Entwurf eines Textes für den KMK-Sammelband „Leistungsmessung in der Schule“
(hrsgg. Von Prof. Dr. Franz E. Weinert)

... mit der Bitte um Anmerkungen zu Inhalt, Form und Verständlichkeit!

Leistungsmessung kontrovers

Eine Zusammenfassung aktueller Argumente in verteilten Rollen

von Hans Brügelmann

ELTN: ElternvertreterIn
LEHR: Mitglied einer Lehrergewerkschaft
MODR: Moderatorin
VERW: BeamteR der Schulverwaltung
WIRT: VertreterIn der Arbeitgeberverbände
WISS: BildungsforscherIn

20 Seiten x 30 Zeilen = 600 Zeilen x 60 Anschläge

15.6. Erstfassung
15.8. Zweitfassung

MODR: In unserem Gespräch geht es um die Frage, wie die Qualität von Schule und Unterricht verbessert werden kann und ob Testprogramme dabei hilfreich sein können, z. B. landesweit wie LAU* in Hamburg, aber auch in Form von internationalen Leistungsvergleichen wie der IEA-Lesestudie, wie TIMSS*, PISA* und IGLU*.

WIRT: Einspruch: es geht nicht nur um die Verbesserung der Qualität von Unterricht, sondern auch um die Pflicht der Schulen zur Rechenschaft. Als Steuerzahler habe ich ein Recht darauf zu wissen, was mit meinem Geld passiert und ob die Schulen ihren gesellschaftlichen Auftrag erfüllen.

LEHR: Allerdings ist festzuhalten, dass die Leistungen, vor allem die fachlichen Leistungen der SchülerInnen, nur einen Teil des Auftrags der Schule ausmachen. Und deren "Qualität" erweist sich nicht nur in kurzfristigen Ergebnissen, sondern auch in der Qualität der Prozesse. Es geht darum, wie junge Menschen in unserer Gesellschaft aufwachsen sollen, und um langfristige Wirkungen, die sich präzise kaum mehr erfassen lassen. Die Qualität von Demokratie würde doch auch niemand nur an ökonomischen Indikatoren messen.

ELTN: Ich verstehe gar nicht, warum in den Schulen so ein Wirbel um Tests gemacht wird. Leistungen werden doch überall bewertet. Und auch Tests gehören zu unserem Alltag. Wenn ich ein neues Auto oder auch nur einen neuen Toaster kaufen will, besorge ich mir einen aktuellen Warentest. Und auch Personen bewertet jedeR von uns tagtäglich, z.B. wenn wir den Friseur wechseln, weil wir unzufrieden sind, oder wenn wir Bekannten unsere Hausärztin empfehlen. Warum tun sich PädagogInnen so schwer mit der Forderung nach Leistungsmessung?

LEHR: So pauschal stimmt das gar nicht. Leistungsmessung gehört auch zu *unserem* Alltag. Wir LehrerInnen erfassen und bewerten Leistungen von SchülerInnen. Die Schulaufsicht entscheidet nach Leistung über die Verbeamtung junger LehrerInnen und über die Besetzung von Funktions-stellen. Das Parlament wiederum bewertet die Leistungen der Verwaltung, z. B. bei Änderungen der Gesetze oder bei der Bewilligung von Mitteln im Haushalt.

WIRT: Genau, und diese Vorgänge müssen transparenter werden. Ich verstehe gar nicht, warum Sie so empfindlich auf den Einsatz von standardisierten Instrumenten reagieren. Sie werden in Zukunft nicth mehr damit rechnen können, dass öffentliche Mittel einfach so fließen, ohne dass sich der Aufwand durch einen entsprechenden Ertrag rechtefrtigen lässt. Die knappen Mittel müssen sparsam und effektiv eingesetzt werden

-- wie in der Industrie und wie in anderen Bereichen der öffentlichen Verwaltung auch. Schulen können sich zukünftig nicht mehr diffus auf ein "gutes Klima" berufen, wenn ihre SchülerInnen nicht anständig lesen, schreiben und rechnen lernen.

VERW: Ihnen geht es um die Messung von Leistungen in einem engen, sozusagen technischen Sinne. Zeugnisse sind aber kein Warentest, und Unterricht ist auch kein Markt, auf dem Kunden zwischen Dienstleistungen wählen -- wie bei Friseur und Ärztin. Schule ist auch kein Betrieb nur zur "Produktion" von Qualifikationen, die auf dem Arbeitsmarkt verwertbar sind, sondern vor allem ein Ort der Bildung. Die Qualität eines Theaters oder eines Krankenhauses können Sie ja auch nicht allein nach dem Verhältnis von Input und Output bewerten.

