

Jürgen Baumert, Eckhard Klieme, Manfred Lehrke & Elwin Savelsbergh:

Konzeption und Aussagekraft der TIMSS-Leistungstests

Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik

1. TIMSS in der Diskussion

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 Diskursebenen 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 Testprogramme 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 kultureller *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-Kor-

relationen 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 vor diesem Hintergrund die Aufgabenkritik Hagemeyers 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.

2. Entwicklung des TIMSS-Leistungstests für die Mittelstufe

2.1 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 (*Per-*

formance 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). Die Kategorien der Leistungserwartungen stellen theoretisch unabhängig Fähigkeitsdimensionen dar, die "quer" zu stofflichen Anforderungen der Curricula definiert sind. Unter dem Gesichtspunkt internationaler Testfairness korrigiert diese kompetenzbezogene Perspektive die primäre Orientierung an curricularer Validität (vgl. Arnold, 1999). Aufgaben mit ein und derselben Leistungserwartung können empirisch unterschiedlich schwierig sein. Ferner können einzelne Testaufgaben mehreren Leistungserwartungen zugleich zugeordnet werden. Die TIMSS-Tests sind also ihrer Konzeption nach im Grunde mehrdimensional angelegt¹.

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. 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*

¹ In der ursprünglichen Fassung des theoretischen Rahmens waren die Kategorien Einzelwissen (*simple information*) und Zusammenhangswissen (*complex information*) als Untergliederung einer einzigen Dimension aufgefaßt worden. Im Laufe der Arbeit mit dem Raster wurden beide Aspekte als selbständige unabhängige Berichtskategorien ausdifferenziert (vgl. Robitaille u.a., 1993; IEA, 1998).

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.

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.

Die Entwicklung der Leistungstests für Mathematik und die naturwissenschaftlichen Fächer war ein kooperatives Unternehmen, in das von Anfang an Fachlehrer, Fachdidaktiker und Fachwissenschaftler einerseits sowie Pädagogen und Psychologen andererseits, die ihren Arbeitsschwerpunkt in international vergleichender Unterrichtsforschung oder Testkonstruktion im Rahmen von *large scale assessment* hatten, eingebunden waren (Garden / Orpwood, 1996). 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 Anforderungsart*

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

*Einschließlich fünf experimenteller Aufgaben (*performance items*).

IEA. Third International Mathematics and Science Study.

2.2 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 Hagemesters (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* nach Fachgebiet und Jahrgangsstufe (in Prozent der maximal erreichbaren Testwerte)

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

* Lehrplanstoff für mindestens 60 Prozent der deutschen Schüler einer Jahrgangsstufe. IEA. Third International Mathematics and Science Study.

Ähnlich sehen die Befunde zur Unterrichtsvalidität aus, in die Lehrerangaben aus allen Bundesländern anteilmäßig 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)

Fach	Im Unterricht behandelte Stoffgebiete				Stoffgebiete insgesamt
	vor der 8. Jahrgangsstufe	vertieft in der 8. Jahrgangsstufe	neu in der 8. Jahrgangsstufe	noch nicht behandelt	
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.

2.3 Maßnahmen zur Sicherung der internationalen Testfairneß

Bei kaum einem anderen internationalen eingesetzten Leistungstest ist soviel 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 Konstanzhaltung 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 ausgesondert².

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.

² Bei der TIMSS-Aufgabe N4, in der Mais-Pflanzen erwähnt werden, vermutet Hagemester, 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 Hagemester für richtig hält.

Die nachweislich hohe Stabilität der 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 3.1).

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 haben Baumert, Klieme und Watermann (1998; 1999) diese Prüfung vorgenommen. Sie konnten 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.³

³ 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.

3. 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 einer einzigen Fähigkeitsdimension, oder wenn der Test mehrdimensional konzipiert ist, mehreren Dimensionen zuordnen 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 3.1 behandeln. Zur Konstruktvalidierung gehört ferner die empirische Abgrenzung des erfaßten Merkmals von anderen, näher oder ferner stehenden Konstrukten (siehe Abschnitt 3.2). Mit diesen Untersuchungen gehen wir auch auf eingangs zitierte Kritik Hagemeisters an der Konstruktvalidität des TIMSS-Tests ein.

3.1 Niveaustufen physikalischer Kompetenz

Das theoretische Rahmenkonzept der TIMSS-Testentwicklung unterscheidet, wie wir in Abschnitt 2 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 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 Facette naturwissenschaftlicher Kompetenz erfassen, besteht in der systematischen Beschreibung von Kompetenzstufen (*proficiency levels*) anhand sogenannter 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, im 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 entspricht⁴. Für jeden Referenzpunkt wählen wir nun diejeni-

⁴ Die Wahl dieser Referenzpunkte ist relativ beliebig. In unserer explorativen post-hoc Analyse wurden sie nach vorläufiger Inspektion der Items bestimmt. Da die 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)

gen 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 Kompetenzniveau 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 2 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.

