

Franz E. Weinert

PÄDAGOGIK

Leistungs- messungen in Schulen

3. Auflage



BELTZ

Weinert (Hrsg.)
Leistungsmessungen in Schulen

Leistungsmessungen in Schulen

Herausgegeben von Franz E. Weinert

3. Auflage

Beltz Verlag · Weinheim und Basel

Dr. Dr. h.c. *Franz E. Weinert* † arbeitete als Professor an den Universitäten Bamberg und Heidelberg. 1981 wechselte er als Gründungsdirektor an das Max-Planck-Institut für psychologische Forschung in München. Seine Hauptarbeitsgebiete sind die Psychologie des menschlichen Lernens, Probleme der differentiellen Entwicklung im Kindes-, Jugend- und Erwachsenenalter sowie Fragen der pädagogischen Psychologie.

Erstellt im Auftrag der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Kultusministerkonferenz).
Herausgeber und Autoren sind für die einzelnen Beiträge allein verantwortlich.

Dieses Buch ist auch als E-Book erhältlich
(ISBN 978-3-407-29318-3).

Alle Rechte, insbesondere das Recht der Vervielfältigung und Verbreitung sowie der Übersetzung, vorbehalten. Kein Teil des Werkes darf in irgendeiner Form (durch Fotokopie, Mikrofilm oder ein anderes Verfahren) ohne schriftliche Genehmigung des Verlages reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt und verbreitet werden.

Druck nach Typoskript
Lektorat: Peter E. Kalb
3., aktualisierte Auflage 2014

© 2001 Beltz Verlag · Weinheim und Basel, und Kultusministerkonferenz, Bonn
www.beltz.de

Druck: DruckPartner Rübelmann, Hemsbach
Umschlaggestaltung: Federico Luci, Köln
Umschlagabbildung: IFA-BILDERTEAM – Photex, Archiv Frankfurt/M.
Printed in Germany

ISBN 978-3-407-25690-4

Vorwort

Seit es Schulen gibt, werden Schulleistungen gemessen. Schulleistungsmessungen haben also eine lange Tradition, die zwar von vielen Reformpädagogen immer wieder in Frage gestellt und von Psychometrikern wegen der Fehleranfälligkeit oft kritisiert wurde, aber für Lehrer, Eltern und Schüler eine manchmal angenehme, manchmal unangenehme Selbstverständlichkeit war und blieb.

In Deutschland hat sich diese Situation seit wenigen Jahren drastisch verändert. Vergleichende Leistungsmessungen in Schulen sind plötzlich zu einem Thema öffentlicher Diskussionen geworden; Bildungspolitiker, Wirtschaftsmanager und Gesellschaftskritiker interessieren sich auf einmal für die Ergebnisse von Schulleistungstests; politische Parteien, Lehrerorganisationen und Elternverbände diskutieren kontrovers über die Bedeutung von Schulleistungsmessungen und deren Rolle für die Qualitätsverbesserung unseres Schulsystems.

Ein wichtiger Grund dafür war TIMSS, eine große internationale Vergleichsstudie über Leistungen in der Mathematik und in den naturwissenschaftlichen Unterrichtsfächern. Deutsche Schüler – so demonstrierten es die empirischen Ergebnisse – konnten im Durchschnitt nur mittelmäßige Leistungen erzielen, waren also weniger erfolgreich als viele geglaubt und gehofft hatten. Wie immer bei schlechten Nachrichten waren die ersten Reaktionen voller Hektik und Widerspruch: Die Untersuchungsbefunde wurden sowohl bagatellisiert als auch dramatisiert, die verwendeten Messinstrumente gerieten – völlig unbegründet – in die Kritik; manche suchten voreilig nach Schuldigen, andere holten ihre angestaubten schulorganisatorischen Vorschläge als Patentrezepte wieder aus der Schublade; es gab aber auch viele nachdenkliche Kommentare über die möglichen Gründe für das schlechte Abschneiden deutscher Schüler und für die notwendigen Reformen im deutschen Bildungswesen. Die kritischen, aber konstruktiven Stimmen gewannen schließlich die Oberhand. Vielen wurde bewusst, dass sich die Qualität des Unterrichts verbessern muss, sollen langfristig die Leistungen möglichst vieler Schüler und damit der Schule insgesamt erhöht werden. Dass eine solche Innovation in die Zukunft des Landes nicht umsonst zu haben ist, wurde in der Diskussion ebenso deutlich wie die Tatsache, dass schnelles Geld allein wenig helfen dürfte.

Besonders auffallend und beeindruckend war, dass die Resultate der TIMS-Studie kaum für parteipolitischen Streit gesorgt haben. Im Gegenteil: Die Konferenz der Kultusmi-

nister bemühte sich über alle Parteien und Koalitionsgruppierungen hinweg nachdrücklich um eine Versachlichung der öffentlichen Diskussion und um eine "Pädagogisierung" der Überlegungen zur Qualitätsverbesserung deutscher Schulen. Diese grundlegende Einstellung führte auch zu der Anregung, ein Buch zur Information der interessierten Lehrer und Eltern, aber auch für eine breite, bildungspolitisch aufgeschlossene Öffentlichkeit herauszugeben. Ziel einer solchen Publikation sollte es sein, eine gewisse Ordnung in die vielen neuen schulischen Leistungsstudien zu bringen, die wichtigen methodischen Grundlagen der Leistungsmessung zu klären, deren Vorteile und Gefahren sachlich zu analysieren und nicht zuletzt auch den pädagogischen Nutzen zu überprüfen. Das vorliegende Buch kann als Ergebnis dieser Bemühungen aufgefasst werden. Zahlreiche prominente Autoren mit jeweils unterschiedlichen bildungspolitischen Profilen, aber mit dem gemeinsamen Bemühen um wissenschaftliche Redlichkeit und pädagogische Sachlichkeit waren bereit, diese Aufgabe zu übernehmen.

Von Anfang an war geplant, in erster Linie ein Buch für die Hand von Lehrerinnen und Lehrern zu schreiben. Die einzelnen Kapitel sollten auf drängende, offene, zum Teil aber noch gar nicht gestellte Fragen möglichst verständliche Antworten geben. Das ist in den Augen der Leser vermutlich nicht immer in der wünschenswerten Weise gelungen. Die schwierigen Probleme bei der Konstruktion von Leistungstests machen methodische und statistische Erläuterungen zwingend notwendig, die Internationalisierung der vergleichenden Schulleistungsmessung fördert den gemeinsamen Gebrauch vieler Fremdwörter, die große Anzahl neuer regionaler, nationaler und internationaler Untersuchungen erschwert selbst Fachleuten den Überblick. Beklagen werden manche Lehrerinnen und Lehrer, dass zu wenig auf die konkreten Nutzungsmöglichkeiten der Resultate solcher Studien in der eigenen Schule und durch den individuellen Lehrer eingegangen wird. Das ist gegenwärtig aber sehr schwierig! Deutschland hat sich jahrzehntelang an internationalen Leistungsvergleichen nicht beteiligt; eine pädagogische Kultur der Qualitätsentwicklung von Schulen auf der Grundlage realistischer Leistungsbewertungen konnte sich also nicht entfalten. Es gehört zu den Aufgaben des vorliegenden Buches, durch die Darstellung der aktuellen Entwicklungen (vor allem der zwei großen internationalen Studien: TIMSS und PISA) und der erkennbaren Perspektiven zu einer Schließung dieser Lücke beizutragen.

Daraus ergeben sich Konsequenzen für die Auswahl und Aufeinanderfolge der Kapitel. Nach einführenden Darstellungen des aktuellen Diskussionsstandes, der Notwendigkeiten und Probleme schulischer Leistungsmessung – auch im Kontext des alltäglichen Unterrichts – folgen Überblickskapitel über die Methoden und Inhaltsbereiche, über nationale und internationale Studien. Abschließend wird aus verschiedenen Perspektiven versucht, die Notwendigkeit, die Problematik und den Nutzen vergleichender Leistungsmessungen in Schulen zu analysieren. Da es sich um kein Lehrbuch im engeren Sinne handelt, kann man sich sehr verschiedene Lesepfade vorstellen. Man kann sich beispielsweise zuerst auf die Darstellung besonders interessierender Studien konzentrieren und sich anschließend mit methodischen und pädagogischen Fragen beschäftigen. Oder man kann mit der aktuellen Diskussion beginnen und dann einzelne Schwerpunkte und Vertiefungen wählen. Viele Varianten erscheinen möglich.

Auch wenn die Lektüre der einzelnen Kapitel gelegentlich schwierig sein dürfte, so müsste der damit verbundene Erwerb professioneller Kompetenzen, fundierten Wissens und eines vielfältigen Problembewusstseins im Bereich der schulischen Leistungsmessung eine hinreichende Kompensation darstellen. Eine sachkundige und engagierte Beteiligung der Lehrerschaft an der öffentlichen wie an der professionellen Diskussion über die weitere Entwicklung der schulischen Leistungsmessung ist nämlich von großer Bedeutung. Nicht empfohlen werden kann das Buch all jenen, die davon überzeugt sind, dass eine Verdoppelung des Bildungsbudgets alle Probleme der deutschen Schulen lösen würde; die sich von einseitigen pädagogischen Heilslehren eine bessere schulische Zukunft versprechen und die vor allem äußere Verhältnisse für die gegenwärtigen unterrichtlichen Defizite verantwortlich machen.

Ein Buch, das in der Unübersichtlichkeit öffentlicher Diskussionen zu einem umstrittenen Thema vielen interessierten Bürgern des Landes Klärung und Orientierung bringen soll, hat notwendigerweise eine komplizierte Entstehungsgeschichte. Zahlreiche Personen waren an der Planung und Realisierung des vorliegenden Bandes beteiligt. Danken möchte ich deshalb vor allem

- vielen praktizierenden Lehrern und Wissenschaftlern, mit denen ich im Anfangsstadium Gespräche über das Projekt führen konnte;
- der Kultusministerkonferenz und ihren Mitgliedern für die anregende, aber völlig neutrale und zurückhaltende Förderung des Buches;
- den Autoren für ihre kompetente, engagierte und tolerante Zusammenarbeit;
- den Mitgliedern einer ad hoc zusammengestellten Expertengruppe – Frau Barbara Basko, Herrn StD Johann Glötzner, Frau Wilma Maier-Michels Päd.M.A. und Herrn Dipl.Soz.Dr. Jürgen Bofinger –, welche die Erstfassungen der einzelnen Kapitel kritisch gelesen und ausführlich kommentiert haben;
- Frau Wilma Maier-Michels, die über alle notwendigen wissenschaftlichen, pädagogischen und redaktionellen Kompetenzen verfügte, um das Buch inhaltlich und formal zu gestalten;
- Frau Heidi Schulze, welche die Druckvorlage in einer sehr professionellen Weise erstellte und
- dem Verlag Beltz, der durch die äußere Form des Buches und durch den sehr knapp kalkulierten Preis zur Verbreitung des Werkes beitragen dürfte.

Danken möchte ich schließlich im Voraus allen Leserinnen und Lesern, die sich anhand des Buches mit einer schwierigen, aber extrem wichtigen Thematik unseres Bildungswesens auseinandersetzen. Auch wenn der Text nicht allen persönlichen Erwartungen gerecht werden kann, so hoffe ich doch sehr, dass er das Interesse Vieler findet und damit für die Entwicklung unserer Schulen von Nutzen ist.

