Baumert, Jürgen; Klieme, Eckhard; Lehrke, Manfred; Savelsbergh, Elwin

Konzeption und Aussagekraft der TIMSS-Leistungstests. Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik [Teil 1]

Die Deutsche Schule 92 (2000) 1, S. 102-115

Jürgen Baumert, Eckhard Klieme, Manfred Lehrke
und Elwin Savelsbergh

Konzeption und Aussagekraft der TIMSS-Leistungstests

Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik

Die *dritte Internationale Mathematik- und Naturwissenschaftsstudie* (TIMSS) hat seit Erscheinen der ersten, auf die Sekundarstufe I bezogenen internationalen und nationalen Berichte (Beaton et al. 1996; Baumert u.a. 1997) Diskussionen in der Öffentlichkeit wie auch in Fachkreisen ausgelöst. Bemerkenswerterweise wurde die Studie in den ostasiatischen Ländern, die im internationalen Vergleich teilweise überragende Ergebnisse erzielten, nur wenig diskutiert, aber auch in der Schweiz, dem europäischen Land mit den höchsten Testresultaten in der Mathematikuntersuchung, eher gelassen aufgenommen, während sie in den USA und Deutschland beträchtliche politische und pädagogische Auseinandersetzungen auslöste. Charakteristisch für die kritische Rezeption der Untersuchung ist die *Vermischung von politischen und fachlichen Argumenten*. Die wünschenswerte analytische Trennung der Diskurs-Ebenen wird selten durchgehalten. Dies macht den Umgang mit der Kritik nicht einfacher. Ein gutes Beispiel dafür ist der in Heft 2/99 dieser Zeitschrift publizierte Aufsatz von Hagemeister, in dem die politische Mission des Autors die fachliche Argumentation durchsetzt. Wenn wir im Folgenden auf diesen Aufsatz eingehen, wollen wir dennoch versuchen, die Argumentationsebenen zu trennen, und uns ausschließlich auf fachliche Gesichtspunkte konzentrieren.

Hagemeisters Kritik des naturwissenschaftlichen TIMSS-Tests für die Mittelstufe *fällt vernichtend aus*. Die Testkonstrukteure haben offenbar kaum einen Fehler ausgelassen, den man bei der Entwicklung eines Leistungstests begehen kann. Die Einwände, die Hagemeister anhand der Inspektion von acht Testaufgaben entfaltet, lassen sich in systematisierter Form folgendermaßen zusammenfassen:

- Der TIMSS-Test sei *curricular invalide*, da über fünfzig Prozent der Aufgaben nicht mit dem Berliner Physik-Rahmenplan konform seien. Die TIMS-Studie lasse auf Grund unzureichender Konstruktvalidität des Leistungstests keine Aussagen über den naturwissenschaftlichen Unterricht in Deutschland zu. Ein Teil der Testaufgaben reduziere die Ziele der Schule auf Faktenwissen, die meisten erfassten ausschließlich Leseverständnis oder allgemeine kognitive Grundfähigkeiten – „Logelei", wie Hagemeister sagt – und nicht naturwissenschaftliche Kompetenz.
- Der TIMSS-Test sei in mehrfacher Hinsicht *kulturell unfair*. (1) Im internationalen Vergleich begünstige er jene Teilnehmerländer, in denen Testpro-

gramme institutionalisiert seien, und zwar insbesondere dann, wenn diese Tests auf Mehrfachwahlantworten (Multiple Choice) beruhten. (2) Auf Grund der Sprachlastigkeit der Aufgaben würden Schüler aus sozial schwächeren Familien benachteiligt, da diese über geringere Sprachkompetenz verfügten. (3) Schließlich sei bei einzelnen Testaufgaben ein kulturelle*r bias* nachweisbar, der auf Übersetzungsmängel zurückgeführt werden könne (wenn zum Beispiel Mais nicht durch eine heimische Kornart ersetzt werde).

- Eine Vielzahl von Testaufgaben weise *fachliche Mängel* infolge der Simplifizierung von komplexen Sachverhalten und der falschen Darstellung von Experimenten auf. Dies weise auf mangelnde Zusammenarbeit mit Fachlehrern hin. Ähnliches wiederhole sich in der Ergebnisdarstellung, bei der gravierende Fehlinterpretationen nachzuweisen seien.
- Bei einer Reihe von Testaufgaben meint Hagemeister, *technische Mängel* entdeckt zu haben, auf deren Überprüfung er allerdings verzichtet. Wenn man die Einwände fachlich korrekt formuliert, gehören dazu: mangelnde Trennschärfen von Aufgaben, erhöhte Ratewahrscheinlichkeit oder positive Part-whole-Korrelationen für Distraktoren, differentielle Itemfunktionen zu Ungunsten von Schülern mit experimentell orientiertem Physikunterricht oder uneindeutige Lösungen bei Mehrfachwahlaufgaben.
- Schließlich vermutet Hagemeister, dass einzelne Testaufgaben Schüler zu fahrlässigem Experimentieren oder liederlichem Sprachgebrauch verführen könnten.

Um die Berechtigung der Kritik zu prüfen, werden wir im Folgenden zunächst die Grundlagen und Methoden der Konstruktion des naturwissenschaftlichen TIMSS-Tests für die Mittelstufe beschreiben, noch einmal die wichtigsten empirischen Belege zur Inhalts- und Konstruktvalidität vorlegen und anschließend (im Teil 2 im nachfolgenden Heft) vor diesem Hintergrund die Aufgabenkritik Hagemeisters einer sorgfältigen Analyse unterziehen. Hauptanliegen unseres Beitrages ist es, insbesondere die Struktur jener Einwände herauszuarbeiten, die häufiger anzutreffen sind und auf Missverständnissen der Grundlagen moderner Testkonstruktion beruhen.