LEHR: Bildung ist eben mehr als eine Addition von Kenntnissen und Fertigkeiten, die man isoliert abprüfen könnte. Tests erfassen doch nur ein Oberflächenverhalten und nicht, was eine Person denkt.
Wenn ich an die Evaluation der Programme kompensatorischer Vorschulerziehung in den 70er Jahren zurückdenke, zweifle ich grudnsätzlich an der Aussagekraft kurzfristiger Effekte. Damals kamen unmittelbar nach Programmende von "Head Start" und anderen Projekten die großen Erfolgsmeldungen aus den Begleitstudien: höherer IQ, bessere Sprachfähigkeit und bessere Leistungen in den anderen trainierten Berichen. Wenige Jahre später hieß es, die Überlegenheit der Versuchsgruppen gehe schon während der Grudnschule verloren, der Aufwand lohne nicht. Wieder einige Jahre später wurden sog. "sleeper"-Effekte festgestellt, d. h. diejenigen, die als Kinder an den Programmen teilgenommen hatten, erreichten bessere Schulabschlüsse, waren im Beruf erfolgreicher und hatten weniger Probleme im privaten und sozialen Umfeld.
Meine Sorge: Kurzfristige Leistungsmessungen erfassen wie in einer Momentaufnahme nur einen kleinen Ausschnitt der Wirkungen, der sich nicht sinnvoll interpretieren, wohl aber gut als Munition im politischen Tagesgeschäft verwenden lässt.

WISS: Diese pädagogische Skepsis gegenüber Tests als Instrument schulischer Leistungsmessung hat eine lange Geschichte. Aber sie ist in anderen Ländern, z. B. im angelsächsichen Raum, viel weniger ausgeprägt als in Deutschland. Dabei hätten gerade LehrerInnen und ihre Verbände allen Grund, selbstkritisch gegenüber ihrer eigenen Urteilsfähigkeit zu sein.
In den 60er und 70er Jahren wurden z. B. reihenweise Untersuchungen zur Notengebung in Klassenarbeiten publiziert. Dabei kam heraus:
- Verschiedene LehrerInnen beurteilen dieselbe Arbeit sehr unterschiedlich, und zwar nicht nur Aufsätze, sondern auch Rechenarbeiten und Diktate.

- Sogar *dieselben* LehrerInnen beurteilen *dieselbe* Arbeit zu verschiedenen Terminen unterschiedlich.
- Und: Urteile über (gleiche) Schülerleistungen hängen mit sozialen Faktoren wie Status der Eltern und persönlicher Beliebtheit der SchülerInnen zusammen.

Wenn man denkt, was alles von Noten abhängt, muss Leistungsmessung stärker objektiviert werden.

LEHR: Theoretisch klingt das gut, aber politisch ist es naiv zu glauben, in der Öffentlichkeit würden Daten aus Leistungsstudien so differenziert wahrgenommen, wie die Befunde es erfordern. Auch quantiative Daten sprechen doch noch für sich selbst. Sie müssen interpretiert wreden. Man müsste die Ergebnisse gleich von verschiedenen Positionen aus kommentieren, um eine einseitige Verwertung zu erschweren, wie wir das z. B. beim TIMSS erlebt haben. So einfach ist nicht mit der Objektivität.

ELTE: Die ist aber auch für uns Eltern wichtig, um die Schullaufbahn unserer Kinder verlässlich planen, und auch, um die Arbeit der LehrerInnen einschätzen zu können. Insofern interessieren mich ganz konkret die Klagen der Kammern über unzureichende, ja sinkende Leistungen der Schule. Was ist dran am sog. "Leistungsverfall"?

WIRT: Wir stützen uns dabei auf Ergebnisse von Tests bei den neu eingestellten Lehrlingen. LehrerInnen-Gewerkschaften, ErziehungswissenschaftlerInnen, aber auch die jeweils betroffenen Regierungsparteien stellen unsere Klagen immer gleich in Frage, ohne sich ernsthaft auf die Kritik einzulassen. Man hat den Eindruck, als ob die Pädagogik und Bildungspolitik unfehlbar seien. Ohne Bereitschaft, sich einer Außenkritik zu stellen, sind Fortschritte doch nicht denkbar.

WISS: Das ist richtig. Andererseits sind den genannten Untersuchungen eine Reihe von Mängeln vorzuwerfen:
- Inhaltlich erfassen die Tests nur kleine Ausschnitte schulischer Ziele (meist: Rechtschreibung im Diktat und Grundrechenarten).
- Methodisch sind die Aufgaben oft problematisch (z. B. eine exotische Auswahl des Wortschatzes oder das Angebot von mehreren falschen Schreibungen, unter denen die richtige auszuwählen ist).
- Für Verallgemeinerungen sind die Stichproben nicht repräsentativ.
- Um historische Vergleiche anstellen zu können, müsste es vergleichbare Bezugsgruppen aus früheren Zeiten geben.