Vorwort zur 3. Auflage

Das von Franz Emanuel Weinert kurz vor seinem Tode im Jahre 2001 herausgegebene und ein Jahr später in 2. Auflage erschienene Werk liegt jetzt in der 3. Auflage vor. Dies zeugt von der anhaltenden Aktualität der Thematik und macht zugleich deutlich, dass dieser Text nach wie vor eine Lücke füllt, die offenbar nicht durch andere Darstellungen zu schließen ist. Das Buch wurde in einer Zeit konzipiert, als Leistungsmessungen in Schulen in Deutschland noch weitaus weniger selbstverständlich waren als heute. Die TIMS-Studie war die erste Studie dieser Art, nachdem sich Deutschland für lange Zeit aus internationalen Vergleichsstudien verabschiedet hatte. Die Erwartung, zu den in Bildung und Forschung herausragenden Nationen zu gehören, wurde damals nachhaltig erschüttert. Das Thema Leistungsmessung ist im Zuge dieser ersten Leistungsvergleiche intensiv diskutiert worden. Vielen interessierten Nicht-Wissenschaftlern fehlten aber die inhaltlichen und methodischen Kenntnisse, um sich fundiert mit den damit verbundenen Fragen auseinandersetzen zu können.

Vor diesem Hintergrund hatte sich F. E. Weinert seinerzeit entschlossen, ein Buch über Leistungsmessungen in Schulen herauszugeben, mit dem insbesondere Lehrerinnen und Lehrer angesprochen werden sollten. Dazu war er in besonderer Weise prädestiniert. Weinert gehörte mit seiner Tätigkeit als Hochschullehrer, Gründer und langjähriger Leiter des Max-Planck-Instituts für psychologische Forschung zu den renommiertesten Vertretern der schulischen Bildungsforschung. Er war selbst als Lehrer ausgebildet und vor seiner wissenschaftlichen Laufbahn viele Jahre im Schuldienst tätig gewesen. Sein Lebenswerk ist geprägt durch die wissenschaftliche Auseinandersetzung mit Schule und Unterricht, die er nicht nur in einen breiteren entwicklungspsychologischen Rahmen stellte, sondern zugleich immer auch auf ihren Wert für die schulische Praxis hin betrachtete.

Ziel des Buches war es seinerzeit, "eine gewisse Ordnung in die vielen neuen schulischen Leistungsstudien zu bringen, die wichtigen methodischen Grundlagen der Leistungsmessung zu klären, deren Vorteile und Gefahren sachlich zu analysieren und nicht zuletzt auch den pädagogischen Nutzen zu überprüfen" (aus dem Vorwort zur 1. Auflage). Zu diesem Zweck umfasst das Buch einführende Darstellungen des aktuellen Diskussionsstandes und der Notwendigkeiten und Probleme schulischer Leistungsmessungen, Überblickskapitel über Methoden und Inhaltsbereiche, über

nationale und internationale Studien; abschließend wird versucht, die Notwendigkeit, die Problematik und den Nutzen vergleichbarer Leistungsmessungen in Schulen zu analysieren. Es ist mit Sicherheit dem herausragenden Renommee von Weinert zu verdanken, dass für dieses Buch die führenden Vertreter der deutschsprachigen Bildungsforschung gewonnen werden konnten. Es sind vielfach die Wissenschaftlerinnen und Wissenschaftler, die für das jeweilige Thema “stehen”, es also maßgeblich mitgeprägt haben. Und es spricht – was bei einem derartigen Projekt keinesfalls selbstverständlich ist – für das editorische Geschick des Herausgebers, dass ein Buch mit Beiträgen entstanden ist, die nicht nur thematisch repräsentativ und ausgewogen sowie wissenschaftlich und fachlich gehaltvoll, sondern auch gut verständlich sind. Dazu hat sicher auch der nicht unerhebliche redaktionelle Aufwand beigetragen, der bei der Konzeption und Produktion des Buches betrieben wurde.

In den folgenden 13 Jahren hat sich im Bereich der Leistungsmessung im deutschen Sprachraum viel getan. Um nur einige Punkte zu nennen:

- Entwicklung und fortlaufende Überprüfung der Zielerreichung von Bildungsstandards in Deutschland durch das IQB, inzwischen auch in Österreich (durch das bifie) und in der Schweiz (Stichworte: HarmoS, Lehrplan 21);
- Ausdehnung der auf den Bildungsstandards basierenden, von der Landauer Forschungsgruppe initiierten Vergleichsarbeiten auf mehrere Klassenstufen und auf alle 16 Bundesländer;
- Einrichtung von Qualitätsagenturen in verschiedenen Bundesländern;
- Fortsetzung der internationalen Leistungsstudien der IEA und OECD, insbesondere TIMSS, PISA und IGLU (PIRLS);
- Ausdehnung der Kompetenzmessung über die Schule hinaus, beispielhaft realisiert im “Programme for the International Assessment of Adult Competencies” (PIAAC-Studie, 2013);
- Entwicklung von Ansätzen zur Vermittlung diagnostischer Kompetenzen durch das Projekt UDiKom der KMK mit Modulen zu Bildungsmonitoring, Vergleichsarbeiten, Individualdiagnostik und Unterrichtsdiagnostik.

Dazu kommen zahlreiche Forschungsprojekte, einschlägige Schwerpunktprogramme der Deutschen Forschungsgemeinschaft (insbesondere “Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen”), Einrichtung universitärer Forschungsschwerpunkte und Programme zur Nachwuchsförderung (Graduiertenschulen wie das von der Deutschen Forschungsgemeinschaft geförderte Graduiertenkolleg “Unterrichtsprozesse” in Landau) und eine Vielzahl von Publikationen zu Themen der Leistungsmessung, die hier auch nicht annähernd skizziert werden können. Der Deutsche Bildungsserver sowie die Portale der Landesinstitute und Qualitätsagenturen informieren umfassend über aktuelle Projekte, Werkzeuge, Material und Publikationen.

Allerdings zeichnet sich inzwischen auch eine Trendwende ab. Es zeigt sich eine gewisse Sättigung, was die Information über Schulleistungen auf der Basis von Lernstandserhebungen und Vergleichsarbeiten anbelangt. Wir wissen immer besser und genauer Bescheid über die Leistungen unserer Schülerinnen und Schüler. Das Interesse der nationalen Bildungspolitik und Bildungsforschung richtet sich demzufolge zunehmend auf die Bedingungen und Prozesse, die für Leistungsunterschiede zwischen Schulen, Klassen sowie Schülerinnen und Schülern verantwortlich sind. Hier war die Hattie-Studie zweifellos ein starker Katalysator. Auch das im deutschsprachigen Raum inzwischen sehr verbreitete, im Auftrag der KMK für die Schulpraxis entwickelte Werkzeug EMU (Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung, siehe www.unterrichtsdiagnostik.info) fokussiert auf die Lehr-Lern-Prozesse, nicht auf die Leistungen als Produkte.

Auf der anderen Seite führt die zunehmende Fokussierung auf “Kompetenzorientierung” in vielen Bundesländern und in der Schweiz – auf dem Umweg über Fragen der Unterrichtsqualität – indirekt wiederum zu den in Weinerts Buch behandelten grundlegenden Fragen der Leistungsmessung.

Obwohl die Akzeptanz von Leistungsmessungen und Kompetenzdiagnostik und vermutlich auch der entsprechende Informationsstand im letzten Jahrzehnt erheblich zugenommen haben, ja inzwischen Alltag geworden sind, stellen sich einige grundlegende Fragen immer wieder, sind gewissermaßen zeitlos. Auf viele dieser Fragen gibt das Buch Auskunft, und zwar aus Sicht der Personen, die die wissenschaftliche und fachliche Diskussion auf dem jeweiligen Gebiet wesentlich mitbestimmt, wenn nicht sogar entscheidend gestaltet haben. Weinert selbst steckt mit dem von ihm verfassten Anfangs- und Schlusskapitel den Rahmen für das Buch ab, wobei er Leistungsmessung als eine in ihren Anfängen (und sicher auch noch zum Zeitpunkt des erstmaligen Erscheinens dieses Buches) durchaus umstrittene, aber auf lange Sicht (und vielleicht schon heute) ganz unumstrittene Selbstverständlichkeit sieht.

Dabei zeigt sich Weinerts Fähigkeit, wissenschaftlich und gesellschaftlich bedeutsame Themen und Trends zu erkennen und aufzugreifen, sie eingängig und gleichzeitig ausgewogen darzustellen, prägnant zu umreißen und einprägsam zu vermitteln. Aus unserer Sicht sind diese Qualitäten auch dem Buch als Ganzem zugute gekommen.

Deshalb sehen wir es nach wie vor als eine hervorragende und in seiner Art einzigartige Quelle, wenn man sich fundiert mit Fragen der Leistungsmessung und ihren Hintergründen auseinandersetzen will.

Inhalt

Vorwort	3
Vorwort zur 3. Auflage	6
Inhaltsverzeichnis	9
Verzeichnis thematisch relevanter Institutionen und Untersuchungen	13
Kapitel 1 Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit <i>Franz E. Weinert</i>	17
Kapitel 2 Kontroversen um die Schulleistungsmessung in Deutschland. Eine fiktive Diskussion über Positionen und Perspektiven in verteilten Rollen <i>Hans Brügelmann</i>	33
Kapitel 3 Alltägliche Leistungsbeurteilung durch Lehrer <i>Friedrich-Wilhelm Schrader & Andreas Helmke</i>	45
Kapitel 4 Bezugsnormen und schulische Leistungsbeurteilung <i>Falko Rheinberg</i>	59
Kapitel 5 Schulleistungen – Leistungen der Schule <u>oder</u> der Schüler? <i>Franz E. Weinert</i>	73

Kapitel 6

Standardisierte Schulleistungsmessungen

Kurt A. Heller & Ernst A. Hany 87

Kapitel 7

Wie misst man Schulleistungen?

Karl Josef Klauer 103

Kapitel 8

Qualitätskriterien für die standardisierte Messung von Schulleistungen.
Kann eine (vergleichende) Messung von Schulleistungen
objektiv, repräsentativ und fair sein?

Karl-Heinz Arnold 117

Kapitel 9

Messung von Schulleistungen im Primar- und Sekundarbereich

Rainer H. Lehmann 131

Kapitel 10

Schulleistungen im Bereich der muttersprachlichen Bildung

Wolfgang Schneider 143

Kapitel 11

Schulleistungen im Bereich der mathematischen Bildung

Elsbeth Stern & Ilonca Hardy 153

Kapitel 12

Schulleistungen im Bereich der naturwissenschaftlichen Bildung

Reinders Duit, Peter Häußler & Manfred Prenzel 169

Kapitel 13

Schulleistungen im moralisch-wertbildenden Bereich.
Das Beispiel Lebensgestaltung-Ethik-Religionskunde (LER)
in Brandenburg

Sabine Gruehn & Kai Schnabel 187

Kapitel 14

Fächerübergreifende Kompetenzen: Konzepte und Indikatoren

Eckhard Klieme, Petra Stanat & Cordula Artelt 203

Kapitel 15 Leistungen im Bereich der beruflichen Bildung <i>Gerald A. Straka</i>	219
Kapitel 16 Jenseits von TIMSS: Messungen sprachlicher Kompetenzen, komplexe Längsschnittstudien und kulturvergleichende Analysen. Ergebnisse und Perspektiven ausgewählter Leistungsstudien <i>Andreas Helmke & Friedrich-Wilhelm Schrader</i>	237
Kapitel 17 Internationale Schulleistungsforschung: Ihre Entwicklungen und Folgen für die deutsche Bildungslandschaft <i>Wilfried Bos & T. Neville Postlethwaite</i>	251
Kapitel 18 TIMSS – Third International Mathematics and Science Study. Dritte internationale Mathematik- und Naturwissenschaftsstudie <i>Olaf Köller, Jürgen Baumert & Wilfried Bos</i>	269
Kapitel 19 PISA – Programme for International Student Assessment. Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren <i>Jürgen Baumert, Cordula Artelt, Eckhard Klieme & Petra Stanat</i>	285
Kapitel 20 Leistungsmessung und die Professionalität des Lehrerberufs <i>Jörg Schlömerkemper</i>	311
Kapitel 21 Die Bedeutung vergleichender Schulleistungsmessungen für die Qualitätskontrolle und Qualitätsentwicklung von Schulen und Schulsystemen <i>Rainer Peek</i>	323
Kapitel 22 Was bringt die vergleichende Leistungsmessung für die pädagogische Arbeit in Schulen? <i>Hans-Günter Rolff</i>	337

Kapitel 23

Perspektiven der Schulleistungsmessung
– mehrperspektivisch betrachtet

<i>Franz E. Weinert</i>	353
Literaturverzeichnis	367
Personenregister	383
Sachregister	387
Autoren	395

Verzeichnis thematisch relevanter Institutionen und Untersuchungen

Institutionen

BLK	Bund-Länder-Kommission
DFG	Deutsche Forschungsgemeinschaft
DIPF	Deutsches Institut für Internationale Pädagogische Forschung
ETS	Educational Testing Service
IAEP	International Assessment of Educational Progress
IBE	International Bureau of Education
IEA	International Association for the Evaluation of Educational Achievement
IPN	Institut der Pädagogik der Naturwissenschaften
KMK	Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (kurz: Kultusministerkonferenz)
MPIB	Max-Planck-Institut für Bildungsforschung
MPIPF	Max-Planck-Institut für Psychologische Forschung
OECD	Organisation for Economic Cooperation and Development (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung)
UNESCO	United Nations Educational, Scientific and Cultural Organisation

Untersuchungen

BIJU

Bildungsverläufe und psychosoziale Entwicklung im Jugendalter

1991 begonnene Längsschnittuntersuchung ab Jahrgangsstufe 7, in der auch *TIMSS*-Aufgaben verwendet wurden; Durchführung: Max-Planck-Institut für Bildungsforschung (MPIB), Berlin.

