Arnold, Karl-Heinz

Schulleistungsstudien und soziale Gerechtigkeit

Zeitschrift für Pädagogik 47 (2001) 2, S. 161-177

Zeitschrift für Pädagogik

Jahrgang 47 – Heft 2 – März/April 2001

Thema: Internationale Schulleistungsvergleichsforschung

Weiterer Beitrag

Diskussion

Besprechungen

Dokumentation

Beilagenhinweis: Dieser Ausgabe der Z.f.Päd. liegen Prospekte des VNR-Verlag für die deutsche Wirtschaft AG, Bonn, und des Verlags der Österreichischen Akademie, Wien, bei.

Topic: International Comparative Research On School Achievement

Further Contributions

Discussion

KARL-HEINZ ARNOLD

Schulleistungsstudien und soziale Gerechtigkeit

Zusammenfassung
Internationale Schulleistungsstudien beanspruchen, relevante Lernergebnisse in den Schulsystemen der beteiligten Staaten zu erfassen, um Qualitätsindikatoren für deskriptive und analytische Vergleiche bereitzustellen. Somit ist die Frage zu beantworten, ob die hier genutzten Verfahren der pädagogischen Diagnostik hinreichend gleiche Untersuchungsbedingungen ermöglichen, die zudem hinreichend sensitiv für die nationalen Eigenarten der Bildungssysteme sind. Im Begriff der Fairness kann diese Fragestellung zusammengefasst werden. Aus schulpädagogischer Sicht zielt Fairness auf die Frage nach der sozialen Gerechtigkeit in der Institution Schule. Im Bereich der psychologischen Diagnostik und insbesondere der Testtheorie stellt Fairness ein instrumentelles Gütekriterium bzw. die Folge hinreichender Testgüte dar. Am Beispiel von TIMSS (*Third International Mathematics and Science Study*) wird aufgewiesen, dass mit den Methoden avancierter pädagogisch-psychologischer Diagnostik ein hohes Maß an Fairness für die international vergleichende Schulleistungsforschung erreicht werden kann.

1. Einleitung

Mit der Veröffentlichung der Ergebnisse von TIMSS (*Third International Mathematics and Science Study*) scheint in Deutschland eine Wendung sowohl der öffentlichen Debatte über Schule als auch der schulsystembezogenen Bildungsforschung eingetreten zu sein. Internationale Schulleistungsstudien mit bundesdeutscher Beteiligung gab es zwar auch zuvor (für einen Überblick vgl. BOS/BAUMERT 1999), deren Ergebnisse wurden jedoch nur wenig beachtet oder gar für obsolet erklärt (vgl. FREUDENTHAL 1975).

Ein bemerkenswertes Beispiel für diese These bietet die RLS (*International Study of Reading Literacy*), die ebenso wie TIMSS von der IEA (*International Association for the Evaluation of Educational Achievement*) durchgeführt und in Deutschland als „Hamburger Lesestudie" von R. LEHMANN und Mitarbeitern (vgl. LEHMANN/PEEK/PIPER/STRITZKY 1995) zu einem eindrucksvollen Schulsystemvergleich der „alten" mit den „neuen" Bundesländern erweitert worden ist. Weder die internationalen Rankingtabellen mit ihrer „schlechten Nachricht" (wie bei TIMSS resultiert bei IEA-RLS für Deutschland eine mittlere Position) noch die zum jahrzehntelangen Ideologiestreit zwischen BRD und DDR „nachgelieferten" Vergleichsdaten, die keineswegs die Überlegenheit des bundesrepublikanischen Schulsystems demonstrieren, verhalfen dieser Studie zu einem ähnlich breiten Medienecho, wie dies fünf Jahre später bei TIMSS und derzeit schon im Vorfeld bei OECD-PISA (*Programme for International Student Assessment*) der Fall ist.

In der Bundesrepublik Deutschland avanciert gegen Ende der neunziger Jahre die wieder entdeckte empirische Bildungs- und Schulleistungsforschung

zu einem machtvollen Instrument der bildungspolitischen Argumentation. Der so genannte Konstanzer Beschluss der Kultusministerkonferenz von 1997 befürwortet ausdrücklich die Teilnahme Deutschlands an internationalen Schulleistungsstudien und beendet somit eine heute schon fast peinlich anmutende, dreißigjährige Epoche der „Abstinenz". Gleichwohl hat die Geschichte der schulsystembezogenen, empirischen Bildungsforschung in der Bundesrepublik Deutschland auch eine besondere Hochphase zu verzeichnen. Das politische Ende des Deutschen Bildungsrates überstand – für einige Jahre – dessen vielleicht bedeutsamstes Forschungsprogramm: die Einrichtung von Gesamtschulen in einem Kontext des empirischen Schulsystemvergleichs. Ende der siebziger und Anfang der achtziger Jahre unternahmen insbesondere die Forschungsgruppen um H. Fend und H. Lukesch (Fend 1982; Haenisch/Lukesch 1980) mehrere große Vergleichstudien, die mit lehrplanorientierten und zugleich schulform- sowie in der Drei-Länder-Studie obendrein lehrplanübergreifend gültigen Leistungstests operierten. Die Ergebnisse wurden – mit großer Medienpräsenz – ebenso kontrovers diskutiert wie die Methodologie dieser Vergleiche. Mit dem empirischen Schulsystemvergleich war die politische Fragestellung verbunden, ob durch Gesamtschulen ein Mehr an sozialer Chancengleichheit im Bildungssystem realisiert und damit in der Gesellschaft mehr soziale Gerechtigkeit erreicht werden kann.

2. *Einwände gegen internationale Schulleistungsstudien*

Internationale Schulleistungsstudien sind immer wieder mit dem Einwand konfrontiert worden, dass die untersuchten Bildungssysteme zu unterschiedlich und somit nicht vergleichbar seien. Die besondere Leistung der IEA besteht auch darin, einen Gutteil der bisherigen Vergleichsstudien als Forschungsvorhaben mit primär wissenschaftlicher Fragestellung realisiert zu haben. Das analytische Potenzial der IEA-Studien, d.h. die Analyse der Bedingungsstruktur von Bildungsleistungen, hat bis in die Berichterstattung hinein durchweg die Funktion des Länderrankings dominiert. Die Präsentation von IEA-TIMSS lässt diese Einschätzung jedoch tendenziell fraglich erscheinen, wohingegen die jetzt nachfolgende OECD-PISA-Studie zwar eine dezidiert politische und bildungsökonomisch akzentuierte Auftraggeberschaft hat, jedoch durch die differenziertere Erfassung der Kontextbedingungen schulischer Lernprozesse ein erheblich größeres analytisches Wissen als TIMSS erschließen wird.

T. Husén – einer der „Gründungsväter" der IEA – hat die immensen Anforderungen einer international fairen Testentwicklung keineswegs unterschätzt oder gar ignoriert. Seine kritische Formulierung, „comparing the outcomes of learning in different countries is in several respects an exercise in comparing the incomparable" (Husén 1983, S. 455), wird jedoch häufig als unlösbares Paradoxon missverstanden (vgl. Freudenthal 1975; Keitel/Kilpatrick 1998) oder auf den zumeist eher wissensschaftsfernen Kontext der bildungspolitischen Ergebnisinterpretation übertragen (vgl. Bracey 1997). Husén sieht in der Entwicklung der IEA-Studien eine zunehmende Annäherung an faire internationale Vergleiche und verweist auf TIMSS als ein exzellentes Beispiel international kooperativer Schulleistungsforschung, die unter Nutzung

avancierter diagnostischer Methoden zu einer beträchtlichen Einlösung fairen Vergleichens gelangt (vgl. HUSÉN 1996, S. 217).