LEHR: Und Interpretationen wie "Leistung zählt nicht mehr" oder "Schulleistungen werden immer schlechter" haben die Aussagekraft der Daten oft weit überzogen, zum Teil auch die Grundsätze der Fairness verletzt.

VERW: Das ist aber doch Vergangenheit. Inzwischen sind forschungsmethodisch versierte Großprojekte in das Geschäft eingestiegen. Wir werden da gar nicht mehr rauskommen. Der ökonomische Druck auf die Schule, Rechenschaft über ihre Arbeit abzulegen, hat in allen Industrieländern zugenommen. Ich sehe die dritte internationale Studie zu den Leistungen in den Naturwissenschaften und Mathematik (TIMSS) in diesem Kontext.

WIRT: Das breite Rauschen im Blätterwald der Tagespresse zeigt, dass diese Form des Systemvergleichs einen empfindlichen Nerv trifft. Die zentrale Frage ist doch: Ist unser Bildungssystem international konkurrenzfähig? Und eine zweite: Investieren wir in den Bildungshaushalten an den richtigen Stellen? Ein Kollege von Ihnen hat es mal treffend mit dem schönen Satz gesagt, notwendig sei eine Haltung, "die die pädagogische Version der protestantischen Rechtefertigungslehre zu überwinden trachtet, wonach es für die Erlösung allein auf den Glauben ankommt, während die Werke keine Bedeutung haben." (H. Lange in: VBE 2000, 48)

LEHR: Das klingt so einfach. Als ob man nur ein Metermaß oder ein Thermoter nehmen müsste und dann daran den Zustand des Bildungswesens ablesen könnte. Aber wir haben doch schon oft die Erfahrung gemacht, dass wissenschaftlichen Untersuchungen nicht zu einem klaren Ergebnis führen. Ich erinnere nur an den Methodenstreit in der Lesedidaktik, der in den 60er Jahren in einem Patt endete. Oder an die Versuche mit Gesamtschulen in den 70er Jahren, die von Verfechtern und von Gegnern bis heute unterschiedlich ausgeschlachtet werden. In den letzten Jahren hat der Streit um die Integration behinderter SchülerInnen umstritten, weil die Untersuchungen keine eindeutigen Ergebnisse erbringen. Mit der "Präzision" und "Objektivität" scheint es da nicht so weit her zu sein. Ich fürchte, mit Tests erwischt man nur ein paar kleine Zipfel von dem, worum es im Unterricht wirklich geht. Schlimmer noch: In den Händen der Messteoretiker wird aus pädagogischen Konzepten wie "Selbstständigkeit" ein mechanisches System von Fähigkeitsmodulen, in dem ich meine Probleme eines partnerschaftlichen Umgangs mit Kindern und Jugendlichen nicht mehr wiedererkenne.

WISS: Zugegeben: Wir stecken da in einer Falle. Versuchen wir, komplexe Fähigkeiten zu erfassen, wirft man uns vor, ihr "Wesen" zu verfehlen; verzichten wir darauf, diese Fähigkeiten einzubeziehen, wird uns vorgehalten, wesentliche Ziele der Schule nicht zu berücksichtigen. Wir müssen pragmatische Lösungen finden.

Aber wenn ich mir anschaue, mit was für primitiven Aufgaben immer noch viele LehrerInnen die Leistungen ihrer SchülerInnen erfassen, dann wünschte ich mir, das von uns in der Testentwicklung erreichte Niveau wäre auch nur halbwegs Standard im Unterrichtsalltag.

WIRT: Ich stimme Ihnen zu. Vom grünen Tisch aus kann man leicht immer neue Forderungen erheben, was ein "vernünftiges" Messinstrument alles zu leisten habe.
Alles, was besser ist als der bisherige Schlendrian, ist einen Versuch wert. LehrerInnen können doch machen was sie wollen. Und die faulen und unfähigen KollegInnen bringen den ganzen Berufsstand in Verruf. Ich verstehe nicht, dass kompetente und engagierte LehrerInnen diese Missstände decken. Sie sollten doch froh sein, wenn ihre Leistung anerkannt wird. Testprogramme sind da unbestechlich.

VERW: Ich sehe das nicht ganz so rosig. Vor allem hiunkt der Vergleich mit den Prüfaufgaben der LehrerInnen. Landesweite Testprogramme haben eine ganz andere Bedeutung.
Und die Nebenwirkungen falsch gewählter oder falsch verstandener Aufgaben können fatal sein. Testergebnisse haben eine hohe Suggestivkraft. Durch ihre scheinbar technische Präzision gewinnen sie das Image einer Unparteilichkeit, der in der öffentlichen Diskussion leicht überschätzt wird. Laien, und dazu zählen auch BildungspolitikerInnen, VerbandsvertreterInnen und JournalistInnen, können im Einzelfall die Annahmen, unter denen ihre Aussagen gelten, und ihre konkrete Leistungsfähigkeit kaum einschätzen. Sie haben eine zu simple Alternative konstruiert: "objektive" Tests hier gegen "subjektive" (Vor-)Urteile da.