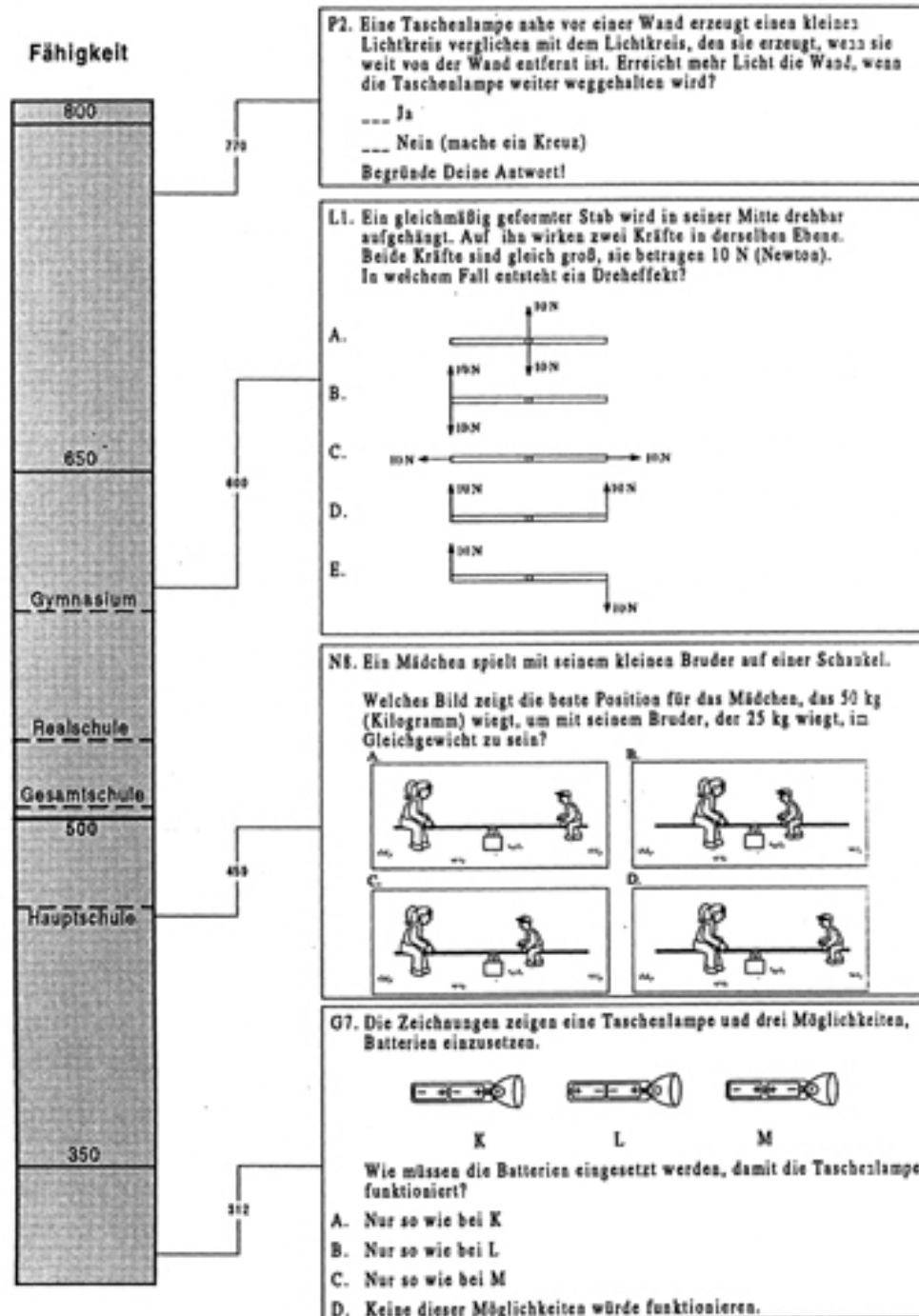
Niveaustufe I: Lebenspraktisches Wissen

Die unterste Kompetenzstufe (Skalenwert 350) wird durch drei physikalische Aufgaben charakterisiert (G7, I16 und B6)⁵. 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 2), 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.

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.

⁵ 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).

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



Niveaustufe II: Anwendung alltagsbezogener naturwissenschaftlicher Konzepte

Die zweite Niveaustufe (Skalenwert 500) lässt sich durch neun physikalische Markieraufgaben beschreiben (D2, C9, M14, N8, A10 und L7). 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 Strahlung, Röntgenstrahlung und Radiowellen) gehen jedoch über ein lebenspraktisches Verständnis von „Strahlung“ hinaus.

Insgesamt vier für diese Stufen 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 Hagemester (1999) ausführlich diskutiert wurden und die wir im Abschnitt 4 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 Neutronen bestehen (E7), oder daß Kräfte als gerichtete Pfeile dargestellt werden⁶ (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 Beleuchtungsstärke 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.

Zusammenfassend lassen sich die vier Kompetenzstufen, die im TIMSS-Test für die Sekundarstufe I identifiziert werden können, charakterisieren:

- Stufe I: Lebenspraktisches Wissen
- Stufe II: Anwendung alltagsbezogener naturwissenschaftlicher Konzepte
- Stufe III: Kenntnis fachlicher Inhalte auf Schulniveau
- Stufe IV: Konzeptuelles Verständnis von Schulphysik.

Der TIMSS-Untertest für Physik erfaßt also eine Facette naturwissenschaftlicher Kompetenz, die zwischen den Polen lebenspraktischer Erfahrung und Fachwissen auf Schulniveau eingespannt ist. Daß diese Tests primär oder gar ausschließlich Faktenwissen erfassen - wie Hagemeister behauptet -, davon kann in der Tat überhaupt keine Rede sein.

⁶ Man beachte, daß ein Verständnis des Vektorkonzepts für die Lösung der Aufgabe *nicht* erforderlich ist.

Im Gegenteil: Die besondere Schwierigkeit der 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 Tests 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 die Aufgabe hat, 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 Kompetenzskala berücksichtigt, vgl. die mathematikdidaktischen Arbeiten von Neubrand / Neubrand / Sibberns (1988), Blum / Wiegand (1998), Wiegand (1998) sowie für die Physikdidaktik Fischer (im Druck).⁷

3.2 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

⁷ 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).

Kompliment, wenn es sich - wie Hagemeister behauptet - zeigen ließe, daß diese Kompetenz mit Lesefähigkeit und schlußfolgerndem Denken zusammenfielen - 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 (Korrelationskoeffizienten) in TIMSS

		(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 Hagemesters Argument hinfällig, daß bei Kontrolle der Deutschleistung Leistungsunterschiede zwischen den Schulformen im Physiktest nicht mehr nachweisbar seien und Haupt- und Gesamtschüler Gymnasialniveau erreichten.⁸ Bemerkenswert ist der Befund, daß der TIMSS-Mathematiktest einen besseren Prädiktor für die Physiknote als der Physiktest selbst darstellt: Im Physik-

⁸ Eine von uns gerechnete Kovarianzanalyse mit dem Faktor Schulformzugehörigkeit und Deutschnote als Kovariate zeigt, daß sich auch bei Kontrollen der Deutschnote die Leistungsunterschiede zwischen den Schulformen praktisch nicht verändern.

unterricht 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 bei $r = .35$ und $r = .43$.

Dennoch bedarf Hagemesters 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 Textkonstruktion 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 Hagemester 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.

4. Analyse der Kritik an den TIMSS-Testaufgaben

4.1 Situationsmodelle als Grundlage der Lösung von Testaufgaben

Will man einen Schulleistungstest beurteilen, ist es, wie in Abschnitt 3.1 ausgeführt, notwendig, die Gesamtheit der Testaufgaben als Indikatoren einer latenten Fähigkeitsverteilung zu berücksichtigen. Es ist absolut unzulässig und geradezu irreführend, Einzelitems ohne Berücksichtigung ihres Schwierigkeitsniveaus und ihrer Funktion im Gesamttest herauszugreifen, um sie vor der Folie willkürlich festgelegter fachlicher Eindringtiefe oder normativer Unterrichtsvorstellungen zu diskutieren - wie dies Hagemeyer bei der Besprechung seiner acht Beispielitems tut. Die Analyse von Einzelaufgaben ist allerdings in hohem Maße sinnvoll und überaus wünschenswert, wenn sie durch die Explikation der zur Lösung von Testaufgaben notwendigen Operationen zur theoretischen Klärung des latenten Kompetenzkonstrukts beiträgt. Glücklicherweise gibt es in der Mathematikdidaktik mittlerweile eine Reihe von Beispielen für diese Form des Umgangs mit Testaufgaben (Neubrand / Neubrand / Sibbers, 1998; Blum / Wiegand, 1998; Wiegand, 1998).