Übersetzt man den oben genannten Einwand in die Terminologie der pädagogisch-psychologischen Diagnostik, so wird die Frage der transnationalen Validität gestellt: Können internationale Schulleistungstests so entwickelt werden, dass ihr Anspruch, in den Schulsystemen der beteiligten Staaten Gleiches zu messen, berechtigt ist, und kann dieses „Gleiche" als hinreichend relevant für die zumeist in Curricula kodifizierten Lehrabsichten der Staaten gelten? Nachzuweisen ist weiterhin, dass die nationalen Testformen äquivalente Übersetzungen bzw. Adaptationen des internationalen Tests darstellen (vgl. VAN DE VIJVER/HAMBLETON 1996). Eine zentrale methodische Vergleichbarkeitsvoraussetzung soll hier nicht näher erörtert werden: die Definition bzw. die Ziehung vergleichbarer Untersuchungsstichproben (vgl. ELLEY 1994, S. 223). Sieht man das Ranking als zentrales Ergebnis internationaler Leistungsvergleiche an, so lassen sich die gestellten Fragen auch auf einen alltagssprachlichen Begriff zuspitzen: Sind die Vergleiche und deren Instrumente – hinreichend – fair? Der vorliegende Beitrag präzisiert diesen Begriff und zeigt auf, dass eine sowohl schulpädagogische als auch diagnostische Definition des Fairnessbegriffs genutzt werden kann, um die erreichbar hohe Qualität internationaler Schulleistungsvergleiche zu belegen. Die Fairnessproblematik schulischer Leistungsmessung kann auch auf die pädagogisch-diagnostischen Möglichkeiten sachnormorientierter unterrichtlicher Leistungsbewertung und binnenschulischer Evaluation bezogen werden – jener Beurteilungsbereiche also, die zum unmittelbaren Handlungsfeld der Lehrer gehören. Fairnessprobleme für den Vergleich von Einzelschulen habe ich am Beispiel des Schulrankings in England (vgl. ARNOLD 1999a, S. 81ff.) dargestellt sowie in übergreifender methodischer Perspektive (vgl. ARNOLD 1999b).

3. *Gerechtigkeit und Fairness als schulpädagogische Leitbegriffe*

Einen schulpädagogischen Zugang zu einem spezifischen Bedeutungsaspekt von Fairness entfaltet A. FLITNER (1985) in seinem Beitrag „Gerechtigkeit als Problem der Schule und als Thema des Bildungswesens". Unter Rückgriff auf die Gerechtigkeitsphilosophie von J. RAWLS, in der Gerechtigkeit als die „oberste Tugend sozialer Institutionen" (RAWLS 1977, S. 35) bezeichnet wird, formuliert A. FLITNER folgende Zielorientierung für die Bildungsreform der sechziger und siebziger Jahre: „Das wichtigste und unüberholte, weil grundlegend demokratische Argument für die Veränderung des Schulwesens war die Forderung nach mehr Gerechtigkeit" (FLITNER 1985, S. 2).

In der Praxis der schulischen Leistungsbeurteilung spiegelt sich auch das Ausmaß der „institutionellen Gerechtigkeit" von Schule. Flitners Argumentation greift zurück auf die von ARISTOTELES im fünften Buch der Nikomachischen Ethik entwickelte Theorie der Gerechtigkeit, die proportional verteilende Gerechtigkeit (iustitia distributiva: „jedem das ihm Gemäße") und gleichsetzende Gerechtigkeit (iustitia commutativa: „jedem das Gleiche") unterscheidet (vgl. RITSERT 1997, S. 22ff.). „Ungerechtigkeit" bedeutet im aristotelischen Verständnis entweder die Nichtbeachtung proportionaler Gleichheit oder die glei-

che Behandlung ungleicher Personen – in der Zuteilung von Gütern oder Lasten.

In dem von J. RAWLS (1958/1977) entwickelten Verständnis von sozialer Gerechtigkeit als Fairness wird für den Zugang zu gesellschaftlichen Positionen „faire Chancengleichheit" (fair equality of opportunity) gefordert, welche jedoch mit dem Differenzprinzip verknüpft sein muss: Ungleiche Leistungsvoraussetzungen, die zu Benachteiligungen führen, können dann gerechtfertigt werden, wenn sich diese Bedingungen „zum größtmöglichen Vorteil für die am wenigsten begünstigten Gesellschaftsmitglieder auswirken" (RAWLS 1998, S. 70/71).

Das in den Gesamtschulstudien von FEND und Mitarbeitern verwendete Konzept der „bedingten Chancengleichheit" (FEND/KLAGHOFER 1980, S. 662) kann als beispielhafte Operationalisierung des RAWLSschen Fairnessbegriffs interpretiert werden. Über das – in pädagogischer Sicht durchaus nicht unproblematische – Merkmal der Intelligenz werden interindividuell unterschiedliche, stabile Lernvoraussetzungen relativierend berücksichtigt: Faire Vergleiche resultieren als Folge einer intelligenzbezogenen adaptierten Vergleichsstrategie. Der Nachweis sozialer Gerechtigkeit bei zugleich bestehenden sozialen Unterschieden in den Bildungsabschlüssen kann im Sinne des RAWLSschen Fairnesskonzeptes dann geführt werden, wenn egalisierende Wirkungen der infrage stehenden „Lerngelegenheiten" (Gesamtschule vs. traditionelles Schulsystem) bestehen – und zwar als besondere Förderung der Schüler mit ungünstigen Lernvoraussetzungen. Ob die in den Gesamtschulstudien nachgewiesenen Effekte in diesem Sinne interpretierbar sind, stellt eine nach wie vor kritisch diskutierte Thematik dar (vgl. BAUMERT/RASCHERT 1985; FEND 1990).

Systembezogene Schulleistungsuntersuchungen thematisieren unausweichlich die moralische Dimension der gesellschaftlichen Institution Schule, indem sozial unterschiedliche Bildungsergebnisse ebenso objektiviert werden können wie die faktischen Wirkungen ausgleichender pädagogischer oder allgemeiner sozialer Förderung. Durch die internationale Perspektive der Studien wird der gesellschaftliche Problemlösungsprozess, der sich in der mehr oder minder großen sozialen Ungleichheit der Schulleistungsergebnisse zeigt, in einen besonderen Zusammenhang gestellt, den FEND (1997, S. 370) als „Maximierung von Kontextvarianz" bezeichnet und damit als Chance, alternative Problemlöseversuche auf der Systemebene erkennbar zu machen (vgl. BOS/BAUMERT 1999, S. 14). Das quasi-kausale Erklärungspotenzial der Studien erhält damit reflexive Qualität: Fairness in der Entwicklung der Studien bildet die methodologische Folie für die Prüfung der sozialen Fairness in den verglichenen Bildungssystemen.

4. *Fairness als Qualitätsmerkmal pädagogischer Diagnostik*

4.1 Fairness als Nebenfolge der Orientierung an Testgütekriterien

Neben den drei Hauptgütekriterien (Objektivität, Reliabilität, Validität) werden in der deutschsprachigen diagnostischen Literatur (LIENERT/RAATZ 1994, S. 7ff.; vgl. FISSENI 1997, S. 66) mehrere Nebengütekriterien (z.B. Normierung oder Nützlichkeit) aufgeführt, zu denen neuere Lehrbücher (vgl. AMELANG/ZIELINSKI 1994; KUBINGER 1995) auch das Merkmal der Fairness zählen.

Das Merkmal der Objektivität – d.h. der „Grad der Unabhängigkeit der Ergebnisse vom Untersucher" (LIENERT/RAATZ 1994, S. 7) – gewährleistet durch eine Standardisierung der Durchführung, der Auswertung und der Interpretation des Tests „Gleichheit" der von den Probanden nicht zu beeinflussenden Testbedingungen und somit aus deren Sicht „faire Prüfungs- bzw. Erfassungsbedingungen", in denen Testpersonen keine andere als die fähigkeitsbezogene Beeinflussung der Ergebnisse ermöglicht wird.

Das Kriterium der Reliabilität – d.h. der „Grad der Genauigkeit, mit der ein Merkmal gemessen wird unabhängig davon, ob dieses Merkmal zu messen beansprucht wird" (LIENERT/RAATZ 1994, S. 9) – sichert eine möglichst geringe Messfehlerbelastung der Testergebnisse. Im Rahmen der Klassischen Testtheorie (KTT) kann das Konzept der testspezifischen Messgenauigkeit nur populationsspezifisch bestimmt werden. Die probabilistischen Testmodelle der IRT (Item Response Theory) hingegen nutzen das Ausmaß der Modellpassung auch als Indikator für die messtechnische Güte des Verfahrens; im Falle von Modellakzeptanz ist dann auf gruppengleiche und somit faire Messeigenschaften zu schließen. Die verwendeten Modellgeltungstests sind in ihrer Interpretierbarkeit jedoch umstritten (vgl. KRAUTH 1995, S. 327).

