

43228 6

Leistungsmessungen in Schulen

Herausgegeben von Franz E. Weinert

2. Auflage



Beltz Verlag · Weinheim und Basel

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 verteufelt 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

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 Ergebnistrückmeldungen 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 I

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 verteufelt 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 Unterrichts 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 Medienutzung) 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 Schulart 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 Lerneterminante 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 zielreicheren 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); erkündende, 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 & Saifanik, 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 Schülensegmenten 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.