Um ein Problem oder eine Aufgabe zu lösen, muß der Bearbeiter ein mentales Situationsmodell der Aufgabenstellung entwerfen, das den Rahmen des Lösungsprozesses definiert (vgl. Reusser, 1996). Mit der Entwicklung des subjektiven Situationsmodells wird entschieden, um was es überhaupt geht, welches Wissen aktiviert, welcher Lösungsweg gewählt und welche Denkoperationen durchgeführt werden. Das Situationsmodell legt auch fest, auf welchem fachlichen Anspruchsniveau eine Aufgabe behandelt wird. Die Grundmerkmale möglicher oder besser: wahrscheinlicher Situationsmodelle werden sowohl durch die Formulierung und Darbietung der Testaufgabe als auch durch die soziale Situation der Testadministration vorgezeichnet. Eine gute Testaufgabe enthält in sparsamer Form die notwendigen Hinweisinformationen, die bei Probanden, die das durch die Aufgabe indizierte Fähigkeitsniveau erreichen oder übertreffen, zur

Bildung eines für die Lösung der Aufgabe adäquaten Situationsmodells führen. Die erforderlichen Hinweisinformationen werden zunächst durch die Aufgabenstellung selbst, dann aber auch durch zusätzlich mitgeteilte Lösungshinweise gegeben. Zusätzliche Lösungshilfen sind zum Beispiel bei Mehrfachwahlaufgaben die in den Distraktoren immer implizit enthaltenen Ausschlußinformationen; bei offenen Aufgabenstellungen können dies präzisierende Hinweise sein, die einen Lösungsansatz oder ein bestimmtes Verfahren nahelegen. Die Schwierigkeit von *multiple choice*-Aufgaben wird nicht zuletzt dadurch bestimmt, inwieweit die Distraktoren die Entwicklung attraktiver, aber nicht adäquater Situationsmodelle nahelegen.

Von nicht zu unterschätzender Bedeutung für die Konstruktion des mentalen Aufgabenmodells ist die soziale Situation der Aufgabenbearbeitung - in der Regel ein schulischer oder schulähnlicher Kontext. In einer solchen sozialen Situation sind generalisierte Vorstellungen über typische Aufgaben eines Schulfachs der Rahmen, der die Wahl möglicher Situationsmodelle von vornherein einschränkt. Dies bedeutet, daß bei erwartungskonformen schultypischen Aufgaben häufig sehr wenige Hinweisinformationen in der Aufgabenstellung genügen, um die Grundzüge des gewünschten Situationsmodells entstehen zu lassen. Umgekehrt heißt dies aber auch, daß bei untypischen Aufgabenstellungen häufig inadäquate Situationsmodelle entwickelt werden. Zwei Beispiele: Ein Ausdruck wie $3x = y - 7 \mid x = -2$ wird von Schülern der 8. Jahrgangsstufe ohne weitere Angaben als Mathematikaufgabe erkannt, bei der x eingesetzt und nach y aufgelöst werden soll. Aber auch eine Aufgabe wie "Johns beste 100-Meter-Zeit ist 17 Sekunden. Wie lange braucht er für 1000 Meter?" (Greer, 1993) führt dann problemlos zu einer Zeitangabe, wenn diese schuluntypische Aufgabe im Rahmen eines Standardsituationsmodells bearbeitet wird, nach dem Schulaufgaben immer eindeutig lösbar sind (Verschaffel / De Corte / Lasure, 1994; Reusser / Stebler, 1997). Auch Testaufgaben sind kontextabhängig. Bei der Analyse der Lösung von Testitems müssen immer auch die generalisierten Aufgabenerwartungen der Probanden mit berücksichtigt werden.

Um die bei der Lösung von TIMSS-Aufgaben ablaufenden Denkprozesse zu rekonstruieren, hat Hagemeister drei Mädchen und vier Jungen aus der 8. und 9. Jahrgangsstufe eines Gymnasiums zur Bearbeitung einer Auswahl von Testaufgaben mit anschließen-

den Interviews zu sich gebeten. Bei allen Probanden handelt es sich um Schüler mit sehr guten Mathematiknoten. Hagemeister hat seinen Partnern eine Auswahl von Aufgaben vorgegeben, jeweils eine Aufgabe bearbeiten lassen und sich anschließend, wie er sagt "über die Aufgabe unterhalten" (Hagemeister, S. 161). Die entscheidende methodische Klippe bei der Rekonstruktion von Denkprozessen oder mentalen Situationsmodellen durch nachträgliche Befragung ist die Konfundierung zwischen der Rekonstruktion eines abgelaufenen Prozesses und der Neukonstruktion von Situationsmodellen in der Interviewsituation. Um dies zu vermeiden, ist seitens des Interviewers größte Abstinenz gegenüber dem Gesprächspartner und ein vorsichtiges, standardisiertes Vorgehen erforderlich. Gleichzeitig müssen Stimuli und Schülerantworten sorgfältig aufgezeichnet werden. Diese Problematik ist in der qualitativen fachdidaktischen Forschung hinreichend präsent. Von all dem ist bei Hagemeister nichts zu sehen. Er unterhält sich munter mit seinen Probanden mit dem Ergebnis, daß mentale Modelle nicht rekonstruiert, sondern neu erzeugt werden: "Na ja, wenn man so fragt, dann kann eigentlich nur C richtig sein" (Hagemeister, S. 164). Dies gelingt umso besser, je intelligenter die Gesprächspartner sind.

Hagemeister legt zu Recht großen Wert darauf, daß bei der Testkonstruktion die Erfassung von Schülervorstellungen eine wichtige Rolle spielen und die Testkonstrukteure sich über die bei der Lösung von Testaufgaben tatsächlich aktivierten Schülervorstellungen auch empirisch vergewissern sollten. Wer allerdings mit der Literatur zu alternativen Schülervorstellungen in den Naturwissenschaften vertraut ist, sieht bei der Durchsicht der TIMSS-Aufgaben auf Anheb, daß dieses Wissen systematisch in die TIMSS-Tests eingegangen ist (vgl. Klieme, im Druck, zu den Aufgaben des TIMSS-Oberstufentests). Wir werden im folgenden anhand der Aufgaben, die Hagemeister der Kritik unterzogen hat, zeigen, wie Aufgabenmerkmale und die schultypische Situation der Testadministration die Konstruktion mentaler Situationsmodelle vorzeichnen und in welcher Weise diese Situationsmodelle die Funktion von Items im Gesamttest bestimmen.