CIVIC EDUCATION STUDY

Studie der *IEA* zur politischen Bildung im Schuljahr 1998/99, Schüler der 8. Jahrgangsstufe betreffend, in insgesamt 23 Nationen. Verantwortlich für den nationalen Teil: MPIB Berlin, für die internationale Koordination: Humboldt-Universität zu Berlin.

DESI

Deutsch-Englisch-Schülerleistungen-International

Als KMK-Projekt deutsche Ergänzungsstudie zu *PISA*, die die aktive Beherrschung der deutschen Sprache und des Englischen als Fremdsprache zum Gegenstand hat (Jahrgangsstufe 9). Datenerhebung 2003 und 2004; koordiniert vom Deutschen Institut für Internationale Pädagogische Forschung (*DIPF*).

Hamburger Aufsatzstudie

Deutscher Beitrag zur *IEA-Written Composition Study*; 1989 in allgemein- und berufsbildenden Hamburger Schulen als Stichprobenerhebung realisiert durch die Universität Hamburg.

Hamburger Lesestudie

Deutsche Teilstudie der *IEA-Reading Literacy Study (IRLS = Internationale Lesestudie)* bei 9- und 13jährigen, im Jahr 1991 durchgeführt in den alten und neuen Bundesländern von der Universität Hamburg; Ergänzungsuntersuchungen bei deutschsprachigen Minderheiten in Europa.

IGLU

Internationale Grundschul-Lese-Untersuchung

Deutsche Teilstudie der *IEA-PIRLS*, ergänzt um Mathematik und einige naturwissenschaftliche Komponenten (*IGLU/E*); für 2001 geplante Erhebung in 4. Klassen (Stichproben) durch die Universität Hamburg. Alle 16 Bundesländer werden sich an *IGLU* beteiligen, 13 an *IGLU/E*.

Internationale Lesestudie (siehe *Hamburger Lesestudie*)

LAU*Lern-Ausgangslagen-Untersuchung*

Längsschnittuntersuchung (Vollerhebung) in Hamburg zum Lernfortschritt in den Fächern Deutsch (Leseverständnis und Rechtschreibung), Mathematik, erste Fremdsprache und in fächerübergreifenden Kompetenzen; durchgeführt seit 1996 (Beginn 5. Klasse) in zweijährigem Abstand von der Humboldt-Universität zu Berlin.

LER-Studie*Lebensgestaltung-Ethik-Religion*

wurde 1996 als verbindliches Unterrichtsfach der Sekundarstufe I in Brandenburg eingeführt und wissenschaftlich begleitet. Untersuchung (Lehrer-, Schüler- und Schulleiterbefragungen) zur Praxis und Wirkung des Unterrichts im Lernbereich *LER* in Brandenburg; durchgeführt 1999 von der Humboldt-Universität zu Berlin.

LOGIK*Longitudinalstudie zur Genese individueller Kompetenzen*

Von 1984 bis 1993 wurde die kindliche Entwicklung (Einstiegsalter 3 bis 4 Jahre) in den Bereichen Intelligenz, Denken, Gedächtnis, Lese- und Rechtschreibleistung, mathematisch-naturwissenschaftliches Verständnis, Motivation, soziale Fähigkeiten und moralisches Urteil/moralische Motivation untersucht. In jedem Untersuchungsjahr wurden die Kinder dreimal beobachtet, befragt und getestet; zeitweise Verquickung mit der ebenfalls vom Max-Planck-Institut für Psychologische Forschung (MPIPF) in München durchgeführten *SCHOLASTIK*-Studie.

MARKUS*Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext*

Vollerhebung im Mai 2000 in Rheinland-Pfalz zu den Mathematikleistungen der Schüler in der 8. Jahrgangsstufe und zu Unterrichtsmerkmalen, Schulqualität sowie zu Lernvoraussetzungen und zum persönlichen Hintergrund der Schüler; durchgeführt von der Universität Koblenz-Landau und dem Zentrum für empirisch-pädagogische Forschung Landau.

Münchener Hauptschulstudie

Erhebung der Mathematikleistungen und leistungsrelevanter Motive und Einstellungen vom Beginn der 5. bis Ende der 6. Jahrgangsstufe; wesentlich erweiterter deutscher Beitrag zu "*Classroom Environment Study*" der *IEA*; vom Max-Planck-Institut für Psychologische Forschung München als Längsschnittuntersuchung 1983–1985 durchgeführt.

PIRLS*Progress in International Reading Literacy Study*

Weiterentwicklung der Internationalen Lesestudie der *IEA* (*IRLS*); deutsches Äquivalent: *IGLU*.

PISA*Programme for International Student Assessment**(Programm zur Internationalen Bewertung von Schülerleistungen)*

Laufende OECD-Studie (1998–2007) zur Lesekompetenz, zur mathematisch/naturwissenschaftlichen Grundbildung und zu fächerübergreifenden Kompetenzen mit vielfältigen Indikatoren für Lernergebnisse und ihre Bedingungen bei 15jährigen Schülern; federführend für die wesentlich erweiterte deutsche Teilstudie: MPIB, Berlin.

QUASUM*Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik*

Eine 1999 in Brandenburg in den Jahrgangsstufen 5 und 9 erfolgte Untersuchung, die auch Merkmale des Unterrichts, den Schulalltag und die Lebens- und Lernumwelt der Schüler einbezog; durchgeführt von der Humboldt-Universität zu Berlin.

SCHOLASTIK*Schulorganisierte Lernangebote und Sozialisation von Talenten, Interessen und Kompetenzen*

Vom Max-Planck-Institut für Psychologische Forschung in München durchgeführte Untersuchung zur Entwicklung während der Grundschulzeit: Mathematik- und Deutschleistungen und motivationale Schülermerkmale.

(Ein wesentlicher Teil von *SCHOLASTIK* wurde in vietnamesischen Grundschulklassen wiederholt.)

TIMSS*Third International Mathematics and Science Study**(Dritte Internationale Mathematik- und Naturwissenschaftsstudie)*

In Deutschland Ende der Schuljahre 1994/95 und 1995/96 in der Sekundarstufe I und II durchgeführte Untersuchung analog der *IEA*-Studie. Unterricht, Lehrer, Schulen und außerschulische Lebensumwelt sowie psychosoziale und individuelle Schülermerkmale waren im Erhebungsdesign enthalten; Querschnittvergleich von 7. und/oder 8. Jahrgangsstufe sowie Federführung: MPIB, Berlin.

WALZER*Wirkungsanalyse der Leistungsevaluation: Zielerreichung, Ertrag für die Bildungsqualität der Schule und die Rückmeldung von Evaluationsergebnissen*

Eine im Anschluss an die *MARKUS*-Studie stattfindende Evaluationsstudie zur Wirkung von Ergebnisrückmeldungen an Lehrkräfte und Schulen und zu den Bedingungen, unter denen diese Rückmeldungen zur Verbesserung der Qualität von Schule und Unterricht genutzt werden. Projekt im Rahmen des auf sechs Jahre angelegten *DFG*-Schwerpunktprogramms "*Bildungsqualität von Schule*". Durchführung: Universität Koblenz-Landau.

KAPITEL 1

Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit

Franz E. Weinert

In bildungspolitischen Diskussionen werden pädagogische Konzepte, Modelle und Positionen oft mit Hilfe eines sehr globalen Maßstabes bewertet und dementsprechend pauschal etikettiert. Besonders prominent geworden ist die häufig nicht diskriminativ, sondern diskriminierend gemeinte Zuschreibung von bewertenden Kategorien wie “traditionell”, “konservativ”, “veraltet” im Gegensatz zu “fortschrittlich”, “progressiv” und “modern”.

Der Erwerb oder gar die Vermittlung von Wissen, die Steuerung und Kontrolle des Unterrichts durch den Lehrer, die automatisierende Einübung von Fertigkeiten, die extrinsische Lernmotivierung der Schüler und die meisten leistungsbezogenen Bildungsziele gelten vielen als konservativ, während die didaktische Moderation autonomer Lerngruppen, die selbständige (nicht selten spielerische) Beschäftigung mit interessanten Lerngegenständen, die intrinsische Motivation, jegliche Form von Gruppenarbeit und die grundlegenden reform-pädagogischen Bildungsziele wie persönliche Autonomie oder soziale Partizipation als modern angesehen werden.

Gegen eine solche Klassifikation ist natürlich nichts einzuwenden, wenn mit der Bezeichnung “modern” gemeint wäre, dass ein didaktisches Konzept relativ neu ist, dass ein Bildungsziel aus den aktuellen Bedürfnissen einer sich wandelnden Gesellschaft herleitbar ist, dass sich neuere Unterrichtskonzepte gegenüber konkurrierenden älteren Modellen als überlegen erwiesen haben. Dem ist aber nicht so! “Traditionelles” wird von den einen als schlecht, als rückständig, als etwas zu Überwindendes und zu Ersetzendes betrachtet, von anderen aber als erhaltenswert, zeitgeistresistent und oft als besonders notwendig angesehen. Ähnliches geschieht mit neuen pädagogischen Ideen, die von vielen als fortschrittlich gefeiert, von anderen als modernistischer Firlefanz verteuelt werden.

Was solchen pädagogischen Kontroversen häufig fehlt, ist das gemeinsame Bewusstsein, dass es für Schulen verschiedene Bildungsziele gibt, die durch jeweils unterschiedliche Lernprozesse, variable didaktische Bedingungskonstellationen und differierende Modi des Lehrer- wie des Schülerverhaltens am besten erreicht werden können. Die

ergebnisoffene Analyse allgemeiner oder spezifischer Zusammenhangsmuster und die Suche nach differentiellen Methoden erfolgreichen Unterrichtens werden nicht selten durch die Selbstgewissheit und den Ausschließlichkeitsanspruch der jeweiligen Anhänger einer bestimmten pädagogischen Position ersetzt, wenn nicht sogar verhindert.

