum erfasst werden können. Die Testteilnehmer müssen keine Experimente vorführen, nicht unter realem Verantwortungsdruck mathematische Probleme lösen und auch nicht mit Kommunikationsproblemen fertig werden. Selbst der Anspruch, mit dem PISA-Test "Lesekompetenz" zu messen, ist fraglich. Da werden die Schüler zu einem Text über "die wissenschaftlichen Waffen der Polizei" nach der Rolle von Gentests als Beweis gefragt. Wer Test-schlau ist, überspringt den Text, liest zuerst die Fragen, "scannt" dann den Text auf mögliche Antworten hin ab, und hält sich dabei strikt an die Instruktion, dafür nur die grau unterlegten (aber falschen!) Zusammenfassungen zu lesen, die vermutlich von einem Redakteur stammen. Das ergibt die "richtige" Antwort: "der Gentest liefert den Beweis". "Käseleicht" sagen die Schüler, die so vorgehen. Wer den Text jedoch "zu verstehen und zu nutzen" versucht, über ihn "reflektiert, um das eigene Wissen und Potential weiter zu entwickeln und am gesellschaftlichen Leben teilzunehmen" (so die PISA-Definition von Lesekompetenz), verliert nicht nur Zeit und damit wertvolle Punkte bei den weiteren Aufgaben, er wird auch die Antwort ankreuzen "der Gentest ist nur einer unter vielen Beweisen", für die er aber keinen Punkte erhält. Es scheint, dass mit solchen Aufgaben eher die Unterwerfung unter die Instruktion von Autoritäten gemessen wird als die Fähigkeit zur kritischen Reflexion, wie es unsere Lehrer heute fördern wollen. Es müssen hohe Anforderungen an die Validität der Testaufgaben und des Erhebungskontextes gestellt werden. Fähigkeitsmessungen reagieren sehr sensibel auf

Beeinträchtigungen wie Zeitdruck. So kann aus einer mathematischen Testaufgabe unter Zeitdruck schnell ein Test zur Lesefähigkeit werden. Die Tatsache, dass die Bedingungen “für alle gleich” sind, macht diesen Tatbestand nicht besser, zumal Schüler in verschiedenen Ländern unterschiedlich gut auf solche Rahmenbedingungen eingestellt werden.

3. Drittens muss Evaluation partizipatorisch sein. Ohne Partizipation bleiben die Bedürfnisse der Beteiligten abstrakte Größen und es entsteht auch kaum Akzeptanz und Motivation für die Konsequenzen, die aus einer Evaluation folgen. Die Beteiligten müssen an allen Phasen – der Durchführung, der Auswertung und der Interpretation – beteiligt werden. Das ist in der Wirtschaft üblich und wird auch durch den Auftrag an PISA nahegelegt. In der PISA-Studie wurden weder Schüler noch Lehrer und Eltern, die wichtigsten Akteure in der Lernumwelt von Kindern, in die Durchführung der Studie einbezogen.

4. Das vierte Kriterium für gute Evaluation ist Transparenz. Transparenz der Durchführung, Auswertung und Darstellung ist die wichtigsten Voraussetzungen dafür, dass unser Bildungssystem von mündigen Bürgern demokratisch gesteuert werden kann und dies nicht wenigen Experten überlassen werden muss. Die Transparenz von PISA lässt dagegen zu wünschen übrig. Was ein “PISA-Testwert” von “490" und ein Unterschied von “25" Punkten bedeutet, bleibt selbst für Experten im Dunkeln. In die komplizierte Berechnung der PISA-Testwerte gehen zu viele strittige Annahmen ein. Die Angabe von Lösungshäufigkeiten würde es jedem ermöglichen, die Bedeutung der Punkte-Unterschiede zwischen den Ländern abzuschätzen. Da trotz sehr großer Stichproben selbst große Punktedifferenzen statistisch insignifikant bleiben, wirken sie merkwürdig aufgebläht. Noch Monate nach der Erhebung blieben die mit Steuermitteln finanzierten PISA-Rohdaten für unabhängige Forscher unter Verschluss. Mit Datenschutz allein kann das nicht erklärt werden.