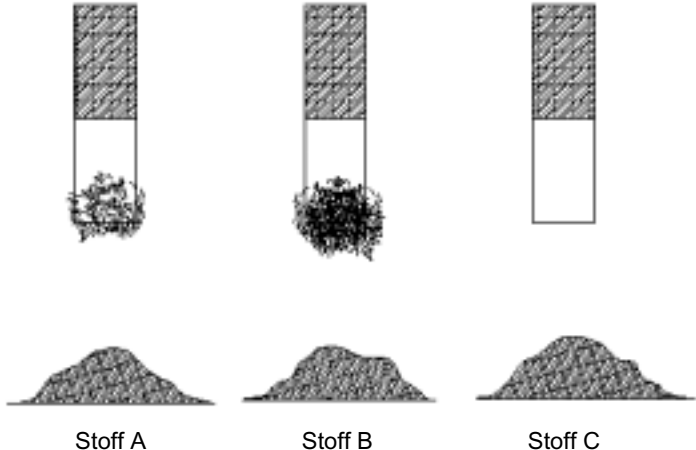
4.2 Überprüfung der Aufgabenkritik

Erstes Beispiel: Testaufgabe D2

Wir wollen mit einem einfachen Beispiel beginnen, an dem sich die Argumentations- und Arbeitsweise Hagemesters gut zeigen läßt:

Abbildung 2: TIMSS-Aufgabe D2

D2. Jeder der drei abgebildeten Magneten ist in den Stoff unter ihm eingetaucht worden. Welcher Stoff könnte Kaffee sein?



A. Nur A
B. Nur B
C. Nur C
D. Nur A und B

Hagemester kritisiert diese Aufgabe aus fachlichen, didaktischen und testtheoretischen Gründen. Seiner Ansicht nach belegt diese Testaufgabe, daß bei der Entwicklung der TIMSS-Tests keine physikalischen Experimente durchgeführt worden seien. Man hätte beim Experimentieren nicht nur bemerkt, daß ein Häufchen Eisenspäne (wie im Bild bei D2 [in den Fällen a und b] dargestellt) an den heute üblichen Dauermagneten vollständig hängenbleibe. Man hätte außerdem zum Beispiel bemerkt, daß die Magnete in dem Stoff, in den sie "eingetaucht worden" sind, Abdrücke hinterlassen hätten. Ferner sei die schematische Darstellung der Eisenspäne mit einer didaktischen Konzeption des Physikunterrichts unverträglich, in dem durch genaues Beobachten gleichzeitig in wissenschaftliche Arbeitsmethoden eingeführt und Achtung vor der Schönheit der Natur vermittelt werde. Hagemesters testtheoretische Einwände besagen, daß durch die seiner Ansicht nach fachlich unrichtige Gestaltung der Distraktoren Schüler mit besonders

guten Physikkenntnissen, Schüler, deren Physikunterricht experimentell ausgerichtet ist, und Mädchen, die gewohnt seien, besonders sorgfältig zu arbeiten, benachteiligt würden (dreifacher Item-*bias*).

Bei der Testaufgabe D2 handelt es sich um eine Aufgabe aus dem internationale TIMSS-Test für die Grundschule, die im Test für die Altersgruppe der 13- und 14jährigen als Anker-Item wiederkehrt. Die Aufgabe wurde bereits in der Zweiten Internationalen Naturwissenschaftsstudie der IEA (SISS) verwendet, so daß schon vor der TIMSS-Testkonstruktion Informationen über die Itemeigenschaften vorlagen. Die Aufgabe wurde ferner in einer international vergleichenden Grundschuluntersuchung zum technischen Problemlösen eingesetzt (Baumert / Evans / Geiser, 1998; Baumert, 1996). Im Rahmen dieser Untersuchung wurde auch für diese Aufgabe in einer qualitativen Vorstudie mit der Methode des *stimulated recall* überprüft, ob die Aufgabe tatsächlich zur Konstruktion des intendierten subjektiven Situationsmodells auf Schülerseite führt.

Im Rahmen der TIMSS-Untersuchung zur Mittelstufenpopulation gehört das Magnet-Item D2 mit einem Schwierigkeits-Kennwert von 434 zu den einfachen Aufgaben, die im untersten Leistungsbereich differenzieren sollen. Die Aufgabe hat eine gute Trennschärfe ($r_{\text{bis}} = .36$) und differenziert dementsprechend zwischen der untersten Kompetenzstufe, die wir als lebenspraktisches Wissen bezeichnet haben, und der zweiten Stufe, die ein Denken in alltagsbezogenen vorwissenschaftlichen Konzepten indiziert. Die relative Lösungswahrscheinlichkeit beträgt für den 8. Jahrgang international 76%, in Deutschland 88%. Die Lösung der Aufgabe fällt deutschen Schülern also im Vergleich zum internationalen Durchschnitt etwas leichter.

Welches mentale Situationsmodell muß vom Schüler konstruiert werden, um die Aufgabe richtig zu lösen? Der subjektive Problemraum wird (1) durch die modellhafte Skizze der Versuchsanordnung, die (2) im Itemstamm verbal erläutert wird, und (3) die gezielte Frage des Aufgabenstammes: "Welcher Stoff könnte Kaffee sein?" (wohlge-merkt: *nicht* "enthalten") vorgezeichnet. Die Aufgabe verlangt die Aktivierung einer einzigen Wissensseinheit, daß Magnete Nichtmetalle nicht anziehen. Dieser Fall wird in

der Lösungsalternative C durch den blanken Magneten symbolisiert - obwohl beim Eintauchen eines Magneten in Kaffeepulver Kaffeereste auch am Magneten hängenbleiben können. Dies ist jedoch – ebenso wie die Abdrücke, die der Magnet nach Hagemeister im jeweiligen Stoff hinterlassen haben müßte - für das zur Lösung erforderliche Situationsmodell unerheblich. Ein angemessenes Situationsmodell ist gerade kein photographisches Abbild einer Versuchsanordnung, sondern ein konzeptueller Rahmen, in dem Situation, Aufgabenstellung und das zur Lösung der Aufgabe notwendige Wissen zusammengebunden werden.

Um welche Substanzen es sich im Fall A und B handelt, läßt die Aufgabe offen, da die Beantwortung dieser Frage für die Entwicklung des adäquaten Situationsmodells ohne Belang ist. Im Vergleich zum Fall C wird jedoch klar, daß es sich nicht (nur) um Kaffee handeln kann. Zur Lösung der Aufgabe wird also nur rudimentäres Wissen über Magnetismus verlangt; eine genauere Kenntnis ferromagnetischer Materialien ist nicht erforderlich. Die guten Meßeigenschaften dieses Items sind nicht zuletzt auf die kluge Wahl der Distraktoren zurückzuführen, die gerade *nicht* den eindeutigen Fall reiner Eisenspäne abbilden. Wem jede Vorstellung von Magnetismus fehlt, kann die Lösungsalternativen A und B ankreuzen, weil er dort Kaffeereste sieht.