Dabei macht eine unvoreingenommene Analyse des empirischen Forschungsstandes mehr als deutlich, dass es kaum ein pädagogisches oder didaktisches Konzept gibt, das für die Erreichung unterschiedlicher Bildungsziele gleichermaßen gut geeignet wäre, dass es kaum eine Unterrichtsmethode oder eine Lehrstrategie gibt, die – einseitig und ausschließlich angewandt – neben den erwünschten Wirkungen nicht auch unerwünschte Nebeneffekte zeigt und dass sich kaum ein pädagogisch-psychologisches Werkzeug (z. B. ein diagnostisches Verfahren) finden lässt, das ohne theoriegeleitete, verfahrenssensible und kritische Interpretation “objektive”, sich selbst auslegende Informationen liefern würde.

Fatalerweise ist die vergleichende Leistungsmessung – mehr noch als das Leistungsprinzip als solches – in die bewertende und pauschalierende pädagogische “Entweder-Oder-Klassifikation” geraten. Für manche, die sich selbst als humanistisch und idealistisch oder aber als anti-bürgerlich und fortschritts-skeptisch verstehen, sind Leistungen, Leistungsanforderungen und Leistungsmessungen kinderfeindliche, antireformpädagogische, ökonomisch instrumentalisierbare Kontrollmechanismen zur Disziplinierung von Schulen, Lehrern und Schülern. Leistungen und ihre methodisch zuverlässige Erfassung werden entweder offen als konservative Relikte einer veralteten Pädagogik abgelehnt, oder es werden Leistungsmessungen gefordert, die weder Leistungen enthalten noch Messungen darstellen. Solche Behauptungen werden von leistungsorientiert eingestellten Pädagogen und Bürgern als romantisch, idealistisch oder ideologisch kritisiert.

Für sie bedeutet Leistung die Manifestation eines menschlichen Grundbedürfnisses, eine Möglichkeit der individuellen Selbstverwirklichung durch Erfahrung eigener Selbstwirksamkeit; Leistungsanforderungen werden als individuelle Herausforderungen und als notwendige Bedingungen des sozio-kulturellen Fortschritts verstanden; Leistungsmessungen und die damit verbundenen Möglichkeiten des Leistungsvergleichs zwischen verschiedenen Schülern, Klassen, Schulen, Schulformen und Schulsystemen stellen für sie wichtige Bedingungen der Möglichkeit zur rationalen Begründung bildungspolitischer Entscheidungen, zur Steuerung der Schulentwicklung, zur Verbesserung der Schulqualität und zur reflexiven Vergewisserung des Verhältnisses von Anspruch und Wirklichkeit bei Politikern, Lehrern, Eltern und Schülern dar.

Beide Positionen verfügen über hinreichend viele gesellschaftlich anerkannte Argumente zur Begründung, Rechtfertigung und auch Immunisierung der eigenen Auffassung und zur Infragestellung der jeweils anderen Perspektive. Der Streit kann deshalb mit Unterstellungen über die “eigentlichen Absichten”, die “unausgesprochenen Interessen” und die zu befürchtenden langfristigen “pädagogischen Nebeneffekte” geführt werden. Hinter solchen Schwarz-Weiß-Diskussionsstrategien verflüchtigen sich leicht die tatsächlichen Ziele, der wahrscheinliche Nutzen und die möglichen Gefahren bestimmter Leistungsanforderungen und Leistungsmessungen.

Es gibt manche Anzeichen dafür, dass sich die unterschiedlichen Einschätzungen gegenüber vergleichenden Leistungsmessungen in Schulen seit kurzem sogar verstärkt haben. Betonen die einen immer nachdrücklicher die reformpädagogischen Ideen einer "guten entschulten Kindheit" (Gardner, 1993), die Vorteile einer stressfreien Persönlichkeitsentwicklung unter anregenden schulischen Bedingungen, die Autonomie des Lernens in kleinen Gemeinschaften von Lernenden als Grundlage einer "höheren" Form von Leistungstüchtigkeit (obwohl oder gerade weil bestimmte Leistungsdispositionen **nicht** erworben werden), so beschwören die anderen wachsende globale wirtschaftliche Konkurrenzkämpfe, die Notwendigkeit des schulischen Leistungsprinzips für das Leben in einer Wissensgesellschaft und die Rolle des kollektiven wie des individuellen Leistungsniveaus für das persönliche wie für das gesellschaftliche Wohlergehen.

Es gibt also bereits seit einiger Zeit scharfe Konturen einer pädagogischen Auseinandersetzung zwischen denen, die Leistung als den primären Auftrag von Schule ansehen und jenen, die zwar auch Leistungen für mehr oder minder bedeutsam halten, ihnen aber keine Priorität im schulischen Aufgabenkatalog einräumen wollen. Auf diese gesellschaftliche und pädagogische Situation traf TIMSS in den späten neunziger Jahren.

Reaktionen auf TIMSS

Die empirischen Ergebnisse der *Third International Mathematics and Science Study* (TIMSS) sowie die bildungspolitischen, pädagogischen und öffentlichen Diskussionen darüber stellen einen markanten Einschnitt in der kontinuierlichen Debatte über die Wirksamkeit von Schulen in der Bundesrepublik Deutschland dar. Natürlich gab es auch früher (genauer gesagt: immer) kontroverse Diskussionen über die Entwicklung des Schulwesens, insbesondere im Zusammenhang mit der äußeren oder inneren Gliederung der Sekundarstufe; es gab Diskussionen über das, was Schüler auf verschiedenen Schulstufen lernen sollen und können müssen; es gab häufig Streit über gute oder weniger gute Lehrmethoden; es gab regelmäßig Kritik von Wirtschaftsverbänden am Leistungsstand deutscher Schüler und Studenten; es gab stets die vielfältigen Sorgen von Eltern und Schülern über Lern- und Leistungsschwierigkeiten, wobei die Lösung ihrer Probleme selten institutionalisiert, sondern vorwiegend privatisiert wurde.

Im Mittelpunkt des öffentlichen Interesses stand über viele Jahrzehnte hinweg aber nur selten der Unterricht, sondern vorwiegend die Schulorganisation. So blieb weithin unstrittig, ja fast selbstverständlich,

- dass in vergleichbaren Klassen ähnliche Leistungsfortschritte erzielt werden; und
- dass deutsche Schüler im internationalen Vergleich relativ gut abschneiden würden.

Systematisch geprüft wurden diese beiden Annahmen nicht. Weder die frühere Bundesrepublik Deutschland noch die ehemalige Deutsche Demokratische Republik beteiligten sich – von einigen wenigen Ausnahmen abgesehen – an den vergleichenden Leistungs-

studien, die von der International Association for the Evaluation of Educational Achievement (IEA) in vielen Schulfächern und in zahlreichen Ländern durchgeführt wurden.

Umso größer war der Schock, den die erste deutschsprachige Publikation über die TIMS-Studie verursachte (Baumert, Lehmann u. a., 1997). Wenigstens vier der für Bildungspolitik, Pädagogik und Öffentlichkeit selbstwertdienliche Überzeugungen wurden durch die Befunde in Frage gestellt:

- (a) Die Leistungsunterschiede innerhalb der gleichen Schulart sind in Mathematik und in den naturwissenschaftlichen Fächern Physik, Chemie, Biologie weitaus größer als bisher unterstellt wurde. Mit anderen Worten: Schulen und Lehrer bewirken bedeutsame Differenzen in den durchschnittlichen Leistungen und Leistungsfortschritten der Schüler.
- (b) Zwischen verschiedenen Bundesländern und zwischen unterschiedlichen, aber vergleichbaren Schulformen gibt es Leistungsdiskrepanzen, die auf Grund des Untersuchungsplanes von TIMSS zwar nicht präzise bestimmbar, wohl aber deutlich vermutbar sind.
- (c) Die Leistungen deutscher Schüler liegen im internationalen Vergleich nur im Mittelfeld. Sowohl im 8. Schülerjahrgang als auch in der 12. Klasse verfügen die deutschen Schüler zum Beispiel über ein durchschnittliches Niveau mathematischer Kompetenzen zur Lösung anspruchsvoller Aufgaben, das weit unter dem Niveau in anderen Ländern, insbesondere in den ost- und südostasiatischen Staaten liegt.
- (d) Die Leistungs- und Kompetenzdefizite deutscher Schüler zeigen sich nicht nur bei den Durchschnittswerten, sondern auch bei ausschließlicher Berücksichtigung der besten fünf oder zehn Prozent der jeweiligen Jahrgangspopulation.

An dieser Stelle braucht nicht detailliert auf die TIMS-Studie und ihre Resultate eingegangen zu werden, weil dieser Untersuchung im vorliegenden Band ein eigenes Kapitel gewidmet ist (vgl. Kap. 18 i. d. Bd.). Festzuhalten bleibt lediglich, dass durch die als spektakulär wahrgenommenen empirischen Befunde viele bildungspolitische, standespolitische, pädagogische und ökonomische Interessen berührt wurden. Das führte in kürzester Zeit zu einer Vielzahl und Vielfalt positiver wie negativer Stellungnahmen.

Die Kritik reicht von weitgehend unbegründeten methodischen Einwänden gegen internationale Vergleichsstudien im allgemeinen bis zur pauschalen Unterstellung, solche Untersuchungen dienten letztlich einer Revitalisierung engmaschiger Kontrollmechanismen über Schüler und Lehrer durch politische, bürokratische, wissenschaftliche oder ökonomische Instanzen. Befürchtungen dieser Art mögen objektiv unbegründet sein, sie signalisieren aber die Sorgen mancher Lehrer und Lehrerverbände über eine vergleichende Evaluation der Leistungen von Schulen durch Messung der Schülerleistungen.

Weniger defensiv, sondern eher offensiv, gelegentlich sogar aggressiv sind jene Einwände gegen die TIMS-Studie, die an Stelle von Leistungsmessungen etwas "ganz

anderes”, besonders Wichtiges und sehr Dringliches zu tun fordern. Statt die kognitive Entwicklung, das Lernen und die Leistungen von Schülern “zu vermessen” und auf dieser Basis Schulsysteme, Schulen und Lehrer zu beurteilen, sollte man sich lieber mit der pädagogischen Förderung von Kindern und Jugendlichen, mit der sozialen Atmosphäre und den außerschulisch verursachten Konflikten im Klassenzimmer, mit einer besseren Qualifizierung von Lehrern oder mit einer größeren Autonomie der einzelnen Schulen beschäftigen. Jede dieser Forderungen ist selbstverständlich berechtigt! Wenig sinnvoll erscheint allerdings die Ablehnung wichtiger Aufgaben durch Hinweise auf angeblich noch Wichtigeres. Natürlich wird ein Schwein durch häufiges Wiegen nicht fatter, um eine im Zusammenhang mit TIMSS oft verwendete Metapher zu zitieren, doch wird niemand auf den Gedanken kommen, das Fressen des Schweines durch Wiegen zu ersetzen. Man sollte das Eine tun, ohne das Andere deswegen zu lassen, wenn es von Nutzen ist. Das gilt auch für die schulische Leistungsmessung, die im Dienste – nicht aber an Stelle – pädagogischer Bemühungen um eine Verbesserung der Schulqualität stehen muss.

Derart grundlegende, aber argumentativ wenig begründete Einwände gegen TIMSS blieben jedoch die Ausnahme. Insgesamt überwogen kritisch-konstruktive Reaktionen.

