Baumert, Jürgen

Vergleichende Leistungsmessung im Bildungsbereich

Oelkers, Jürgen [Hrsg.]: Zukunftsfragen der Bildung. Weinheim : Beltz 2001, S. 13-36. - (Zeitschrift für Pädagogik, Beiheft; 43)

Zeitschrift für Pädagogik
43. Beiheft

Zeitschrift für Pädagogik
43. Beiheft

Zukunftsfragen der Bildung

Herausgegeben von Jürgen Oelkers

Beltz Verlag · Weinheim und Basel

Herstellung: Klaus Kaltenberg
Satz: Mediapartner Satz und Repro GmbH, Hemsbach
Druck: Druckhaus „Thomas Müntzer“, Bad Langensalza
Printed in Germany
ISSN 0514-2717

Bestell-Nr. 41144

Inhaltsverzeichnis

Teil IV: Neue Medien

Jürgen Baumert

Vergleichende Leistungsmessung im Bildungsbereich

Schulentwicklung und die Sichtbarkeit von Steuerungsproblemen

Qualitätsentwicklung ist seit vielen Jahren ein wichtiges Thema im Bildungsbereich, das Thema Qualitätssicherung durch vergleichende Leistungsmessung dagegen relativ neu. Noch vor wenigen Jahren konnten in Deutschland internationale Vergleichsuntersuchungen nahezu ohne jede öffentliche Aufmerksamkeit durchgeführt werden, obwohl deren Ergebnismuster sich von den Befunden von TIMSS kaum unterschieden (Lehmann/Peek/Pieper/v. Stritzky 1995). Bei TIMSS und PISA scheint alles anders zu sein. Wie ist dieser Wandel der öffentlichen Aufmerksamkeit zu erklären?

Für den Aufmerksamkeitswandel ist wahrscheinlich ein Zusammenspiel vieler Faktoren verantwortlich. Dazu gehören allgemeine Globalisierungstrends in Ökonomie und Kommunikation, die auch die Bildungssysteme nicht unberührt lassen. Die Entwicklung eines internationalen bildungsbezogenen Indikatorensystems durch die OECD ist sichtbarster Ausdruck eines universalisierten komparativen Interesses. Die OECD legt ihrem Indikatorenprogramm ein breit gefasstes Qualitätskonzept zu Grunde, in dem Ausstattungs- und Strukturmerkmale nur Randbedingungen für die erreichbare Güte von Prozessen und Ergebnissen darstellen. Dieses Konzept nimmt die zentralen ökonomischen Themen der Qualitätsentwicklung und des Qualitätsmanagements auf, in deren Kontext in den 80er- und 90er-Jahren Antworten auf den unter anderem durch neue Informations- und Kommunikationssysteme hervorgerufenen wirtschaftlichen Strukturwandel gesucht wurden. Diese Diskussion schärft das allgemeine Bewusstsein für Qualitätsstandards von Arbeitsprozessen und Dienstleistungen. Sie strahlt auch auf das Bildungssystem aus, wie die Debatte über Bildungscontrolling und die Zertifizierung von Ausbildungs- und Weiterbildungsangeboten zeigt. Schließlich wird auch die angespannte öffentliche Haushaltslage die Aufmerksamkeit für einen effizienten Mitteleinsatz erhöht haben. Im Zusammenwirken dieser zahlreichen Faktoren wird allerdings häufig die interne Dynamik des Bildungssystems übersehen. Die neue Aufmerksamkeit für Qualitätssicherung scheint mir auch – wenn man so will – ein paradoxes Ergebnis des Erfolges der Schulentwicklungsbewegung zu sein. Dieser Erfolg kommt am deutlichsten in der Institutionalisierung von Schulentwicklung als gesetzlich vorgeschriebener Aufgabe der Einzelschule zum Ausdruck. Mit der verpflichtenden Vorgabe an die Einzelschule, Schulprogramme zu entwerfen, werden in formalisierter Weise Steuerungsprobleme sichtbar gemacht und auf Dauer ge-

stellt, die im traditionellen Modell einer input-orientierten Bildungsverwaltung nicht gelöst, sondern nur verdeckt gehalten werden konnten. Was heißt dies?

Die Funktionsprinzipien einer traditionellen Bildungsverwaltung sind relativ einfach und ihr Steuerungsinstrumentarium ist begrenzt. Drei typische Steuerungsinstrumente sind zu unterscheiden: regulative Programme, die Allokation von Mitteln (hauptsächlich von Personalmitteln) und die Ausführungskontrolle durch die Schulaufsicht, die sich im Wesentlichen auf Personalbeurteilung an Gelenkstellen von Karrieren und Krisenintervention beschränkt. Zentrale Bedeutung haben die regulativen Programme, deren Kernbestand – trotz der Flut von Einzelvorschriften, auf die man vielfach verzichten kann – gering ist. Dazu gehören die gesetzliche Regelung der Schulpflicht, die Stundentafeln, die Lehrpläne, die Vorschriften zur Leistungsbeurteilung und die Grundzüge der Lehrerdienstvorschriften. Herausgehobene Bedeutung haben Stundentafeln und Lehrpläne. Die Stundentafeln bestimmen die Grundstruktur des Bildungsprogramms, Anzahl und Abgrenzung der großen Domänen, die Portionierung und Kumulativität der Gegenstandsbereiche und deren interne Hierarchie. Mit der Lehrplanarbeit erhält die Bildungsverwaltung eine Ziel- und Programmorientierung, die sie der Struktur nach als »moderne« Verwaltung von der Ordnungsverwaltung unterscheidet, die überwiegend nach Konditionalprogrammen – also Wenn/dann-Vorschriften – arbeitet. Die potenzielle Dynamik der Programmorientierung wird jedoch weitgehend durch die Ausgestaltung der Lehrplanarbeit als Lizenzierungsprozess stillgelegt. Lehrpläne sind wie Hopmann (1998), Künzli (1999), Biehl/Hopmann/Ohlhaver (1996) und Biehl/Ohlhaver/Riquarts (1999) gezeigt haben, die nachträgliche Lizenzierung einer sich langsam verändernden Schulpraxis. Dabei hat der Einfluss der Praktiker in den letzten 20 Jahren eher zu- als abgenommen. Die formale Lizenzierung von vorgängigen Entwicklungsprozessen der Praxis stärkt wahrscheinlich das Vertrauen in die Lehrpläne und deren Versprechen auf einschlägige Bildungsresultate. Die für selbstverständlich gehaltene Übereinstimmung von Vorgabe und Ergebnis und deren scheinbare Sicherung durch die Schulaufsicht bei Abweichungen im Einzelfall ist die Grundfiktion der Verwaltung des Bildungswesens, die, solange sie Glaubwürdigkeit besitzt, sowohl die Autonomie des Einzellehrers im Klassenraum als auch die Paritätsvorstellungen innerhalb der Lehrerschaft sichert und gleichzeitig den Schulbetrieb vor öffentlicher Rechenschaftslegung schützt. In diesem System ist die Thematisierung von Bildungsergebnissen nicht nur überflüssig, sondern sogar ein Fremdkörper. Allein die Frage nach ihnen ist schon eine implizite Verwaltungsbeschwerde, die durch die Struktur des Systems stillgelegt wird. Dementsprechend haben Schulleistungsstudien, auch wenn sie international durchgeführt wurden, praktisch kein öffentliches Interesse gefunden. Umso bemerkenswerter ist der Strukturwandel der Aufmerksamkeit, der sich in den letzten Jahren vollzogen hat und mittlerweile von den Schulverwaltungen selbst mitgetragen wird.

Zur Verwaltungsphilosophie der Ergebnissicherung durch regulative Programme gehört auch die Vorstellung der Gleichförmigkeit institutioneller Bedingungen, sodass Unterschiede in den Ergebnissen individuell zugerechnet werden können: Es ist gleichgültig, welche Primarschule ein Kind besucht – die institutionellen Opportunitätsstrukturen sind vergleichbar –, und deshalb sind

auch die Leistungen Resultate individueller Begabung und Anstrengung. Diese Figur ist auch die Rechtsgrundlage für die Aufrechterhaltung des Sprengelprinzips. Schulunterschiede sind Webfehler im System, die, wenn sie bekannt werden, im Grunde ein Eingreifen der Schulaufsicht verlangen. Nun sind aber die Erkenntnis und Akzeptanz von Schulunterschieden Ausgangspunkt der Schulentwicklungsbewegung. Die Entdeckung der Qualität der Einzelschule, zu der HELMUT FEND in Deutschland maßgeblich beigetragen hat, ist im Grunde aus administrativer Sicht ein unerhörter Vorgang. Dass dieser Tatbestand auf der Basis von nur zwei Untersuchungen, nämlich der RUTTER- und der FEND-Studie auch von den Bildungsverwaltungen – wenn auch zunächst nur in einzelnen Ländern Deutschlands – anerkannt wurde, ist ein Vorgang, der alles andere als selbstverständlich ist (RUTTER 1980; FEND 1982, 1986, 1988, 1998). In dieser Entwicklung begegneten und verstärkten sich die Schuleffektivitätsforschung (GRAY/REYNOLDS/FITZ-GIBBON/JESSON 1996; SAMMONS 1999; SAMMONS/HILLMAN/MORTIMORE 1997; SCHEERENS 1992; SCHNABEL 1998), reformpädagogische Vorstellungen von der Unverwechselbarkeit der Einzelschule und der Autonomie des Pädagogischen (BECKER 1954; VON HENTIG 1993; SPRANGER 1927) und demokratie- und verwaltungstheoretisch begründete Bemühungen um eine Dezentralisierung des Schulsystems (Deutscher Bildungsrat 1973; Deutscher Juristentag 1981; DASCHNER/ROLFF/STRYCK 1995; AVENARIUS/BAUMERT/DÖBERT/FÜSSEL 1998). Mit der Schulentwicklungsbewegung und der Schulentwicklungsforschung wurden Schulunterschiede unter dem Gesichtspunkt der Optimierung von Qualität hoffähig. Über welche Unterschiede reden wir?