Da sowohl Item- als auch Personenparameter Elemente der IRT-Modelle bilden, geschieht Modellanpassung durch Selektion geeigneter Elemente – mit dem entscheidenden Unterschied, dass zur Konstruktion eines Rasch-konformen Tests nicht nur die Ausscheidung modellunverträglicher Items, sondern auch die Aussonderung „modellunverträglicher Personen" gehört bzw. gehören kann (vgl. ROST 1996, S. 381). Letztgenannte Vorgehensweise ist im Hinblick auf die Testoptimierung bzw. eine realistische Geltungsbereichseinschränkung unvermeidlich; in einer politisch-moralischen Betrachtungsweise erscheint diese jedoch als bedenklich oder unfair.

Das Kriterium der Validität – d.h. der „Grad der Genauigkeit, mit der ein Merkmal gemessen bzw. vorhergesagt wird, das zu messen beansprucht wird" (LIENERT/RAATZ 1994, S. 10) – sichert eine hohe, allerdings inhaltlich zu spezifizierende „Möglichkeit von Generalisierungen, d.h. Rückschlüsse aus dem Verhalten in der Testsituation auf Merkmalsunterschiede außerhalb davon" (AMELANG/ZIELINSKI 1994, S. 136). Wenn gruppenspezifische Validitätsunterschiede bestehen, resultieren erhebliche Fairnessprobleme.

Das auf Korrelationsmaßen basierende und somit populationsabhängige Konzept der externen Validität ermöglicht auch die Prüfung der Prognosegültigkeit bzw. der Bewährung in Selektionsentscheidungen. Das IRT-Konzept enthält diesen Validitätsaspekt jedoch nicht. „Fairness" resultiert in diesem Konzept dadurch, dass eine einheitliche, psychologische Fähigkeit, d.h. eine „latente Dimension", gemessen wird. Üblicherweise werden Testergebnisse über die Anzahl gelöster Aufgaben berechnet, was jedoch das Problem aufwirft, dass unterschiedliche Antwortmuster, die zu gleichen Aufgabensummenwerten führen, auch gleiche Leistungen darstellen. Der maximal unfaire, zugleich aber auch völlig unrealistische Fall träte dann ein, wenn Person A die 50% leichteren Aufgaben löste und Person B die 50% schweren Testaufgaben. Im Modell der KTT besteht „Verrechnungsfairness" (KUBINGER 1995, S. 67; 1996, S. 514), wenn alle Testaufgaben gleich schwer sind. Diese Forderung wird nur selten erfüllt, und sie ist auch wenig nützlich, da sie zu suboptimalen Item-

auswahlen führt. Unter Geltung des Rasch-Modells sind Testsummenwerte hingegen „erschöpfende Statistiken", worin LUKESCH (1998, S. 86) eine besondere „Fairnesseigenschaft" dieses Modells sieht. Grundsätzlich muss für probabilistische Verfahren hingegen die Frage, „was" gemessen wird, theoretisch – und zwar „ex post", d.h. nach Item- und Personenselektion – beantwortet werden. Die größere Rationalität in der Itemselektion mündet somit in das Folgeproblem, den im Sinne der Modellverträglichkeit resultierenden Itempool als Operationalisierung der gewünschten latenten Variablen begründen zu können (vgl. BERGAN 1990, S. 134). Ein ähnliches Theorieproblem ergibt sich für die messinhaltliche Bestimmung der Faktoren einer Faktoranalyse, da auch hier die Menge der einem Faktor zuzuordnenden Items oftmals nicht präzise vorhergesagt werden kann.

Im Modell der probabilistischen Testtheorie bildet die Prüfung der Gruppenfairness des Testverfahrens ein zentrales Element. Modellgeltung kann nur dann beansprucht werden, wenn keine bedeutsamen Subgruppenunterschiede für die Modellparameter feststellbar sind. Damit ist jedoch keineswegs ausgeschlossen, dass Gruppenunterschiede für die Verteilungseigenschaften der Personenparameter, z.B. in Form von geschlechtsspezifischen Mittelwertsunterschieden, bestehen. An dieser Konstellation setzt das im nächsten Abschnitt behandelte Kriterium der „Testfairness" an, das eng mit politischen und moralischen Wertsetzungen und Einschätzungen verbunden und nur bedingt mit statistisch-methodischen Verfahren prüfbar ist.

Eine andere Konstellation ergibt sich jedoch auf der Messebene des einzelnen Items. Diverse statistische Verfahren sind entwickelt worden, um verzerrend messende und in diesem Sinne unfaire Testaufgaben identifizieren zu können (für einen Überblick vgl. ANGOFF 1993). R.J. ADAMS/K.J. ROWE (1990) sprechen von „item bias" und verwenden als Oberbegriff ebenso wie W.H. ANGOFF das eher psychometrisch-technische Konzept des „Differential Item Functioning" (DIF). DIF liegt dann vor, wenn Personen gleicher Fähigkeit, die unterschiedlichen sozialen Gruppen zugehörig sind, nicht hinreichend gleiche Wahrscheinlichkeit zur richtigen Beantwortung des Items besitzen (vgl. HAMBLETON/ROGERS 1989, S. 292). Diese Definition basiert auf der immer wieder ignorierten Voraussetzung, dass der Test selbst fair ist und somit kein „external test bias" besteht, was keineswegs leicht nachweisbar ist. Vieles spricht zudem für die von G. CAMILLI (1993, S. 409) diskutierte Annahme, dass in einem insgesamt fairen Test die gruppenbezogene Bilanz von DIF ausgeglichen sein muss: Den Effekten benachteiligender Items müssen gleiche Effekte anderer, bevorzugender Items gegenüberstehen. Damit wird die Zielorientierung fragwürdig, durch die Eliminierung von DIF-Items die Fairness des Tests insgesamt erhöhen zu können. Auch „innerhalb" eines Items kann eine solche Konstellation auftreten: Für den Spezialfall des „non-uniform DIF" ist die Fairnessproblematik des Items nicht klar abschätzbar, da der gruppenbezogenen Benachteiligung in dem einen Teil des gemessenen Fähigkeitsbereichs eine Bevorzugung im komplementären Fähigkeitsbereich gegenübersteht (vgl. ARNOLD 1999a, S. 36).

Derzeit scheint es so zu sein, dass DIF eine primär analytische Kategorie darstellt. Die zur Identifikation von DIF genutzten Verfahren führen zu Resultaten, die durch direkte Iteminspektion zumeist nicht vorhergesagt werden

können, was gleichbedeutend ist mit der Tatsache, dass sich mangelnde Fairness einzelner Items quasi naturwüchsig einstellt und durch rationale Entwicklungsstrategien kaum vermeidbar ist, wobei allerdings berücksichtigt werden muss, dass in einem IRT-konformen Test per definitionem DIF-Items relativ selten sein müssen, da andernfalls die vorausgesetzte Modellgeltung fragwürdig wird (vgl. BAUMER/KLIEME/WATERMANN 1998, S. 310). R.K. HAMBLETON (1996, S. 919) resümiert die bisherigen Erfahrungen mit DIF-Analysen wie folgt: Die Ursache dieser Itemprobleme kann zumeist nicht identifiziert werden – womit auch die Schlussfolgerung nahe gelegt wird, dass das Rationale einer fairen Testentwicklung auf der Item-Ebene nicht vollständig auszubuchstabieren ist (vgl. ADAMS/ROWE 1990, S. 139; BAUMERT 1998, S. 222).

4.2 Testfairness als eigenständiges Testgütekriterium

Einen zusammenfassenden Überblick über die Möglichkeiten, „Testfairness" als eigenständiges Gütekriterium zu operationalisieren, geben für den deutschen Sprachraum C. MÖBUS (1983) und in kürzerer Form M. AMELANG/W. ZIELINSKI (1994, S. 130–136) sowie L. TENT/I. STELZL (1993, S. 134–142). Insbesondere im angelsächsischen Sprachraum gilt die von A.R. JENSEN (1980) verfasste Monographie mit dem Titel „Bias in mental testing" als Standardwerk. Eine akzentuiert kritische Übersicht über Verfahren zur Einschätzung von „test bias" findet sich bei N.S. COLE/P.A. MOSS (1989).