1. Entwicklung des TIMSS-Leistungstests für die Mittelstufe

1.1. Testkonzeption

Die Leistungstests von TIMSS streben in der Mittelstufe und im voruniversitären Mathematik- und Physikunterricht *transnationale curriculare Validität* auf einem pragmatischen Niveau an. Die theoretische Grundkonzeption der Testentwicklung lehnt sich an Vorarbeiten der zweiten Internationalen Mathematik- und Naturwissenschaftsstudien der IEA (SIMS und SISS) an, entwickelt diese jedoch weiter. Heuristisches Werkzeug der Testentwicklung für SIMS und SISS war eine Ordnungsmatrix, bei der die Zeilen durch die zentralen Stoffgebiete des Faches oder Fachgebietes und die Spalten durch hierarchisch angeordnete Stufen kognitiver Operationen bestimmt wurden (Inhalt x kognitiver Anspruch). Die kognitiven Operationen wurden im Anschluss an die Taxonomien Bloom's (1956) und Wilson's (1971) definiert. Für TIMSS wurde – und dies ist eine entscheidende Änderung – die Vorstellung *hierarchisch geordneter kognitiver Operationen* zu Gunsten eines

kategorialen Rasters typischer Leistungserwartungen („Performance Expectations") aufgegeben. Die endgültige Matrix, die den Rahmen für Aufgabenanalysen bildete und als Grundlage der TIMSS-Berichterstattung dienen sollte, unterscheidet *vier Dimensionen der Leistungserwartungen*: *Einzelwissen* („understanding simple information"), *Zusammenhangswissen* („understanding complex information"), *Konzeptualisieren und Anwenden* („theorizing, analyzing, and solving problems") sowie *Experimentieren und Beherrschung von Verfahren* („science processes and investigating the natural world") (Robitaille u.a. 1993; IEA 1998). In bewusster Abgrenzung zur Inhalt x kognitiver Anspruch-Matrix der früheren IEA-Studien sind die *Hauptkategorien* („reporting categories") der Leistungserwartungen in TIMSS gerade nicht hierarchisch aufgebaut, sondern werden *theoretisch als orthogonale Testdimensionen verstanden*. „Understanding simple information" und „understanding complex information" gelten als solche Berichtskategorien für Leistungserwartungen (IEA 1998). Einzelwissen („simple information") und Zusammenhangswissen („complex information") haben theoretisch nichts mit dem Schwierigkeitsgrad eines Items zu tun. Eine Aufgabe, die Einzelwissen erfasst, kann sehr schwer sein – wie zum Beispiel Item R2, das wir weiter unten diskutieren – und eine Aufgabe, die Zusammenhangswissen prüft, wiederum sehr leicht – wie die unten vorgestellte Testaufgabe D2. Ferner können einzelne Testaufgaben mehreren Leistungserwartungen zugleich zugeordnet werden. Die TIMSS-Tests sind also ihrer Konzeption nach im Grunde mehrdimensional angelegt.

Hagemeister verwechselt aufgrund eines Übersetzungsfehlers Testdimensionen und Aufgabenschwierigkeit. Er moniert bei einer Reihe von eher schwierigen Aufgaben, dass die Testkonstrukteure tatsächlich angenommen hätten, dass es lediglich um das Verstehen simpler Informationen gehe (Hagemeister, S. 166). Wer „simple information" in Abgrenzung zu „complex information" mit „simpler Information" ins Deutsche übersetzt, hat nicht nur einem jener im Englischen häufig anzutreffenden „false friends" die Hand gereicht, sondern er zeigt, dass er die Grundstruktur des gesamten Klassifikationssystems der TIMSS-Items nicht verstanden hat.

Zwei weitere Gesichtspunkte haben bei der Beurteilung und Auswahl insbesondere der naturwissenschaftlichen Aufgaben eine nicht unerhebliche Rolle gespielt. Unter den an der Testentwicklung beteiligten Fachdidaktikern wurde weitgehend eine Auffassung vom Erwerb naturwissenschaftlicher Kompetenzen geteilt, in der die *naturwissenschaftlichen Alltagsvorstellungen von Kindern und Jugendlichen* zentrale Bedeutung haben. Dies sollte auch in den Testaufgaben Berücksichtigung finden, insofern ein Teil der Aufgaben bereits auf der Grundlage lebenspraktischer Erfahrung oder mit Hilfe qualitativer naturwissenschaftlicher Konzepte auf Alltagsniveau lösbar sein und ein anderer Teil der Aufgaben bekannte Schülervorstellungen systematisch als Falschlösungen (Distraktoren) anführen sollte. Ferner bestand unter den Naturwissenschaftsdidaktikern Einvernehmen, dass der TIMSS-Test für die Mittelstufe *primär ein qualitatives Verständnis von naturwissenschaftlichen Konzepten und Prozessen* erfassen solle und keinen Schwerpunkt auf der Mathematisierungsfähigkeit oder der Beherrschung von Rechenroutinen haben dürfe.

Nach der Felduntersuchung im Frühjahr 1994 wurden 135 naturwissenschaftliche Testaufgaben für die Hauptuntersuchung in der Mittelstufe ausgewählt. Ein Blick auf Tabelle 1 belegt, dass im endgültigen TIMSS-Test eine annähernde Gleichverteilung der Testaufgaben über drei Dimensionen der Leistungserwartungen erreicht werden konnte, die Dimension des Konzeptualisierens und Anwendens aber unterrepräsentiert blieb – am wenigsten allerdings in der Physik.

Tabelle 1: Naturwissenschaftliche Testaufgaben nach Sachgebiet und Leistungserwartung (einschließlich fünf experimenteller Aufgaben, „performance items"):

Sachgebiet	Leistungserwartung				Insgesamt
	Einzelwissen (*Understanding simple information*)	Zusammenhangswissen (*Understanding complex information*)	Konzeptualisieren und Anwenden (*Theorizing, analyzing and solving problems*)	Experimentieren, Beherrschung von Verfahren (*Science processes and investigating the natural world*)	
Biologie	19	13	1	8	41
Chemie	10	5	0	6	21
Physik	14	11	6	9	40
Earth Sciences	8	7	0	8	23
Environmental Issues	4	3	2	6	15
Insgesamt	55	39	9	37	140

IEA. Third International Mathematics and Science Study.

1.2. Überprüfung der Lehrplan- und Unterrichtsvalidität

Bei der Analyse der Validität von Messinstrumenten unterscheidet man Inhalts- und Konstruktvalidität. *Fragen der inhaltlichen Gültigkeit* – d.h. der Repräsentativität für Lernziele und Lerninhalte, wie sie im Curriculum verankert und im Unterricht realisiert werden – sind für Schulleistungstests offensichtlich zentral. Um die *Lehrplanvalidität* zu sichern bzw. zu prüfen, wurden in TIMSS unterschiedliche Wege begangen. Während der Phase der Testkonstruktion wurden die Aufgaben des Feldtests anhand der Lehrplanbank des Instituts für die Pädagogik der Naturwissenschaften (IPN) differenziert nach Fächern, Jahrgangsstufen und Ländern auf Lehrplankonformität überprüft. Gleichzeitig haben Fachdidaktiker des IPN die Aufgaben unter fachlichen Gesichtspunkten einer Kontrolle unterzogen. Anschließend haben Lehrplanexperten aus zwei großen Bundesländern die Testaufgaben sowohl auf landesspezifische Lehrplangültigkeit als auch auf Unterrichtsangemessenheit überprüft. Die Ergebnisse dieser beiden Validitätsprüfungen wurden bei der Auswahl der Aufgaben für die Hauptuntersuchung berücksichtigt (Garden/Orpwood 1996). Nach Abschluss der Hauptuntersuchung haben wir im Rahmen der internationalen *Test-Curriculum-Matching-Analyse* eine Expertenbefragung zur curricularen Validität der tatsächlich eingesetzten Testaufgaben für die 7. und 8. Jahrgangsstufe durchgeführt (Beaton et al. 1996; Beaton/Gonzalez 1997). Ziel dieser Befragung war es, jene Testaufgaben zu ermitteln, die für mindestens 50 Prozent der Schüler einer Jahrgangsstufe zum Lehrplan gehörten.