LEHR: Zusätzlich sehe ich Probleme in der Passung von Test und Unterricht. Vor allem die Beteiligung an internationale Vergleichen erzwingt oft problematische Kompromisse. Da wird der Lehrplan eines Bundeslandes (von 16!) und der Aufgabenpool eines Schulbuchs zum Maßstab für "das Curriculum der BRD".
Und der Zeitdruck, unter dem die Testaufgaben entwickelt worden sind, erschwert es, solchen Schwächen auf die Spur zu kommen oder gar bessere Alternativen zu entwickeln. Wir haben ja schon bei TIMSS gesehen, dass eine Reihe von Aufgaben inhaltlich oder sprachlich problematisch war.

WISS: Es ist sicher richtig, dass Paper-and-Pencil- Tests im Bereich Lesen und Schreiben schon vom Medium her näher an den untersuchten Fähigkeiten sind als etwa in den Naturwissenschaften. Aber in Zusatzstudien haben sich die Antworten in den Auswahlfragen der Tests als gute Annäherung an die Leistungen in Experimenten erwiesen. Die Tests sollen ja nur als Indikatoren dienen, sie müssen nicht die angezielten Fähigkeiten selbst erfassen.

LEHR: In der Theorie stimmt das. Aber sie müssen wieder an die Wahrnehmung durch die Betroffenen denken. SchülerInnen oder LehrerInnen, die in Tests gut abschneiden wollen, werden sich an den Indikatoren orientieren. Was bringt
es z. B. noch, im Unterricht zeitlich aufwendige Experimente durchzuführen, wenn SchülerInnen Test auch bestehen können, indem man einfach solche Testaufgaben mit ihnen übt?
Diese Umorientierung kennen wir von den japanischen Jukus, vom englischen General Certificate of Secondary Education und von den A-Levels am Ende der Sekundarstufe, aber auch von den Repetitorien, die deutsche StudentInnen vor dem juristischen Staatsexamen besuchen."Teaching to the test" lässt sich kaum vermeiden, wenn von den Ergebnissen der Tests persönliche Vor- und Nachteile für die Betroffenen abhängen. Denken Sie nur an die Nebenwirkungen von Provisionsregelungen für VertreterInnen in der Wirtschaft. Da wird auch ohne Rücksicht auf langfristige Folgen für die Firma oder gar für die Kunden auf rasche Erfolge gesetzt.

WISS: Ich finde, hier gehen verschiedene Dinge durcheinander. Wir müssen uns über den konkreten Zweck der Leistungsmessungen verständigen. Ich sehe da drei ganz verschiedene Funktionen:
Landesweite Testprogramme können einmal dazu dienen, die Vergabe von Berechtigungen auf einen einheitlichen Maßstab zu beziehen. Das Zentralabitur in den süddeutschen Ländern, aber beispielsweise auch in Frankreich, hat diese Funktion. Ein ganz anderes Ziel ist es, die Arbeit einzelner Schulen oder gar LehrerInnen zu bewerten. Vor allem in den angel-sächsischen Ländern werden standardisierte Tests zu diesem Zweck eingesetzt. Bei Veröffentlichung der Daten kommt es dann zu dem viel diskutierten "ranking", einer Rangliste von Schulen. In einigen Kantonen der Schweiz wird auch überlegt, die Bezahlung der LehrerInnen an Ergebnisse in solchen Tests zu knüpfen.
Eine wieder andere Funktion haben die aktuell diskutierten länderübergreifenden Stichprobenuntersuchungen, die auf einen Vergleich von Schulsystemen zielen.

ELTN: Und worum geht es bei den anstehenden Untersuchungen? Wir Eltern haben ja ein Interesse an Informationen über die Qualität des Unterrichts in der Klasse unseres Kindes.

VERW: Wir müssen aktuell zwei Typen von Untersuchungen auseinanderhalten. Die KMK hat in ihren Konstanzer Beschlüssen vom Oktober 1997 einerseits länderübergreifende Untersuchungen wie TIMSS und PISA oder jetzt IGLU für die Grundschulen gefordert. Da geht es gar nicht um einzelne
LehrerInnen und Schulen, sondern um eine vergleichende Untersuchung von Bildungssystemen. Insofern greift der Vorwurf des Ranking hier nicht.
Der zweite Ansatz zielt auf "qualitätsverbessernde Maßnahmen" innerhalb

der einzelnen Bundesländer, und da gibt es unterschiedliche Ansätze. Landesweite Testprogramme wie in Hamburg sind da die Ausnahme. Und auch da werden keine Ranglisten veröffentlicht. Schulbezogene Daten erhalten nur
die Schulen selbst, als Anstoß, über ihre Arbeit im Vergleich zu anderen Schulen nachzudenken.