AMELANG/ZIELINSKI (1994, S. 131) weisen darauf hin, dass es „nicht den fairen Test oder das faire Selektionsverfahren [gibt], sondern nur Fairness im Hinblick auf Handlungs- und Entscheidungsaspekte", wobei diese jedoch expliziert werden müssen. JENSEN (1980, S. 375) versucht diese Problematik durch folgende Unterscheidung zu strukturieren. Als „test bias" bezeichnet er die systematische Fehlerhaftigkeit in der Vorhersage- oder Konstruktvalidität, deren Prüfung beansprucht, gänzlich objektive, empirische und statistische Verfahren zu nutzen. Die Bestimmung von „test fairness" hingegen ist an soziale und damit auch politische sowie ethische Wertentscheidungen gebunden und betrifft den entscheidungsorientierten Einsatz von Testverfahren, wobei diese sowohl unverzerrt als auch verzerrt messen können. „Test bias" kann somit – zunächst im Sinne einer wertfreien Definition – als gruppenspezifische, differenzielle Validität eines Testverfahrens gelten.

S. MESSICK (1989) hat in dem von ihm formulierten umfassenden Konzept der Validität die von JENSEN vorgeschlagene Trennung aufgehoben. Der entscheidungsbezogene Nutzen von Testverfahren stellt ebenso eine Facette der Validität dar wie die sozialen Konsequenzen des Testeinsatzes (vgl. MESSICK 1995). Die Testgütekriterien („measurement principles") fungieren nicht nur als methodische Standards, sondern sie selbst repräsentieren soziale Werte, die immer in Bewertungs- und Auswahlprozessen involviert sind. Aus der Perspektive der Testentwicklung erscheint jedoch die Forderung, dass die Folgen und Nebenfolgen aller denkbaren, testbasierten Selektionsentscheidungen berücksichtigt werden sollten, als überzogen (vgl. COLE/MOSS 1989, S. 217).

Die Unterscheidung unterschiedlicher Validitätsarten, die wiederum in unterschiedlicher Weise von sozialen Wertentscheidungen beeinflusst werden, hat

im Bereich der anwendungsorientierten Diagnostik wenig Relevanz. Für Schulleistungsmessungen gilt, dass Konstruktvalidität, kriterienorientierte Validität und prognostische Validität immer in Kontexten sozialer Selektionsprozesse untersucht werden müssen, da das Schulsystem selbst selektive Strukturen hat bzw. diese nicht hinreichend kompensieren kann. Das Konzept der Fairness wird gleichwohl zumeist mit dem Konzept der prognostischen Validität verknüpft. J.E. HUNTER/F.L. SCHMIDT (1976) haben drei ethische Grundpositionen für die Praxis von Selektionsentscheidungen formuliert, die knappe soziale Ressourcen betreffen wie Arbeits- oder Ausbildungsplätze (vgl. Tab. 1). Die politikwissenschaftlichen Bezeichnungen von G. GUITON/J. OAKES (1995) können diesen zugeordnet werden. K.R. HOWE (1994) bezeichnet die drei Positionen als „formal", „compensatory" und „democratic framework of equality". H.-G. ROLFF (1993) hat politisch akzentuiertere Definitionen für Chancengleichheit gewählt, indem er Position (2) als „liberal-demokratisch" und Position (3) als „radikal-demokratisch" benennt. S. MESSICK (1989, S. 81) sowie A.R. JENSEN (1980, S. 383) haben darauf hingewiesen, dass diese Definitionen unverzerrte Kriteriumsmaße unterstellen und damit in einen Zirkelschluss geraten.

Tabelle 1: **Ethische bzw. politische Grundpositionen im Rahmen der prognostischen Validität von Selektionsentscheidungen**

Ethische Grundpositionen (HUNTER/SCHMIDT 1976)	**Politische Grundpositionen** (GUITON/OAKES 1995)
(1) *Unqualified Individualism* Gruppenunterschiede werden bei der Eignungsvorhersage beachtet. Diese werden entweder als zusätzliche Vorhersagemerkmale in die Entscheidungsregel einbezogen oder gruppenspezifische Entscheidungsregeln werden formuliert (Fairness als gruppenspezifische Kriteriumsschätzung).	(1) *The Libertarian Position* Der Zugang zu Lerngelegenheiten erfolgt aufgrund leistungsrelevanter Merkmale (Meritokratie), wobei Gruppenzugehörigkeit ein akzeptables Merkmal darstellt. Staatliche Eingriffe bzw. Regelungen bleiben minimal (Fairness als uneingeschränkter Wettbewerb).
(2) *Qualified Individualism* Gruppenunterschiede werden strikt ignoriert, sowohl für die Verteilung in der Grundgesamtheit als auch für die Verteilung nach der Entscheidung. Die Auswahl erfolgt strikt „kriteriumsbezogen", wie immer auch dieses Kriterium definiert sein mag (Fairness als gruppengleiche Kriteriumsschätzung).	(2) *The Liberal Position* Der Zugang zu Lerngelegenheiten erfolgt ausschließlich aufgrund inhaltlich als leistungsrelevant begründbarer Merkmale, wobei Gruppenzugehörigkeit als Merkmal inakzeptabel ist. Staatliche Eingriffe bzw. Regelungen sorgen für gruppengleiche Lerngelegenheiten (Fairness als gruppenspezifisch adaptierter Wettbewerb).
(3) *Fair Share* Die Auswahl der Geeigneten soll die Gruppenanteile in der Population widerspiegeln (gruppengleiche Repräsentation).	(3) *The Democratic Liberal Position* Der Zugang zu Lerngelegenheiten erfolgt gruppenspezifisch, wobei staatliche Eingriffe und Regelungen für das Erreichen gruppenübergreifend gleicher Lernergebnisse sorgen (Fairness als gruppeninterner Wettbewerb bei Ausschaltung des Wettbewerbs zwischen Gruppen).

Die Position des „Faire Share" („Quotenmodell") gründet auf der Setzung, dass unter den testbasiert ausgewählten Bewerbern die Proportion der miteinander verglichenen Gruppen dieselbe ist wie in der Gesamtpopulation. Eine testpraktische Realisierung dieses Modells besteht in separierten Testnormen für z.B. weibliche bzw. männliche Personen oder für Gruppen, für die die Testsprache Primärsprache bzw. Fremdsprache ist. Ebenso werden „Altersnormen" begründet. AMELANG/ZIELINSKI (1994, S. 132) verweisen auf die methodische Unstimmigkeit, dass „von einer Normierung für die verschiedenen sozioökonomischen Schichten in den allermeisten Fällen abgesehen wird, obwohl gerade im Hinblick darauf die Mittelwertunterschiede gravierend sind". Für den Bereich soziostruktureller Analysen von Bildungssystemen erfordert das Quotenmodell eine kritische Festlegung: Fairness als „equal access to outcomes" kann nur bestimmt werden für Standardqualifikationen, die von allen Schülern erreicht werden sollen, und nähert sich damit dem umstrittenen „minimum competence testing" (ARNOLD 1999a, S. 151). J.P. KEEVES/C. MORGENSTERN/L.J. SAHA (1991, S. 78) halten die Position der gruppengleichen Repräsentation für ein unerreichbares Ziel. Sie können jedoch anhand eines Ergebnisvergleichs der ersten und zweiten IEA-Science-Studie nachweisen, dass international in einem Zeitraum von 15 Jahren die soziale Benachteiligung im Zugang zu höheren sekundären Bildungseinrichtungen reduziert worden ist. Diese Konstellation bezeichnen sie als eine Zunahme von Chancengerechtigkeit („equity" im Sinne von „fairness in access by an individual to an opportunity"). H.-P. BLOSSFELD/Y. SHAVIT (1993) sowie zusammenfassend H. DITTON (1995) gelangen jedoch zu pessimistischeren Schlussfolgerungen.

In den USA ist das Kriterium der Testfairness bereits seit geraumer Zeit zu einem professionellen Standard psychologischer und erziehungswissenschaftlicher Testanwendung geworden. So werden in den bekannten „Standards for Educational and Psychological Testing" (American Psychological Association [APA] 1985) das Ausweisen von „differential prediction" ebenso verlangt wie Testanpassungen für Minoritäten. Diese Standards sind erst 1998 ins Deutsche übersetzt und publiziert worden (vgl. HÄCKER/LEUTNER/AMELANG 1998). Die 1986 vom Testkuratorium Deutscher Psychologenvereinigungen veröffentlichten „Kriterien für die Testbeurteilung" beinhalten das Kriterium der Fairness.