Die Organisatoren und Autoren der deutschen TIMS-Studie resümieren drei Jahre nach der ersten Publikation ihre Erfahrungen und Eindrücke so: “Der unseres Erachtens wohl wichtigste bisherige Beitrag von TIMSS liegt in der Neustrukturierung der öffentlichen und professionellen Aufmerksamkeit. Nach TIMSS finden Bildungsthemen größeres Interesse, der Unterricht selbst ist als Kernaufgabe der Schule in das Zentrum gerückt. Dieser Aufmerksamkeitswandel ist nicht folgenlos geblieben. Auf politischer Ebene ist damit begonnen worden, Gräben einzuebnen und sich gemeinsam der Aufgabe der Qualitätsentwicklung und Qualitätssicherung zu widmen. Auf wissenschaftlicher Ebene ist anwendungsbezogene Anschlussforschung mit einem neuen Schwerpunktprogramm der Deutschen Forschungsgemeinschaft (DFG) zur Qualität von Schule und Unterricht auf den Weg gebracht. Sowohl in der Erziehungswissenschaft als auch in den Fachdidaktiken hat TIMSS der empirischen Forschung und der Entwicklung innovativer Lehr- und Lernformen Anregungen gegeben. Auf praktischer Ebene – und hier liegt der entscheidende Prüfstein – sind die Wirkungen am vielfältigsten. In der Aus- und Weiterbildung von Lehrern haben die Fragestellungen und Befunde von TIMSS starken Widerhall gefunden. Prozesse und Inhalte des kumulativen fachlichen Lernens, der intelligente Umgang mit Aufgaben und die Verbindung von motivierenden und kognitiv aktivierenden Unterrichtsformen gehören zu den Themen, die Aufmerksamkeit gefunden haben. In fast allen Ländern und länderübergreifend sind Initiativen begonnen worden, um die Rolle der Einzelschule als Ausgangspunkt und Zentrum der Qualitätsentwicklung zu stärken und die Leistungen der die Schule unterstützenden Einrichtungen zu verbessern. Mit dieser Stoßrichtung gehen die Maßnahmen, auch wenn sie zunächst dem mathematisch-naturwissenschaftlichen Unterricht gewidmet sind, weit über diesen hinaus. In der Konzentration auf den Unterricht in der *einzelnen* Schule versprechen die Initiativen auch Nachhaltigkeit, da ein großes Entwicklungspotential unserer Schulen in der fachlichen Qualifikation der Lehrkörper und der schulinternen Zusammenarbeit liegt” (Baumert, Bos & Lehmann, 2000a, Bd. 1, S. 17 f.).

Die insgesamt positive Bewertung der Reaktionen auf TIMSS durch die Projektleiter entspricht dem Tenor des größten Teils der zahlreichen wissenschaftlichen, bildungspolitischen und schulpraktischen Publikationen, die in den letzten drei Jahren erschienen sind. Die allgemeine Wirksamkeit und die spezifischen Wirkungsweise der TIMS-Studie wurden zusätzlich durch einige längerfristige internationale und nationale Entwicklungen in der Schulpädagogik begünstigt:

- (a) Die erziehungswissenschaftliche Forschung beschäftigt sich immer intensiver mit den Wirkungen der Schule auf die Entwicklung der Schüler und mit der Wirksamkeit bestimmter schulischer Faktoren auf das Erreichen unterschiedlicher Bildungs- und Lernziele. Dabei hat sich das theoretische wie praktische Interesse zunehmend von der Ebene des Schulsystems und der Schulorganisation auf die Ebene der einzelnen Schule und des Unterrichts verlagert. Fachliche Leistungen, fächerübergreifende Kompetenzen, gelegentlich auch Merkmale der motivationalen, personalen, sozialen und moralischen Entwicklung dienen in den Studien zur Wirksamkeit von Schulen als häufig verwendete "Erfolgskriterien" (Creemers, im Druck; Fraser et al., 1987; Helmke & Weinert, 1997a; Scheerens, im Druck; Reynolds et al., 1994). Versucht man die kaum überschaubare Fülle relevanter Ergebnisse dieses Forschungsansatzes zusammenzufassen, so erscheint es zweckmäßig, zwischen wirkungsnahen, proximalen Einflussfaktoren (Merkmale der einzelnen Schüler; Zusammensetzung von Schulklassen; Quantität und Qualität von Lerngelegenheiten; unmittelbare Kontexte des Lernens in der Klasse und in der Schule etc.) und wirkungsfernen, distalen Einflussfaktoren (Schulsystem; außerschulische Bedingungen etc.) zu unterscheiden. Zu allen diesen Faktoren verfügen wir zur Zeit über sehr viele Einzelbefunde, doch mangelt es noch immer an wissenschaftlichen Erkenntnissen über das Zusammenspiel verschiedener Faktoren, über kumulative Effekte sowie über die begrenzten Möglichkeiten der Substitution oder Kompensation defizitärer Einflussvariablen.
- (b) In engem Zusammenhang mit der schulischen Wirkungs- und Wirksamkeitsforschung steht die Frage nach der pädagogischen Zurechenbarkeit und damit nach der Verantwortlichkeit für Erfolge und Misserfolge, für Defizite und Vorzüge, für Stärken und Schwächen in den Leistungen von Schulen und Schülern (vgl. Leithwood, Edge & Jantzi, 1999). Da Erfolge bekanntlich viele Väter haben, der Misserfolg aber in der Regel vaterlos ist, weil auch die Zuschreibung von Verantwortlichkeit bei unbefriedigenden Lernergebnissen der Schüler schnell zu wechselseitigen Schuldzuweisungen und vielfältigen Rechtfertigungen oder Immunisierungsreaktionen führt, wird dieser Forschungsansatz neuerdings in eine konstruktive Verbindung mit Fragen der qualitativen Schulentwicklung gebracht (Creemers, im Druck).
- (c) Schulentwicklung und Schulqualität, Qualitätsmanagement, Qualitätssicherung und Qualitätsentwicklung von Schulen stellen gegenwärtig große bildungspolitische und pädagogische Herausforderungen dar, die immer noch viele ungelöste Aufgaben und Probleme enthalten. Nichtsdestoweniger kann man den Eindruck gewinnen, dass diese Begriffe momentan in Gefahr sind, zu allgemein benutzten Schlagwörtern zu werden, weil jeder damit meint, was er für richtig und wichtig hält. Trotz-

dem: Ein gewisser Konsensus über Minimalstandards zur Charakterisierung “guter Schulen” hat sich inzwischen herausgebildet. Dazu gehören:

- eine differenzierte Schulkultur im Sinne gemeinsamer Ziele, Wertorientierungen und sozialer Verhaltensregeln;
- ein schulspezifisches Profil von curricularen und extracurricularen Ansprüchen und Aktivitäten;
- ein positives soziales Klima und ein Verantwortungsstil, der dem Schulleiter, dem Lehrerkollegium, den einzelnen Lehrern, den Eltern und den Schülern bestimmte soziale Rollen mit individuellen Spielräumen zuweist;
- eine ausreichende schulische Lerneffektivität und deren regelmäßige interne und gegebenenfalls externe Evaluation;
- eine prospektive Schulentwicklungsplanung, welche die pädagogischen Qualitätsansprüche definiert und zu Verbesserungen auf der Ebene der Schule, der Schulklassen, des Unterrichts und zusätzlicher Lerngelegenheiten in Schulen führt (Fend, 1988, 1995, 1998)¹.

(d) In einer großen Anzahl empirischer Untersuchungen wurden in den letzten Jahrzehnten Zusammenhänge zwischen Merkmalen des Schulsystems, der Schule, des Unterrichts und der Schüler auf der einen Seite und Unterschieden bzw. Veränderungen in den fachlichen Leistungen, in fachübergreifenden Kompetenzen (z. B. Problemlösen, kritisches Denken, Teamfähigkeit, kooperatives und selbständiges Lernen, kompetente Mediennutzung) und einigen nicht-kognitiven Entwicklungsmerkmalen auf der anderen Seite analysiert. Die dabei gewonnenen wissenschaftlichen Erkenntnisse begünstigten eine sachgerechte Interpretation und Nutzung der TIMS-Studie wie auch anderer vergleichender Leistungsmessungen. Im Einzelnen:

- Schulorganisatorische Bedingungen sind im Vergleich zu unterrichtlichen Faktoren weniger wichtig als lange Zeit geglaubt wurde. Solche Rahmenbedingungen wirken sich in erster Linie erschwerend oder erleichternd auf den Unterricht und das Lernen aus.
- Der Qualifikation von Lehrern kommt für die Qualität des Unterrichts und für die Entwicklung der Schüler größere Bedeutung zu als dies einige modernistische Strömungen wahrhaben wollen. Lehrer spielen nicht nur als Lehrende eine wichtige, oft unterschätzte Rolle, sondern auch als Moderatoren von Lerngruppen, als Tutoren einzelner Schüler, als Gestalter anregender Lernumwelten, als Verhaltensmodelle und als verantwortliche Garanten für die simultane Erreichung unterschiedlicher Bildungsziele durch möglichst viele Schüler.
- Anspruchsvolle Lernziele, kognitiv herausfordernde Lernaufgaben und “objektive”, aber faire Leistungsbeurteilungen mit Hilfe verschiedener Maßstäbe (indi-

¹ In diesem Zusammenhang wird gegenwärtig häufig der Begriff “Benchmark(ing)” verwendet. Von der ursprünglichen englischen Wortbedeutung her ist er nicht sinnvoll ins Deutsche zu übersetzen. Gemeint ist damit ein in der Wirtschaft, in der Politik, in der Verwaltung, aber eben auch im Bildungsbereich von den Betroffenen selbst festgelegter oder mit anderen vereinbarter Vergleichsstandard zur Bestimmung der zu erreichenden Bildungsziele.

viduale, kriteriale und soziale Bezugssysteme; vgl. Rheinberg i. d. Bd.) begünstigen die Lernfortschritte aller Schüler, wenn der Unterricht effektiv, adaptiv und variabel gestaltet wird.

- Gelernt wird im Prinzip nur, was Inhalt des aktiven Lernens der Schüler ist. Das gilt auf wenigstens drei Ebenen: Erstens handelt es sich dabei um die festgelegten Curricula. Welche Fächer für wie viele Jahre mit welcher Stundenzahl und welchen Lerninhalten in einer bestimmten Schularart vorgesehen sind, ist von großer Wichtigkeit für das durchschnittlich erreichbare Leistungsniveau. Die Substitution oder Kompensation fehlender schulischer Lerngelegenheiten durch außerschulische Lernmöglichkeiten hat sich als relativ begrenzt erwiesen. Lehrpläne sind aber nur das eine, die Repräsentation der dafür relevanten Aufgaben im Unterricht ist das andere. Diese zweite Ebene ist deshalb so bedeutsam, weil im Unterricht entschieden wird, was, wieviel und wie anspruchsvoll gelernt wird. Schließlich ist drittens für den individuellen Lernerfolg maßgebend, wieviel Zeit der einzelne Schüler mit der aktiven Bearbeitung welcher Aufgaben verbringt (aktive Lernzeit, "time on task").
- Die meisten kognitiven Leistungsziele werden durch kumulatives, aufeinander aufbauendes und miteinander verbundenes Lernen erreicht. Im Verlauf dieser Lernprozesse verstärkt sich der Einfluss inhaltspezifischer Vorkenntnisse auf die Verarbeitung neuer Informationen und die Nutzung des bereits Gelernten, während die Rolle der allgemeinen Intelligenz als Lerndeterminante abnimmt. Dabei erweist sich nicht nur die Menge, sondern vor allem die Qualität des Vorwissens (hierarchische Organisation, Vernetztheit der Wissenskomponenten innerhalb eines Inhaltsbereiches und zwischen verschiedenen Domänen, die Verbindung zwischen Wissen und Können, der leichte Zugriff auf das verfügbare Wissen und die flexiblen Nutzungsmöglichkeiten) als wichtig.
- Alles, was Schüler wissen und können (sollen), muss – unabhängig vom allgemeinen Fähigkeitsniveau – gelernt werden. Die Qualität der Schule und die Effektivität des Unterrichts spielen für die Lern- und Leistungsfortschritte eine wichtige, für den häufigen Fall, dass keine geeigneten außerschulischen Lerngelegenheiten verfügbar sind, sogar die entscheidende Rolle. Völlig anders verhält es sich allerdings bei der Veränderung, vor allem bei der Verringerung fachlicher Kompetenzunterschiede und intellektueller Fähigkeitsdifferenzen auf hohem Niveau. Da Schüler mit besseren Lernvoraussetzungen im Vergleich zu solchen mit schlechteren Lernvoraussetzungen von vergleichbaren Lerngelegenheiten im Durchschnitt mehr oder mindestens gleich viel profitieren, ist die Egalisierung von individuellen Unterschieden bei anspruchsvollen Leistungen nur in sehr begrenztem Maße möglich. Egalisierungsbemühungen der Schule müssen sich deshalb auf elementare Lernziele beschränken. Dafür sind Methoden des zielreichenden Lehrens und Lernens geeignet (vgl. Kap. 5 i. d. Bd.).
- Geht es um den Erwerb fachübergreifender Kompetenzen und Handlungsbereitschaften, so hängt viel davon ab, dass Schüler und Schulklassen (oft durch den Lehrer angeregt, angeleitet und/oder unterstützt) tatsächlich das tun, was sie lernen sollen. In vielen Fällen überlappen sich also die Ziele und Methoden des Lernens (z. B. beim Erwerb von Problemlösekompetenz, beim "Lernen Lernen",

bei der Entwicklung sozialer Kompetenz). Lerninhalte, Lernziele und Lernmethoden verschmelzen. Die damit verbundenen didaktischen Aufgaben werden häufig unterschätzt und die erzielbaren Lernfortschritte werden ebenso oft überschätzt.