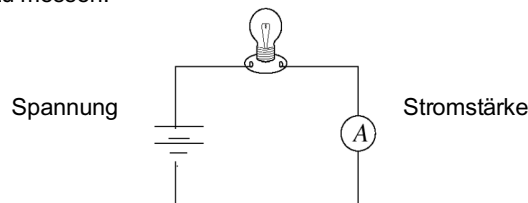
Wenn Hagemeister schließlich von Modellskizzen in Testaufgaben die Wiedergabe ästhetisch faszinierender Naturerscheinungen oder gar eine Erziehung zur Achtung vor der Natur erwartet, werden seine Einwände abwegig. Man kann diese Argumentation überhaupt nur verstehen, wenn nicht zwischen den Funktionen, die Aufgaben in einem didaktisch anspruchsvollen Unterricht haben, und der spezifischen Indikatorfunktion von Testaufgaben im Rahmen einer Leistungsdiagnostik unterschieden wird. Aufgaben im Unterricht haben komplexe, für Lernprozesse strukturbildende Funktionen, die sich aus dem didaktischen Modell der Unterrichtsstunde ergeben. Dabei können ästhetische und moralische Aspekte auch im naturwissenschaftlichen Unterricht eine wichtige Rolle spielen. Aufgaben eines Leistungstests haben jedoch spezifische Indikatorfunktionen im Rahmen der spezifizierten Dimensionen des Tests und sie müssen insgesamt in der Lage sein, die Leistungsverteilung der Untersuchungsgruppe abzubilden, auch wenn das Ergebnis unter normativ-didaktischen Gesichtspunkten enttäuschend ist.

Es bleiben zwei Einwände, die der empirischen Prüfung bedürfen. Hagemester vermutet, daß elaborierte Physikkenntnisse zur Konstruktion eines komplexeren, aber für die Aufgabe unangemessenen mentalen Situationsmodells verleiten könnten, gerade weil offen bleibt, um welche Substanzen es sich bei den Alternativen A und B handele. Ähnliches nimmt er für Personen an, in deren Unterricht regelmäßig experimentiert wird und die möglicherweise mit der Trennung von eisenpulverhaltigen Gemischen durch Magnete vertraut sind. Wenn die beide Einwände stimmen, muß sich dies in einer unbefriedigenden Trennschärfe des Items und positiven biserialen Korrelationen zwischen Distraktoren und Gesamtestwert niederschlagen. Ferner muß sich für Schüler, die einen experimentell ausgerichteten Unterricht genießen, unter Konstanthaltung der sonstigen Testleistung eine höhere Itemschwierigkeit nachweisen lassen (*Differential Item Functioning*) (Camilli / Shepard, 1994; Baumert / Klieme / Watermann, 1998). Von beidem kann jedoch keine Rede sein. Die Aufgabe differenziert hervorragend zwischen Schülern auf den unteren Kompetenzniveaus. Ebenso haben die Distraktoren mit Korrelationen zwischen $r_{bis} = -.10$ und $r_{bis} = -.25$ befriedigende Qualität. Selbst wenn im Einzelfall leistungsstarke Schüler irritiert sein sollten, ist dies für die Brauchbarkeit des Items ohne Belang. Da wir in TIMSS auch die Häufigkeit von Demonstrations- und Schülerexperimenten erfragt haben, läßt sich die differentielle Itemschwierigkeit prüfen. Hier zeigt sich keinesweges ein Benachteiligung von Schüler mit experimentell orientiertem Unterricht: der differentielle Schwierigkeitsindex ist nicht signifikant und im Vorzeichen überdies negativ (DIF = $-.07$; SE = $.09$).

Zweites Beispiel: Test-Aufgabe M12

Abbildung 3: TIMSS-Aufgabe M12

M12. Einige Schüler benutzen ein Amperemeter A, um die Stromstärke im Stromkreis bei verschiedenen Spannungen zu messen.



Die Tabelle gibt einige Ergebnisse wieder. Vervollständige die Tabelle.

Spannung (Volt)	Stromstärke (Milliampere)
1,5	10
3,0	20
6,0	

Die Aufgabe verlangt die Vervollständigung einer Meßwertetabelle; der Lösungsschlüssel gibt den Wert 40 als richtige Lösung vor.

Hagemeister bringt verschiedene Einwände gegen die Brauchbarkeit der Testaufgabe M12 vor, die auf unterschiedlichen Ebenen liegen. Der zentrale Einwand betrifft die fachliche Korrektheit der Aufgabe. Er besagt, daß es keinen Glühlampentyp gebe, bei dem die Stromstärke linear mit der Spannung im Bereich von 1,5 bis 6 Volt zunehme; in einer eigenen Versuchsanordnung hat er den Wert von 22 Milliampère als realistisch ermittelt. Die übrigen Kritikpunkte liegen auf testtheoretischer Ebene. Hagemeister vermutet, daß diese Aufgabe gerade jenen Schülerinnen und Schülern besondere Schwierigkeiten bereiten könnte, in deren Unterricht experimentell gezeigt wurde, daß beim Betrieb von Glühlampen die Beziehung zwischen Spannung und Strom nichtlinear verläuft. Umgekehrt könnten schwache Physikschrüler bevorzugt werden, wenn sie die Vervollständigung der Meßwerttabelle als einfache Mathematik- oder Denkaufgabe behandeln. Insgesamt wünscht sich Hagemeister komplexere Testaufgaben, die "der Realität keinen simplen linearisierenden Ansatz überstülpen" (Hagemeister, S. 163).

Nach unserer Definition der Fähigkeitsstufen liegt die Aufgabe M12 mit einem internationalen Schwierigkeitsindex von 571 zwischen der Anwendung alltagsbezogener vorwissenschaftlicher Konzepte und der Kenntnis fachlicher Inhalte, die Standardschulstoff entsprechen. Die Aufgabe weist mit $r_{\text{bis}} = .38$ eine sehr gute Trennschärfe auf. Die internationale Lösungswahrscheinlichkeit für Schüler der 8. Jahrgangsstufe liegt bei 54 Prozent; in Deutschland liegt die Lösungswahrscheinlichkeit für diese Schülergruppe bei 69 Prozent.