Tabelle 1 zeigt die Ergebnisse der Zerlegung der Varianz der Mathematikleistungen in individuelle und institutionelle Komponenten an einer Zufallsstichprobe von 147 Klassen in 68 Schulen über einen Zeitraum von vier Jahren. Zerlegt man die Varianz der Mathematikleistungen am Anfang der 7. Jahrgangsstufe, erhält man im Wesentlichen einen Eindruck von den Effekten der Übergangsselektion am Ende der Grundschule in die Sekundarstufe I. 14 Prozent der Leistungsvarianz entfallen auf Unterschiede zwischen Schulen derselben Schulform. Darin kommt im Wesentlichen der nicht systemkonforme Effekt regionalspezifischer Verteilungsdisparitäten zum Ausdruck. Mit der folgenden Spalte wird der Entwicklungszeitraum eines einzigen Schuljahres in den Blick genommen. Die Ergebnisse zeigen wiederum eine beträchtliche institutionelle Variabilität. Schulformen und einzelne Schulen innerhalb einer Schulform können ganz unterschiedliche Leistungszuwächse erreichen. Betrachtet man einen Zeitraum von vier Schuljahren, sieht man, dass sich diese großen institutionellen Unterschiede nicht ausbalancieren. Die Schulformen stellen auch bei Kontrolle der Eingangsleistungen unterschiedliche akademische Entwicklungsumwelten dar (BAUMERT/KÖLLER/SCHNABEL 2000), aber auch die einzelne Schule derselben Schulform bildet ein jeweils spezifisches Lern- und Leistungsmilieu aus. Während im Leistungsbereich – zumindest im Fach Mathematik – die institutionellen Effekte auch bei der Betrachtung der Sekundarstufe I insgesamt erheblich sind, reduzieren sie sich im motivationalen Bereich – exemplifiziert am Beispiel des Mathematikinteresses – auf einen Bruchteil. Dies gilt sowohl für die Schulform als auch für die Einzelschule. Auch die Schulfreude ist kein Merkmal, das besonders sensibel auf institutionelle Bedin-

gungen reagierte. Herauszustellen ist, dass die Schulformen unterschiedliche akademische Entwicklungsmilieus darstellen, aber für die motivationale Entwicklung jeweils eigene, äquivalente Referenzräume bilden.

Tab. 1: **Zerlegung der Varianz der Mathematikleistungen, des Mathematikinteresses und der Schulfreude in individuelle und institutionelle Komponenten* (Angaben in Prozent der Gesamtvarianz auf individueller Ebene)**

Kriterien	Quellen individueller Unterschiede			
	Schulform	Schule	Klasse	Schüler (+ Fehler)
Mathematikleistung am Anfang der 7. Jahrgangsstufe	28	14	3	55
Leistungszuwachs im 7. Jahrgang**	15	16	6	63
Leistungszuwachs bis zum Ende der 10. Jahrgangsstufe**	14	20	4	62
Interesse an Mathematik am Anfang der 7. Jahrgangsstufe	1	8	4	87
Entwicklung des Interesses von der 7.–10. Jahrgangsstufe**	1	6	6	87
Entwicklung der Schulfreude von der 7.–10. Jahrgangsstufe**	1	8	4	87

* Basis: 147 Klassen in 68 Schulen eines Bundeslandes
** Varianz der Residuen

Schulentwicklung geht von der Existenz institutioneller Differenzen im Schulsystem aus und gibt damit die Gleichförmigkeitsvorstellungen der traditionellen Ordnungsverwaltung auf. Mit der Akzeptanz von Schulunterschieden werden unter dem Gesichtspunkt der Optimierung pädagogischer Prozesse aber auch die Ergebnisse schulischer Arbeit in den Mittelpunkt der Aufmerksamkeit gerückt. Damit stellt sich die Frage nach der Rechenschaftslegung für die Qualität von Prozessen und deren Resultaten. Das geschwisterliche Verhältnis von Schulentwicklung und Rechenschaftslegung war vermutlich vielen Schulentwicklern in der Anfangsphase nicht klar. Anfänglich konzentrierten sich die Reformmaßnahmen auch weitgehend auf die Verbesserung des Lebensraumes Schule, ohne den Unterricht als Kernauftrag von Schule systematisch und fachspezifisch in den Blick zu nehmen. Es scheint so, als ob das Autonomie-Paritätsparadigma der Lehrerschaft noch nicht wirklich in Frage gestellt werden sollte. Mit der Formalisierung der Schulentwicklung als eines gesetzlichen Auftrags der Einzelschule werden jedoch Fragen der Qualitätssicherung auch bezüglich der Erträge von Unterricht unabweisbar.

Damit erhöht sich die Komplexität des Geschäfts der politischen Steuerung sprunghaft. Dies gilt für die Struktur der Aufsicht, stärker aber noch für die Aufgaben der Qualitätssicherung und Evaluation. Die systematischen Reibungen zwischen Schulentwicklung und einer verstärkten Verantwortung der Einzelschule einerseits und den Durchgriffsrechten der Schulaufsicht andererseits

wurden als Erste sichtbar. Sie wurden unter dem Stichwort der Trennung von Beratung und Kontrolle oft diskutiert (SCHRATZ 1993; FISCHER/ROLFF 1997). Notwendige Konsequenzen für die Erweiterung der dienstrechtlichen Befugnisse der Schulleitung, um bei verstärkter Selbstständigkeit von Schulen die parlamentarische Legitimationskette zu sichern, werden allerdings erst allmählich und keineswegs immer zur Freude der Lehrerschaft erkennbar (HÖFLING 1998). In Fragen der Qualitätssicherung und Evaluation bewegen sich die Schulverwaltungen und Schulen noch weitgehend auf Neuland. In dieser ungeklärten Situation wird mit jedem Einzelschritt der Gesamtzusammenhang zwischen Qualitätsentwicklung, Qualitätssicherung, politischer Steuerung und administrativer Kontrolle in einer Weise thematisiert, wie dies vorher nicht der Fall war. Evaluationsmaßnahmen werden sowohl unter der Perspektive der implizierten Steuerungsmodelle als auch unter dem Gesichtspunkt ihrer bildungstheoretischen Begründung beobachtet. Dies macht die Diskussion über Qualitätsentwicklung und Qualitätssicherung in der Öffentlichkeit, aber auch zwischen Schulverwaltung und Lehrerschaft schwierig.

Gleichzeitig wird die Multikriterialität von Schule in ihrer gesamten Komplexität zum Problem. Ein Abarbeiten von Teilproblemen durch sachliche und zeitliche Trennung, wie dies etwa in der Lehrplanarbeit der Fall war, wird schwieriger. Die Schule wird gleichzeitig unter den Gesichtspunkten von Qualifikation, erzieherischen Wirkungen und Handlungsprozessen als eigenen Zielen sowie unter Kriterien der Verteilungsgerechtigkeit betrachtet. Damit stehen Selektionsentscheidungen bei Evaluationsmaßnahmen unter erheblicher Begründungspflicht. Diese Begründungspflicht erhöht sich, wenn gleichzeitig die Verfügungsberechtigung über die Ergebnisse evaluativer Maßnahmen geklärt werden muss. Abbildung 1 stellt den komplexen Zusammenhang zwischen Zielbestimmungen, Gesichtspunkten der Verteilungsgerechtigkeit und involvierten Handlungsebenen grafisch dar.

Angesichts der Komplexität dieser Sachverhalte überrascht es nicht, wenn gerade in der Schulentwicklung engagierte Lehrkräfte, die primär ihre eigene Schule im Blick haben, und Lehrergewerkschaften, die Standesprivilegien verteidigen, überrascht und besorgt auf die scheinbar paradoxen Ergebnisse des Strukturwandels reagieren, bei dem Autonomie Rechenschaftspflicht erzeugt. Umso notwendiger sind Klärungen sowohl hinsichtlich wünschenswerter und praktikabler Steuerungsmodelle als auch bezüglich der Zusammenhänge und der Verträglichkeit von unterschiedlichen Zielsetzungen der Schule. Als Beitrag zu einer solchen Klärung möchte ich im Folgenden zuerst eine Typisierung von Evaluationsmaßnahmen versuchen und mich anschließend den inhaltlichen Fragen der adäquaten Erfassung von Zielkriterien zuwenden.

Typisierung von Evaluationsmaßnahmen

Eine Reihe von Streitpunkten, die in der Diskussion über Qualitätssicherung an Schulen immer wieder auftauchen, ist darauf zurückzuführen, dass weder ausreichend zwischen unterschiedlichen Evaluationsmaßnahmen und deren spezifischen Funktionen noch zwischen unterschiedlichen Instrumenten und deren spezifischer Eignung unterschieden wird.

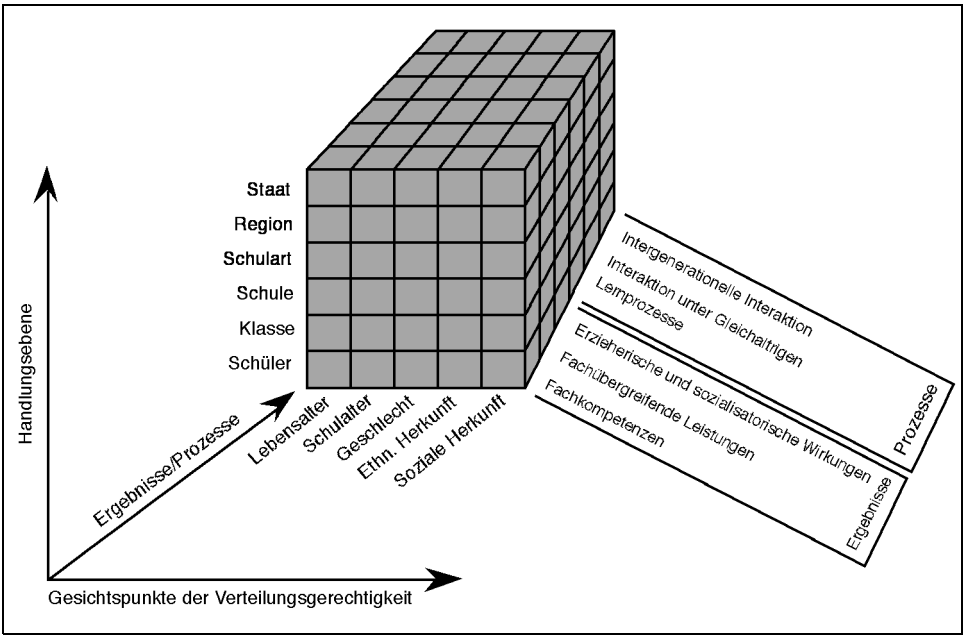

Abb. 1: **Multiperspektivität von Qualitätsentwicklung und Qualitätssicherung** (nach: Arnold 1999).