Eine grundlegende Strategie zur konzeptionellen Minderung von Testunfairness besteht darin, die Testkonstruktion ausschließlich auf Aufgabenformen und -bereiche zu richten, von denen angenommen werden kann, dass diese „nur solche Erfahrungen für die Lösung ... voraussetzen, die verschiedenen Kulturen gemeinsam ist" (ANASTASI 1964 in: SIMONS/MÖBUS 1978, S. 191). Damit ist das einerseits viel versprechende, andererseits stark kritisierte Modell der „kulturfairen" bzw. „kulturfreien Tests" (insbesondere in dieser Weise konstruierter Intelligenztests) thematisiert. W.J. POPHAM (1990, S. 181f.) beschreibt die Entwicklungsbemühungen um entsprechende Intelligenztests, deren bekanntester – der *Culture-Fair Intelligence Test* (CFT) – in den USA von CATTELL konstruiert worden ist und unter der Bezeichnung „Grundintelligenztest" (CFT1, CFT20, CFT3; vgl. WEISS 1997) auch in der Bundesrepublik Deutschland große Verbreitung gefunden hat. H. SIMONS/C. MÖBUS (1978, S. 191) erläutern die validitätsbezogenen Einwände gegen diesen Ansatz (Ignorieren sprachlicher, kognitiver Leistungen) und folgern, dass in dieser Weise konstru-

ierte Tests immer „nur bis zu einem gewissen Grade kulturfrei" sein können. S. HEGARTY (1990, S. 368) wählt deshalb die Bezeichnung „culture reduced test". J. GUTHKE/K. WIEDL (1996, S. 226) vertreten die radikale Position, dass die Entwicklung „kulturfairer Tests" prinzipiell nicht möglich sei, da Lernprozesse immer in kulturellen Zusammenhängen stattfinden und somit kulturell geprägt werden.

5. *Fairnessaspekte in der Auswertung internationaler Schulleistungsstudien: Das Konzept der Test-Curriculum Matching Analysis (TCMA) von TIMSS*

Die Entwicklung der TIMS-Studie erfolgte in einem Prozess intensiver wissenschaftlicher Kooperation der bedeutendsten nationalen Bildungsforschungsinstitute aus über 40 Ländern unter Leitung des IEA-TIMSS-Studienzentrums am Boston College (zuvor Universität Vancouver). Die Erstellung des Item-Pools unterlag strengen und alle beteiligten Nationen in gleicher Weise betreffenden Anforderungen zur Sicherung der inhaltlichen Validität (vgl. GARDEN/ORPWOOD 1996; zusammenfassend: ARNOLD 1999a, S. 94–100). Dieser Gleichheitsaspekt der Testfairness wurde ergänzt durch den differenziellen Fairnessaspekt: Die kulturelle Akzeptanz und Passung der Iteminhalte sowie die Erstellung optimaler Itemübersetzungen wurden durch spezifische nationale Itemadaptionen realisiert (vgl. MULLIS/KELLY/HALEY 1996; zusammenfassend: ARNOLD 1999a, S. 103–104). Aus diagnostischer Perspektive kann die Testaufgabenübersetzung nicht ausschließlich an dem Prinzip linguistischer Gleichwertigkeit orientiert werden, da formal korrekte Übersetzung gleichwohl zu Verzerrungen des Messinhaltes führen kann. R.J. MISLEVY (1995, S. 424) bezeichnet die validitätsoptimierte Übersetzungsstrategie deshalb als „functional rather than literal equivalence".

Die TIMSS-Fachleistungstests basieren auf dem Konzept der kriterienorientierten Leistungsmessung. Ein mehrstufiges Verfahren wurde verwendet, um die Validität der Tests zu maximieren. Dieses Qualitätsmerkmal gilt für die internationale Testversion, deren Itemmenge von allen Nationen verbindlich zu administrieren war. Im Folgenden soll als Beispiel für die Leistungsfähigkeit moderner Testmethodik ein Verfahren vorgestellt werden, das die Fairnessqualität eines in dieser Weise optimalen internationalen Testverfahrens nochmals durch eine differenzielle Passungsstrategie zu steigern vermag, ohne gegen den Grundsatz gleicher Testbedingungen zu verstoßen.

In der der Itementwicklung vorausgehenden TIMSS-Curriculumstudie wurde eine Gesamtanalyse der von den fast 50 beteiligten Staaten entwickelten Curricula in den Fächern Mathematik und Naturwissenschaften sowie eine Untersuchung der in diesen Fächern gebräuchlichen Lehrbücher (vgl. SCHMIDT/MCKNIGHT/VALVERDE/HOUANG/WILEY 1997) erarbeitet. Ergänzt wurde die Untersuchung durch eingehende Befragungen von nationalen Curriculumexperten. Um die Curriculumanalysen weitgehend unabhängig von den vorhandenen, kulturellen Färbungen mathematischer Leistungsvollzüge („performance processes") zu halten und damit einen Aspekt von Fairness zu wahren, wurden „Leistungserwartungen" („expectations for student mathematical performan-

ces") spezifiziert: „This decision focused on more generic, less culture-bound task expectations and demands" (SCHMIDT u.a. 1997, S. 185). Die international übereinstimmenden Curriculumelemente („commonly intended mathematics topics", SCHMIDT u.a. 1997, S. 97) bezeichnen J. BAUMERT/R. LEHMANN (1997, S. 187) als „internationales Kerncurriculum mit konsensueller Jahrgangsstufenbehandlung". Auf dieses Kerncurriculum ist der Prozess der Itementwicklung ausgerichtet worden, dessen Ergebnis, d.h. die TIMSS-Leistungstests, eine hohe, nicht jedoch eine perfekte auch ebenso wenig eine für alle beteiligten Nationen gleiche Inhaltsvalidität beanspruchen kann.

Eine bedeutsame Validierungsstrategie für die entwickelten Untersuchungsinstrumente besteht darin, deren Lehrplanangemessenheit empirisch zu prüfen. D. SPEARRITT (1990, S. 452) fasst unter diesem Aspekt die Ergebnisse der Second International Science Study (SISS) wie folgt zusammen: „Schulleistung zeigt eine signifikant positive Korrelation mit dem Ausmaß der Gelegenheit, den Unterrichtsstoff zu lernen." Während bei allgemeiner Betrachtung diese Feststellung trivial erscheint, gewinnt sie für die Suche nach Optimalitätsbedingungen im Rahmen von Schulsystemvergleichen eine entscheidende Bedeutung.

In Erweiterung des bereits in SISS entwickelten Konzeptes ist für TIMSS ein Wirkungsmodell für Schulleistungen formuliert worden, das die zentralen Systemebenen einbezieht und zugleich den Begriff des Curriculums in den Mittelpunkt der Analyse stellt (vgl. Abb. 1). Die Erfassung „unterrichtlicher Lerngelegenheiten" wird erweitert zu einer kontextorientierten Erfassung von Lernerfahrungen („potential educational experiences"). In der Darstellung soll

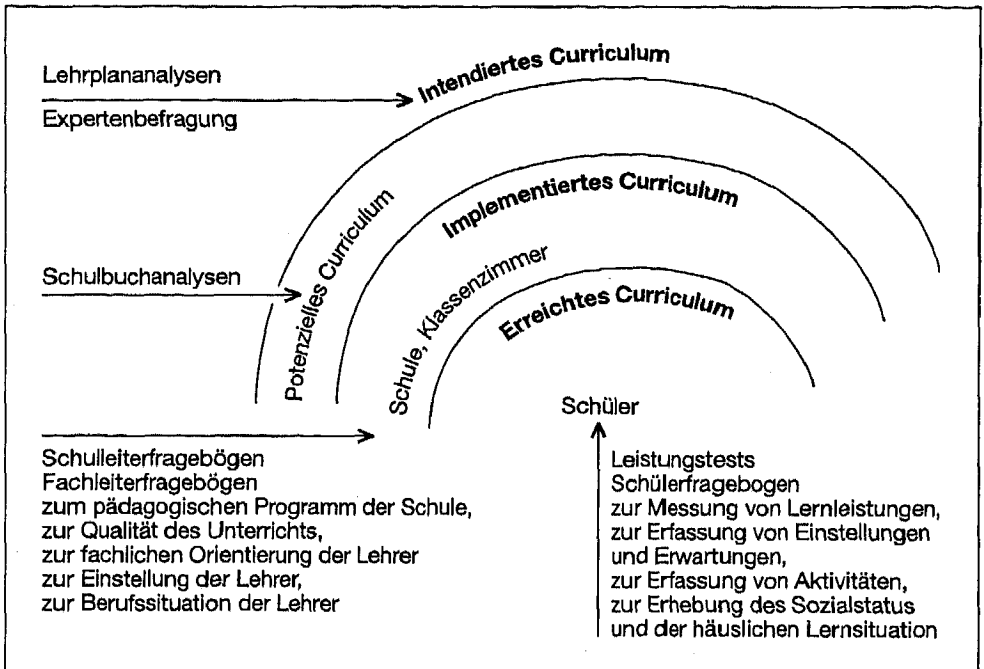

Abbildung 1: **TIMSS-Curriculummodell und Modell der Lerngelegenheiten**

jedoch der Begriff der „Lerngelegenheiten" beibehalten werden, da der Begriff der „Erfahrung" insbesondere für sozioökonomische Merkmale eine psychische Wirkungsqualität unterstellt, die operational nicht einfach zu beschreiben ist.