5. Schließlich sollen Evaluationen der Beurteilung von bildungspolitischen Maßnahmen und didaktischen Konzepten dienen, aber nicht der Maßregelung von Menschen. Werden sie dafür eingesetzt, werden sie “unscharf” und für die

Qualitätsentwicklung nutzlos oder sogar gefährlich, wie das Beispiel USA zeigt (Amrein & Berliner, 2002). Zu Sanktionen zählen nicht nur materielle Strafen in Form von Mittelentzug oder Entlassungen ganzer Kollegien, wie dies in den USA und England bereits praktiziert wird. Auch eine Rangreihung verbunden mit öffentlicher Belobigung oder Herabwürdigung ("Nur im unteren Bereich sind wir einsame Spitze!") stellt eine Sanktion dar. Rangreihen tragen zum Unterhaltungswert des Sports bei, an dem sich Menschen freiwillig beteiligen, haben aber in der Qualitätsentwicklung nichts zu suchen. Bildung wird nicht zur Belustigung der Öffentlichkeit betrieben. Hier möchte man vielmehr wissen, mit welchen Methoden und Maßnahmen wir unsere Kinder am besten fördern.

Das Phänomen der Unschärfe ist in der Physik spätestens seit Heisenberg bekannt: Jede Messung kann das Objekt, das gemessen wird, verändern, im extremsten Fall so stark, dass eine eindeutige Messung nicht mehr möglich ist. Das gilt auch für Schulleistungstests. Sie messen mit der Zeit immer weniger die Schulleistungen, dafür aber immer mehr die gemeinsame Anstrengung von Schülern, Eltern, Lehrern und Schulverwaltungen, in der Öffentlichkeit gut auszusehen und Sanktionen zu vermeiden. In England und den USA führte die Einführung von Sanktionen dazu, dass immer mehr Schüler, Lehrer und Schuladministrationen beim Testen allerlei Tricks bis hin zum Betrug anwenden, um ihre Testwerte nach oben zu puschen. Amrein und Berliner (2002) weisen nach, dass zwar die Punktwerte in den Sanktionstests oft steigen und einen Erfolg der Sanktionen vorspiegeln, aber die tatsächlichen Leistungen, die in sanktionsfreien Schultests gemessen werden, kontinuierlich sinken. "In den Vereinigten Staaten hat der verbreitete Einsatz von external standardisierten Tests Evaluationen auf fast jeder Klassenstufe [. . .] sehr ernsthafte Zerstörungseffekte gehabt", glaubt denn auch Alan Schoenfeld, renommierter Mathematikdidaktiker und ehemaliger Präsident der *American Educational Research Association.*

Dazu passt, dass in den USA und England über ein Drittel der Schulen die Teilnahme an der PISA-Studie verweigert hatte! Damit der PISA-Vergleich dennoch möglich war, mussten Schulen "nachgezogen" werden. Es wäre interessant zu wissen, wie

stark diese selektive Verzerrung der Stichprobe den Rangplatz dieser Länder nach oben getrieben hat.

In anderen Ländern wie Korea und Japan haben Tests das Bildungssystem auf andere Weise verändert. Eltern schicken ihre Kinder jeden Tag nach der Schule noch bis zu vier Stunden in Paukschulen, die diese auf Tests hin trimmen. Wer die “guten” PISA-Werte in diesen Ländern den auf Video gebannten Unterrichtsmethoden dort zuschreibt, führt sich und uns in die Irre. Schließlich hat der Einsatz von Schulleistungstests dort, wo er Belohnung und Bestrafung statt Methodenevaluation im Blick hat, zu einer stetigen Verengung des Fach- und Fähigkeitsspektrums bei den Schülern geführt. Es wird oft nur noch gelehrt und gelernt, was Punktgewinne in Mathematik-, Deutsch- und Naturkundetests bringt. In den USA distanzieren sich daher inzwischen immer mehr Hochschulen von Hochschuleingangstests. Wie gesagt, nicht der Einsatz von Tests per se steht hier in Frage, sondern ihr einseitiger Einsatz als Sanktionsinstrument.

Wollen wir solche Entwicklungen auch in Deutschland? Oder wollen wir es eher wie Finnland oder Schweden halten, wo man Schulleistungstests – mit gutem Erfolg – vor allem zur Verbesserung der Schulqualität einsetzt und auf Sanktionen und Rangreihung weitgehend verzichtet? Natürlich gibt es in unseren Schulen einiges zu verbessern und dafür werden Evaluationsstudien dringend benötigt. Aber Schulleistungsvergleiche mit Sanktionen sind dafür ebenso wenig hilfreich wie blinder Reformaktionismus, der jede Idee flächendeckend einführt, ohne sie vorher unvoreingenommen erprobt zu haben. Wer unsere Schule verbessern will, kann sich auf die Befunde vieler sorgfältig angelegter Studien über unser Bildungssystem stützen (nur einige Forscher seien hier genannt: Aurin, Bargel, Fend, Ingenkamp, Klemm, Lukesch, Rolff, u.v.a.m.). Allerdings sind viele Befunde nicht mehr aktuell, weil die Bildungsforschung in den Universitäten an den Rand gedrängt wurde. Immer öfter müssen Befunde aus dem Ausland bezogen werden.