LEHR: Diese Unterscheidung kann man doch nur auf dem Papier so klar durchhalten. Wenn bei einem Systemvergleich herauskommt, dass einzelne Schulformen -- sagen wir mal Gymnasien mit interner Differenzierung nach Leistungsniveaus in den Hauptfächern -- nach bestimmten Kriterien schlechter abschneiden, dann hat das Auswirkungen auf die Wahrnehmung der einzelnen Schulen, die nach diesem Ansatz arveiten, und damit indirekt auch auf die Arbeit in ihnen.

ELTN: Aber das ist doch gut! Wo sollen die denn sonst die Information herbekommen, was dieses System taugt oder nicht?

VERW: Na, so einfach ist das auch nicht. Wenn bei einer Studie wie TIMSS im *internationalen* Teil herauskommt das Schulsysteme mit integrierten Sekundarschulen vergleichsweise gut abschneiden, im *Deutschland-internen* Ländervergleich aber, dass Gesamtschulen eher schlechter als dreigliedrige Systeme -- was folgt daraus? Sollen wir die Gesamtschule flächendeckend einführen, weil sie anscheinend nur dann ihre Qualitäten voll entfalten kann? Müssen wir die deutschen Gesamtschulen besser ausstatten, ihre LehrerInnen besser aus- und fortbilden, damit sie das Potential der integrierten Schulform besser ausnutzen können? Oder sollten wir die Gesamtschulen ganz abschaffen, weil sie in unserem gesellschaftlichen Umfeld nicht so gut passen wie in anderen Ländern? Oder zeigt der Widerspruch zwischen internationalem und nationalem Vergleich vielleicht auch nur, dass hier Äpfel mit Birnen verglichen werden, dass deutsche Gesamtschulen unter ganz anderen Randbedingungen arbeiten müssen, z. B. weil ihnen durch die Konkurrenz mit Gymnasien und Realschulen ein wichtiger Teil der SchülerInnenschaft verlorengeht?

WISS: Möglicherweise werden hier die falschen Fragen gestellt. Die Organisationsform von Schule scheint nicht der entscheidende Faktor für die Qualität des Unterrichts zu sein. Umso wichtiger ist es, dass wir die Wirkungen des Unterrichts selbst erfassen -- und dass wir unterschiedliche Leistungen auf konkrete Bedingungen beziehen können, um diese dann zu verbessern.

LEHR: Ich habe Probleme mit diesem Ansatz. Mal wird behauptet, die Programme zielten nur auf die Ebene des Gesamtsystems. Da frage ich mich, ob die BildungspolitikerInnen überhaupt in der Lage sind, Konsequenzen aus den Befunden zu ziehen. Was machen die z. B., wenn sich ein Ergebnis

aus neueren us-amerikanischen Studien bestätigt, nämlich dass kleinere Klassen am Schulanfang förderlich sind und vor allem leistungsschwachen SchülerInnen zugute kommen? Werden die dann wirklich mehr LehrerInnen einstellen?
Andererseits wird gesagt, die Testprogramme sollten den Schulen helfen, ihre Arbeit vor Ort zu verbessern. Aber dann gibt es Konflikte mit dem Datenschutz und mit der fehlenden Dichte der Daten, um auf der Prozessebene wirklich hilfreich zu sein.

ELTN: Dazu habe ich auch noch eine Frage. Landesweite Testprogramme werden ja immer wieder mit dem Argument begründet, man müsse die Arbeit der Schulen besser erfassen, um ihre Qualität steigern zu können. Warum schneidet dann aber ein Land, in dem solche Programme schon lange an der Tagesordnung sind, nämlich die USA, in internationalen Studien nicht besser ab als z. B. Deutschland? Dort mehren sich doch sogar die Stimmen, die eher eine schulnahe Aufsicht und Beratung der Schulen fordern -- wie wir sie
schon lange haben.

LEHR: Da möchte ich mich gleich anhängen: Bei TIMSS hat sich doch gezeigt, dass Länder mit zentralen Prüfungen nicht besser abschneiden als Bildungssysteme ohne. Woher rührt die Sicherheit, externe Leistungsmessungen führten zu mehr Qualität als schulinterne Prüfungen?