Im Sinne einer spezifischen Testauswertungsstrategie kann geprüft werden, ob – präzise formuliert – ein hinreichend großer Teil der Testaufgaben, bezogen auf die Lehrpläne eines Landes, auch eine hohe „Unterrichtsvalidität" (BAUMERT/LEHMANN 1997, S. 188) aufweist. Schon diese operationale Definition macht deutlich, dass hier ein indirektes Nachweisverfahren genutzt wird, das in den internationalen Berichten die Bezeichnung „Test-Curriculum Matching Analysis" (TCMA) trägt. Die TCMA kehrt den Prozess der Skalenentwicklung in gewisser Weise um und unterschreitet das hohe Informationsniveau, das in der Gesamtstudie erreicht wird (vgl. BEATON/GONZALES 1997, S. 187–188). Das Verfahren der TCMA erfragte von den Curriculumexperten der Länder pro Item der TIMSS-Tests, ob dieses mit mehr als 50-prozentiger Wahrscheinlichkeit als „dem Lehrplan der jeweiligen Klassenstufe zugehörig" bezeichnet werden kann. Diese Quotierung berücksichtigt, dass ein Item für bestimmte, nicht aber für alle Schulformen eines Landes curriculare Geltung haben kann. Die Ergebnisse dieser Analyse der „curricularen Passung" zeigen beträchtliche Unterschiede zwischen den beteiligten Ländern auf. Abweichend vom internationalen Kriterium wurde für TIMSS-Deutschland eine 60-Prozent-Quote festgelegt, die jedoch weit unter dem auffindbaren Passungsniveau liegt (vgl. BAUMERT/LEHMANN 1997, S. 185).

„Zur Kontrolle der transkulturellen Fairness", wie BAUMERT/LEHMANN (1997, S. 188) formulieren, wurden für jedes Teilnehmerland die mittleren Prozentanteile richtig gelöster Aufgaben („average percent correct") unter drei Voraussetzungen berechnet: (a) Lösungsquote für die gesamte Aufgabenmenge des Tests, (b) Lösungsquote für die national-curricular passende Aufgabenmenge, (c) Lösungsquoten für alle alternativ von anderen Nationen gewählten, curricular passenden Aufgabenmengen. BAUMERT/LEHMANN (1997, S. 188) bezeichnen die unter (b) genannten Aufgabenteilmengen als „Tests optimaler nationaler Validität", was zwar unter der Voraussetzung, dass das IRT-Modell gültig ist, zulässig ist, da „item sampling" innerhalb gewisser Grenzen zu identischen Parameterschätzungen führt, allerdings zugleich das Missverständnis begünstigt, die TIMSS-Leistungstests seien suboptimale, internationale Testverfahren. BAUMERT (1998, S. 220) bezeichnet diese Konstellation als „moderate Validitätsmängel". Die Ergebnisse der TCMA zeigen im paarweisen Ländervergleich kaum Änderungen der Leistungspositionen, was als ein sehr bedeutsames Argument für die Fairness des gesamten Testverfahrens gewertet werden kann.

Zusammenfassend kann festgestellt werden, dass die Identifikation eines internationalen Kerncurriculums den Gleichheitsaspekt der Untersuchungsfairness sichert (equal opportunity to learn), während das Verfahren der TCMA unterhalb dieser Gleichheitsebene individuelle, d.h. nationale Adaptation zulässt und damit für ausgleichende Passung (equity) der Testuntersuchung sorgt. Beide Merkmale gelten jedoch nur unter der Voraussetzung, dass die Lehrpläne (intendiertes Curriculum) eine hinreichende Steuerungswirkung für den Unterricht in den Klassen (implementiertes Curriculum) haben, was aufgrund der Untersuchungsergebnisse von W. VOLLSTÄDT/K.-J. TILLMANN/U. RAUIN/

U. HÖHMANN/A. TEBRÜGGE (1999) eher fraglich erscheint. Und schließlich ist Fairness der internationalen Vergleiche auch auf der Ebene des erreichten Curriculums, d.h. bei den tatsächlichen Unterrichtsmerkmalen sowie den Lernvoraussetzungen und lernunterstützenden, außerschulischen Erfahrungen der Schüler zu sichern. Der in TIMSS genutzte Schülerfragebogen umfasst jedoch nur wenige Items zu diesem Bereich; die PISA-Studie wird diese Kontextualisierungsmerkmale sehr viel differenzierter erfassen.

6. *International faire Schulleistungsmessung als multiperspektivische Systemanalyse*

Internationale Schulleistungsstudien, die das Differenzierungsniveau von TIMSS oder PISA erreichen, sollten in Bezug auf jede der drei in Tabelle 1 genannten ethischen bzw. politischen Grundpositionen ausgewertet werden. Diese Schlussfolgerung am Ende eines Beitrags über Untersuchungsfairness mag überraschen; sie hat jedoch eine besondere Berechtigung. TIMSS kann aufgrund der Aufgabenbindung an das internationale Kerncurriculum nur in begrenztem Maße „Bestenidentifkation" betreiben, da lehrplanorientierte Tests immer auch die Idee des „mastery learning" und damit die egalisierende Zielerreichung für möglichst viele Schüler einschließen; ein uneingeschränkter internationaler Leistungswettbewerb (libertäre Position) findet somit nicht statt. Für PISA hingegen wird das internationale Ranking dieses Merkmal eher aufweisen können, da die Kompetenzorientierung der Skalen eine enge Lehrplanbindung meidet (vgl. Deutsches PISA-Konsortium 2000) und unklar bleibt, ob diese Kompetenzen in allen beteiligten Nationen gleichermaßen ausgebildet werden sollen oder können.

Die liberale Position wird in TIMSS durch die Orientierung an einem für alle beteiligten Nationen für verbindlich gehaltenen Kerncurriculum bzw. dessen national optimierter Variante (TCMA) eingenommen. PISA wird hingegen für diese Position in weitaus stärkerem Maße die Berücksichtigung der Kontextvariablen nutzen, die eine Berechnung adjustierter Länderkennwerte ermöglicht.

Die liberal-demokratische Position kann in TIMSS und umfassender noch in PISA dargestellt werden, indem mit typischen Maßen der Chancengleichheit (z.B. sozialstatusbezogene Bildungsbeteiligung bzw. Varianzanteile an Leistungstestergebnissen) gerechnet und diese mit dem Zielkriterium gruppengleicher Quoten verglichen werden. PISA geht auch hier einen Schritt über TIMSS hinaus und nähert sich der schon von J.S. COLEMAN (1968, S. 22) vertretenen Position, dass Chancengleichheit nicht nur materiell und inhaltlich gleiche Bildungsangebote voraussetzt, sondern sich auch auf die relative Intensität der schulischen Wirkung im Vergleich zu konkurrierenden externen Einflüssen auf die Lernleistung der Schüler erstreckt. Diese „Kontextualisierung" von Schulleistung bildet die Grundidee der modernen Schuleffizienzforschung (vgl. SCHEERENS/BOSKER 1997).