Wenn alles, was über den kommunikativen Austausch im Kollegium hinausgeht, unter dem Oberbegriff standardisierter Testuntersuchungen zusammengefasst wird, ist es schwierig, sich über Evaluationsmaßnahmen und mögliche ungeplante Nebenfolgen zu verständigen. Als erster Schritt zu einer Systematisierung soll im Folgenden eine grobe Unterscheidung hinsichtlich der Lokalisation der Verantwortung für eine Evaluationsmaßnahme und der Verfügbarkeit von Evaluationsergebnissen getroffen werden. In die eine Kategorie fallen Maßnahmen, die von der Einzelschule initiiert werden und deren Ergebnisse innerhalb der Schule verbleiben; zur zweiten Kategorie gehören zentral durchgeführte Untersuchungen, an denen eine größere Anzahl von Schulen beteiligt ist und die in der Regel in der einen oder anderen Form die Schulaufsicht involvieren. Innerhalb beider Kategorien sind wiederum unterschiedliche funktionale Ebenen zu unterscheiden. Dezentrale Evaluationsmaßnahmen, dienen entweder der individuellen Bewertung von Schülerinnen und Schülern, der schulinternen Selbstvergewisserung zur Optimierung von Schulentwicklung oder der externen Bewertung von Programmen, wobei die Überprüfung entweder von der Schule initiiert und mit einem außenstehenden Partner ausgehandelt werden kann oder im Rahmen der schulaufsichtlichen Tätigkeit als Prozess- und Ergebnisevaluation erfolgt. In beiden Varianten ist die Verbesserung der Praxis eigentliches Anliegen. Abbildung 2 systematisiert dezentrale Evaluationsmaßnahmen nach Funktion und Beurteilungskriterium.

1.	Individuelle Bewertung und Zertifizierung von Schülerinnen und Schülern
1.1	Bewertung nach dem Grad der Zielerreichung (kriteriale Bezugsnorm)
1.2	Bewertung des individuellen Lernfortschritts (ipsative Bezugsnorm)
1.3	Bewertung im Vergleich innerhalb der Lerngruppe (soziale Bezugsnorm)
2.	Schulinterne Evaluation zur Optimierung von Schulentwicklung
2.1	Selbst gesetzte Ziele als Vergleichskriterium (Schulprogramm)
2.2	Vorgegebene Standards als Kriterium (Lehrpläne)
2.3	Soziale Normwerte als Kriterium (Vergleichsschulen, normierte Tests)
3.	Externe Evaluation von Programmen
3.1	Ausgehandelte Vergleichskriterien
3.2	Vorgegebene Vergleichskriterien

Abb. 2: **Funktionale Ebenen dezentraler Evaluationsmaßnahmen**

Bei der Betrachtung von Abbildung 2 sei daran erinnert, dass es keine öffentliche Einrichtung gibt, in der regelmäßiger und häufiger evaluiert wird, als die Schule. Die Bewertung von Schülerleistungen ist so selbstverständlich, dass sie bei der Diskussion um Evaluation in der Regel vergessen wird. Die Bewertung der Schülerleistung hat in der Regel alle Merkmale, die bei Systemevaluationen gerade im höchsten Maße umstritten sind. Lehrkräfte bilden Rangreihen von Schülern innerhalb des Referenzrahmens einer Lernergruppe (Ranking anhand von Noten oder Punkten), die Note gibt keine inhaltlichen Auskünfte über die verfügbaren Kompetenzen, und das mit einer Note verbundene Fähigkeitsniveau kann von Lerngruppe zu Lerngruppe erheblich schwanken. Individuelle Lernfortschritte spielen – wenn überhaupt – bei der Notenvergabe nur eine ergänzende Rolle. Lernentwicklungsberichte sind insofern ein korrigierender Ansatz, als sie die soziale Bezugsnorm in den Hintergrund drängen möchten und die Rückmeldung individueller Profile betonen. Die Nutzung individueller Beurteilungsnormen gerät allerdings auch wiederum schnell an ihre Grenzen, wenn Konflikte mit meritokratischen Gesichtspunkten distributiver Gerechtigkeit sichtbar werden.

Zum Kernbereich der dezentralen Erfassung von Ergebnis- und Prozessmerkmalen gehören alle Maßnahmen der schulinternen Evaluation, die der Optimierung der Schulentwicklung dienen. Häufig wird die schulinterne Evaluation in Gegenüberstellung zur externen Evaluation von Einzelschulen, aber auch in Gegenüberstellung zum Systemmonitoring, das später vorgestellt werden soll, als wünschenswerte und – romantisierend – auch als problemarme Form der Ergebnissicherung verstanden. In dem Augenblick aber, in dem Unterricht in die schulinterne Evaluation einbezogen wird, ist sie ein konfliktreiches Unterfangen, das einen Sprengsatz für die traditionelle Organisationsstruktur der Schule darstellt, insofern das Autonomie-Paritätsparadigma der Lehrerschaft ausgehebelt wird. Hier wird ein Prüfstein für die professionelle Entwicklung des Lehrerberufs liegen. Gleichzeitig wird für die Lehrerfortbildung eine dauerhafte Herausforderung durch die Aufgabe, Unterrichtsentwicklung kooperativ zu unterstützen, entstehen. Denn schulinterne Evaluation setzt Schulentwicklung voraus (Gray u.a. 1996).

Dies gilt gleichermaßen für die externe Evaluation von Schulen und Schulprogrammen, insbesondere dann, wenn ausgehandelte Vergleichskriterien Referenzpunkte darstellen sollen. Einen kritischen Punkt stellt die externe Evaluation von Einzelschulen nach *vorgegebenen* Vergleichskriterien dar. Dabei sind insbesondere die Rolle der Schulaufsicht und die Reichweite ihrer Eingriffsrechte strittig. Allerdings deutet sich in der einschlägigen Literatur ein Verständnis darüber an, dass die Schulaufsicht sich zunächst auf eine Überprüfung der Einhaltung der Gütestandards der Schulprogrammarbeit und der internen Evaluation beschränken sollte (Prozessevaluation). Erst beim Vorliegen von Mängelrügen sollte die Schulinspektion folgen (Posch/Altrichter 1997; Strittmatter 1999; Rolff 1995; Grogger/Specht 1999). Erfolgt die externe Evaluation nach vorgegebenen Kriterien durch Dritte – etwa Wissenschaftler –, wird es immer vorherige Absprachen zwischen den beteiligten Schulen und den Partnern geben müssen, bei denen sich jede Schule in spezifischer Weise zu den Vergleichskriterien positioniert. Zwei Beispiele aus jüngerer Zeit liegen in publizierter Form vor (Klieme/Baumert/Schwippert 2000; Köller/Trautwein, in Vorbreitung). In diesen Fällen wollten Schulen ihre Arbeit zu TIMSS-Normen in Beziehung setzen, um eine generelle Niveauorientierung zu erhalten – allerdings sehr wohl in dem Bewusstsein, dass die in TIMSS erfassten Leistungsaspekte keineswegs den Kern ihrer pädagogischen Bemühungen treffen.

Insgesamt scheint die Verpflichtung auf interne und externe Evaluation bei einer verstärkten Selbstständigkeit von Schulen unstreitig zu sein. Verfahrensstandards sind in Deutschland dagegen noch weitgehend unklar, auch wenn es in der Literatur zur Evaluation gute Vorlagen gibt (Joint Commitee 1994). Hoch umstritten sind dagegen zentrale Ergebnis- und Prozessevaluationen. Unter anderem wohl auch deshalb, weil nicht zwischen funktionalen Ebenen zentraler Evaluationsmaßnahmen unterschieden wird. Umso dringender ist die Klärung. Abbildung 3, in der die funktionalen Ebenen zentraler Evaluationsmaßnahmen unterschieden werden, stellt das Pendant zur Kategorisierung dezentraler Evaluationsmaßnahmen, wie sie in Abbildung 2 vorgestellt wurden, dar.

1. Individuelle Zertifizierung durch zentrale Tests oder Prüfungen mit berechtigender Wirkung 1.1 Abschlussprüfungen 1.2 Zugangsprüfungen
2. Flächendeckende Evaluation von Einzelschulen bzw. ausgewählten Jahrgängen an Einzelschulen 2.1 Wettbewerbsmodell a) ohne Berücksichtigung von Ausgangsbedingungen b) mit Berücksichtigung von Ausgangsbedingungen 2.2 Modell professioneller Qualitätsentwicklung und Qualitätssicherung a) Separierung von Beratung und Kontrolle/Steuerung b) Verbindung von Beratung und Kontrolle/Steuerung
3. Systemmonitoring auf Stichprobenbasis
4. Internationale Vergleichsstudien auf Stichprobenbasis

Abb. 3: **Funktionale Ebenen zentraler Evaluationsmaßnahmen**

Voraussetzung einer sachgerechten Diskussion über Funktionen und Nebenwirkungen zentraler Evaluationsmaßnahmen ist die Unterscheidung zumindest der folgenden vier Ebenen:

– individuelle Zertifizierung,
– flächendeckende Evaluation von Einzelschulen,
– nationales Systemmonitoring und
– internationale Vergleichsuntersuchungen.

Bei der individuellen Zertifizierung durch zentrale Tests oder Prüfungen, mit denen Zugangsberechtigungen erteilt werden, ist zwischen Abschlussprüfungen, die von der abgebenden Institution veranstaltet werden, und Zugangsprüfungen, für die Abnehmer verantwortlich sind, zu unterscheiden. In fast allen europäischen Ländern sind Abschlussprüfungen am Ende der Schulzeit, die den Übergang auf weiterführende Bildungseinrichtungen regeln, gängiger Standard (Frankreich, England, Norwegen, Schweden, Dänemark, Niederlande). Die durchgängig dezentrale Organisation zum Beispiel der Matura-Prüfungen in der Schweiz ist eher eine Ausnahme. Diese Abschlussprüfungen haben *intentionale* Rückwirkungen auf die abgegebene Institution: Auf individueller Ebene sollen sich Schüler ausreichend auf die Prüfungen vorbereiten und auf institutioneller Ebene haben die Schulen auf entsprechende Standards zu achten. Gleichzeitig erhalten die Schulverwaltungen Rückmeldungen über die Qualität der Arbeit von Einzelschulen. *Teaching to the test* oder *Testcoaching* finden selbstverständlich statt, auch wenn sie nicht mit diesen Namen belegt werden, sondern als Lehrplantreue oder wünschenswerte Prüfungsvorbereitung gelten. In Deutschland sind die zentralen Abiturprüfungen der Länder Bayern, Baden-Württemberg, Sachsen, Mecklenburg-Vorpommern und des Saarlands Beispiele diesen Evaluationstyps. Die Wirkung der zentralen Prüfungen geht in einigen Fächern deutlich über die Länder hinaus, in denen sie abgenommen werden. Die veröffentlichten Prüfungsaufgaben wirken standardisierend auch auf die Abiturprüfungen in Ländern, die eine dezentrale Prüfungsorganisation haben.