VERW: Man muss doch sehen, dass kein System ohne Kontrolle auskommt. Manche Länder kontrollieren den "input" stärker, z. B. die Qualifikation von LehrerInnen oder die Vorgaben der Lehrpläne (so die meisten kontinentaleuropäischen Staaten). Andere geben in diesem Bereich mehr Freiräume und kontrollieren dafür den "output" stärker, z. B. durch zentrale Prüfungen (so traditionell die angelsächsischen Staaten).
Das deutsche System hat sich in den letzten Jahren in Richtung auf mehr Schulautonomie bewegt, d. h. aber auch die Aufsicht über "input" und Arbeitsabläufe verringert. Also müssen wir Rechenschaft und Kontrolle an anderer Stelle verstärken.

LEHR: Und dazu greifen Sie auf Instrumente zurück, deren Wert und Auswirkungen äußerst umstritten sind, ganz zu schweigen von dem Problem, dass nicht einmal die Kriterien klar sind.

WIRT: Über grundlegende Anforderungen wie Lesen, Schreiben und Rechnen dürfte es aber doch keinen Streit geben.

ELTN: Naja, die Frage ist doch, welches Niveau in diesen Bereichen angesetzt wird: Was soll als Ziel für das Ende der Pflichtschule oder der Grundschulzeit gelten?

WIRT: Dazu kann man z. B. unsere Betriebe befragen, welches Leistungsniveau im Berufsalltag erforderlich ist.

ELTN: Meines Wissens gibt es in dieser Hinsicht erhebliche Unterschiede zwischen verschiedenen Branchen, aber auch zwischen Betrieben innerhalb einer Branche.

LEHR: Zudem setzen die staatlichen Vorgaben keine klaren Prioritäten. Während jeder Lehrplan seine fachlichen Ansprüche favorisiert, stehen in den allgemeinen Richtlinien fachunabhängige Schlüsselqualifikationen wie Kooperationsfähigkeit oder Selbstständigkeit im Vordergrund. Für
LehrerInnen bedeutet das, sie müssen ihre begrenzten Ressourcen auf konkurrierende Ziele verteilen. Da sind unterschiedliche Kompromisse denkbar und legitim. Die Testprogramme setzen ihrerseits eigene Prioritäten. Ich befürchte eine Uniformierung des Unterrichts, der nicht mehr auf die Besonderheiten vor Ort Rücksicht nehmen kann.

WISS: Ich muss noch einmal darauf hinweisen, dass es bei LAU, PISA usw. nicht um die Arbeit der einzelnen Schulen geht. Es ist aber doch wichtig, über eine solche Bestandsaufnahme zu sehen, wo ein System insgesamt seine Stärken und seine Schwächen hat. Da geht es nicht um LehrerIn A oder LehreIn B.

VERW: Indirekt doch. Und ich denke, das ist auch so gewollt. Durch eine solche Bestandsaufnahme wird ja ein öffentlicher Druck erzeugt. Die Begriffe "Stärken" und "Schwächen" machen das deutlich. Mit den Aufgaben werden bestimmte Maßstäbe gesetzt. Die geprüften Leistungen markieren, welche inhaltlichen Bereiche besonders wichtig sind, also auch, welches Niveau in diesen Bereichen erwartet wird. Es muss deutlich werden, dass dies eine politische Frage ist. Solche Entscheidungen kann man nicht WissenschaftlerInnen überlassen.

LEHR: Meine Sorge ist, dass grundlegende Anforderungen nicht nur für den Schulabschluss, sondern auch für Zwischenschritte festgelegt werden. In England gibt es ja ein so gestuftes "national curriculum". Nehmen Sie den aktuellen Fall der auch in Deutschland bekannten Reformschule "Summerhill": Die SchülerInnen haben die Freiheit, wann sie was lernen, würden also bei Zwischentests eher schwach abschneiden. Am Ende der Schulzeit erreichen sie aber überdurchschnittliche Ergebnisse.
Pädagogische Konzepte, die unterschiedliche Lernwege fördern und das individuelle Lerntempo respektieren, werden also durch ein solches System gleich geschaltet.
Und wenn sie dann noch an die unterschiedlichen Lernvoruassetzungen der SchülerInnen denken...

WISS: Wenn Sie sich Studien wie PISA, IGLU oder LAU ansehen, gilt dieser Einwand nicht mehr. Dort werden auch sozialstrukturelle Daten erhoben, um den Einfluss unterschiedlicher Randbedingungen zu berechnen.

ELTN: Mir erscheint das aber nicht so einfach. Einzelne Schulen haben oft unterschiedliche Einzugsgebiete. So können zwei Parallelklassen in derselben Schule ganz unterschiedlich günstige Voraussetzungen haben. Aber auch aus demselben Einzugsbereich berichten LehrerInnen für aufeinander folgende Kohorten ganz unterschiedliche Zusammensetzungen.