Werden hoch kontextualisierte Leistungskennwerte als Indikatoren verwendet, so ergeben sich für die Rechenschaftslegung der Einzelschule und des Lehrpersonals durchaus faire Vergleiche: Jene Leistungsvarianz, die auf päd-

agogisch wenig beeinflussbare Merkmale zurückgeht, wird aus den Indikatoren herausgenommen. Dennoch verschleiert gerade diese Adaptivität die Faktenlage sozial ungleicher Bildungsergebnisse. Das ist der „Preis", der für die Zugrundelegung eines differenzierten Schuleffizienzmodells zu zahlen ist: Schulsysteme, die Schüler mit sehr geringen Lernvoraussetzungen und ungünstigen familiären Hintergrundbedingungen fördern und ihnen, bezogen auf die niedrige Lernausgangslage, zum Ende der Schulzeit eine beträchtliche Menge an Wissen und Können vermittelt haben, müssen als hochgradig effektiv bezeichnet werden, auch wenn ihre Absolventen kaum an jenes Niveau heranreichen, das andere Systeme für Schüler aus günstigeren Verhältnissen erreichen. Anders formuliert, bedeutet dies: Die soziale Selektivität eines Bildungswesens zeigt sich ganz nur in der Perspektive der libertären Position.

Würde nicht so viel Unfug und politische Effekthascherei mit den Ergebnissen der diagnostisch exzellenten Schulleistungsstudien betrieben, so könnte R.J. MISLEVY nur zugestimmt werden: Das volle Bild erschließt sich erst unter mehrfachen Perspektiven:

„My answer to people who want comparative standings is to give them comparative standings – lots of them: in different topics, at different ages, with different kinds of tasks, both unadjusted and adjusted for factors such as national curricula and proportion of students in school. Recognizing that no single index of achievement can tell the full story and that each has its own limitations, we increase our understanding of how nations compare by increasing the breadth of our vision. Even so, however, simply ascertaining nations' relative standing tells us little about how to set educational policy or improve instructional practice." (MISLEVY 1995, S. 419)

Internationale Schulleistungsstudien verstärken auch eine gesamtgesellschaftliche Dynamik. Da sich derzeit auch in unserem Land der nationale Blick weitet und die Bildungsergebnisse anderer Nationen mit großem Interesse in den Vergleich genommen werden, so entfaltet sich damit zugleich ein Bemühen um Multikulturalität und internationale Kooperation. Wird hier nicht – auf dem Umweg über die pädagogische Diagnostik – eines der von W. KLAFKI postulierten „epochaltypischen Schlüsselprobleme" angegangen: die „Problematik des Nationalitätsprinzips, mit anderen Worten: Die Frage nach Kulturspezifik und Interkulturalität" (KLAFKI 1995, S. 12)?

Literatur

ADAMS, R.J./ROWE, K.J.: Item bias. In: H. WALBERG/G.D. HAERTEL (Hrsg.): The international encyclopedia of educational evaluation. Oxford (Pergamon) 1990, S. 133–139.

AMELANG, M./ZIELINSKI, W.: Psychologische Diagnostik und Intervention. Berlin 1994.

American Psychological Association (APA): Standards for educational and psychological testing. Washington, DC (APA) 1985.

ANGOFF, W.H.: Perspectives on differential item functioning methodology. In: P.W. HOLANND/H. WAINER (Hrsg.): Differential item functioning. Mahwah/NJ (Erlbaum) 1993, S. 3–23.

ARNOLD, K.-H.: Fairness bei Schulsystemvergleichen: Diagnostische Konsequenzen von Schulleistungsstudien für die unterrichtliche Leistungsbewertung und binnenschulische Evaluation. Münster 1999 (a).

ARNOLD, K.-H.: Schulen im Vergleich. Probleme des Ranking und Chancen eines Monitoring. In: Deutsche Schule 91 (1999), S. 218–231 (b).

BAUMERT, J.: Internationale Schulleistungsvergleiche. In: D.H. ROST (Hrsg.): Handwörterbuch der Pädagogischen Psychologie. Weinheim 1998, S. 219–225.

BAUMERT, J./KLIEME, E./WATERMANN, R.: Jenseits von Gesamttest und Untertestwerten: Analyse differentieller Itemfunktionen am Beispiel des mathematischen Grundbildungstests der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie der IEA (TIMSS). In: H.-J. HERBER/F. HOFMANN (Hrsg.): Schulpädagogik und Lehrerbildung. Festschrift zum 60. Geburtstag von Josef Thonhauser. Innsbruck (Studienverlag) 1998, S. 301–324.

BAUMERT, J./LEHMANN, R.: TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen 1997.

BAUMERT, J./RASCHERT, J.: Gesamtschule. In: K.-P. HEMMER/H. WUDTKE (Hrsg.): Erziehung im Primarschulalter. Enzyklopädie Erziehungswissenschaft. Bd. 7. Stuttgart 1985, S. 229–269.

BEATON, A.E./GONZALES, E.J.: TIMSS test-curriculum matching analysis. In: M.O. MARTIN/D.L. KELLY (Hrsg.): Third International Mathematics and Science Study (TIMSS) Technical Report. Vol. II: Implementation and analysis (Population 1 and Population 2). Chestnut Hill/MA (Boston College) 1997, S. 187–194.

BERGAN, J.R.: Contributions of behavioral psychology to school psychology. In: T.B. GUTKIN/C.R. REYNOLDS (Hrsg.): The handbook of school psychology. New York (Wiley) ^{2}1990, S. 126–142.

BLOSSFELD, H.-P./SHAVIT, Y.: Dauerhafte Ungleichheiten. Zur Veränderung des Einflusses der sozialen Herkunft auf die Bildungschancen in dreizehn industrialisierten Ländern. In: Zeitschrift für Pädagogik 39 (1993), S. 25–32.

BOS, W./BAUMERT, J.: Möglichkeiten, Grenzen und Perspektiven internationaler Bildungsforschung: Das Beispiel TIMSS/III. In: Das Parlament 35-36 (1999), S. 3–15.

BRACEY, G.W.: On comparing the incomparable: A response to Baker and Stedman. In: Educational Researcher 26 (1997), S. 19–26.

CAMILLI, G.: The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In: P.W. HOLLAND/H. WAINER (Hrsg.): Differential item functioning. Mahwah/NJ (Erlbaum) 1993, S. 397–418.

COLE, N.S./MOSS, P.A.: Bias in test use. In: R.L. LINN (Hrsg.): Educational measurement. 3rd ed. New York (Macmillan) 1989, S. 201–219.

COLEMAN, J.S.: The concept of equality of educational opportunity. In: Harvard Educational Review 38 (1968), S. 7–22.

Deutsches PISA-Konsortium (Hrsg.): Schülerleistungen im internationalen Vergleich. Eine neue Rahmenkonzeption für die Erfassung von Wissen und Fähigkeiten (Original: Measuring Student Knowledge and Skills. A New Framework for Assessment. Paris: OECD). (Berlin: Max-Planck-Institut für Bildungsforschung) 2000.

DITTON, H.: Ungleichheitsforschung. In: H.-G. ROLFF (Hrsg.): Zukunftsfelder der Schulforschung. Weinheim 1995, S. 89–124.

ELLEY, W.B. (Hrsg.): The IEA study of reading literacy: Achievement and instruction in thirty-two school systems. Oxford (Pergamon) 1994.

FEND, H.: Gesamtschule im Vergleich. Weinheim 1982.

FEND, H.: Bilanz der empirischen Bildungsforschung. In: Zeitschrift für Pädagogik 36 (1990), S. 687–709.

FEND, H.: Schulleistung und Fähigkeitsselbstbild: Literaturüberblick. In: F.E. WEINERT/A. HELMKE (Hrsg.): Entwicklung im Grundschulalter. Weinheim 1997, S. 361–371.

FEND, H./KLAGHOFER, R.: Durchlässigkeit und Chancengleichheit in unterschiedlichen Schulsystemen. Dargestellt am Beipiel des Flächenversuchs Wetzlar. In: Zeitschrift für Pädagogik 26 (1980), S. 653–671.

FISSENI, H.: Lehrbuch der psychologischen Diagnostik. Göttingen ^{2}1997.

FLITNER, A.: Gerechtigkeit als Problem der Schule und als Thema des Bildungswesens. In: Zeitschrift für Pädagogik 31 (1985), S. 1–26.

FREUDENTHAL, H.: Schülerleistungen im internationalen Vergleich. In: Zeitschrift für Pädagogik 21 (1975), S. 889–910.

GARDEN, R.A./ORPWOOD, G.: Development of the TIMSS achievement tests. In: M.O. MARTIN/D.L. KELLEY (Hrsg.): Third International Mathematics and Science Study (TIMSS) Technical Report. Vol. I: Design and Development. Chestnut Hill/MA (Boston College) 1996, S. 1–20.