Unterschiedlich ist die Sachlage bei Zugangsprüfungen, die von Abnehmern veranstaltet werden. Je nach Schulnähe der Tests gehen unterschiedliche Rückwirkungen auf die abgebende Institution aus. Musterbeispiele für diesen Testtyp sind der *Scholastic Aptitude Test* (SAT) oder die Berufseingangsprüfungen der Industrie- und Handelskammern in Deutschland. Der SAT ist ein zentraler Hochschulzugangstest der USA, der aber weitgehend curriculumunspezifisch ist (das gilt auch für den Konkurrenten ACT oder den Englischtest für Ausländer, den TOEFL). Bei diesen standardisierten zentralen Tests, die regelmäßig eingesetzt werden, ist ein Testcoaching nicht zu verhindern. Die einzige Möglichkeit, damit rational und auch einigermaßen fair umzugehen, besteht darin, ein *Testcoaching* für möglichst alle Testteilnehmer anzubieten. Dies ist für die Validität des Tests relativ unproblematisch, da bekannt ist, dass ein Coaching relativ schnell an die Obergrenze seiner Wirksamkeit kommt (Powers/Rock 1999). In dieser Weise ist auch bei dem Test zur Zulassung für die medizinischen Studiengänge in Deutschland verfahren worden (Trost u.a. 1998). Rückwirkungen auf die Schulpraxis durch Tests dieser Art sind relativ

gering, wenn sie überhaupt auftreten. Umgekehrt sind diese Tests aber auch nicht geeignet, Schulprogramme zu evaluieren. Die Kritik POPHAM's (1999) an dem Verfahren, Schulprogramme mit standardisierten Tests zu evaluieren, bezieht sich exakt auf Tests diesen Typs. Ganz anders liegen die Dinge möglicher Rückwirkungen der Berufseingangstests der Kammern. Diese Tests sind curriculumnah, beschränken sich aber auf basale, überwiegend sogar technische Fertigkeiten, deren Validität für die Berufsausübung ebenso strittig ist wie für den Unterricht der abgebenden Schulen. Wahrscheinlich wird nicht zu Unrecht eine Einschränkung des curricularen Spektrums im Abschlussjahrgang der betroffenen Schulform durch *teaching to the test* befürchtet (BLUM, im Druck).

Von der individuellen Zertifizierung ist die flächendeckende Evaluation von Einzelschulen bzw. ausgewählter Jahrgänge an Einzelschulen abzusetzen. Zwei Varianten sind hier zu unterscheiden: Die erste Variante ist das Wettbewerbsmodell englischer oder schottischer Prägung. In einem Fall werden die Rohwerte der Evaluationsergebnisse von Schulen in den so genannten *League Tables* veröffentlicht (ein Überblick über die Kritik gibt SAMMONS 1999). In der schottischen Spielart werden beim Bericht der Ergebnisse unterschiedliche Eingangsbedingungen in Rechnung gestellt (*Value Added Approach*) (MACPHERSON 1992; WOODHOUSE/GOLDSTEIN 1988). Das Rationale beider Verfahren ist die Annahme einer nachfragegesteuerten Qualitätsentwicklung. Dieses Wettbewerbsmodell hat in den deutschsprachigen europäischen Ländern meines Erachtens keine Anhänger.

Davon zu unterscheiden ist die flächendeckende Evaluation von Einzelschulen, die in ein Modell professioneller Qualitätsentwicklung und Qualitätssicherung eingebettet ist. Auch hier sind zwei Spielarten zu erkennen. In einem Fall werden Beratung und Kontrolle konsequent getrennt, insofern die Evaluationsergebnisse ausschließlich den betroffenen Schulen ohne Einschaltung der Schulaufsicht zur Verfügung stehen. Im zweiten Fall gibt es keine klare Unterscheidung von Beratung und Kontrolle: Die Evaluationsdaten gehen an die Einzelschule und an die zuständige Behörde. Für beide Verfahren gibt es in Deutschland erste Beispiele: Brandenburg setzt auf die Trennung von Beratung und Kontrolle, Hamburg und Rheinland-Pfalz folgen dem Mischmodell. Welche Auswirkungen diese Evaluationsmaßnahmen in den Einzelschulen haben, inwieweit sie ein Instrument professioneller Entwicklung darstellen können und welcher systematischen Unterstützung ein solcher Prozess bedarf, ist weitgehend unklar. Ebenso wenig ist ausgelotet, wie weit die Verwendungsmöglichkeiten dieser Evaluationsergebnisse für Zwecke der Systemsteuerung reichen. Sicher ist, dass die flächendeckende Evaluation von Einzelschulen, wenn sie über die Erfassung einer Baseline hinausgeht und regelmäßig wiederholt wird, nur Sinn macht, wenn Schulentwicklungsmaßnahmen intensiver Art vorgeschaltet sind. Die Vorstellung, Veränderungen über die Rückmeldung von Evaluationsergebnissen per se einleiten zu können, ist wenig begründet (FITZ-GIBBON 1996). Nutzt man die flächendeckende Evaluation von Einzelschulen als Steuerungsinstrument, sind an Tests hohe Ansprüche bezüglich ihrer Verträglichkeit mit politisch und fachlich gewünschten didaktischen Konzeptionen zu stellen. Denn auch hier wird sich *Testcoaching* und möglicherweise auch *teaching to the test* einstellen.

In einer Reihe von Bundesstaaten der USA sind obligatorische Testprogramme eingerichtet worden, welche die Funktion der individuellen Zertifizierung und der Evaluation von Einzelschulen, Klassen und indirekt auch von Lehrern verbinden (z.B. STAR in Kalifornien oder ISAT in Illinois). Sie sind nicht nur als Abschlusstests konzipiert, sondern werden in der Regel auch Jahrgangsweise administriert. Vergleichbar sind in gewisser Weise die in Frankreich schulbezogen veröffentlichten Ergebnisse des Abiturs. Vielfach werden die Ergebnisse in so genannten *School Report Cards* veröffentlicht. Die Tests sollen schulübergreifende Standards etablieren. Auf Grund der dezentralen Schulorganisation sind sie jedoch weitgehend curriculumunspezifisch. Testtraining findet in diesen Programmen selbstverständlich statt. Ob die Tests die Lehrkräfte in ihrer Unterrichtsgestaltung einschränken und sich das Spektrum der Unterrichtsgegenstände auf die abgetesteten Stoffe reduziert, ist schwer zu beurteilen (SHEPARD 1990). Vorliegende Berichte weisen eher darauf hin, dass die Lehrkräfte das Testtraining von ihrem eigentlichen Unterricht abkoppeln und vor der Testadministration zwei- bis dreiwöchige Übungsphasen einlegen mit der Folge, dass die verfügbare Unterrichtszeit faktisch verkürzt wird.

Von der zentralen Evaluation von Einzelschulen sind nationale und internationale Studien zum Systemmonitoring auf Stichprobenbasis abzusetzen. Diese beiden Evaluationsformen haben keine direkt steuernden Funktionen für die Einzelschule, sondern primär die Aufgabe, Steuerungswissen auf Systemebene zur Verfügung zu stellen. Inwieweit von solchen Untersuchungen normative Rückwirkungen auf Unterricht und Schule ausgehen, ist ungeklärt. Direkte Rückwirkungen auf Einzelschulen kann man jedoch praktisch ausschließen, ebenso wie ein *Testcoaching*. Wenn Rückwirkungen zu verzeichnen sind, müssen diese über Vermittlungsinstitutionen mit politischer Legitimation transportiert werden. Die institutionalisierte Lehrplanarbeit oder Lehrerfortbildung könnten solche Transmissionsriemen sein. In diesem Fall sind die Rückwirkungen nicht nur legitim, sondern in der Regel auch in ein didaktisches Konzept eingebettet.

Abbildung 4 (S. 24) ordnet noch einmal Beispiele von Untersuchungen mit unterschiedlichem Evaluationsanspruch.

Probleme der Selektivität von Evaluationsmaßnahmen

Die traditionelle Lehrplanarbeit gewinnt ihre Stabilität durch die enge Rückkoppelung an vorgängige Veränderungen in der Praxis und durch eine sachlich und zeitlich gegliederte Abarbeitung von Problemen. Auch bei der Entwicklung von Schulprogrammen wird man davon ausgehen können, dass sie vielfach der Explikation und Fortschreibung vorgängiger Praxis dienen werden. Insofern wird auch hier in gewisser Weise das Lizenzierungsprinzip der Lehrplanarbeit greifen. Zum Schwur wird es jedoch bei der schulinternen Evaluation kommen, wenn ernsthafte Zielklärung betrieben werden muss. Hier ist mit schwierigen Auswahl- und Prioritätsentscheidungen zu rechnen, die zu Kontroversen führen und beträchtlichen Begründungsaufwand erfordern können.

Arbeitsform		Ziele	Beispiele		
Kontrollierte Interventionsstudien	Wissenschaft	Erklärung von Bedingungen der Wirksamkeit pädagogischer Maßnahmen	DFG-Schwerpunkt »Bildungsqualität«		
Längsschnittstudien		Analyse von Entwicklungsverläufen im Zusammenwirken von institutionellen und psychosozialen Faktoren	Scholastik (Weinert) BIJU (Baumert)		TIMSS (Baumert/ Lehmann) u.a.
Systemmonitoring		Beschreibung von (a) Rahmenbedingungen (b) Schule/Unterricht (c) Kompetenzen/Einstellungen der Schüler sowie (d) Zusammenhängen	Markus (Helmke) LAU (Lehmann)	PISA (Baumert u.a.)	
Evaluation von Schulen und Programmen		Zielbezogene, handlungsorientierte Feststellung und Bewertung von Stärken und Schwächen	Quasum (Lehmann)	Netzwerk selbstwirksamer Schulen (Edelstein u.a.)	
Begleitung von Reforminitiativen	Praxis	Feststellung der Ausgangsbedingungen, Praxisberatung, Prozessevaluation	BLK-Modellversuchsprogramm SINUS (Prenzel u.a.)		

Abb. 4: **Unterstützung der Qualitätsentwicklung und Qualitätssicherung durch empirische Bildungsforschung.**

Dies gilt verstärkt für zentrale Evaluationsmaßnahmen, wenn sich diese flächendeckend auf Einzelschulen oder auf das System insgesamt beziehen. Dann unterliegen Auswahlentscheidungen weitaus größeren Begründungspflichten, da mit jeder Selektionsentscheidung gleichzeitig ein bildungstheoretischer Gesamtzusammenhang thematisiert wird.

In diesem Zusammenhang steht auch die Mehrzahl der Vorbehalte gegen standardisierte Leistungsmessungen. Diese Vorbehalte lassen sich im Wesentlichen auf zwei Basiseinwände zurückführen. Der erste Einwand behauptet einen Widerspruch zwischen Ganzheitlichkeit von Bildungsprozessen und den eingeschränkten Fragestellungen von Evaluationsmaßnahmen, insbesondere wenn sie mittels standardisierter Leistungserhebungen erfolgten. Der zweite Einwand besagt, dass standardisierte Untersuchungen oftmals den ihnen zu Grunde liegenden Bildungsbegriff oder das vorausgesetzte Fachverständnis nicht explizierten und Bildungsqualität letztlich mit dem in eins setzten, was ein Test erfasse.