WISS: Das Problem löst sich auf, wenn wir Längsschnitte anlegen, also dieselben SchülerInnen über mehrere Jahre begleiten, wie bei der LAU in Hamburg in der 5., in der 7. und in der 9. Klasse. Da vergleichen wir nicht nur die Leistung zu einem Punkt X, sondern wir können Lernzuwächse messen.

ELTN: Aber ein Test kann doch nie umfassend die Fähigkeiten einer Person messen, sondern nur ein Verhalten in einer bestimmten Situation, und das kann wechseln.

WISS: Ihr Einwand ist grundsätzlich richtig. Darum werden ja auch Vertrauensintervalle angegeben, um mögliche Schwankungen bei der Interpretation zu bedenken. Man muss aber zwischen Einzelfalldiagnose und Aussagen über Stichproben unterscheiden.
Der Einfluss von Tagesform, von Missverständnissen einzelner Aufgaben usw. gleichen sich in größeren Gruppen aus. Es ist also etwas Anderes, ob wir die Leistung eines Schülers oder die Leistung einer *Gruppe* von SchülerInnen, z. B. eines Bundeslandes oder einer Schulform miteinander vergleichen.

VERW: Hier haben die Verantwortlichen ja auch dazu gelernt. Wie schon gesagt: Die Daten werden nicht genutzt, um Urteile über einzelne Schulen, LehrerInnen oder gar SchülerInnen zu fällen.

MODR: Lassen Sie uns eine Zusammenfassung versuchen.
Unsere Diskussion hat erstens gezeigt, dass es sich bei Leistungsvergleichen um ein komplexes Problem handelt. Wir müssen bei der Evaluation von Unterricht drei Ebenen unterscheiden:

-- die *politische*: Wer bestimmt die Kriterien, wer kontrolliert die Umsetzung -- zentrale oder dezentrale Entscheidungsträger ?

- die *forschungsmethodische*: Welches Instrumentarium ist das aussagekräftigste und verlässlichste -- die standardisierte Messung oder das personengebundene Urteil ?

-- die *reformstrategische*: Auf welchen Wegen ist Schulentwicklung am ehesten in Gang zu bringen -- durch eine interne oder durch eine externe Evaluation ?

Zum zweiten gibt es ja nach Standpunkt unterschiedliche Einschätzungen der Funktion und der Wirkungen des Instruments "Leistungstests". Können Sie dazu die Ihnen wichtigsten Punkte noch einmal kurz auf den Nenner bringen?

VERW: Wir brauchen ein differenziertes Rechenschaftssystem, in dem Leistungsmessungen einen wichtigen, aber auch nur einen begrenzten Beitrag leisten können. Zudem ist zu bedenken, dass dieser Beitrag unterschiedlich ausfällt, je nachdem, ob sie den Lernerfolg von SchülerInnen, den Lehrerfolg von LehrerInnen oder die Wirksamkeit des Schulsystems insgesamt erfassen sollen. Mir kommt es darauf an, dass Fragen untersucht werden, zu denen es auch klare Handlungsoptionen gibt.

WIRT: Aus meiner Sicht sollten Leistungsmessungen in allen drei Funktionen eingesetzt werden:
- zur Beurteilung einzelner SchülerInnen,
  - um Fortschritte, aber auch Schwierigkeiten genauer erkennen und entsprechend gezielt fördern zu können und
  - um die formelle Berechtigung für den Besuch weiterführender Schulen vergleichbar und nachprüfbar zu erfassen;
- zur Bewertung des Lehrerfolgs einer LehrerIn, indem die Wirkungen verschiedener Ansätze unter vergleichbaren Bedingungen überprüft werden;
- zur Einschätzung von Stärken und Schwächen des Systems insgesamt bzw. einzelner Teilsysteme, z. B. Schularten.

ELTR: Ich wünsche mir, dass zunächst einmal die Ziele der Schule in Abstimmung aller beteiligten Gruppen geklärt und klarer umrissen werden: Wo liegen die Prioritäten für Allgemeinbildung heute?
Außerdem finde ich es wichtig, dass die Interpretation der Ergebnisse nicht in der Hand eines einzelnen Forscherteams liegt, sondern dass durch konkurrierende Interpetationen offen gelegt wird, wo die Daten in der Tat eindeutig sind und wo unterschiedliche Deutungen möglich sind. Es geht doch nicht an, dass die einen aus TIMSS folgern, Japan habe wegen seines problemlösenden Unterrichts so gut abgeschnitten und deshalb müsse eine anspruchsvollere "Lernkultur" entwickelt werden, während andere aus derselben Studie folgern, Japan habe so gut abgeschnitten, weil die privaten Jukus nachmittags durch schlichtes Pauken den Lernerfolg gesichert hätten.
Ich habe aus der Diskussion den Eindruck gewonnen, dass die die Testprogramme keine Selbstläufer sind und dass deshalb der sozialen

Kontrolle von Entscheidungen und Interpretationen eine Schlüsselfunktion zukommt.