Aber auch die Lehrplangültigkeit der Testaufgaben garantiert noch keine *Unterrichtsvalidität*, da die bindende Wirkung der curricularen Vorgaben durchaus ungewiss ist. Wir haben deshalb die an TIMSS teilnehmenden Fachlehrkräfte anhand von Beispielaufgaben gebeten anzugeben, inwieweit die durch die Testaufgaben abgedeckten Stoffgebiete im Unterricht der untersuchten Klassen tatsächlich behandelt wurden.

Vor dem Hintergrund der bei Baumert u.a. 1997 zu findenden ausführlichen Darstellung dieser Maßnahmen überrascht die Behauptung Hagemeisters (S. 173), dass der TIMSS-Naturwissenschaftstest curricular nicht valide sein könne, da in Berlin nur weniger als 50 Prozent der Physikaufgaben rahmenplankonform seien. Zur Klärung seien die zentralen Befunde hier noch einmal knapp rekapituliert. Als internationales *Validitätskriterium der Test-Curriculum-Matching-Analyse* wurde festgelegt, dass eine Aufgabe dann als curricular valide gelten solle, wenn mindestens 50 Prozent der Schüler eines Landes bis zur 8. Jahrgangsstufe die Gelegenheit hatten, sich mit dem zugehörigen Stoff auseinanderzusetzen. Wir haben die kritische Schwelle in Deutschland auf 60 Prozent erhöht. Bei der Aufgabenbeurteilung zeigte sich allerdings, dass diese Validitätsgrenze praktisch unbedeutend war. Die Expertenübereinstimmung war insgesamt sehr hoch. Leichtere Abweichungen traten in Mecklenburg-Vorpommern und Brandenburg, größere in Berlin auf. Die Berliner Sondersituation war uns bekannt. Sie ist auch leicht erklärbar, da das Land Berlin bei einer Stundentafelkürzung im Jahr 1991 den Physikunterricht in der 7. Jahrgangsstufe ausgesetzt hatte. Für die Bewertung der Lehrplanvalidität der Aufgaben auf *nationaler* Ebene spielen diese regionalen Einschränkungen quantitativ jedoch überhaupt keine Rolle.

Tabelle 2 zeigt, dass der Naturwissenschaftstest nach dem Expertenurteil für die 8. Jahrgangsstufe als weitgehend lehrplanvalide gelten kann. Die Differenz zwischen der 7. und 8. Jahrgangsstufe ist beabsichtigt und notwendig, um Leistungszuwächse zwischen beiden Jahrgangsstufen erfassen zu können.

Tabelle 2: Lehrplanvalide Aufgaben (Lehrplanstoff für mindestens 60 Prozent der deutschen Schüler einer Jahrgangsstufe) nach Fachgebiet und Jahrgangsstufe; in Prozent der maximal erreichbaren Testwerte:

Fachgebiete	Jahrgangsstufe	
	7. Jahrgang	8. Jahrgang
Mathematik	80	95
Naturwissenschaften	60	88

IEA. Third International Mathematics and Science Study.

Ähnlich sehen die *Befunde zur Unterrichtsvalidität* aus, in die Lehrerangaben aus allen Bundesländern anteilsmäßig eingegangen sind. Tabelle 3 zeigt, dass die Stoffgebiete, aus denen die TIMSS-Testaufgaben entnommen worden sind, nach den Angaben der Fachlehrkräfte im Durchschnitt auch zu 77 bis 89 Prozent bis zur 8. Jahrgangsstufe unterrichtet wurden.

Tabelle 3: Behandlung der in den Fachleistungstests repräsentierten Stoffgebiete im Unterricht nach Fächern und Zeitraum (in Prozent der Stoffgebiete der einzelnen Unterrichtsfächer)

	Im Unterricht behandelte Stoffgebiete				
Fach	vor der 8. Jahrgangsstufe	vertieft in der 8. Jahrgangsstufe	neu in der 8. Jahrgangsstufe	noch nicht behandelt	Stoffgebiete insgesamt
Mathematik	34	27	29	11	100
Biologie	42	19	22	17	100
Physik	29	15	33	23	100

IEA. Third International Mathematics and Science Study.

1.3. Maßnahmen zur Sicherung der internationalen Testfairness

Bei kaum einem anderen internationalen eingesetzten Leistungstest ist so viel Mühe darauf verwandt worden, kulturübergreifende Testfairness herzustellen, wie dies bei TIMSS der Fall war. Dass die Entwicklung kulturell äquivalenter Testitems ein schwieriges und oft nicht perfekt zu lösendes Problem darstellt, ist bekannt (z.B. van de Vijver/Tanzer 1998; van de Vijver/Hambleton 1996). Gerade deshalb ist im Rahmen von TIMSS versucht worden, durch drei Maßnahmen ein Optimum zu erreichen:

(1.) Alle Aufgaben wurden durch die nationalen Projektgruppen auf einen möglichen semantischen kulturellen *bias* überprüft. Dabei wurde eine größere Anzahl von Aufgaben ausgesondert.

(2.) In Deutschland und Österreich wurden die Testaufgaben kooperativ ausschließlich von Fachlehrern übersetzt, die insbesondere darauf zu achten hatten, dass die im jeweiligen nationalen Unterricht akzeptierten fachlichen Sprachkonventionen beachtet wurden.

(3.) Es wurde für jede Testaufgabe geprüft, ob sie – bei Konstanthaltung der Gesamttestleistung – in allen Ländern eine vergleichbare Lösungswahrscheinlichkeit aufweist. Aufgaben, die eine nennenswerte Wechselwirkung zwischen Aufgabenschwierigkeit und Land *(item-by-country-interaction)* aufwiesen, wurden korrigiert oder ausgesondert[1].