Der vermeintliche Widerspruch zwischen Ganzheitlichkeit von Bildungsprozessen und dem Reduktionismus empirischer Forschung wird mit folgenden Argumenten begründet:

– In standardisierten Schuluntersuchungen würden Aufgaben der Schule selektiv berücksichtigt und die schulischen Ziele implizit auf messbare Bereiche eingeengt. Musterbeispiel für dieses Argument ist die Gegenüberstellung fachlicher Schulleistungen und fachübergreifender Qualifikationen.

- Innerhalb des Fächerspektrums würden wiederum Fächer primär kognitiver Rationalität bevorzugt, sodass selbst innerhalb des obligatorischen Fächerspektrums der Schule bestimmte Formen des Weltverstehens unbegründet privilegiert würden. Unberücksichtigt blieben fast immer der historisch-sozialwissenschaftliche und insbesondere der ästhetisch-expressive Bereich.
- Innerhalb der Fächer würde die Substanz der Fächer auf Faktenwissen, das allein durch Tests erfassbar sei, eingeschränkt. Theoretisches und begriffliches Verständnis, methodisches Können und die selbstständige Auseinandersetzung mit einem Sachverhalt sowie fächerübergreifende Perspektiven würden nicht zu ihrem Recht kommen.

Diese Kritik geht davon aus, dass Leistungstests nicht in der Lage seien, anspruchsvolle Aspekte fachlichen Verständnisses zu erfassen, geschweige denn fachübergreifende Leistungen, und dass es sich bei dem Verhältnis von fachlichen Leistungen und fachübergreifenden Qualifikationen um ein didaktisches Optimierungsproblem handle, bei dem widersprüchliche Ziele auszubalancieren seien. In diesem Zielkonflikt seien ganz unterschiedliche Akzentsetzungen denkbar und auch mit guten Gründen zu rechtfertigen.

Richtig ist zunächst, dass eine einzelne empirische Evaluationsmaßnahme nicht das gesamte Spektrum schulischer Ziele abbilden kann. Es sind notwendigerweise Auswahlen zu treffen. Damit wird auch immer Aufmerksamkeit in eine bestimmte Richtung gelenkt. Daraus ist aber zunächst nur der Schluss zu ziehen, dass einzelne Leistungstests jeweils spezifische Dimensionen von Schule thematisieren und kein Gesamturteil weder über eine Schule noch über ein Schulsystem erlauben.

Zum Beispiel haben die an PISA teilnehmenden Staaten sich mit gutem Grund dafür entschieden, die Prüfung der Lesekompetenz und die Erfassung der mathematischen und naturwissenschaftlichen Grundbildung in den Mittelpunkt zu stellen und dieses Programm schrittweise um die Erfassung fächerübergreifender Kompetenzen zu ergänzen (OECD 1999). Leseverständnis ist in modernen Gesellschaften eine zentrale Schlüsselqualifikation. Sie ist nicht nur Voraussetzung dafür, in allen schulischen Fächern den Anschluss zu halten, sondern Grundlage für die aktive Teilnahme am gesellschaftlichen Leben und notwendige Voraussetzung für jede Form selbstständigen Weiterlernens. Sie ist Voraussetzung und Teil sprachlich-literarischer Bildung, aber selbstverständlich nicht mit dieser identisch. Ein hinreichendes mathematisch-naturwissenschaftliches Verständnis, das deutlich über einfache Rechenfertigkeiten und die Anwendung von Formeln hinausgeht, ist der Schlüssel zu Kernbereichen moderner Kulturen. Man kann sich mit guten Gründen eine Erweiterung der Erhebungsgegenstände wünschen – Fremdsprachenkenntnisse wären wahrscheinlich ein wichtiger Kandidat für eine Erweiterung. Die für PISA getroffene Auswahl ist jedoch nicht beliebig. Allerdings ist PISA auch keine Studie, die generelle Aussagen über das erreichte Allgemeinbildungsniveau von Schülerinnen und Schülern erlaubt. Gleichwohl ist kaum zu bestreiten, dass mit der Auswahl von Untersuchungsbereichen immer auch eine Entscheidung über die Bedeutung von Fächern oder Domänen verbunden ist. In der Regel sind diese Präferenzen – was häufig übersehen wird – bereits aus den Stundentafeln und

Vorschriften über die Leistungsfeststellungen in der Schule zu ersehen, die für die Institutionalisierung eines Bedeutungsgefälles von Bildungsgegenständen sorgen.

Berechtigt ist auch, dass Testkritiker darauf insistieren, dass die zu erfassenden Konstrukte theoretisch expliziert werden. Dieses Argument verliert auch nicht an Gültigkeit durch den Hinweis, dass die Leistungsmessung in der Schule diesen Anspruch gerade nicht erfülle. Bei internationalen Vergleichsstudien hat man sich mit unterschiedlicher Intensität und in der Regel mit begrenztem Erfolg um die Entwicklung theoretischer Rahmenkonzeptionen bemüht. Im Rahmen der Testkonstruktion von TIMSS ist auch deutlich geworden, wie konfliktträchtig ein solcher Versuch beim Aufeinandertreffen unterschiedlicher didaktischer Ansätze sein kann. Über die Qualität der erreichten Kompromisse wird man streiten können. Dies gilt besonders dann, wenn man hinzufügt, dass sich die *frameworks* bislang empirisch nur begrenzt bewährt haben.

Dieser Mangel lässt sich in gewissem Umfang post hoc durch Maßnahmen der Konstruktvalidierung kompensieren. Im Rahmen von TIMSS haben einige Länder, darunter auch die Schweiz und Deutschland, entsprechende Schritte unternommen (Ramseier/Keller/Moser 1999; Klieme 2000; Klieme/Baumert/Köller/Bos 2000; Watermann/Baumert 2000). Die wichtigste Maßnahme ist die Definition von Fähigkeitsniveaus und deren Operationalisierung durch ausgewählte Test-Items, die bei einem gegebenen Niveau mit hinreichender Sicherheit bearbeitet werden können. Durch dieses so genannte *Proficiency Scaling* kann man auch leicht landläufigen Einwänden entgegentreten, standardisierte Tests erfassten nur Faktenwissen. Im Rahmen von PISA haben die Expertengruppen neue Anläufe unternommen, die theoretischen Konstrukte Lesekompetenz, mathematische und naturwissenschaftliche Literalität a priori theoretisch zu konzipieren. Mittlerweile liegen relativ explizierte Testkonzeptionen vor (OECD 1999; Baumert/Artelt/Klieme/Stanat 2001).

Die Abbildungen 5 und 6 zeigen zwei Mathematikaufgaben aus dem Mittelstufen- und Oberstufentest von TIMSS, die stellvertretend für jene Gruppe von Aufgaben stehen, die mathematisches Problemlösen erfassen. Beide Aufgaben sind Indikatoren für das jeweils oberste Fähigkeitsniveau. Die Aufgaben sind entsprechend schwer. Ihre Schwierigkeit ist jedoch nicht darauf zurückzuführen, dass exotische Stoffe oder komplizierte Algorithmen abgeprüft werden, sondern darauf, dass Standardstoffe der Mittelstufe in einem Kontext präsentiert werden, in dem die schlichte Anwendung von Routinen versagt und die Situation mathematisch rekonstruiert werden muss. In einem Fall wird ein erstes Verständnis nichtlinearen Wachstums und im anderen die Fähigkeit der geometrischen Exploration einer alltäglichen Situation erfasst. Die extrem unterschiedlichen Lösungswahrscheinlichkeiten in den ausgewählten Ländern geben erste Hinweise auf eine differenzielle Verständnisorientierung des Mathematikunterrichts (Baumert/Klieme/Watermann 1998; Baumert/Bos/Watermann 2000).

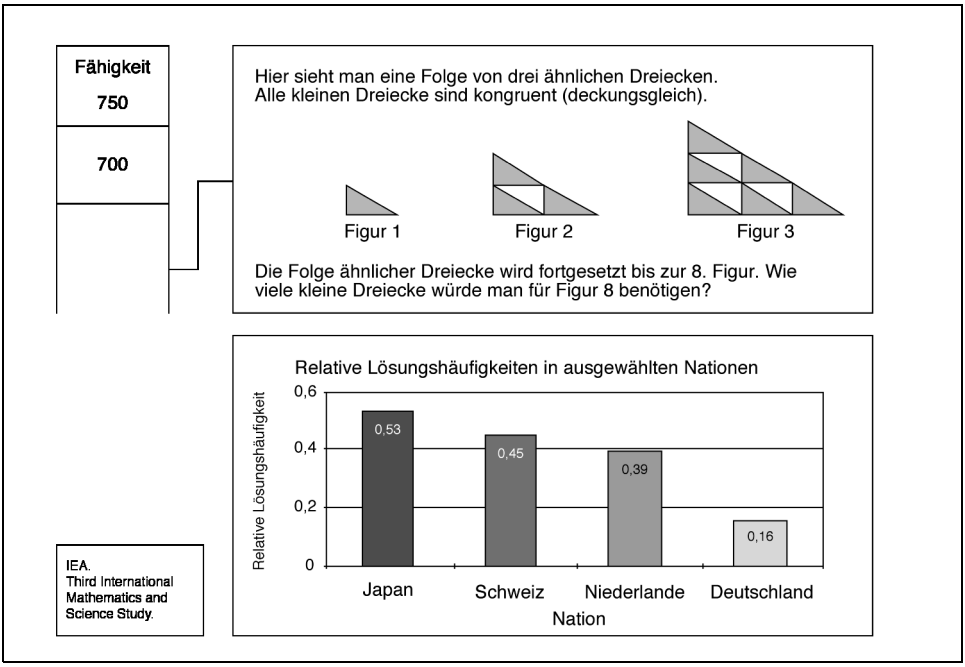

Abb. 5: **Mathematisches Problemlösen** (8. Klasse).