Der Aufgabenstamm gibt als Versuchsanordnung einen einfachen schematisch dargestellten Stromkreis mit einer Batterie, einer Lampe und einem Amperemeter vor. Kontext und Darstellung des Aufgabenstamms bilden in idealisierter Form eine Schulsituation, keinen Alltagszusammenhang ab. Der Aufgabenstamm skizziert eine relativ offene Situation, die der weiteren Präzisierung durch Handlungsanweisungen bedarf, um zu einer Aufgabe zu werden. Mit der nachfolgenden Wertetabelle, in der bereits zwei Meßwerte vorgegeben sind, wird die Situation hinreichend bestimmt: Gelten soll der lineare Fall der direkten Proportionalität zwischen Spannung und Stromstärke unter der idealisierten Annahme eines konstanten elektrischen Widerstandes. Damit ist auch die mit dieser Aufgabe implizit erfaßte physikalische Wissensseinheit - das Ohmsche Gesetz - bezeichnet. Mit der Vorgabe der zwei Meßwerte werden auch alle Fragen stillgelegt, die der Aufgabenstamm aufwirft: Handelt es sich bei den vorgegebenen Spannungen um Leerlauf- oder Klemmenspannungen? Welches ist die UI-Kennlinie der Glühlampe? Gibt es Glühlampen, die in dem angegebenen Spannungsbereich eine annähernd lineare Kennlinie aufweisen? Werden Einschalt- oder Betriebsströme gemessen? Diese Fragen werden implizit beantwortet: Es sollen Bedingungen gegeben sein, unter denen das Ohmsche Gesetz gilt. Die Aufgabe hat also nur eine richtige Antwort - und zwar jene, die der Lösungsschlüssel vorgibt.

Dennoch hat auch Hagemeyer Recht, wenn er feststellt, daß es keinen gebräuchlichen Glühlampentyp gebe, bei dem die Stromstärke linear mit der Spannung im Bereich von 1,5 bis 6 Volt zunehme. Selbst wenn man eine Glühlampe auftriebe, deren UI-Kennlinie im Bereich dieser schwachen Spannungen annähernd linear verlief, wäre die Aufgabe fachlich nicht weniger unglücklich. Denn in jedem Physikbuch der Mittelstufe wird der

nicht-lineare Zusammenhang zwischen Spannung und Stromstärke bei Temperaturabhängigkeit des elektrischen Widerstandes anhand der in Aufgabe M12 wiedergegebenen Versuchsanordnung eingeführt. Der Versuch dient in der Schule in der Regel dazu, die Gültigkeitsbedingungen des linearen Zusammenhanges zu klären, die im Betriebszustand einer Glühlampe gerade nicht erfüllt sind. Gelegentlich wird allerdings mit derselben Versuchsanordnung auch gezeigt, daß die Linearität im Kaltzustand der Lampe, also bei Einschaltströmen, sehr wohl gilt (Walz, 1997; Dorn / Bader, 1992). Dies macht die Aufgabe aber ebenfalls nicht besser, denn in diesem Fall ergibt sich unmittelbar die Frage, ob unter diesen nicht explizierten Voraussetzungen die Aufgabe gerade für jene Schülerinnen und Schüler eine ganz erhebliche Klippe darstellen könnte, denen mit derselben Versuchsanordnung gezeigt wurde, daß beim Betrieb von Glühlampen das Ohmsche Gesetz nicht direkt anwendbar sei.

Dementsprechend formuliert Hagemeister zwei Vermutungen. Schüler, die Experimente zum nicht-linearen Fall gesehen hätten, seien mit besonderen Schwierigkeiten bei der Aufgabe M12 konfrontiert. Gleichzeitig würden jene Personen, die bar jeden Physikwissens seien, bevorzugt, da sie die Vervollständigung der Meßwertetabelle als Mathematik-oder Denkaufgabe behandelten. Wir haben die Vermutungen empirisch überprüft. Geht man davon aus, daß die Vervollständigung einer einfachen Wertetabelle unabhängig von physikalischem Wissen ist und praktisch allen Schülern der Mittelstufe gelingt, wird man mit sehr geringer Aufgabenschwierigkeit zu rechnen haben. Falls Schüler mit komplexerem Wissen mit dem erwarteten Situationsmodell der Aufgabe M12 besondere Schwierigkeiten haben, erwartet man eine niedrige - möglicherweise sogar negativer - Trennschärfe des Items. Ein Blick auf die empirisch ermittelten Schwierigkeits- und Diskriminationsindizes zeigt, daß beides nicht zutrifft.

Ein etwas anderes Bild ergibt sich, wenn man die differentielle Itemschwierigkeit für Schülerinnen und Schüler prüft, deren Unterricht experimentell orientiert ist. Für die Aufgabe M12 läßt sich in der Tat eine signifikante differentielle Itemfunktion nachweisen, die besagt, daß diese Aufgabe auch bei Kontrolle der Gesamtestleistung für experimentell erfahrene Schüler schwieriger ist ($DIF = .24$; $SE = .11$).

Wir wollen es allerdings bei dieser technischen Prüfung der Einwände nicht belassen, sondern versuchen, ihnen auch sachlich näher auf den Grund zu gehen - auch um zu prüfen, ob eine Lösung, die "der Realität (k)einen simplen, linearisierenden Ansatz überstülpt" (Hagemeister, S. 163), zu einer besseren Aufgabe führt. Wir haben deshalb die Aufgabe M12 dreifach variiert und einem 8. und 9. Schuljahrgang, der mit zwei bzw. drei Klassen besetzt war, vorgegeben.

In der ersten Variante haben wir die Aufgabe M12 aus dem Zusammenhang des Physikunterrichts herausgenommen. In dieser dekontextualisierten Form gleicht sie einer einfachen Mathematikaufgabe.

Abbildung 4: TIMSS-Aufgabe M12-1

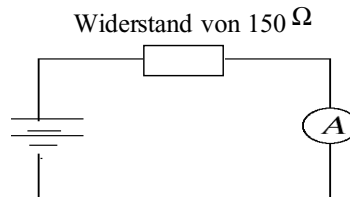
Bitte vervollständige die Wertetabelle:

x	y
2,5	20
5,0	40
10,0	

In der zweiten Variante wurde die Aufgabe so gestaltet, daß aus Zeichnung und Text klar hervorgeht, daß die Gültigkeitsbedingungen des Ohmschen Gesetzes als erfüllt gelten sollen. Gleichzeitig wurde mit dem Verzicht auf Vorgabe eines zweiten Meßwertes der unmittelbare Hinweis auf direkte Proportionalität beseitigt.

Abbildung 5: TIMSS-Aufgabe M12-2

Einige Schüler benutzen ein Amperemeter A , um die Stromstärke im Stromkreis bei verschiedenen Spannungen zu messen. Sie verwenden Materialien, deren elektrischer Widerstand sich während des Versuchs nicht ändert.



Die Tabelle gibt einige Ergebnisse wieder. Vervollständige die Tabelle.

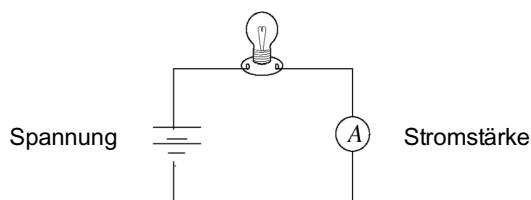
Spannung (Volt)	Stromstärke (Milliampere)
1,5	10
3,0	
6,0	

Begründe Deine Wahlen.