Schließlich wurde post-hoc für jedes Land und Sachgebiet eine *Skala optimaler nationaler curricularer Validität* konstruiert, in die ausschließlich jene Testaufgaben eingingen, die durch Experten des jeweiligen Landes als lehrplanvalide beurteilt worden waren. Auf der Grundlage jeder dieser Skalen wurden die Ländervergleiche wiederholt. Die nachweislich hohe Stabilität der

1 Bei der TIMSS-Aufgabe N4, in der Mais-Pflanzen erwähnt werden, vermutet Hagemeister, daß deutsche Kinder im Vergleich zu Schülern aus den USA, die mit Maispflanzen vertrauter wären, benachteiligt würden. Wir haben diese Aufgabe nachträglich noch einmal anhand der Hauptstichprobe auf kulturellen *bias* geprüft. Es ist keine differentielle Itemfunktion nachweisbar; die Übersetzer haben gut daran getan, keine „grundlegende Transformation" der Aufgabe vorzunehmen, wie es Hagemeister für richtig hält.

Rangreihen ist ein guter Beleg für die erreichte kulturelle Fairness des Gesamttests innerhalb eines Fachgebietes (Beaton et al. 1996; Baumert, Lehmann u.a. 1997; Beaton/Gonzalez 1997; Arnold 1999). Dies schließt nicht aus, dass es auf der Ebene von Einzelitems durchaus auch beträchtliche Abweichungen geben kann (Schmidt u.a. 1997; 1998); sie sind jedoch auf der Ebene von Kompetenzschätzungen zu vernachlässigen (vgl. dazu unten Abschnitt 2.1).

Ein Wort sei noch zu dem in Deutschland immer wieder zu hörenden und auch von Hagemeister vorgetragenen Einwand gesagt, nach dem *Mehrfachwahlantworten* jene Länder bevorzugten, in denen Testprogramme mit MC-Aufgaben institutionalisiert seien. Da die TIMSS-Tests Aufgaben mit gebundenen und offenen Antwortformaten enthalten, lässt sich dieser Einwand prüfen. Am Beispiel des TIMSS-Grundbildungstests der Population 3 konnten Baumert/Klieme/Watermann (1998; 1999) zeigen, dass deutsche Schüler bei der Bearbeitung von Multiple-Choice-Aufgaben *keineswegs benachteiligt* sind – auch nicht im Vergleich zu Schülern aus den USA. Dieser Befund korrespondiert mit den vergleichbaren Ergebnissen, die Ramseier (1997, S. 28 ff.) für den TIMSS-Mittelstufentest berichtet. Eine multivariate varianzanalytische Prüfung, ob sich die differentiellen Itemschwierigkeiten von Multiple-Choice-Aufgaben zwischen den USA und Deutschland unterscheiden, bestätigt Ramseiers Ergebnisse. Von einer Benachteiligung deutscher Schülerinnen und Schüler kann keine Rede sein.[2]

2. Konstruktvalidierung des Physiktests

Im Unterschied zu einer Übungsarbeit in der Schulklasse, bei der jede Aufgabe der Lehrkraft Auskunft über die Beherrschung eines durchgenommenen Stoffelements gibt, haben Aufgaben in einem standardisierten Leistungstest *Indikatorfunktion für eine latente, im Hintergrund stehende Fähigkeit* oder Kompetenz, deren individuelle Ausprägung für die jeweilige Testleistung einer Person verantwortlich ist. Dies verlangt, dass sich die Testaufgaben auf einer einzigen Fähigkeitsdimension, oder wenn der Test mehrdimensional konzipiert ist, mehreren Dimensionen anordnen lassen. Die inhaltliche Qualität eines Tests hängt nicht zuletzt davon ab, inwieweit es gelingt, die zu erfassende latente Kompetenz als theoretisches Konstrukt vorab zu definieren oder zumindest post hoc zu rekonstruieren. Dies ist die erste Aufgabe der Konstruktvalidierung, die wir in Abschnitt 2.1 behandeln. Zur Konstruktvalidierung gehört ferner die em-

2 Auch Hagemeisters Einwand, daß in bestimmten Ländern (z.B. Japan) ein Testtraining zu besonders guten TIMSS-Ergebnissen geführt haben könnte, kann man kaum ernst nehmen. Ein *coaching* für spezifische der Struktur nach bekannte Tests, die regelmäßig wiederholt werden, hat testleistungssteigernde Effekte, die jedoch sehr schnell eine Obergrenze erreichen. In den USA gibt es eine breite Forschungsliteratur zu den begrenzten Auswirkungen von Test-Coaching. Im Manual zu einem weit verbreiteten Trainingsprogramm zur Vorbereitung auf den TOEFL-Test (Rymniak/Kurlandski/Smith, 1997) wird dementsprechend für den Fall eines erfolglosen kurzen Trainings auch empfohlen, zunächst noch einmal systematisch Englisch zu lernen. In Deutschland liegen Coaching-Untersuchungen mit ähnlichen Resultaten vor, die im Rahmen des Zulassungstests für medizinische Studiengänge durchgeführt wurden (Klieme / Maichle, 1990). Bei repräsentativen Untersuchungen, die auf Zufallsstichproben beruhen, unbekannte Tests benutzen und keinerlei Folgen für die Untersuchungsteilnehmer haben, ist ein Coaching zu vernachlässigen.

pirische Abgrenzung des erfaßten Merkmals von anderen, näher oder ferner stehenden Konstrukten (siehe Abschnitt 2.2).

2.1. Niveaustufen physikalischer Kompetenz

Das theoretische Rahmenkonzept der TIMSS-Testentwicklung unterscheidet, wie wir in Abschnitt 1 dargestellt haben, vier Klassen von Leistungserwartungen, die als potentiell unabhängige Facetten eines Kompetenzkonstrukts verstanden werden. Die Leistungserwartungen waren als Berichtskategorien konzipiert worden, die sich zu einem Kompetenzprofil verbinden lassen sollten. Analysen der internen Teststruktur zeigten jedoch, *dass die Dimensionen der Leistungserwartungen hoch interkorreliert waren*, so dass entgegen den theoretischen Ausgangsannahmen (s.o.) eine *unidimensionale Rasch-Skalierung vertretbar* erschien, die dann Grundlage der internationalen Berichterstattung wurde (Adams/Wu/Macaskill 1997).