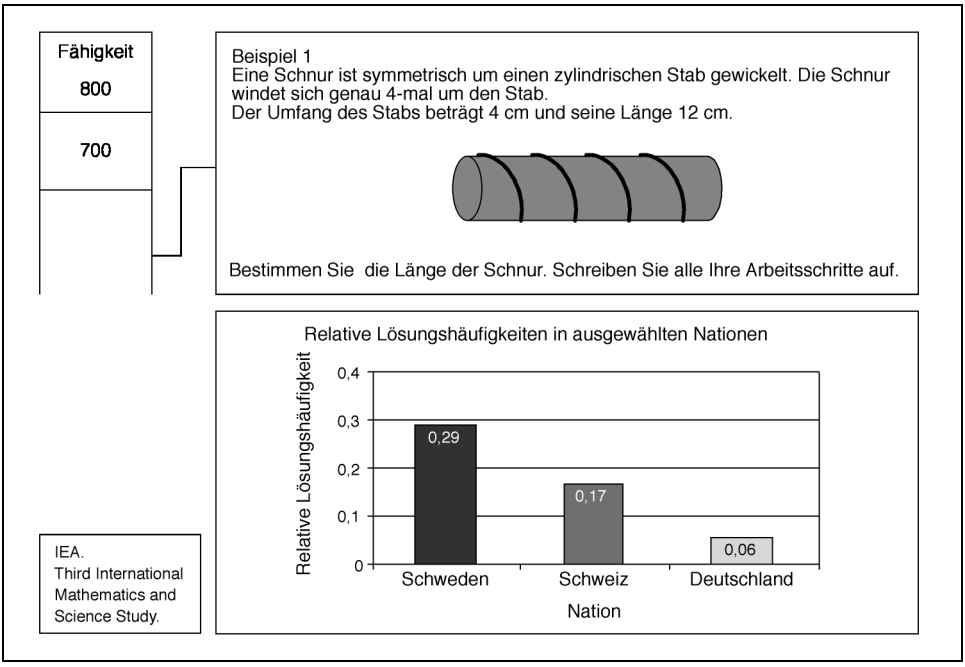

Abb. 6: **Mathematisches Problemlösen** (Gymnasiale Oberstufe).

Fragen der Inhalts- und Kriteriumsvalidität

Ein weiteres Problem, das insbesondere bei internationalen Vergleichen, aber nicht nur dort, auftritt und häufig gegen standardisierte Leistungsmessung ins Feld geführt wird, ist die Festlegung der inhaltlichen Kriterien für die Aufgabenauswahl eines Tests. Bei Schulleistungstests lassen sich zwei grundsätzlich unterschiedliche Vorgehensweisen unterscheiden: In einem Fall wird curriculare Validität in der Regel auf Lehrplan- und Unterrichtsebene angestrebt. Dies ist das übliche Vorgehen, wenn es sich um Unterrichtsgegenstände handelt, für deren Vermittlung die Schule weitgehend verantwortlich ist (also Mathematik, Physik usw.). Im Vorfeld der Testkonstruktionen werden Unterrichtsstoffe identifiziert, die in möglichst vielen Ländern curriculare Validität beanspruchen können und dann den inhaltlichen Rahmen der Itementwicklung bestimmen. In der Regel wird bei diesem Verfahren ein Kompromiss zwischen kleinstem gemeinsamen Nenner und möglichst breiter Erfassung von einschlägigen Stoffgebieten gesucht. Leitendes Prinzip ist dabei, in ähnlicher Weise unfair gegenüber allen beteiligten Ländern zu sein. Wie gut dieses Verfahren gelingt, hängt vor allen Dingen von dem Grad der latenten internationalen Standardisierung eines Unterrichtsgebietes ab. Es setzt also ein – zumindest bereichsspezifisch – kulturübergreifend geteiltes Verständnis einer modernen Schule, nicht aber die Annahme transkultureller Universalien voraus, wie Eckensberger und Römhild (2000) anzunehmen scheinen. In der letzten Generation der IEA-Studien zur Mathematik und den Naturwissenschaften ist dieses Vorgehen für die Untersuchungen in der Sekundarstufe I und den vorakademischen Bildungsgängen gewählt worden. Im Rahmen von TIMSS ist die internationale curriculare Validierung in Deutschland und den Niederlanden durch eine Unterrichtsvalidierung ergänzt worden. In diesen beiden Ländern wurden Lehrkräfte anhand von Testaufgaben befragt, inwieweit die in den TIMSS-Tests repräsentierten Stoffe tatsächlich im Unterricht unterrichtet worden waren (Baumert/Lehmann u.a. 1997; Klieme 2000). Um die Fairness der Kompromissentscheidung gegenüber den teilnehmenden Ländern zu prüfen, kann man darüber hinaus das Curriculum als variierendes Systemmerkmal betrachten. Dieser Weg ist erstmalig in TIMSS beschritten worden. Nachdem in jedem beteiligten Land die ausgewählten Testaufgaben einer Lehrplanvalidierung unterzogen worden waren, wurden national angepasste Tests konstruiert, die nur die für das jeweilige Land curricular validen Aufgaben enthielten. Anschließend wurden die internationalen Vergleiche mit jeder der nationalen Testversionen wiederholt und die Stabilität der Rangreihen geprüft. Bleiben die Rangreihen stabil, ist dies ein starkes Argument für relativ große interkulturelle Fairness (Beaton/Mullis u.a. 1996; Beaton/Martin u.a. 1996; Arnold 1999).

Ein zweiter Ansatz geht von einem normativ-didaktischen Entwurf als Kriterium der Aufgabenauswahl aus. Dies setzt voraus, dass es eine internationale Verständigung über das normative Konzept gibt. Mit diesem Ansatz wird explizit ein internationales Benchmarking (Vergleichsnormierung) angestrebt. Dieses Vorgehen bietet sich immer dann an, wenn Kompetenzen erfasst werden sollen, die über die Schule hinausreichen, nicht allein in der Schule erworben werden und funktionale Bedeutung im Rahmen der Bewältigung allgemeiner

Lebenssituationen haben. Lesekompetenz als Kulturwerkzeug ist ein solches Beispiel. Hier werden in der Regel Anwendungssituationen vorgegeben, an deren Bewältigung die verfügbare Kompetenz abgelesen wird. Im Rahmen von PISA wurde allerdings auch ein normativ-didaktischer Ansatz für den Mathematik- und Naturwissenschaftstest gewählt. Im Falle der Mathematik ist das Testkonzept weitgehend an holländische Vorstellungen von *realistic mathematics* angelehnt, die auf Hans Freudenthal (1977) zurückgehen (vgl. NCTM 2000). In ähnlicher Weise ist der Naturwissenschaftstest funktional angewandt orientiert. Er folgt den von AAAS entwickelten Standards für naturwissenschaftliche Grundbildung (AAAS 1993). Im Rahmen der deutschen PISA-Konzeption wird der internationale normativ-didaktische Entwurf als variierendes Systemmerkmal betrachtet. Es wurde deshalb ein zusätzlicher Subtest konstruiert, der die für den Deutsch- und Mathematikunterricht charakteristische innerfachliche Ausrichtung mit stärker algorithmischer Akzentsetzung zur Geltung bringt (Neubrand u.a. 1999). In ähnlicher Weise verfuhren die Niederlande, als sie im Rahmen von TIMSS einen Zusatztest entwickelten, der der holländischen Vorstellung des realistischen Mathematikunterrichts entsprach (Kuiper/Bos/Plomp 2000). Abbildung 7 fasst die unterschiedlichen Perspektiven der Inhalts- und Kriteriumsvalidität noch einmal in einem Überblick zusammen.

I. Curriculum

Variationsgrad	Ebenen: Lehrplan	Ebenen: Unterricht
Konstant gehalten	TIMSS-International	TIMSS – Deutschland/ Niederlande
variierendes Systemmerkmal	TIMSS-International	

II. Normativ-didaktischer Entwurf (Benchmark)

Variationsgrad	Reichweite: innerfachlich	Reichweite: angewandt/funktional
Konstant gehalten		PISA – International
variierendes Systemmerkmal	PISA – Deutschland	TIMSS – Niederlande

Abb. 7: **Inhaltliche und kriteriale Bezugsnormen der Aufgabenauswahl.**

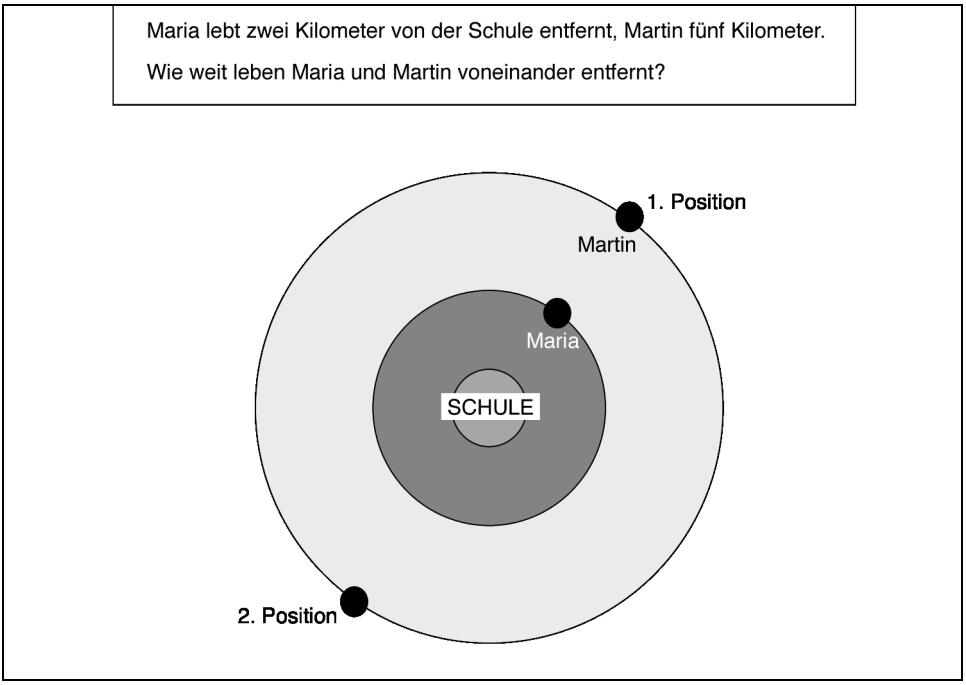

Abb. 8: **Möglicher Ansatz zur Exploration der Mathematikaufgaben.**

Die funktionale Orientierung des internationalen PISA-Mathematiktests wird vermutlich eine interessante Herausforderung für eingebürgerte epistemologische Überzeugungen über Mathematik und Mathematikunterricht in Deutschland ergeben. An einem Beispiel soll dies verdeutlicht werden. Die in der internationalen PISA-Rahmenkonzeption als Beispiel für eine mittlere Kompetenzstufe wiedergegebene Aufgabe: »Maria lebt 2 km von der Schule entfernt, Martin 5 km. Wie weit leben Maria und Martin voneinander entfernt?«, wurde von einer Referentin während einer PISA-Konferenz als unbrauchbares Mathematik-Item bezeichnet, da die Aufgabe keine eindeutige Lösung habe (Demmer 2000). Hierin kommt eine epistemologische Vorstellung von Mathematik und Mathematiktreiben zum Ausdruck, die Grigutsch (1996) und Törner und Grigutsch (1994) als Schemaorientierung bezeichnet haben. Sie ist für den deutschen Mathematikunterricht selbst in der gymnasialen Oberstufe charakteristisch (Köller/Baumert/Neubrand 2000). Im Rahmen dieser Konzeption hat die mathematische Exploration einer Situation keinen Platz. Gerade diese Fähigkeit erfasst aber das PISA-Item auf einem noch niedrigen mathematischen Anspruchsniveau (vgl. Abbildung 8).