In der dritten Variante haben wir versucht, eine Mehrfachwahlaufgabe mit Begründungspflicht zu entwickeln, um das Verständnis des nicht-linearen Zusammenhang von Spannung und Stromstärke zu prüfen.

Abbildung 6: TIMSS-Aufgabe M12-3

Maria, Peter und Jan benutzen ein Amperemeter (A), um die Stromstärke im Stromkreis bei verschiedenen Spannungen zu messen.



Die Tabelle gibt ihre erste Messung wieder.

Spannung (Volt)	Stromstärke (Milliampere)
1,5	10
3,0	
6,0	

Sie überlegen, welche Stromstärke sie bei einer zweiten Messung mit einer Spannung von 3,0 Volt erwarten können.

Maria sagt: "Ungefähr 20 Milliampere."

Peter sagt: "Deutlich weniger als 20 Milliampere."

Jan erwidert: "Es könnten auch deutlich mehr sein."

Wer hat Recht? _____

Begründe Deine Entscheidung.

Die in Abbildung 6 wiedergegebene Aufgabe ist trotz Mehrfachwahlformat komplexer und offener als das TIMSS-Item M12. Als Wissensinheit werden die lineare Beziehung zwischen Spannung und Strom und deren Gültigkeitsbedingungen sowie das Konzept des elektrischen Widerstands und dessen Temperaturabhängigkeit vorausgesetzt. Es gibt keine Hilfe beim Lösungsansatz. Allein auf Grund der Tatsache, daß mehrere Wissensinheiten thematisiert werden, dürfte diese Aufgabe erheblich schwieriger als das TIMSS-Pendant sein.

Die Aufgaben wurden zwei Gesamtschulklassen der 8. Jahrgangsstufe in Ost-Berlin und drei Parallelklassen der 9. Jahrgangsstufe eines Gymnasiums in einem anderen Bundesland vorgegeben. Alle Schüler bearbeiten M12/1 sowie anschließend die Magnet Aufgabe D2. Dann wurde jeweils der Hälfte der Schüler M12/2 bzw. M12/3 vorgelegt. Die Gymnasialklassen wurden ausgewählt, da zum Zeitpunkt der Untersuchung das Ohmsche Gesetz im Physikunterricht entweder bereits behandelt worden war oder gerade behandelt wurde. In einer dieser Klassen war der Unterricht zum einfachen elektrischen Stromkreis etwa vier Monate vor der Untersuchung durchgeführt worden. Nachdem das Ohmsche Gesetz anhand des üblichen Konstantendraht-Versuchs eingeführt worden war, war der nicht-lineare Fall durchgenommen worden. In der Parallelklasse wurde das Ohmsche Gesetz gerade zum Zeitpunkt der Testvorgabe behandelt. Die Temperaturabhängigkeit des Widerstandes war jedoch noch nicht eingeführt. In der dritten Klasse schließlich hatte der Physiklehrer das Thema mit dem nicht-linearen Fall am Beispiel des Glühlampenversuchs eröffnet. Die UI-Kennlinie war eingeführt, aber noch nicht der Begriff des Widerstandes. Wie sehen die Ergebnisse aus?

Die Aufgabe können in dekontextualisierter Form (M12/1) fast alle Probanden - 125 von 129 oder 97 Prozent - lösen (in drei Fällen wird die Regel: "Addiere 20" angewandt). Damit unterscheidet sie sich deutlich von der kontextualisierten Physikaufgabe M12. Bei unserer zweiten Aufgabe, in der die Anwendung des Ohmschen Gesetzes verlangt wird (M12/2), sinkt die Lösungswahrscheinlichkeit auf 69 Prozent. Dabei erreicht auch die Klasse, die gerade das Ohmsche Gesetz durchgenommen hatte, kein besseres Ergebnis. Diese Aufgabe ist konzeptuell etwas schwerer als das TIMSS-Pendant. Nicht nur, weil ein möglicherweise noch nicht eingeführtes Symbol für den technischen Widerstand auftaucht, sondern vor allem, weil der zusätzliche Hinweis auf direkte Proportionalität fehlt, der durch den zweiten Meßwert gegeben wird. Als Begründung für die eingetragene richtige Lösung wird in 38 von 44 Fällen die direkte Proportionalität von Spannung und Stromstärke genannt. Sechs Schüler wählen die formalisierte Schreibweise: $U \sim I$ ($R = \text{konst.}$).

Ganz anders sieht das Ergebnismuster im nicht-linearen Fall aus (M12/3). 70 Prozent der Befragten gehen bei ihren Antworten von einem Zusammenhang direkter Proportionalität aus und begründen dies mit der Gültigkeit des Ohmschen Gesetzes. Selbst in der Klasse, die gerade den nicht-linearen Zusammenhang von Spannung und Stromstärke bei Glühlampen erarbeitet hatte, nehmen immer noch 60 Prozent der Schüler Linearität an. Nur acht Schüler von 65 (12 Prozent) wählen die richtige Lösung, die allerdings nur drei Schüler begründen können. Die übrigen Befragten kreuzen überwiegend ohne Begründung eine Lösung an.

Die Ergebnisse der Itemanalysen und die Befunde der Nachuntersuchung zeigen folgendes: Entgegen den Vermutungen Hagemesters dekontextualisieren Schüler, wenn ihnen Aufgaben vom Typus M12 vorgelegt werden, diese nicht auf ihren logischen Kern, und zwar auch dann nicht, wenn sie mit dem physikalischen Gegenstand nicht vertraut sind. Die Aufgabe bleibt eine Physikaufgabe; sie wird weder in den Kontext des Mathematikunterrichts übertragen, noch als einfache Denkaufgabe behandelt.

Diese Physikaufgabe wird jedoch durchweg, auch von leistungsstärkeren Schülern, unter der idealisierenden Annahme der Gültigkeit des Ohmschen Gesetzes bearbeitet - sogar wenn man sie so umformuliert, daß der nicht-lineare Fall explizit thematisiert wird. Deshalb kann die TIMSS-Aufgabe M12 meßtechnisch auch den eingangs beschriebenen Zweck erfüllen: zu überprüfen, ob Schüler das Ohmsche Gesetz korrekt anwenden können. Die Selbstverständlichkeit, mit der Schüler ein Situationsmodell mit dieser idealisierenden Annahme aufbauen, ist aus fachdidaktischer Sicht in der Tat diskussionswürdig. Eine Testaufgabe, die in Hagemesters Sinn das Ergebnis eines erfolgreichen offenen Experimentierens im Physikunterricht erfaßt, liegt jenseits des physikalischen Horizonts fast aller Mittelstufenschüler und zwar auch dann, wenn der entsprechende Stoff im Unterricht unmittelbar vorher durchgenommen wurde. Eine Aufgabe zum nicht-linearen Zusammenhang von Spannung und Stromstärke, wie wir sie unserer Stichprobe vorgelegt haben, differenziert nur im Bereich der obersten 5 Prozent der Leistungsverteilung. Verhältnisse dieses Leistungsbereichs generalisieren zu wollen, ist

gewiß eine wünschenswerte fachdidaktische Vision; in diagnostischer Hinsicht führt dies zu einem unbrauchbaren Leistungstest.