Die beste Methode, um – jenseits der statistischen Prüfung der Modellanpassung des Gesamttests und der Modellverträglichkeit von einzelnen Items – festzustellen, ob die eingesetzten Aufgaben tatsächlich eine inhaltlich identifizierbare Dimension naturwissenschaftlicher Kompetenz erfassen, besteht in der *systematischen Beschreibung von Kompetenzstufen* („proficiency levels") anhand so genannter Markieritems. Hierzu haben Beaton und Allen (1992) ein Verfahren entwickelt, auf das wir zurückgreifen wollen. Im Folgenden werden wir uns auf die Entwicklung einer physikalischen Kompetenzskala beschränken, die den Ausgangspunkt für eine Analyse der von Hagemeister vorgetragenen Aufgabenkritik bilden wird.

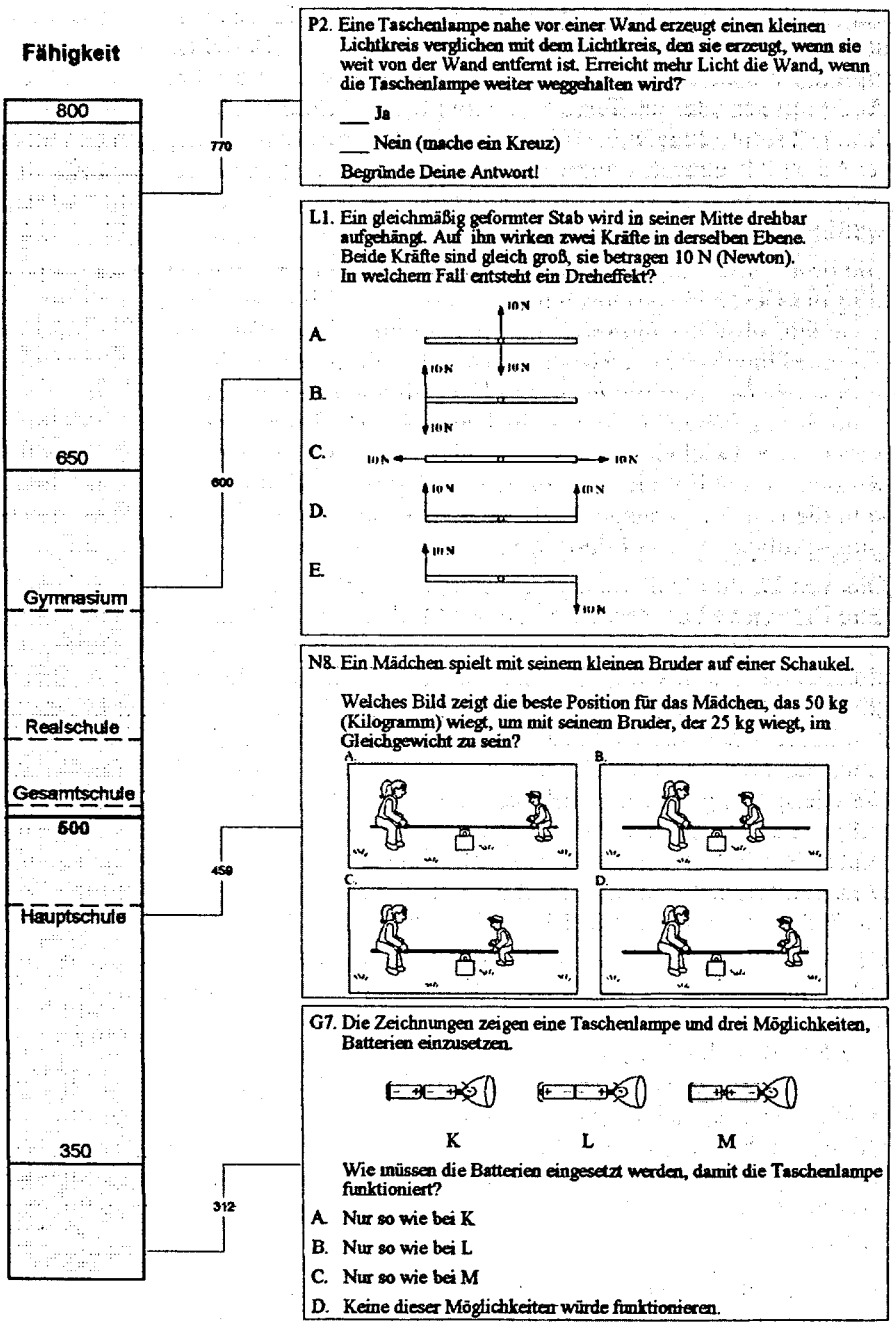
Das von Beaton und Allen vorgeschlagene Verfahren macht sich eine besondere Eigenschaft des testtheoretischen Modells einer Rasch-Skala zunutze, die es erlaubt, die Fähigkeitskennwerte der Bearbeiter und die Schwierigkeits-Kennwerte der Testaufgaben auf derselben Skala anzuordnen. Die Wahrscheinlichkeit, ein bestimmtes Item korrekt zu lösen, steigt mit der Fähigkeit des Bearbeiters an; das Rasch-Modell beschreibt diesen Zusammenhang mit einer bestimmten mathematischen Funktion (logistische Funktion). Den Schwierigkeitsgrad einer Aufgabe kennzeichnet man nun durch jenen Punkt auf der Fähigkeitsskala, bei dem sie mit einer Wahrscheinlichkeit von 65 Prozent richtig beantwortet wird. Je schwieriger eine Aufgabe ist, desto höher liegt dieser Referenzpunkt auf der Skala (zur Rasch-Skalierung vgl. Baumert u.a. 1997; Adams/Wu/Macaskill 1997, Knoche/Lind in Druck).

Die Skala wurde bei TIMSS so gewählt, dass der Wert 500 in der internationalen Testpopulation der Schüler des 7. und 8. Jahrgangs ein genau durchschnittliches Fähigkeitsniveau anzeigt, während die Werte 400 und 600 jeweils eine Standardabweichung unter bzw. über dem Mittelwert liegen. Um Kompetenzstufen naturwissenschaftlicher Bildung zu definieren und inhaltlich zu beschreiben, betrachten wir im Folgenden die Punkte 350, 500, 650 und 800 auf dieser Skala genauer. Wir wollen beschreiben, welche Leistungen ein Schüler oder eine Schülerin erbringen kann, deren Fähigkeit dem betreffenden Skalenwert entspricht[3]. Für jeden Referenzpunkt wählen wir nun diejenigen

3 Die Wahl dieser Referenzpunkte ist relativ beliebig. In unserer explorativen post-hoc Analyse wurden sie nach vorläufiger Inspektion der Items bestimmt. Da die

Aufgaben aus, die (a) mit hinreichender Sicherheit, das heißt mit einer Wahrscheinlichkeit von über 65 Prozent von Probanden mit den entsprechenden Fähigkeitswerten gelöst werden, und (b) auf dem nächstniedrigeren Kompe-