Mögliche Erträge von Large Scale-*Schulleistungsstudien*

Häufig wird kritisch gefragt, welchen Nutzen die Schulen und insbesondere die beteiligten Schulen von Untersuchungsprogrammen wie TIMSS oder PISA hätten. In der Regel steht hinter dieser Frage die mehr oder minder explizite

Vorstellung, pädagogische Forschung und allemal empirische Forschung habe nicht nur schulspezifische Diagnosen sondern gleichzeitig auch die Anleitung zur Therapie mitzuliefern. Und nur unter dieser Voraussetzung seien Leistungsstudien zu rechtfertigen. Die Frage scheint mir in doppelter Hinsicht falsch gestellt zu sein. Einmal handelt es sich bei Studien wie TIMSS oder PISA nicht um die Evaluation von Einzelschulen, sondern um Untersuchungen, die dem Systemmonitoring dienen und in erster Linie Wissen über Systemzusammenhänge erzeugen. Dieses Wissen wird in der Regel nur vermittelt über politisch administrative Entscheidungen, die Lehrplanarbeit, die Revision von Lehrbüchern oder die Lehrerfortbildung, gelegentlich vielleicht auch durch die Orientierung von Einzelschulen zur Schulentwicklung beizutragen, in keinem Fall ersetzt es aber die konstruktive Fantasie, die allemal Grundlage konkreter Handlungsentwürfe ist. Diese aber fallen in den Kernbereich der professionellen Zuständigkeit des Lehrers und der Lehrerin. Aber selbst unter der Perspektive der (wenigen) Schulen, die an einer solchen Stichprobenuntersuchung teilnehmen, sollte die Frage umformuliert werden und lauten: Welchen Gebrauch machen Einzelschulen in professioneller Verantwortung von den Ergebnissen aus Schulvergleichen, die sie als Schulrückmeldungen erhalten? Mit dieser Frage betreten wir weitgehend unbekanntes Land. Erste Studien, die untersuchen, wie Schulen mit rückgemeldeten Ergebnissen dezentraler Evaluationsmaßnahmen umgehen, zeigen, dass es keine automatische Verbindung von Rückmeldung und Schulentwicklung gibt (Specht/Altrichter/Soukup-Altrichter 1998). PISA könnte Anlass sein, der Informationsnutzung in Schulen systematisch nachzugehen. Bei bisherigen internationalen Vergleichsstudien haben die beteiligten Schulen Rückmeldungen über Schulmerkmale und Ergebnisprofile im internationalen, nationalen und regionalen Vergleich sowie unter Berücksichtigung ihrer spezifischen Rahmenbedingungen erhalten.

In jüngster Zeit mehren sich die Vorschläge, diese Schulrückmeldungen auszubauen und systematisch in Schulentwicklungsprozesse zu integrieren (Rolff 1999; Helmke 2000). Umso wichtiger wird es, gleichzeitig auf die Grenzen der Aussagefähigkeit und Belastbarkeit solcher Schulrückmeldungen hinzuweisen. Die schulbezogenen Ergebnisse – auch wenn sie ein breites Spektrum von Merkmalen umfassen – basieren in der Regel auf Stichproben einzelner Jahrgänge, Klassen und Fächer, sodass sie keine generalisierten Aussagen über die gesamte Schule erlauben. Denn die Erträge der Arbeit einer Schule können sich von Klasse zu Klasse, von Jahrgangsstufe zu Jahrgangsstufe und von Fach zu Fach unterscheiden (Scheerens/Bosker 1997; Sammons 1999). Darüber hinaus handelt es sich – von wenigen Ausnahmen abgesehen – fast immer um querschnittlich angelegte Survey-Untersuchungen, bei denen eine ausreichende Kontrolle der individuellen Leistungsvoraussetzungen der Schülerinnen und Schüler sowie der institutionellen Kontextbedingungen praktisch kaum möglich ist, sodass man nicht mit hinreichender Sicherheit entscheiden kann, inwieweit die Befunde einer Schule ein Ergebnis der Übergangsselektion oder ihrer pädagogischen Arbeit sind.

Aufklärung – Handlungsorientierung – Rechenschaftslegung	
I. für unmittelbar beteiligte Schulen bzw. Organisationseinheiten Rückmeldung über Schulmerkmale und Ergebnisprofile (Schulleistungen, übergreifende Kompetenzen, Motivation, Einstellungen)	
im internationalen, nationalen bzw. regionalen Vergleich	
unter Berücksichtigung von Rahmenbedingungen (Selektivität, Eingangsbedingungen, soziales Umfeld, Ausstattung usw.)	
unter Berücksichtigung von Grenzen der Aussagefähigkeit der Daten	
auf der Basis von Freiwilligkeit und Vertraulichkeit	
II. für andere Akteure im Bildungssystem Beschreibung und Analyse der Bedingungen, Prozessmerkmale und Ergebnisse schulischer Arbeit	
Ebene	Beispiele aus TIMSS
Schüler	Problemlöse- und Denkfähigkeiten bauen auf fachlichem Wissen auf. Überwindung von Fehlvorstellungen ist größte Kompetenzschwelle. Einfache Anwendungsaufgaben sind Leistungsschwerpunkte deutscher Schüler.
Lehrer	Lehrer können die Effizienz des eigenen Unterrichts und den Anforderungsgehalt von Aufgaben schlecht einschätzen.
Klasse	Schülerorientierter Unterricht fördert die Motivation der Schüler; es lässt sich jedoch kein positiver Zusammenhang mit der Leistungsentwicklung nachweisen. Ein mittleres didaktisches Komplexitätsniveau ist optimal für Leistungsentwicklung; effiziente Klassenführung ist Voraussetzung für didaktisch anspruchsvollen Unterricht. Es lassen sich keine strukturellen Unverträglichkeiten zwischen kognitiven, affektiven und sozialen Zielen des Unterrichts nachweisen.
Schule	Geschlechtsspezifische Unterschiede bleiben innerhalb der Schulen bedeutsam.
Schulform	12- und 13-jähriges Gymnasium sind bzgl. math.-nat. Leistungen gleichwertig.
Bildungssystem	Gymnasialquote hat keinen bedeutsamen Einfluss auf durchschnittliches Leistungsniveau. Zentralabitur sichert möglicherweise Standards in Kursen mit geringer Selektivität. Lehrpläne für Grundkurse werden nicht stringent umgesetzt.

Abb. 9: **Nutzen von Large Scale-Schulleistungsstudien.**

Der eigentliche Nutzen von zentralen Schulleistungsstudien auf Stichprobenbasis liegt nicht in den einzelschulbezogenen Informationen sondern in den mehr oder minder generalisierbaren deskriptiven und analytischen Befunden, die sich bei Studien wie TIMSS oder PISA in charakteristischer Weise auf fast alle Ebenen des Schulsystems beziehen. Abbildung 9 stellt einige Befunde auf unterschiedlichen Systemebenen beispielhaft dar. Charakteristisch für jeden dieser Befunde ist, dass keiner direkte Entscheidungshilfen liefert, sondern eher die Komplexität von Entscheidungssituationen vergrößert. Die Ergebnisse eröffnen Raum für konstruktive Entwürfe. Dies heißt aber auch, dass Evaluation

auf System – ebenso wie auf Schulebene nicht per se praktisch nützlich ist. Nutzen wird erst im reflexiven Gebrauch der Ergebnisse erzeugt. Die eigentliche Arbeit beginnt in der Schule, den sie unterstützenden Einrichtungen und der Politik erst nach der Untersuchung.

Literatur

American Association for the Advancement of Science (Ed.): Benchmarks for science literacy. Project 2061. New York (Oxford University Press) 1993.

Arnold, K.-H.: Fairness bei Schulsystemvergleichen: diagnostische Konsequenzen von Schulleistungsstudien für die unterrichtliche Leistungsbewertung und binnenschulische Evaluation. Münster 1999.

Avenarius, H./Baumert, J./Döbert, H./Füssel, H.-P. (Hrsg.): Schule in erweiterter Verantwortung. Positionsbestimmungen aus erziehungswissenschaftlicher, bildungspolitischer und verfassungsrechtlicher Sicht. Neuwied 1998.

Baumert, J./Lehmann, R. et al.: TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen 1997.

Baumert, J./Artelt, C./Klieme, E./Stanat, P.: PISA (Programme for International Student Assessment) – Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren. In: F.E. Weinert (Hrsg.): Leistungsmessungen in Schulen – Eine Zwischenbilanz. Weinheim 2001.

Baumert, J./Klieme E./Watermann, R.: Jenseits von Gesamttest- und Untertestwerten: Analyse differenzieller Itemfunktionen am Beispiel des mathematischen Grundbildungstests der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie der IEA (TIMSS). In: F. Hofmann (Hrsg.): Schulpädagogik und Lehrerbildung. Festschrift zum 60. Geburtstag von Josef Thonhauser-Herber. Innsbruck, Wien 1998, S. 301–324.

Baumert, J./Köller, O./Schnabel, K.-U.: Schulformen als differenzielle Entwicklungsmilieus – Eine ungehörige Fragestellung? In: Bildungs- und Förderungswerk der Gewerkschaft Erziehung und Wissenschaft im DFG e.V., Messung sozialer Motivation. Eine Kontroverse, Nr. 14/2000, S. 28–68.

Baumert, J./Bos, W./Watermann, R.: Mathematisch-naturwissenschaftliche Grundbildung im internationalen Vergleich. In: J. Baumert/W. Bos/R. Lehmann (Hrsg.): Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Kapitel IV in Band I: TIMSS – Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen 2000, S. 135–197.

Becker, H.: Die verwaltete Schule. In: Recht der Jugend und des Bildungswesens, 41/1954 (2), S. 130–147.

Biehl, J./Hopmann, S./Ohlhaver, F.: Wie wirken Lehrpläne? Modelle, Strategien, Widersprüche. Pädagogik, 48/1996 (5), S. 32–35.

Biehl, J./Ohlhaver, F./Riquarts, K.: Sekundäre Lehrplanbindungen: Vergleichende Untersuchungen zur Entstehung und Verwendung von Lehrplanentscheidungen. Endbericht zum DFG-Projekt. Kiel 1999.

Blum, W.: Was folgt aus TIMSS für Mathematikunterricht und Mathematiklehrerausbildung? In: E. Klieme/J. Baumert (Hrsg.): Mathematik und Naturwissenschaften im Schulunterricht – Bestandsaufnahme und pädagogische Konsequenzen auf der Basis von TIMSS. Bonn (im Druck).