Wer unsere Schulen durch Evaluation verbessern will, muss daher dafür sorgen, dass die Lehr-Lernforschung an den Universitäten wieder belebt und die seit Jahren verfolgte Politik der Auszehrung auf diesem Gebiet gestoppt wird. Wenn

Schulforschung immer mehr in kommerzielle Institute ausgelagert wird, fehlt uns bald das notwendige kritische Potential für gute Evaluation.

Wer unsere Schule wirklich verbessern will, muss also für den Bildungsbereich eine neue Evaluationskultur entwickeln: Größtmögliche Ziel-Validität, Partizipation aller Beteiligten in allen Phasen der Evaluation, Transparenz aller Schritte, vor allem der Berichterstattung, und ein Fokus auf Maßnahmen und Methoden aber nicht auf (oder gegen) Personen und Personengruppen (in Abwandlung von Poppers Diktum: Lasst Theorien sterben und nicht Menschen!). Dazu müssen Messinstrumente valide sein. Sollen praktische Fähigkeiten untersucht werden, sind Verhaltenstests in realitätsnahen Kontexten notwendig, die nicht durch noch so wortreiche Fragebogen ersetzt werden können, denen man das Etikett "authentisch" aufgedrückt hat. Aber es genügt nicht, valide Messinstrumente zu haben. Sie müssen auch richtig eingesetzt und interpretiert werden. Schließlich garantiert auch eine gute Evaluation noch lange keine Qualitätsverbesserung. Hier sind kompetente Planung und ehrliche Erprobungen mit hohen Evaluationsstandards gefragt, und nicht überhastet eingeführte, flächendeckende Reformdiktate, die unter großer Geheimhaltung vorbereitet werden. Solche Erprobungen müssen mit geeigneten Designs evaluiert werden, die als Mindestanforderung Vor- und Nachtests und Kontrollgruppen (oder zumindest mit Querschnittgruppen) aufweisen. PISA kann noch nicht einmal sehr lockeren Evaluationsstandards genügen. Zudem müssten diese Anforderungen auch mit allen Betroffenen diskutiert werden, also mit Eltern, Lehrern, Schülern und Bildungsexperten. Schließlich muss eine neue Evaluationskultur ihr Augenmerk auf die Unterscheidung von guten und schlechten Maßnahmen, guter und schlechter Politik, effektiver und nicht-effektiver Unterrichtsmethoden richten, und nicht auf die Abwälzung der Verantwortung auf die Opfer (Schüler, Lehrer, Eltern) schlecht vorbereiteter Maßnahmen, undurchdachter politischer Entscheidungen und falscher pädagogischen Theorien. Das Thema Evaluation und Qualitätsentwicklung sollte auch verstärkt in die Ausbildung von SchulpädagogInnen und LehrerInnen eingebracht werden, womit aber nicht die Kunst gemeint ist, aus nichtigen Datenmengen mittels komplexer Daten eindrucksheischende Rangreihen zu bilden, sondern eine solide Einführung in die Mindeststandards dieser Wissenschaft.

Dies wäre ein Weg, eine neue Evaluationskultur zu erreichen, die auf allen Ebenen nachhaltig zur Verbesserung unseres Bildungssystems beiträgt, das im übrigen den Vergleich nicht scheuen muss, wenn er nur richtig angestellt wird: “Wir haben engagierte Lehrerinnen und Lehrer. Wir haben kompetente Ausbilderinnen und Ausbilder. Und wir haben eine lernbereite, motivierte Jugend”, stellt Bundespräsident Rau zu Recht fest. Hierauf lässt sich gut aufbauen. Wir benötigen keine Anpassung der Schule an Tests (“Kerncurricula”), sondern die Anpassung der Evaluationen an die Erfordernissen der Bildung.

Professor Dr. Georg Lind, Jahrgang 1947, lehrt an der Universität Konstanz Pädagogische Psychologie. Er befasst sich seit langem mit der Förderung moralisch-demokratischer Kompetenzen und mit der Messung von Bildungseffekten.
Seine Email: Georg.Lind@uni-konstanz.de; Web-Seite: http://www.uni-konstanz.de/ag-moral/.