GUITON, G./OAKES, J.: Opportunity to learn and conceptions of educational equality. In: Educational Evaluation and Policy Analysis 17 (1995), S. 323–336.

GUTHKE, J./WIEDL, K.: Dynamisches Testen. Zur Psychodiagnostik der intraindividuellen Variabilität. Göttingen 1996.

HÄCKER, H./LEUTNER, D./AMELANG, M. (Hrsg.): Standards für pädagogisches und psychologisches Testen. Bern (Huber) 1998.

HAENISCH, H./LUKESCH, H.: Ist die Gesamtschule besser? Gesamtschulen und Schulen des gegliederten Schulsystems im Leistungsvergleich. München 1980.

HAMBLETON, R.K.: Advances in assessment models, methods, and practices. In: D.C. BERLINER/

R. C. CALFEE (Hrsg.): The handbook of educational psychology. New York (Macmillan) 1996, S. 899–925.

HAMBLETON, R.K./ROGERS, H.J.: Die Anwendung von Item-Response-Modellen in nationalen Lernerfolgsmessungen. In: K. INGENKAMP/W. SCHREIBER (Hrsg.): Was wissen unsere Schüler? Weinheim 1989, S. 267–310.

HEGARTY, S.: Culture-fair assessment. In: H. WALBERG/G.D. HAERTEL (Hrsg.): The international encyclopedia of educational evaluation. Oxford (Pergamon Press) 1990, S. 367–368.

HOWE, K.R.: Standards, assessment, and conceptions of equality. In: Educational Researcher 23 (1994), S. 27–33.

HUNTER, J.E./SCHMIDT, F.L.: Critical analysis of the statistical and ethical implications of various definitions of test bias. In: Psychological Bulletin 83 (1976), S. 1053–1071.

HUSÉN, T.: Are standards in US schools really lagging behind those in other countries? In: Phi Delta Kappan 64 (1983), S. 455–461.

HUSÉN, T.: Lessons from the IEA studies. In: International Journal of Educational Research 25 (1996), S. 207–218.

JENSEN, A.R.: Bias in mental testing. New York (Free Press) 1980.

KEITEL, C./KILPATRICK, J.: The rationality and irrationality of international comparative studies. In: G. KAISER/E. LUNA/I. HUNTLEY (Hrsg.): International comparisons in mathematics education. London (Falmer) 1998, S. 241–257.

KEEVES, J.P./MORGENSTERN, C./SAHA, L.J.: Educational expansion and equality of opportunity: Evidence from studies conducted by IEA in ten countries in 1970–71 and 1983–84. In: International Journal of Educational Research 15 (1991), S. 61–80.

KLAFKI, W.: „Schlüsselprobleme" als thematische Dimensionen eines zukunftsorientierten Konzeptes von „Allgemeinbildung". In: W. MÜNZINGER/W. KLAFKI (Hrsg.): Schlüsselprobleme im Unterricht. Thematische Diskussionen einer zukunftsorientierten Allgemeinbildung (Deutsche Schule, 3. Beiheft). Weinheim 1995, S. 9–14.

KRAUTH, J.: Testkonstruktion und Testtheorie. Weinheim 1995.

KUBINGER, K.D.: Einführung in die Psychologische Diagnostik. Weinheim 1995.

KUBINGER, K.D.: Objektive Diagnostik. In: K. PAWLIK (Hrsg.): Grundlagen und Methoden der Differentiellen Psychologie. Göttingen 1996, S. 508–544.

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

LIENERT, G./RAATZ, U.: Testaufbau und Testanalyse. Weinheim ^{5}1994.

LUKESCH, H.: Einführung in die pädagogisch-psychologische Diagnostik. Regensburg ^{2}1998.

MESSICK, S.: Validity. In: R.L. LINN (Hrsg.): Educational measurement. New York (McMillan) ^{3}1989, S. 13–103.

MESSICK, S.: Validity of psychological assessment: Validation of inferences from persons' responses and performance as a scientific inquiry into score meaning. In: American Psychologist 50 (1995), S. 741–749.

MISLEVY, R.J.: What can we learn form international assessments? In: Educational Evaluation and Policy Analysis 17 (1995), S. 419–437.

MÖBUS, C.: Die praktische Bedeutung der Testfairness als zusätzliches Kriterium zu Reliabilität und Validität. In: R. HORN/K. INGENKAMP/R.S. JÄGER (Hrsg.): Tests und Trends. 3. Jahrbuch der Pädagogischen Diagnostik. Weinheim 1983, S. 155–204.

MULLIS, I.V.S./KELLY, D.L. /HALEY, K.: Translation verification procedures. In: M.O. MARTIN/I.V.S. MULLIS (Hrsg.): Third International Mathematics and Science Study (TIMSS): Quality assurance in data collection. Chestnut Hill/MA (Boston College) 1996, S. 1–14.

POPHAM, W.J.: Modern educational measurement: A practioner's perspective. Englewood Cliffs/NJ (Prentice Hall) 1990.

RAWLS, J.: Gerechtigkeit als Fairness. In: O. HOEFFE (Hrsg.): Gerechtigkeit als Fairness. Freiburg 1977, S. 34–83. (Original: Justice as fairness. The Philosophical Review 67 [1958], S. 164–194.)

RAWLS, J.: Politischer Liberalismus. Frankfurt a.M. 1998.

RITSERT, J.: Gerechtigkeit und Gleichheit. Münster 1997.

ROLFF, H.-G.: Chancengleichheit. In: D. LENZEN (Hrsg.): Pädagogische Grundbegriffe. Bd. 2. Reinbek 1993, S. 293–298.

ROST, J.: Testtheorie, Testkonstruktion. Bern (Huber) 1996.

SCHEERENS, J./BOSKER, R.J.: The foundations of educational effectiveness. Oxford (Pergamon) 1997.

SCHMIDT, W.H./MCKNIGHT, C.C./VALVERDE, G.A./HOUANG, R.T. (Hrsg.): Many visions, many aims. A cross-national investigation of curricular intentions in school mathematics. Boston/MA (Kluwer) 1997.

SIMONS, H./MÖBUS, C.: Testfairness. In: K. KLAUER (Hrsg.): Handbuch der pädagogischen Diagnostik. Bd. 1. Düsseldorf 1978, S. 187–197.

SPEARRITT, D.: Evaluation of national comparisons. In: T. HUSÉN/T.N. POSTLETHWAITE (Hrsg.): The international encyclopedia of education: Research and studies. Oxford (Pergamon) 1990, S. 447–455.

TENT, L./STELZL, I.: Pädagogisch-psychologische Diagnostik. Bd. 1: Theoretische und methodische Grundlagen. Göttingen 1993.

VIJVER, F. VAN DE/HAMBLETON, R.K.: Translating Tests: Some Practical Guidelines. In: European Psychologist 1 (1996), S. 89–99.

VOLLSTÄDT, W./TILLMANN, K.-J./RAUIN, U./HÖHMANN, K./TEBRÜGGE, A.: Lehrpläne und Schulalltag. Eine empirische Studie zur Akzeptanz und Wirkung von Lehrplänen in der Sekundarstufe I. Opladen 1999.

WEISS, R.H.: Grundintelligenztest Skala 2 (CFT 20) mit Wortschatztest (WS) und Zahlenfolgentest (ZF). Handanweisung für die Durchführung, Auswertung und Interpretation. Braunschweig ^{4}1997.

Abstract

International studies on school achievement claim to record relevant learning results in the school systems of the countries participating in order to provide quality indicators for descriptive and analytic comparisons. Thus, the question arises whether the methods of pedagogical diagnosis employed here suffice to allow for similar research conditions which are at the same time sufficiently sensitive regarding the national peculiarities of the educational systems. This issue may be summarized in the concept of fairness. From the school-pedagogical point of view, fairness aims at the question of social justice within the institution ‚school'. In the field of psychological diagnosis and in that of test theory, in particular, fairness represents an instrumental value criterion or, rather, the result of an adequate test quality. Taking TIMSS (Third International Mathematics and Science Study) as an example, the author shows that, with the methods of advanced pedagogical-psychological diagnosis, a high degree of fairness may be achieved in international comparative research on school achievement.

Anschrift des Autors

Prof. Dr. Karl-Heinz Arnold, Technische Universität Berlin,
Fachbereich Erziehungs- und Unterrichtswissenschaften,
Franklinstr. 28/29 (FR 4–5), 10587 Berlin