Daschner, P./Rolff, H.-G./Stryck, T.: Schulautonomie – Chancen und Grenzen. Weinheim 1995.

Demmer, M.: PISA-INFO 05/2000: Qualitätsdebatte. Informationen der Gewerkschaft Erziehung und Wissenschaft, Vorstandsbereich Schule, 13.4.2000.

Deutscher Bildungsrat: Verstärkte Selbstständigkeit der Schule und Partizipation der Lehrer, Schüler und Eltern. Bonn 1973

Deutscher Juristentag: Schule im Rechtsstaat. Entwurf für ein Landesschulgesetz. (Vol. 1). München 1981.

Eckensberger, L./Römhild, R.: Kulturelle Einflüsse. 12. Kapitel. In: M. Amelang (Hrsg.): Determinanten individueller Unterschiede. Band 4 der Enzyklopädie der Psychologie. Göttingen u.a. 2000.

Fend, H.: Gesamtschule im Vergleich. Bilanz der Ergebnisse des Gesamtschulversuchs. Weinheim 1982.

Fend, H.: »Gute Schulen – schlechte Schulen«. Die einzelne Schule als pädagogische Handlungseinheit. In: Die Deutsche Schule, 78/1986 (3), S. 275–293.

Fend, H.: Schulqualität. Die Wiederentdeckung der Schule als pädagogische Gestaltungsebene. In: Neue Sammlung, 28/1988 (4), S. 537–547.

Fend, H.: Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung. Weinheim u.a. 1998.

Fischer, D./Rolff, H.G.: Autonomie, Qualität von Schulen und staatliche Steuerung. Chancen und Risiken von Schulautonomie. Zeitschrift für Pädagogik, 43/1997 (4), S. 537–549.

Fitz-Gibbon, C.: Monitoring school effectiveness: Simplicity and complexity. In: J. Gra/D. Reynold/C. Fitz-Gibbon/D. Jesson (Hrsg.): Merging traditions: The future of research on school effectiveness and school improvement. London (Cassell) 1996.

Freudenthal, H.: Mathematik als pädagogische Aufgabe (Bd. 1, Bd., 2). Stuttgart 1977.

Gray, J./Reynolds, D./Fitz-Gibbon, C./Jeeson, D. (Hrsg.): Merging traditions: The future of research on school effectiveness and school improvement. London (Cassell) 1996.

Grigutsch, S.: Mathematische Weltbilder von Schülern: Struktur, Entwicklung, Einflussfaktoren. Unveröffentlichte Dissertation vom Fachbereich 11/Mathematik der Gerhard-Mercartor-Universität – Gesamthochschule Duisburg 1996.

Grogger, G./Specht, W.: Evaluation und Qualität im Bildungswesen. Problemanalyse und Lösungsansätze am Schnittpunkt von Wissenschaft und Bildungspolitik. Dokumentation eines internationalen Workshops in Blumau/Steiermark, 18. bis 21. Februar. Graz 1999.

Helmke, A.: TIMSS und die Folgen. Der weite Weg von der externen Leistungsevaluation zur Verbesserung des Lehrens und Lernens. In: U.P. Trier (Hrsg.): Bildungswirksamkeit zwischen Forschung und Politik. Bern 2000.

Hentig, H. v.: Die Schule neu denken. München 1993.

Höfling, W.: Die Bedingungen für eine Schule in erweiterter Verantwortung nach deutschem Verfassungsrecht. In: H. Avenarius/J. Baumert/H. Döbert/H.-P. Füssel (Hrsg.): Schule in erweiterter Verantwortung. Positionsbestimmungen aus erziehungswissenschaftlicher, bildungspolitischer und verfassungsrechtlicher Sicht. Neuwied 1998, S. 51–66.

Hopmann, S.: Der Lehrplan als Maßstab öffentlicher Bildung. In: J. Oelkers/F. Osterwalder/H. Rhyn (Hrsg.): Bildung, Öffentlichkeit und Demokratie. Weinheim 1998, S. 165–188

Joint Commitee of Standards for Educational Evaluation. The program evaluation standards. How to assess evaluations of educational programs. Thousand Oaks (Sage) 1994.

Klieme, E.: Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In: J. Baumert/W. Bos/R. Lehmann (Hrsg.): Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Kapitel II in Band II: TIMSS – Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. Opladen 2000, S. 57–128.

KLIEME, E./BAUMERT J./SCHWIPPERT, K.: Schulbezogene Evaluation und Schulleistungsvergleiche. Eine Studie im Anschluss an TIMSS. In: H.G. ROLFF/W. BOS/K. KLEMM/H. PFEIFFER/R. SCHULZ-ZANDER (Hrsg.): Jahrbuch der Schulentwicklung, Band 11. München 2000, S. 387–438.

KLIEME, E./BAUMERT, J./KÖLLER, O./BOS, W.: Mathematisch-naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In: J. BAUMERT/W. BOS/R. LEHMANN (Hrsg.): Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Kapitel III in Band I: TIMSS – Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen 2000, S. 85–133.

KÖLLER, O./TRAUTWEIN, U.: Mehr als nur eine Momentaufnahme: Möglichkeiten, TIMSS als Basis für die Schuldiagnostik und Schulentwicklung zu nutzen (in Vorbereitung).

KÖLLER, O./BAUMERT, J./NEUBRAND, J.: Epistemologische Überzeugungen und Fachverständnis im Mathematik- und Physikunterricht. In: J. BAUMERT/W. BOS/R. LEHMANN (Hrsg.): Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Kapitel VI in Band II: TIMSS – Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. Opladen 2000, S. 229–269.

KÜNZLI, R.: Lehrplanpolitik. Regelungs- und Steuerungsleistungen eines alten Instrumentes. In: Bildungsforschung und Bildungspraxis, 2/1999, S. 140–160.

KUIPER, W.A.J.M./BOS, K.T./PLOMP, T.: The TIMSS National Option Mathematics Test. In: Studies in Educational Evaluation, 26/2000, S. 43–60.

LEHMANN, R. H./PEEK, R./PIEPER, I./STRITZKY, R. VON: Leseverständnis und Lesegewohnheiten deutscher Schüler und Schülerinnen. Weinheim 1995.

MACPHERSON, A.: Measuring added values in schools. London (National Commission on Education) 1992.

National Council of Teachers of Mathematics (NCTM): Principles and standards for school mathematics. Reston, VA (NCTM) 2000.

NEUBRAND, M. u.a.: Grundlagen der Ergänzung des internationalen PISA-Mathematik-Tests in der deutschen Zusatzerhebung: Framework zur Einordnung des PISA-Mathematik-Tests in Deutschland. Berlin 1999.

OECD: Measuring student knowledge and skills. A new framework for assessment. Paris 1999.

POPHAM, W.J.: Why standardized test scores don't measure educational quality. In: Educational Leadership, 56/1999 (6), S. 8–15.

POSCH, P./ALTRICHTER, H.: Möglichkeiten und Grenzen der Qualitätsevaluation und Qualitätsentwicklung im Schulwesen. Innsbruck/Wien 1997.

POWERS, D. E./ROCK, D. A.: Effects of coaching on SAT I: Reasoning test scores. In: Journal of Educational Measurement, 36/1999 (2), S. 93–118.

RAMSEIER, E./KELLER, C./MOSER, U.: Bilanz Bildung. Eine Evaluation am Ende der Sekundarstufe II auf der Grundlage der Third International Mathematics and Science Study. Zürich 1999.

ROLFF, H.-G.: Steuerung, Entwicklung und Qualitätssicherung von Schulen durch Evaluation. In H.-G. Rolff (Hrsg.): Zukunftsfelder von Schulforschung. Weinheim 1995.

ROLFF, H.-G.: PISA Initial Overall Evaluation. Gutachten. Erstellt im Auftrag der OECD für Treffen des Board of Participating Countries vom 13. bis 15. März 2000 in Melbourne, 1999.

RUTTER, M. u.a.: Fünfzehntausend Stunden. Schulen und ihre Wirkung auf die Kinder. Weinheim/Basel 1980.

SAMMONS, P.: School effectiveness. Coming of age in the twenty-first century. Lisse u.a. 1999.

SAMMONS, P./HILLMAN, J./MORTIMORE, P.: Key characteristics of effective schools: a review of school effectiveness research. In: M. BARBER/J. WHITE (Hrsg.): Perspectives on school ef-

fectiveness and school improvement. London, UK (University of London, Institute of Education) 1997.

Scheerens, J.: Effective schooling: Research, theory and practice. London (Cassell) 1992.

Scheerens, J./Bosker, R.: The Foundations of Educational Effectiveness. Oxford (Elsevier) 1997.

Schnabel, K.U.: Schuleffekte. In: D.H. Rost (Hrsg.): Handwörterbuch Pädagogische Psychologie. Weinheim 1998.

Schratz, M.: Autonomie und Schulaufsicht – ein Widerspruch? In: Schul-Management, 24/1993 (4), S. 8–15.

Specht, W./Altrichter, H./Soukup-Altrichter, K.: Qualitätsentwicklung mit Programm. Endbericht über die Begleitevaluation der Pilotphase. Report Nr. 41 des Zentrums für Schulentwicklung. Graz 1998.

Spranger, E.: Die wissenschaftlichen Grundlagen der Schulverfassungslehre und Schulpolitik. 1927. Abgedruckt in Klinkhardts Pädagogische Quellentexte. Bad Heilbrunn 1963.

Strittmatter, A.: Qualitätsevaluation und Schulentwicklung. In: J. Thonhauser/J.-L. Patry (Hrsg.): Evaluation im Bildungsbereich. Innsbruck-Wien 1999.

Törner, G./Grigutsch, S.: »Mathematische Weltbilder« bei Studienanfängern – eine Erhebung. In: Journal für Mathematik-Didaktik, 15/1994, S. 211–251.

Trost, G./Blum, F./Fay, E./Klieme, E./Maichle, U./Meyer, M./Nauels, H.-U.: Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse. Bonn 1998.

Watermann, R./Baumert, J.: Mathematisch-naturwissenschaftliche Grundbildung beim Übergang von der Schule in den Beruf. In: J. Baumert/W. Bos/R. Lehmann: Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Kapitel V in Band I: TIMSS – Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen 2000, S. 199–259.

Woodhouse, G./Goldstein, H.: Educational Performance Indicators and LEA League Tables. In: Oxford Review of Education, 14/1988 (3), S. 301–320.

Anschrift des Autors
Prof. Dr. Jürgen Baumert, Max-Planck-Institut für Bildungsforschung,
FB Schule und Unterricht, Lentzeallee 94, D-14195 Berlin-Dahlem.