LEHR: Mir liegt daran, dass die Ministerien jetzt nicht nur auf das Pferd "zentrale Leistungsmessung" setzen. Was nutzt es, wenn wir wissen, dass wir schlecht sind -- oder jedenfalls schlechter als andere? Vom wiederholten Wiegen wird die Sau auch nicht fetter...
Wir brauchen als erstes Mittel und Unterstützung für schulinterne Bestandsaufnahmen und vor allem für die Verbesserung der alltäglichen Arbeit und ihrer Bedingungen. Und dazu brauchen wir Fallstudien von erfolgreichen Schulen, vor allem von Schulen, die unter schwierigen Bedingungen, also wider Erwarten, erfolgreich sind. Von deren Praxis können andere Schulen am ehesten lernen, was sie konkret besser machen können.
Sorgen macht mir auch, ob mit den Ergebnissen zentraler Studien behutsam und differenziert genug umgegangen wird, ob allen Beteiligten der experimentelle Charakter dieser Testprogramme bewusst ist und ob sie ihren begrenzten Beitrag zur Aufklärung des Schulalltags recht einschätzen können. Diese kostspieligen Programme entfalten ja eine eigene Dynamik. Ich fürchte,
-- dass die WissenschaftlerInnen, um die Aufträge überhaupt zu bekommen, mehr versprochen haben, als sie halten können
-- dass die PolitikerInnen mehr erwarten, als solche Untersuchungen leisten können;
-- dass die InteressenvertreterInnen Einzelbefunde einseitig nutzen werden; und
-- dass gut gemeinte Aufgaben eine oberflächliche Normierung entfalten (Buchstaben statt Geist der definierten Anforderungen).

WISS: Akzeptiert man, dass menschliche Erkenntnis immer beschränkt, dass menschliches Handeln immer unzulänglich ist, dann erreichen die neuen Leistungstests ein erfreulich hohes Niveau des Möglichen. Deshalb verdienen sie eine faire Chance -- als Versuch und als ein Element in einem größeren Verbund. Denn schon in der IEA-Studie, mehr noch aber in PISA nähern wir uns Prüfungsformen, die im schulischen Alltag in der Breite und Differenziertheit oft nicht erreicht werden:
- ein großes Spektrum anspruchsvoller Leistungen;
- offene neben geschlossenen Aufgaben;
- zwei BewerterInnen statt nur einer Person (wie im Unterricht).

Ertragreich werden diese Studien aber nur sein, wenn die Programme langfristig angelegt sind, so dass Entwicklungen erkennbar werden. Und nur dann können die ForscherInnen auch aus den ersten Studien lernen und ihre Instrumente kontinuierlich verbessern.

MODR: Ich halte mal als Minimalanforderung fest:

Aussagen über Leistungen sind nur dann sinnvoll, wenn
a] die Maßstäbe klar definiert sind (z. B. als fachliche Anforderungen in Lehrplänen);
b] Vergleichswerte vorliegen (z. B. aus anderen Regionen oder Schulsystemen);
c] Leistungen auf die jeweiligen Voraussetzungen (z. B. fachliche Leistungen zu Beginn eines Schuljahres) bezogen werden;
d] Prozessmerkmale (z. B. methodischer Ansatz der Lehrperson) und
e] Randbedingungen (z. B. Anteil der ausgefallenen Stunden) erfasst werden.
Untersuchungen werden umso aussagekräftiger, je mehr Faktoren sie in diesen Dimensionen einbeziehen.
Andererseits hat dieser Anspruch auch Grenzen: zeitliche, finanzielle, aber auch den Schutz der Betroffenen vor einer Überlastung durch Befragungen und Tests. Insofern müssen auch pragmatische Lösungen erprobt werden können.
Umgekehrt heißt das: Bewertungen und Folgerungen aus solchen Studien können nur dann sachgerecht und fair diskutiert werden, wenn man sich der forschungsmethodischen Einschränkungen der Kompromisslösungen bewusst ist.

Zehn Empfehlungen für eine vertiefende Lektüre

[Titel mit * sind als Einführung bzw. Überblick gedacht, Publikationen mit # können LehrerInnen und Schulen helfen, ihre Arbeit selbst zu evaluieren]

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

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

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

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

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

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

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

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

Schratz, M., u. a. (1999): Qualitätsentwicklung. Vorschläge, Methoden, Instrumente. Manuskript Innsbruck/ Wien (erscheint 2000 im Beltz Verlag: Weinheim).

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