Abbildung 1. Physikalische Beispielaufgaben für die vier Niveaustufen des TIMSS-Naturwissenschaftstests



tenzniveau überwiegend falsch beantwortet werden, das heißt, eine Lösungswahrscheinlichkeit von unter 50 Prozent besitzen. Die Markieritems, die beispielsweise dem Skalenwert 500 (= Niveaustufe II) zugeordnet sind, beinhalten demnach jene Kompetenzen, durch die sich Schüler der Niveaustufe II von Bearbeitern auf Niveaustufe I (= Skalenwert 350) unterscheiden. Von den insgesamt 135 naturwissenschaftlichen Aufgaben des TIMSS-Tests für die Mittelstufe gehört knapp die Hälfte zu diesen für die vier Kompetenzstufen charakteristischen Markieritems; darunter befinden sich 20 Physikaufgaben. Ein Teil von ihnen ist in Abbildung 1 wiedergegeben und auf der TIMSS-Skala verankert.

Niveaustufe I: Lebenspraktisches Wissen

Die unterste Kompetenzstufe (Skalenwert 350) wird durch drei physikalische Aufgaben charakterisiert (G7, I16 und B6)[4]. Diese drei Aufgaben kann man ohne fachliches physikalisches Wissen ausschließlich auf der Basis lebenspraktischer Erfahrung beantworten. Wie Batterien in eine Taschenlampe eingelegt werden (vgl. Aufgabe G7 in Abbildung 1), dass Metall sich schneller erwärmt als Plastik (I16) und dass weiße Flächen mehr Licht reflektieren als rot, rosa oder schwarz angestrichene (B6) – diese Kenntnisse gehören im Sinne einer umfassenden naturwissenschaftlichen Grundbildung durchaus zur naturwissenschaftlichen Kompetenz, wie sie in TIMSS erfasst werden soll. Sie stehen allerdings nur für die unterste, lebenspraktische Stufe dieser Kompetenz.

Niveaustufe II: Anwendung alltagsbezogener naturwissenschaftlicher Konzepte

Die zweite Niveaustufe (Skalenwert 500) lässt sich durch neun physikalische Markieraufgaben beschreiben (A8, A10, C9, D2, I5, L7, M14, N8 und R1). Eine Inspektion dieser Aufgaben macht unmittelbar deutlich, was die zweite Kompetenzstufe von der ersten unterscheidet: Lebenspraktische Erfahrungen reichen allein nicht aus, um die Testaufgaben zu lösen, sondern man muss – wenn auch auf Alltagsniveau – qualitative physikalische Konzepte einbringen. Beispielsweise wird danach gefragt, welche Art von Sonnenstrahlung Sonnenbrand verursacht (J5). Auch hier wird ein alltagsnaher Kontext eingeführt; die angebotenen Lösungsalternativen (sichtbare, ultraviolette, infrarote Strah-

---

Item-Charakteristik-Kurven bei den naturwissenschaftlichen Aufgaben relativ flach verlaufen – die Naturwissenschaftsaufgaben sind also etwas weniger trennscharf als die mathematischen Textaufgaben – wählen wir hier Abstände von mehr als einer Standardabweichung, um Kompetenzstufen deutlich gegeneinander abgrenzen zu können. Bereits in den deskriptiven Befunden zur TIMSS-Mittelstufenstudie (Baumert u.a., 1997, S. 82–84) wurden Beispielaufgaben beschrieben und mit Hilfe der Schwierigkeitskennwerte auf der TIMSS-Fähigkeitsskala verankert. Dort wurden jedoch mit Abständen von jeweils 50 Punkten feinere Abstufungen vorgenommen. Der Nachteil einer solchen feineren Untergliederung ist, daß die Kompetenzstufen weniger trennscharf gegeneinander abgegrenzt werden können.

4 Die Aufgaben mit den Kennungen I bis Z sind inzwischen in deutscher Form publiziert, so daß auch über die hier abgedruckten Beispiele hinaus unsere Interpretationen nachgeprüft werden können (vgl. Baumert u.a., 1998).

lung, Röntgenstrahlung und Radiowellen) gehen jedoch über ein lebenspraktisches Verständnis von „Strahlung" hinaus.

Insgesamt vier für diese Stufe charakteristische Aufgaben beschäftigen sich mit den einfachen optischen Phänomenen der Spiegelung und Reflexion. So ist Aufgabe A10 dann lösbar, wenn man weiß, dass ein Objekt sichtbar wird, indem es Licht reflektiert oder streut. Bei zwei Aufgaben (C9 und M14) muss das Spiegelbild eines Gegenstandes in einer Zeichnung erkannt oder eingetragen werden, wobei ein Gitternetz und somit eine Art Koordinatendarstellung vorgegeben ist. Spiegelbilder stellen alltägliche Phänomene dar, aber ihre Darstellung mit Hilfe eines Koordinatensystems und einer perspektivischen Zeichnung erfordert mehr als nur lebenspraktische Erfahrung. Ähnliches gilt, wenn der Gang eines Lichtstrahls bei Reflexionen in einem Spiegel zu erkennen ist (R1) oder die richtige Position von zwei Kindern zur Ausbalancierung einer Wippe gesucht wird (siehe Aufgabe N8 in Abb. 1). Auch hier sind rudimentäre, noch alltagsgebundene Konzepte von Reflexion und Hebelwirkung erforderlich. Wenn schließlich herausgefunden werden soll, dass eine stark zusammengedrückte Feder mehr „gespeicherte Energie" enthält als eine gleiche, aber nur leicht zusammengedrückte Feder (A8), muss ein qualitatives Vorverständnis von Energie verwendet werden. Hervorzuheben ist, dass Schüler der 8. Jahrgangsstufe im Allgemeinen weder das Hebelgesetz noch den Energiebegriff in Anwendung auf Federn aus dem Physikunterricht kennen. Die Kompetenzstufe II indiziert daher für diese Gruppe kein physikalisches Fachwissen, sondern – wie beschrieben – ein Denken in alltagsbezogenen vorwissenschaftlichen Konzepten. Charakteristisch für diese Kompetenzstufe sind auch zwei Aufgaben (L7 und D2), die bei Hagemeister (1999) ausführlich diskutiert wurden und die wir im Abschnitt 3 wieder aufnehmen werden.