- Erziehungswissenschaftler und pädagogische Psychologen sind als Forscher in der Regel Spezialisten. Ihre wissenschaftlichen Überzeugungen können deshalb in der Praxis einseitig sein und damit den Voreingenommenheiten vieler Laien entsprechen. Das führt nicht selten dazu, dass in der einschlägigen Literatur bestimmte Bedingungs- und Wirkungsfaktoren des Unterrichts über- oder untergewichtet werden. “Entscheidend ist doch letztlich die Motivation”, “Nur wer sich im Klassenzimmer wohl fühlt, kann unverkrampft lernen” oder “Am Ende zählt ausschließlich die Leistung”, sind finale (oft triviale), sehr extreme Überzeugungen, die den komplexen Bedingungs-Wirkungsgefügen des Unterrichts und des Lernens nicht gerecht werden. Neuere Längsschnittstudien (z. B. Weinert & Helmke, 1997) bestätigen nämlich ältere Ergebnisse, die zum Beispiel Fend (1988) so zusammenfasst: “Selbstverständlich schließen sich eine hohe Motivation, ein hohes Wohlbefinden und eine hohe Leistungsbereitschaft sowie ein hohes Niveau der Lernleistungen nicht aus. Im Gegenteil, wir konnten in der großen Leistungsstudie ... nachweisen, dass die positiven Zusammenhänge überwiegen, also jene, dass gleichzeitig eine hohe Motivation und ein hohes Wohlbefinden sowie ein hohes Leistungsniveau in der gleichen Schule kumulieren” (S. 545 f.).
- Im Vergleich zu diesen einflussreichen Bedingungskonstellationen erfolgreichen Lernens haben sich in der neueren Forschung andere, bisher ebenso hoch bewertete Faktoren als nicht ganz so bedeutsam herausgestellt, obwohl ihre Wirksamkeit in einem Teil der Literatur bis heute überschätzt wird. Dazu gehören zum Beispiel die intrinsische Motivation (thematischer Gleichklang von Lernmotiv und Lerninhalt) im Vergleich zu extrinsischer Motivierung (z. B. Hoffnung auf Erfolg und Belohnung; Furcht vor Misserfolg und negativen Sanktionen); erkundende, erfindende und entdeckende Tätigkeiten im Verhältnis zu didaktisch vorstrukturierten und strukturierten Lernprozessen; verschiedene Formen der sozialen Organisation des Lernens.

Die verfügbaren Resultate von theoretisch fundierten und praktisch nutzbaren Ergebnissen der Schul-, Unterrichts- und Lernforschung haben zweifellos dazu beigetragen, dass die bildungspolitisch unbefriedigenden und pädagogisch enttäuschenden Befunde der deutschen TIMS-Studie in der Regel weder resignativ-ignorierend noch aggressiv-abwehrend, sondern kritisch-konstruktiv aufgenommen wurden. Eine Welle neuer, zum Teil methodisch modifizierter Leistungsmessungen, eine große Anzahl didaktischer Reform-Vorhaben (vor allem, aber nicht nur im mathematisch-naturwissenschaftlichen Bereich), interessante wissenschaftliche Untersuchungsprogramme zur Verbesserung von Unterrichts- und Lernprozessen, neue Anstrengungen zu einer Veränderung der Lehrerbildung, ungezählte Initiativen von Lehrerorganisationen, Schulbehörden, Schulen und Lehrern zur Verbesserung der pädagogischen Expertise von Pädagogen waren und sind zu registrieren, – alles in allem sehr positiv einzuschätzende Folgen von TIMSS.

Warum muss die pädagogische Wirksamkeit von Schulen wissenschaftlich überprüft werden?

So pädagogisch ermutigend viele der Reaktionen auf TIMSS waren und sind, so sorgfältig muss aber auch die Kritik am vergleichenden Untersuchungsansatz der Leistungsmessung geprüft werden. Zu klären ist vor allem, ob überhaupt eine Notwendigkeit besteht, solche Studien in Schulen durchzuführen. Diese Frage lässt sich nur beantworten, wenn es eine minimale gesellschaftliche Übereinstimmung oder eine deutliche politische Mehrheitsentscheidung darüber gibt, welcher Stellenwert dem Leistungsprinzip im Allgemeinen und welche Rolle den Schulleistungen innerhalb des Kanons von Bildungszielen im Besonderen zukommen soll. Bei einigen aktuellen Diskussionen über die Leistungsmessung drängt sich nämlich der Eindruck auf, dass es den Kontrahenten weniger um die speziellen Evaluationsverfahren als um die Bedeutung der zu messenden Leistung geht.

Geht man davon aus – und die öffentliche Diskussion auf die TIMSS-Ergebnisse enthält dafür signifikante Hinweise – dass Leistung zwar nicht das einzige, aber ein unverzichtbar wichtiges Bildungsziel der Schule ist, dann braucht man über die Notwendigkeit ihrer methodisch soliden und pädagogisch begründeten Erfassung nicht mehr zu streiten.

Auch wenn bei der Frage nach den Zusammenhängen zwischen dem durchschnittlichen Bildungsniveau eines Landes und seiner ökonomischen Prosperität im internationalen Vergleich noch große Empirie- und Theoriedefizite bestehen, so sprechen doch alle verfügbaren Daten für einen starken Einfluss verschiedener Bildungsindikatoren auf die wirtschaftliche Entwicklung, – und zwar sowohl generell als auch segmental. Insofern spielt der Faktor “Humankapital” in den neueren makroökonomischen wie mikroökonomischen Modellen eine wichtige Rolle (vgl. z. B. Gundlach, 2000; Weiß, 2000). Wurden in früheren Studien in erster Linie Investitionen in das nationale Bildungssystem berücksichtigt (z. B. Bildungsbudget, Größe von Schulklassen, Lehrerausstattung), so interessiert man sich jetzt mehr und mehr auch für die effektive Nutzung dieser Ressourcen und den damit verbundenen Bildungserträgen (häufig reduziert auf die Qualität von Schulleistungen). Internationale Vergleichsuntersuchungen sind dafür wichtige Datenquellen.

Selbstverständlich sind die Zusammenhänge zwischen Merkmalen des Bildungssystems und der wirtschaftlichen Prosperität nicht einfach und direkt, sondern werden durch wichtige zusätzliche Faktorenbündel beeinflusst (z. B. Kapitalausstattung, typische Formen der Unternehmensverfassungen, Arbeitsrecht, internationale Verflechtungen, Konjunkturzyklen). Es erscheint deshalb wissenschaftlich völlig verfehlt, wenn zum Zwecke einer Bagatellisierung der TIMSS-Resultate darauf verwiesen wird, dass Japan und Südkorea trotz des hervorragenden Abschneidens ihrer Schüler bei den Mathematiktests von einer schweren Wirtschaftskrise heimgesucht wurden. Die empirischen Zusammenhangsmuster zwischen Bildung und Wirtschaft sind kompliziert; die verfügbaren bildungsökonomischen Theorien entsprechen dieser Komplexität; die Bedeutung der Schule und der von ihr erbrachten Leistungen sind wissenschaftlich unstrittig.

Ähnliches gilt für den durchschnittlichen Einfluss der Schulleistungen auf die berufliche Entwicklung und das zu erwartende Lebenseinkommen auf der Individualebene. Auch auf dieser Analyseebene bedarf es allerdings komplexer Modelle zur Beschreibung der Realität. Die vorhandenen Resultate zeigen – zumindest in den industrialisierten Ländern – die große Bedeutung der Schulbildung für die berufliche Entwicklung, wobei es im Einzelfall allerdings zahlreiche Faktoren (z. B. sozialer Status des Elternhauses, Möglichkeiten eines schnellen Expertiseerwerbs im Beruf, Rolle bildungsunabhängiger persönlicher Merkmale für eine berufliche Karriere) zur Kompensation mangelnder Bildungsabschlüsse gibt (vgl. z. B. Weinert & Hany, 2000).

Je wichtiger Schulleistungen als Bildungsindikatoren für die Gesellschaft wie für das Individuum werden, umso weniger können sie als eine Art "Privatgeschenk" von Schulen gelten, umso wichtiger wird ihr Informationswert für Politiker und Schuladministratoren, für die Lehrerschaft, die Lehrer, die Eltern und für alle, die öffentliche Verantwortung tragen oder übernehmen wollen. Ihren Informationswert erhalten Schulleistungen natürlich durch Maßstäbe, die Vergleiche erlauben. Leistungsmessungen in Schulen müssen diese Voraussetzung erfüllen. Um innerhalb von internationalen, nationalen und landesweiten Studien zu solchen objektiven Maßstäben und den dafür geeigneten Messverfahren zu kommen, ist viel wissenschaftliche Kompetenz und harte Arbeit erforderlich.

Was aber soll eigentlich gemessen werden? Fachliche oder fachübergreifende Leistungen, Wissen oder Können, kognitive Kompetenzen oder Handlungskompetenzen? An der Beantwortung dieser Frage hat sich in jüngster Zeit viel Streit entzündet. Dabei geht es vor allem um die Behauptung, dass fachlichen Leistungen in Zukunft eine immer geringere, fachübergreifenden Kompetenzen aber eine ständig wachsende Bedeutung zukommen wird. Diese These ignoriert die gut belegte Tatsache, dass Fächer nicht beliebige Wissenskonglomerate darstellen, sondern sachlogische Systeme, die Schüler aktiv und konstruktiv erwerben müssen, wollen sie schwierige inhaltliche Phänomene und Probleme tiefgründig verstehen und soll zukünftiges Lernen durch Transferprozesse erleichtert werden. Das gilt auch für die Bearbeitung von Aufgaben, die zwar in nicht-fachlichen Kontexten eingebettet, aber ihrem Wesen nach fachlicher Natur sind. Um ein mathematisches Problem, das in einem sozio-ökonomischen Kontext situiert ist, erfolgreich zu lösen, braucht man das notwendige mathematische Wissen und zugleich die fachübergreifenden Kompetenzen, um die Aufgabe aus dem übergeordneten sachlichen Zusammenhang überhaupt herauslösen und das Ergebnis sinnvoll für die Lösung des nicht-mathematischen Problems nutzen zu können.

Daneben gibt es natürlich auch Probleme, die inhaltsunspezifisch und im engeren Sinne fächerübergreifend sind. Ihre Bewältigung hängt in erster Linie von der Verfügbarkeit allgemeiner Problemlösestrategien ab. Die OECD hat in diesem Zusammenhang mehrfach vorgeschlagen, den vieldeutigen Leistungsbegriff generell durch das Konzept der Kompetenz zu ersetzen (vgl. Rychen & Salganik, im Druck). Dabei versteht man unter Kompetenzen die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um

die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können (Weinert, im Druck).