Niveaustufe III: Kenntnis fachlicher Inhalte auf Schulniveau

Auf der dritten Stufe naturwissenschaftlicher Kompetenz (Skalenwert 650) haben wir vier Markieraufgaben identifiziert (L1, Q12, E7 und D1). Hier ist nun erstmals explizit fachliches Wissen erforderlich, das die meisten Schüler nur im Unterricht gewinnen können. Die Fragen zur Optik beispielsweise lassen sich auf dieser Stufe nicht mehr allein mit vorwissenschaftlichen Konzepten über die Lichtreflexion am Spiegel oder gar mit lebenspraktischen Erfahrungen beantworten. Man muss hier schon die Lichtbrechung an einer Sammellinse darstellen (D1) und eine physikalische Begründung für die unterschiedliche Helligkeit von gebündeltem und gestreutem Licht angeben (Q12) können. Wissen muss man auch, dass Atomkerne aus Protonen und Neuronen bestehen (E7), oder dass Kräfte als gerichtete Pfeile dargestellt werden[5] (L1; vgl. Abb. 1).

Niveaustufe IV: Konzeptuelles Verständnis der Schulphysik

Die vierte Kompetenzstufe (Skalenwert 800) ist schließlich durch Aufgaben charakterisiert, die ein konzeptuelles Verständnis in einzelnen Gebieten der Schulphysik erfordern. In Aufgabe P2 etwa (vgl. Abb. 1) geht es – in physikalischen Fachbegriffen gesprochen – darum, dass die Lichtstärke I, das heißt die übertragene Energie, eine Eigenschaft der Lichtquelle ist, während die Beleuchtungs-

5 Man beachte, daß ein Verständnis des Vektorkonzepts für die Lösung der Aufgabe *nicht* erforderlich ist.

stärke E, die auf dem beleuchteten Objekt hervorgerufen wird, vom Abstand r zwischen Quelle und Objekt abhängt ($E = I/r^2$). Um die Aufgabe zu lösen, muss man allerdings weder diese Begriffe noch die genannte Größengleichung kennen. Als richtig wird eine Antwort gewertet, wenn sie die Aussage enthält, dass bei größerem Abstand gleich viel oder (aufgrund von Absorption) weniger Licht die Wand erreicht. Es kommt also auf ein qualitatives Verständnis der „Beleuchtung" als Übertragung von Energie an. Ähnlich bei Aufgabe Y2, wo nach der Temperatur im Inneren eines schmelzenden Schneeballs gefragt wird. Auch hier muss man keine Fachbegriffe oder Größengleichungen anwenden, sondern verstehen, dass der Schmelzpunkt von Wasser bei 0° Celsius liegt und dass Wärmeenergie von außen nach innen weitergeleitet wird.

Die dritte charakteristische Aufgabe (B3) erfasst das Verständnis eines weiteren grundlegenden Konzeptes der Physik. Die Schüler haben aus einer Tabelle, die vier Gegenstände mit unterschiedlichen Massen und Volumina anführt, denjenigen mit der höchsten Dichte herauszusuchen. Die Schwierigkeit kann nicht allein in der Berechnung der Verhältnisse liegen (im nächsten Abschnitt werden wir am Beispiel der Aufgabe M12 zeigen, dass dies der Mehrheit der Schüler der 8. Jahrgangsstufe gelingt). Das Problem liegt hier – wie aus psychologischen und didaktischen Untersuchungen bekannt ist (z.B. Bassok 1990) – im Verständnis der Dichte als einer „intensiven Größe", das heißt einer Verhältnisgröße.

Die Tatsache, dass derartige Verständnisaufgaben das höchste Kompetenzniveau charakterisieren, bestätigt jene in der Fachdidaktik bekannte und anhand der TIMSS-Oberstufentests belegte Erkenntnis (Klieme, im Druck), dass ein *qualitatives* Verständnis von Konzepten und die Überwindung von typischen Alltagsvorstellungen ein Merkmal hoher physikalischer Kompetenz von Schülern ist.

Der TIMSS-Untertest für Physik erfasst also eine Facette naturwissenschaftlicher Kompetenz, die zwischen den Polen lebenspraktischer Erfahrung und konzeptuellem Verständnis auf Schulniveau eingespannt ist. Dass dieser Test primär oder gar ausschließlich Faktenwissen erfasst – wie Hagemeister behauptet –, davon kann in der Tat überhaupt keine Rede sein. Im Gegenteil: Die besondere Schwierigkeit des TIMSS-Tests für deutsche Schülerinnen und Schüler ergibt sich gerade aus dem Umstand, dass die Abfrage von auswendig gelernten Begriffen und der Vollzug rechnerischer Routinen nicht im Mittelpunkt dieser Aufgaben stehen. Deshalb konnte auch Ramseier (1997; 1998; 1999) zeigen, dass die relativen Stärken der schweizer Schüler bei naturwissenschaftlichen Aufgaben deutlich werden, die ein tieferes Verständnis naturwissenschaftlicher Sachverhalte prüfen, ohne spezifische Fachterminologie abzurufen.

Wer den Anforderungsgehalt und die Aussagekraft, also die Validität von TIMSS-Aufgaben untersuchen will, muss sich stets klarmachen, *welche Kompetenzstufe mit welchen Aufgaben angezeigt wird.* Wenn ein Leistungstest dazu dient, im gesamten Fähigkeitsbereich der Zielpopulation zu differenzieren, müssen seine Aufgaben auch *die ganze Spannweite der Kompetenzverteilung abbilden.* Daraus ergibt sich trivialerweise, dass nicht jede Testaufgabe didaktischen Wunschvorstellungen entsprechen kann, die – wenn überhaupt – in der Regel nur von den leistungsstärksten Schülern eines Jahrgangs erfüllt werden. Dies gilt auch für den TIMSS-Physiktest der Mittelstufe: Aufgaben, die fachdidaktischen Zielvorstellungen für den Physikunterricht genügen, findet man am ehesten auf den Niveaustufen III und IV. Für Beispiele eines differenzierten Umgangs mit TIMSS-Aufgaben, der deren Stellung im Rahmen der Kom-

petenzskala berücksichtigt, vgl. die mathematikdidaktischen Arbeiten von Neubrand/Neubrand/Sibberns (1988), Blum/Wiegand (1998), Wiegand (1998) sowie für die Physikdidaktik Fischer (im Druck).[6]

2.2. *Physikleistungen, Leseverständnis und kognitive Grundfähigkeiten*

Die TIMSS-Testaufgaben repräsentieren einen Aspekt physikalischer Kompetenz, der sich von Alltagsvorstellungen bis zum qualitativen Verständnis naturwissenschaftlicher Konzepte auf Mittelstufenniveau erstreckt. Es wäre für den Physikunterricht kein großes Kompliment, wenn es sich – wie Hagemeister behauptet – zeigen ließe, dass diese Kompetenz mit Lesefähigkeit und schlussfolgerndem Denken zusammenfiele – der Unterricht also belanglos wäre. Die nachfolgende Korrelationstabelle (Tab. 4) zeigt, dass davon auch nicht ernsthaft gesprochen werden kann.