Als Erträge des schulischen Unterrichts kann man zwischen folgenden Kompetenzen unterscheiden:

- fachliche Kompetenzen (z. B. physikalischer, fremdsprachlicher, musikalischer Art),
- fachübergreifende Kompetenzen (z. B. Problemlösen, Teamfähigkeit),
- Handlungskompetenzen, die neben kognitiven auch soziale, motivationale, volitionale und oft moralische Kompetenzen enthalten und es erlauben, erworbene Kenntnisse und Fertigkeiten in sehr unterschiedlichen Lebenssituationen erfolgreich, aber auch verantwortlich zu nutzen.

Es ist unbestritten, dass diese Klassen von Kompetenzen für ein gutes und erfolgreiches Leben innerhalb wie außerhalb der Schule notwendig sind. Prioritätssetzungen zwischen diesen Kompetenzen oder gar die Ablehnung einzelner Kompetenzbereiche (z. B. der fachlichen Kenntnisse) haben sich im Lichte des kognitionspsychologischen Erkenntnisstandes als höchst problematisch erwiesen. Insofern scheinen sich gegenwärtig alle in einer Art Minimalkonsens darüber einig zu sein, dass Schulen weder eine Generation von Fachidioten noch eine Generation geschwätziger Dilletanten ausbilden sollten. Was das genau heißt, bleibt umstritten.

Die Notwendigkeiten und Aufgaben der vergleichenden Messung von Schulleistungen bleiben von diesem grundsätzlichen Streit allerdings unberührt. Es geht um die Entwicklung und ständige Verbesserung von objektiven Maßstäben und fairen Messinstrumenten zur Erfassung wichtiger fachlicher, fachübergreifender und handlungsrelevanter Kompetenzen. Dabei ist die Entwicklung fachspezifischer Tests auf curricularer Grundlage sehr viel weiter gediehen als der Entwicklungsstand von fachübergreifenden und handlungsrelevanten Kompetenzen. Schon PISA wird diese missliche Situation zumindest leicht verändern. Es wird allerdings national wie international noch großer Anstrengungen bedürfen, um die Messstandards unterschiedlicher Klassen von Kompetenzen nachhaltig zu verbessern.

Die Differenzen im Entwicklungsstand verschiedener Kompetenzmessungen können allerdings kein Grund sein, um den Aussagewert der verfügbaren Messinstrumente für fachliche Leistungen zu negieren. Diese bieten schon jetzt solide Voraussetzungen für die Fundierung bildungspolitischer und pädagogischer Entscheidungen. Es muss stets die beabsichtigte Nutzung der Daten sein, d. h. welche Ziele mit der Leistungsmessung verfolgt werden, die darüber entscheidet, welche Verfahren eingesetzt werden müssen. Dabei sind verschiedene Ebenen des Schulsystems zu unterscheiden:

- (a) *Bildungspolitische Entscheidungen auf der Ebene des Staates oder eines Bundeslandes*: Internationale (z. B. TIMSS und PISA), nationale (z. B. ebenfalls PISA) Vergleichsuntersuchungen schaffen Orientierungswissen, das viele praktisch relevante Informationen enthält, wichtige Vergleiche erlaubt und zur Bildung von Pla-

nungshypothesen beiträgt, in der Regel aber nicht geeignet ist, bildungspolitische Entscheidungen – seien sie schulorganisatorischer, curricularer oder schullaufbahntypischer Art – direkt zu fundieren oder zu steuern (Trier, 1995). Dazu bedarf es spezifisch geplanter Untersuchungen, häufig in Form von evaluierten Modellversuchen (denen bisher allerdings oft die erforderliche Strenge der Untersuchungsplanung, die Qualität der Messverfahren und die notwendigen Vergleichsstichproben fehlen). Beobachtet man die Entwicklung in Ländern, die im Vergleich zu Deutschland über eine längere Tradition bei der Messung von Schulleistungen verfügen (z. B. England, Australien, USA) und berücksichtigt man die Reaktionen auf TIMSS in unserem Land, so darf man erwarten, dass sich die Begründungen bildungspolitischer Entscheidungen gegenüber der Öffentlichkeit künftig stark verändern werden. Populäre Ideen, erhoffte Bildungswirkungen und suggestive Behauptungen werden in künftigen Diskussionen über Veränderungen von Schulstrukturen nicht mehr reichen. Man wird zumindest eine Expertise über den empirischen Forschungsstand und den internationalen Erfahrungsbestand sowie Aussagen über die geplante Begleitforschung verlangen. Dafür muss das inzwischen erreichte methodische Niveau der vergleichenden Leistungsmessung obligatorisch sein. Darüber hinaus wird sich der interessierten Öffentlichkeit wie der Bildungspolitik mehr als bisher erschließen, dass es keine Patentrezepte zur Verbesserung des Schulwesens gibt, dass Schulen gewachsene Institutionen sind, die man von außen nicht beliebig beeinflussen kann, dass die meisten schulorganisatorischen Entscheidungen erwünschte Wirkungen und unerwünschte Nebenwirkungen haben, und dass Reformen nur dann die beabsichtigten Effekte erzielen, wenn Schulleiter und Lehrer dafür weitergebildet sowie Eltern und Schüler darauf eingestellt werden.

- (b) *Landesweite oder regionale Orientierungsstudien*: Ziel dieser Art von Forschung ist es, allen am Bildungsgeschehen Beteiligten, den davon Betroffenen oder den daran Interessierten möglichst repräsentative (im Extremfall: vollständige) Informationen über den Entwicklungsstand bestimmter Kompetenzen in Schulen oder Schulsegmenten zu geben. Für die Nutzung solcher Daten ist es günstig, wenn die Leistungserhebungen nicht nur einmal (Ein-Punkt-Messung), sondern zwei- oder mehrfach (Mehr-Punkt-Messungen) in Form von Längsschnittuntersuchungen durchgeführt werden und/oder wenn möglichst viele leistungsrelevante Variablen der Schulen, der Schüler, der Schulklassen, des Unterrichts und der Unterrichtskontexte zusätzlich erhoben werden. Ein typisches Beispiel für diese Untersuchungsform ist MARKUS, eine flächendeckende Studie in den Schulen von Rheinland-Pfalz (Ministerium für Bildung, Wissenschaft und Weiterbildung des Landes Rheinland-Pfalz, 2000)².
- (c) *Schulische und unterrichtliche Qualitätsentwicklung*: Selbstverständlich kann in einer pädagogisch angemessenen Rückmeldung von schul- und klassenrelevanten Ergebnissen aus landesweiten Leistungsmessungen viel Potential zur Verbesserung der Schul- und Unterrichtsqualität stecken. Die Mehrzahl der Schulklassen in

² Die Erläuterung von Abkürzungen entnehmen Sie bitte den Seiten 13 bis 16.

Deutschland verfügt bei Kompetenz- oder Leistungsanalysen kaum über zuverlässige Vergleichsdaten zur Beurteilung der eigenen Arbeit. Das ist ein notorischer Mangel des Systems, dessen Überwindung wettbewerbsfreie Vergleichsstudien auf Klassen- und Schulebene erfordert. Ziel solcher Untersuchungen ist nicht in erster Linie die Registrierung von Leistungen und Leistungsunterschieden, sondern die Beförderung des Unterrichts und des Lernens durch regelmäßige Überprüfung der Lernfortschritte (Veränderungsmessung) bei gleichzeitiger gezielter Veränderung der Lernbedingungen. Es geht also dabei nicht um spektakuläre Studien im grellen Licht der öffentlichen Aufmerksamkeit, sondern um kleine, möglichst häufig durchgeführte empirische Arbeiten auf lokaler schulischer Ebene. In neuen Testmodellen werden dafür bestimmte traditionelle Gütekriterien der Leistungsmessung verändert, um der didaktischen Funktion solcher Studien besser gerecht werden zu können (Dochy, im Druck; Dochy & McDowell, 1997). Testverfahren für die Hand des Lehrers, diagnostische Ausbildung aller Pädagogen, hinreichend viele kleine Studienzentren in Verbindung mit Stätten der Lehrerbildung und der lokalen Lehrerweiterbildung sind mittelfristig dafür erforderlich.

Um einer einseitigen Legendenbildung von vornherein vorzubeugen: Jede dieser Formen vergleichender schulischer Leistungsmessung ist notwendig und erfüllt spezifische Zielsetzungen. Deutschland braucht internationale Vergleichsstudien ebenso dringend wie nationale und regionale Untersuchungen über Indikatoren der Realisierung von Bildungszielen in den verschiedensten Schulfächern mit ihren fachlichen und fachübergreifenden Kompetenzsystemen. Dazu kommt der dringende Bedarf an kleiner, aber sorgfältiger diagnostischer Forschung für Reformen des Unterrichts und zur Verbesserung des Lernens.

Anforderungen an schulische Leistungsmessungen

Kein Zweifel: Leistungsmessungen in Schulen werden künftig eine größere Rolle als bisher spielen. Ihren Ergebnissen kommt für die Fundierung und Überprüfung bildungspolitischer Entscheidungen wie für die Qualitätsentwicklung der einzelnen Schulen und des Unterrichts eine wachsende Bedeutung zu. Das gilt allerdings nur unter der Voraussetzung, dass die Studien wissenschaftlich und praktisch halten, was sie versprechen. Methodisch unzuverlässige, gegen variable Kontextbedingungen unsensible und die eigentlichen pädagogischen Fragestellungen verfehlende Untersuchungen können für die Entwicklung des Schulwesens eher schädlich sein und bei den interessierten Politikern und betroffenen Lehrern große Glaubwürdigkeitseinbußen in die Wissenschaft als einer neutralen und nützlichen Instanz verursachen.

Insofern bedarf es stets eines erheblichen, je nach Fragestellung unterschiedlich großen Aufwandes, um zielspezifische, maßgeschneiderte und qualitativ anspruchsvolle Untersuchungsprogramme sowie die dafür erforderlichen Messverfahren zu erstellen. Dafür

ist sowohl testmethodische, statistische und untersuchungstechnische Kompetenz erforderlich als auch schulorganisatorische, pädagogische und didaktische Expertise. Teamarbeit ist also geboten.

In vielen Kapiteln des vorliegenden Buches wird auf die Gütestandards der schulischen Leistungsmessung ausführlich eingegangen werden. Das von J. R. Sanders herausgegebene und in deutscher Übersetzung vorliegende "Handbuch der Evaluationsstandards" (2000) bietet einen guten Überblick über die verschiedenen Nützlichkeits-, Durchführbarkeits-, Korrektheits- und Genauigkeitsstandards, die bei vergleichenden Leistungsmessungen in Schulen zu beachten sind.

An dieser Stelle genügt es deshalb, noch einmal daran zu erinnern und zu unterstreichen, dass einzelne gute Untersuchungen mit ihren öffentlichen Signalwirkungen und ihren zwar ausschnitthaften, aber wichtigen Informationen über den Stand des Bildungswesens schon jetzt für weiterführende bildungspolitische und pädagogische Überlegungen von großem Nutzen sind, dass diese Studien aber erst als Teil einer mittelfristig sich entwickelnden schulischen Evaluationskultur ihr volles Gewicht erhalten.

KAPITEL 2

Kontroversen um die Schulleistungsmessung in Deutschland

Eine fiktive Diskussion über Positionen und Perspektiven in verteilten Rollen

Hans Brügelmann

ModeratorIn: Kürzlich habe ich über die Medien folgenden Vorschlag der Krankenkassen mitbekommen: Ärzte sollten – ähnlich wie im alten China – danach bezahlt werden, wie gut und wie rasch sie Krankheiten heilen. Nur so könne die Qualität des Gesundheitswesens verbessert und der finanzielle Aufwand erfolgversprechend konzentriert werden.