Tabelle 4: Zusammenhänge zwischen ausgewählten Testleistungen, Noten und kognitiven Grundfähigkeiten in TIMSS; Korrelationskoeffizienten:

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Testleistungen	Physik (1)	1.00						
	Mathematik (2)	.53	1.00					
Noten	Physik (3)	-.25	-.27	1.00				
	Mathematik (4)	-.15	-.26	.51	1.00			
	Deutsch (5)	-.12	-.18	.33	.35	1.00		
Kognitive Grundfähigkeiten	Figural (6)	.35	.49	-.20	-.20	-.15	1.00	
	Verbal (7)	.45	.59	-.22	-.14	-.21	.60	1.00

Die Korrelationen von r = .35 bzw. r = .45 zwischen Physikleistungen und kognitiven Grundfähigkeiten fallen im Vergleich zu den üblicherweise berichteten Zusammenhängen zwischen Schulleistungen und schlussfolgerndem Denken eher niedrig aus. Bei einer durch die Intelligenzleistungen erklärten Varianz von knapp 20 Prozent wird kein Sachverständiger behaupten wollen, dass der Physiktest primär oder gar ausschließlich verbale oder figurale Intelligenz erfasse. Erwartungsgemäß sind die ebenfalls in Tab. 4 ausgewiesenen Korrelationen zwischen Mathematik- und Intelligenzleistungen höher. Ferner weist die Tabelle 4 auch differentielle Zusammenhänge zwischen Testleistung und einzelnen Fachnoten aus, wobei die Deutschnote von der Testleistung in Physik weitgehend abgekoppelt ist. Damit ist auch Hagemeisters Argument hinfällig, dass bei Kontrolle der Deutschleistung Leistungsunterschiede zwischen den Schulformen im Physiktest nicht mehr nachweisbar seien und Haupt- und Gesamtschüler Gymnasialniveau erreichten.[7] Bemerkenswert ist der Befund,

6 Anspruchsvollere TIMSS-Aufgaben machen sich auch die von einer Arbeitsgruppe des nordrhein-westfälischen Kultusministeriums entwickelten Handreichungen zum mathematisch-naturwissenschaftlichen Unterricht zu Nutze (vgl. dazu auch Fischer, im Druck).

7 Eine von uns gerechnete Kovarianzanalyse mit dem Faktor Schulformzugehörigkeit und Deutschnote als Kovariate zeigt, daß sich auch bei Kontrolle der Deutschnote die Leistungsunterschiede zwischen den Schulformen praktisch nicht verändern.

dass der TIMSS-Mathematiktest einen besseren Prädiktor für die Physiknote als der Physiktest selbst darstellt: Im Physikunterricht werden offensichtlich – und dies ist eine in der Naturwissenschaftsdidaktik häufig vorgetragene Kritik – zu einem nicht unerheblichen Teil mathematische Leistungen bewertet, die der TIMSS-Physiktest gerade nicht erfasst und konzeptuell auch nicht erfassen soll. Um ein weiteres zu tun, haben wir in der Stichprobe der Längsschnittuntersuchung „Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU)" (Baumert/Köller 1998) den Zusammenhang zwischen Physikleistungen, die durch einen Test erfasst wurden, der durch gemeinsame Anker-Aufgaben mit dem TIMSS-Test verbunden ist, und dem durch schulformspezifische Tests gemessenen Leseverständnis überprüft. Auch diese Korrelation liegt – je nach Schulform – zwischen r = .35 und r = .43.

Dennoch bedarf Hagemeisters *Kritik an der vermeintlichen Sprachlastigkeit* der Naturwissenschaftsaufgaben von TIMSS eines gesonderten Kommentars, da darin eine didaktische Position sichtbar wird, die dem mathematisch-naturwissenschaftlichen Unterricht Schaden zufügt. Wenn man an den TIMSS-Testaufgaben – und das gilt sowohl für die Mathematik als auch die Naturwissenschaften – Kritik üben will, wird man an der *geringen kontextuellen Einbettung vieler Aufgaben* ansetzen können. Bei den TIMSS-Items handelt es sich überwiegend um „kleine", spracharme Schulaufgaben, die von komplexeren und realitätsnäheren Anwendungssituationen weitgehend abstrahieren. Auf diesen Mangel der TIMSS-Aufgaben ist bereits in der Phase der Testkonstruktion von den beteiligten Fachdidaktikern hingewiesen worden. Der Mangel konnte jedoch in der für die Testentwicklung verfügbaren Zeit nicht behoben werden. (Er stellte die Ausgangsherausforderung für die internationale Expertengruppe dar, die für die *Testkonstruktion im PISA-Programm* verantwortlich ist (OECD 1999).)

Eine stärkere Kontextualisierung von Testaufgaben heißt jedoch immer auch stärkere Sprachgebundenheit. Wenn Hagemeister nun bei den weitgehend dekontextualisierten und spracharmen TIMSS-Aufgaben Sprachlastigkeit bemängelt, kommt darin *eine falsch verstandene Fürsorge für Schüler aus sozial schwächeren Familien* zum Ausdruck, die vermeintlich über geringere Sprachkompetenz verfügten und deshalb mit sprachlichen Anforderungen verschont werden sollten. Die Zurückhaltung des naturwissenschaftlichen Unterrichts gegenüber der Verschriftlichung komplexer Gedankengänge ist, wie Nieswandt (1998) überzeugend gezeigt hat, gerade einer seiner Schwachpunkte.

Teil 2 mit Analysen der Kritik an den Test-Items folgt in Heft 2/00. Auch die Literaturliste wird dort publiziert. Sie kann in der home-page des MPI für Bildungsforschung (www.mpib-berlin.mpg.de) eingesehen werden.

Jürgen Baumert, geb. 1941, Dr. phil., Professor für Erziehungswissenschaften
Eckhard Klieme, geb. 1954, Dr. phil., Dipl.-Mathematiker und Dipl.-Psychologe
Manfred Lehrke, geb. 1942, Dr. phil., Dipl.-Psychologe
Elwin Savelsbergh, geb. 1968, Dr. phil., Dipl.-Physiker und Dipl.-Psychologe
Anschrift: Max-Planck-Institut für Bildungsforschung, Lentzeallee 94, 14195 Berlin
e-mail: baumert@mpib-berlin.mpg.de