Dieses Beispiel wirft die Frage auf, wie eigentlich der Lehrerfolg des Schulsystems und einzelner LehrerInnen erfasst werden kann und welche Rolle dabei landesweite Testprogramme wie LAU in Hamburg, aber auch internationale Leistungsvergleiche wie die IEA-Lesestudie, wie TIMSS, PISA und PIRLS bzw. IGLU spielen können.

ArbeitgeberIn: Genau, als Steuerzahler habe ich ein Recht darauf zu wissen, was mit meinem Geld passiert und ob die Schulen ihren gesellschaftlichen Auftrag erfüllen.

LehrerIn: Ich finde diese Testerei problematisch. Die Leistungen, vor allem die fachlichen Leistungen der SchülerInnen, machen doch nur einen Teil des Auftrags der Schule aus? “Qualität” erweist sich nicht nur in kurzfristigen Ergebnissen, sondern auch in den langfristigen Wirkungen des persönlichen Umgangs der LehrerInnen mit den SchülerInnen und des inhaltlichen Anspruchs von Aufgaben und Aktivitäten und im methodischen Niveau ihrer Bearbeitung. Prozesse haben einen Eigenwert. Es ist wichtig, wie junge Menschen in unserer Gesellschaft aufwachsen sollen, und es geht damit um langfristige Wirkungen, die sich messtechnisch kaum mehr erfassen lassen. Auch den Wert der Demokratie würde doch niemand nur an ökonomischen Indikatoren festmachen wollen, oder?

Eltern: Ich verstehe gar nicht, warum in den Schulen so ein Wirbel um Tests gemacht wird. Leistungen werden doch überall bewertet. 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?

LehrerIn: Tun wir doch gar nicht! Leistungsmessung gehört auch zu *unserem* Alltag. Wir LehrerInnen erfassen und bewerten Leistungen von SchülerInnen, und unsere Vorgesetzten entscheiden nach einer Unterrichtsbeurteilung über die Verbeamtung junger LehrerInnen und über die Besetzung von Funktionsstellen.

ArbeitgeberIn: Aber diese Vorgänge müssen transparenter werden! Warum reagieren Sie so empfindlich auf den Einsatz von standardisierten Instrumenten? Sie werden in Zukunft nicht mehr damit rechnen können, dass öffentliche Mittel einfach so fließen, ohne dass sich der Aufwand durch einen entsprechenden Ertrag rechtfertigen lässt. Die knappen Mittel müssen sparsam und effektiv eingesetzt werden – wie in der Industrie und wie in anderen Bereichen der öffentlichen Schulaufsicht auch. Schulen können sich zukünftig nicht mehr diffus auf “guten Unterricht” berufen, wenn ihre SchülerInnen nicht anständig lesen, schreiben und rechnen lernen.

Schulaufsicht: Sie sehen das zu technisch. Die Bewertung von Lern- und Lehrerfolgen ist 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 ein Ort der *Bildung*. Die Qualität eines Theaters oder eines Krankenhauses können Sie ja auch nicht allein nach dem Verhältnis von finanziellem Input und Output bewerten. Bildung ist 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.

LehrerIn: Genau. Denken Sie nur an die Evaluation der Programme kompensatorischer Vorschulerziehung in den 70er Jahren. 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 Bereichen. Wenige Jahre später hieß es, die Überlegenheit der Versuchsgruppen gehe schon während der Grundschulzeit 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 ohne Kenntnis der Erfahrungen im Unterricht nicht sinnvoll interpretieren, wohl aber gut als Munition im politischen Tagesgeschäft verwenden lässt.

BildungsforscherIn: Ihre Skepsis gegenüber Tests als Instrument schulischer Leistungsmessung hat eine lange Geschichte. Seit jeher wehrt sich die Schule gegen Urteile von außen und verweist auf ihre ständigen eigenen Leistungsüberprüfungen. Dabei

hätten gerade LehrerInnen 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 bedenkt, was alles von Noten abhängt, muss Leistungsmessung stärker objektiviert werden.

LehrerIn: Theoretisch klingt das plausibel, aber auch standardisierte Tests haben ihre Schwächen: Dieselbe Leistung kann Ergebnis angelernten Wissens oder selbstständigen Denkens sein. Und politisch ist es naiv zu glauben, in der Öffentlichkeit würden Daten aus Leistungsstudien so differenziert wahrgenommen, wie die Befunde es erfordern.

Schulaufsicht: Vor allem: Auch quantitative Daten sprechen nicht für sich selbst. Sie müssen interpretiert werden. Man müsste die Ergebnisse gleich von verschiedenen Positionen aus kommentieren, um eine einseitige Verwertung zu erschweren, wie wir sie z. B. bei TIMSS erlebt haben. So einfach ist das nicht mit der "Objektivität".

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

ArbeitgeberIn: Wir stützen uns dabei auf Ergebnisse von Tests bei den neu eingestellten Lehrlingen. Gewerkschaften, ErziehungswissenschaftlerInnen, aber auch die jeweils betroffenen Regierungsparteien stellen unsere Klagen immer gleich in Frage. Pädagogik und die Bildungspolitik verstehen sich wohl als unfehlbar. Aber ohne Bereitschaft, sich einer Außenkritik zu stellen, sind Fortschritte nicht denkbar.

BildungsforscherIn: Das sehe ich auch so. Andererseits leiden die genannten Untersuchungen unter einer Reihe von Mängeln:

- Inhaltlich erfassen die Tests nur kleine Ausschnitte schulischer Ziele (meist nur die Rechtschreibung, und die nur im Diktat, und die Grundrechenarten).
- Methodisch sind die Aufgaben oft problematisch (z. B. nicht selten exotische Wörter oder das alltagsferne und irritierende Angebot von mehreren falschen Schreibungen, unter denen die richtige auszuwählen ist).
- Für Verallgemeinerungen sind die Stichproben nicht repräsentativ.

- Für historische Vergleiche fehlen zudem vergleichbare Bezugsgruppen aus früheren Zeiten.

LehrerIn: Und Zeitungsberichte wie “Leistung zählt nicht mehr” oder gar “Neue Max-Planck-Studie widerlegt die Reformpädagogik” haben die Aussagekraft der Daten oft weit überzogen, ihre methodischen Grenzen nicht beachtet und zum Teil die Grundsätze einer fairen Interpretation empirischer Daten verletzt.

Schulaufsicht: Das ist aber doch Vergangenheit. Mit TIMSS, LAU usw. sind forschungsmethodisch versierte Institute in das Geschäft eingestiegen. Meine Sorge ist eher die Eigendynamik, die solche Testprogramme entfalten. Ob die ForscherInnen der Politik nicht doch zu viel versprechen? Um auf das Ärzte-Beispiel zurückzukommen: Wer kann eigentlich sagen, wann ein Patient “gesund” ist? Analog ist unser Problem: Kann man *Bildung* (und nicht nur Kenntnisse und Fertigkeiten) verlässlich messen? Und dann auch noch belegen, wo die Gründe für unterschiedliche Lernerfolge liegen?

ArbeitgeberIn: Wir haben doch dasselbe Problem bei der Beurteilung unserer MitarbeiterInnen in den Betrieben. Warum wollen Sie sich vor dieser Anforderung drücken? 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 Rechtfertigungslehre zu überwinden trachtet, wonach es für die Erlösung allein auf den Glauben ankommt, während die Werke keine Bedeutung haben.” (Lange in VBE, 2000, S. 48)

LehrerIn: Das klingt so einfach. Als ob man nur ein Metermaß oder ein Thermometer nehmen müsste und dann daran den Zustand des Bildungswesens ablesen könnte. Aber wir haben doch schon oft die Erfahrung gemacht, dass wissenschaftliche 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 ist 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 Messtheoretiker wird aus pädagogischen Konzepten wie “Selbstständigkeit” ein mechanisches System von Fähigkeitsmodulen, in dem ich meine Idee eines partnerschaftlichen Umgangs mit Kindern und Jugendlichen nicht mehr wiedererkenne.

BildungsforscherIn: 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. JedeR von uns muss pragmatische Lösungen finden. Und wenn ich mir anschau, mit was für primitiven Aufgaben immer noch viele Lehre-

rInnen Leistungen überprüfen, dann wünschte ich mir, dass das von uns in der Testentwicklung erreichte Niveau auch nur halbwegs Standard im Unterrichtsalltag wäre.

Eltern: 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. Heute können LehrerInnen 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.

Schulaufsicht: So einfach ist das nicht. Vor allem hinkt der Vergleich mit den Prüfaufgaben der LehrerInnen. Landesweite Testprogramme haben eine ganz andere Funktion. 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, die in der öffentlichen Diskussion leicht überschätzt wird. Laien, und dazu zählen auch BildungspolitikerInnen, VerbandsvertreterInnen und JournalistInnen, können im Einzelfall die Annahmen, unter denen Aussagen von Tests gelten, und ihre konkrete Leistungsfähigkeit kaum einschätzen. Sie haben eine zu simple Alternative konstruiert: "objektive" Tests hier gegen "subjektive" (Vor-)Urteile da.

LehrerIn: Zusätzlich sehe ich Probleme in der Passung von Test und Unterricht. Vor allem die Beteiligung an internationalen Vergleichen erzwingt oft problematische Kompromisse. Da wird der Lehrplan *eines* Bundeslandes (von 16!) und der Aufgabenpool *eines* Schulbuchs zum Maßstab für "das deutsche Curriculum". 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.

BildungsforscherIn: Es ist sicher richtig, dass Papier- und Bleistift-Aufgaben im Bereich Lesen und Schreiben 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.

LehrerIn: In der Theorie mag das stimmen. 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 Tests auch bestehen können, indem man einfach solche Testaufgaben mit ihnen übt? Wir kennen das ja vom privaten Repetitor, den StudentInnen vor dem juristischen Staatsexamen besuchen, um Prüfungsaufgaben zu pauken, statt zu studieren. "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 Nebenwirkun-

gen von erfolgsabhängigen Verträgen 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, um die Provision zu kassieren.

BildungsforscherIn: 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 anderes Ziel ist es, die Arbeit einzelner Schulen oder gar LehrerInnen zu bewerten. Vor allem in den angelsächsischen 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 ganz andere Funktion haben schließlich die aktuell diskutierten länderübergreifenden Stichprobenuntersuchungen, die lediglich auf einen Vergleich von Schulsystemen zielen.

Eltern: Wieso das? Wir Eltern haben doch ein berechtigtes Interesse an Informationen über die Qualität des Unterrichts in der Klasse unseres Kindes.

BildungsforscherIn: 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, 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 LAU in Hamburg sind da die Ausnahme. Und auch dort werden keine Ranglisten veröffentlicht. Schulbezogene Daten erhalten nur die Schulen selbst, als Anstoß, intern über ihre Arbeit im Vergleich zu anderen Schulen nachzudenken. Wo liegt da die Gefahr?

LehrerIn: Diese Unterscheidung kann man doch nur auf dem Papier so klar durchhalten. Wenn bei einem Systemvergleich herauskommt, dass bestimmte Schulformen – sagen wir mal Gymnasien mit Gruppierung nach Leistungsniveaus in den Hauptfächern – nach bestimmten Kriterien schlechter abschneiden, dann hat das Auswirkungen auf die Wahrnehmung der Eltern in Schulen, die nach diesem Ansatz arbeiten, und damit indirekt auch auf die Arbeit in diesen Schulen.

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

Schulaufsicht: Na, so einfach ist das auch nicht. Wenn bei einer Studie wie TIMSS im *internationalen* Teil herauskommt dass Schulsysteme mit integrierten Sekundarschulen vergleichsweise gut abschneiden, im *Deutschland-internen* Ländervergleich aber, dass Gesamtschulen eher schlechter sind als dreigliedrige Systeme – was folgt daraus? Sol-