Zusammenfassend: Der Ansatz der TIMSS-Aufgabe M12, das Ohmsche Gesetz anhand des Glühlampenversuchs erschließen zu lassen, ist wenig glücklich, auch wenn die üblichen Itemparameter auf keinerlei Mängel hinweisen. Hauptschwäche der Aufgabe ist, daß sie zu einem nachweisbaren, allerdings geringen *bias* gegenüber experimentell erfahrenen Schülern führt. Es bleibt die Frage, weshalb die Autoren der Aufgabe an Stelle der Glühlampe nicht einen technischen Widerstand verwendet haben: sicherlich nicht um Lebenswirklichkeit zu suggerieren - die Aufgabe zielt bewußt auf Schulwissen und nicht auf Anwendung im Alltag -, sondern eher, weil der technische Widerstand im Physikunterricht dieser Alltagsgruppe nicht in allen TIMSS-Ländern behandelt wird. Bei einer Weiterentwicklung des TIMS-Tests sollte die Aufgabe M12 ausgeschlossen werden. Ob eine Aufgabengestaltung, die auf den Standardversuch mit einem Konstantdraht zurückgreift, zu einem besseren internationalen Testitem führt, ist offen.

Drittes Beispiel: TIMSS-Aufgabe R2

Abbildung 7: TIMSS-Aufgabe R2

- R2. Wenn weißes Licht auf Peters Hemd fällt, sieht es blau aus. Warum sieht das Hemd blau aus?
- A. Es nimmt das ganze weiße Licht in sich auf und verwandelt das meiste davon in blaues Licht.
 - B. Es strahlt den blauen Teil des Lichts zurück und nimmt den größten Teil des Restes in sich auf.
 - C. Es nimmt nur den blauen Teil des Lichts in sich auf.
 - D. Es gibt sein eigenes blaues Licht von sich.

Hagemeister schreibt zu dieser Aufgabe: "Auch bei dieser Aufgabe wird vor allem Textverständnis und ein wenig Logik benötigt. Wer sich den Text in Ruhe durchliest, wird sich sagen, daß eigentlich nur die Alternativen A oder B richtig sein können. Damit wäre die Wahrscheinlichkeit, die richtige Lösung anzukreuzen, immerhin schon auf 50 Prozent angestiegen. Ob nun A oder B richtig ist, können deutsche Achtklässler nicht entscheiden. Daß die Variante A mit dem Energiesatz nicht vereinbar ist, (...) wird bei

uns in der 8. Klasse in der Regel nicht mitgeteilt." Er fährt dann fort, daß der in der Variante B als richtig angenommene Fall für die Farben unserer Umwelt nicht relevant sei, da es sich bei der Farbe von Peters Hemd um ein spektralreines Blau handeln solle. Dabei werde "wiederum der bunten farbenfrohen Natur ein simples monokausales Schema übergestülpt" (Hagemeister, S. 167).

Ein Blick auf den Schwierigkeitsindex und die Verteilung der Schülerantworten über die vorgegebenen Alternativen zeigt, daß die Vorstellungen, die Mittelstufenschüler von Licht und Farben haben, offenbar wenig mit dem zu tun haben, was nach Hagemeisters Ansicht in den Köpfen von Schülern vorgeht, die über Textverständnis und ein wenig Logik verfügen. Der internationale Schwierigkeitsindex dieser Aufgabe liegt bei 653. Die Aufgabe läßt sich als Indikator für die dritte Stufe naturwissenschaftlicher Kompetenz heranziehen, auf der explizites fachliches Wissen verfügbar ist, das fast alle Schüler nur im Schulunterricht erwerben können. Der Stoff ist in den meisten Bundesländern - Ausnahmen sind Mecklenburg-Vorpommern, Brandenburg und Berlin - lehrplankonform. Die Aufgabe ist aber für Achtklässler schwierig: Die internationale relative Lösungshäufigkeit liegt bei 39%, in Deutschland bei 32%. Das Item besitzt befriedigende Trennschärfe ($r_{\text{bis}} = .27$). Die Lösungswahrscheinlichkeiten unterscheiden sich zwischen den Schulformen im wesentlichen erwartungsgemäß (Gesamtschule: 16%, Hauptschule: 24%, Realschule: 31%, Gymnasium: 42%).

Die fehlerhaften Antworten verteilen sich gleichmäßig über die Distraktoren ($A = .20$, $C = .19$, $D = .23$) bei einer nur wenig höheren Antwortwahrscheinlichkeit für die richtige Lösung ($B = .32$). Diese wünschenswerte und für die Qualität des Items sprechende Verteilung der Antworten über die Distraktoren wird selten erreicht. Wer ferner mit den (wenigen) fachdidaktischen Arbeiten über Schülervorstellungen zu Licht und Farbe vertraut ist, weiß nicht nur, daß es sich bei diesem Thema um ein schwieriges Unterrichtsgebiet handelt, da hartnäckige Alltagsvorstellungen sich dem fachlichen Verständnis entgegenstellen, sondern sieht auch, daß bei der Konstruktion der Distraktoren Schülervorstellungen systematisch berücksichtigt wurden.

Prüfen wir zunächst, welches mentale Situationsmodell zur Wahl der als richtig bezeichneten Antwort B führt. Die Alternative B verlangt, daß die Wahrnehmung der Farbe undurchsichtiger Körper als Folge selektiver Absorption und Streuung von Spektralfarben verstanden wird. Danach nehmen wir undurchsichtige Körper in der (Misch-)Farbe der gestreuten farbigen Lichter wahr. Das Verständnis dieses Konzepts setzt voraus, daß weißes Licht als Kombination von Spektralfarben aufgefaßt und die Sichtbarkeit von Objekten auf von ihnen abgestrahltes, auf unsere Retina treffendes Licht zurückgeführt wird. Das Konzept der selektiven Absorption und Streuung ist also selbst voraussetzungsvoll. Selbst wenn die theoretischen Voraussetzungen im Unterricht behandelt worden sind, ist keineswegs sicher, daß sie von den Schülern akzeptiert und verstanden wurden und Alltagsdeutungen ersetzt haben (Wiesner 1994).

Das zur Antwort B führende mentale Situationsmodell verlangt also ein Grundverständnis der selektiven Absorption und Streuung, jedoch nicht die Kenntnis genauer Fachterminologie. Ebenso wenig werden weiterführende Kenntnisse über die Regeln von Komplementärfarben oder der additiven und subtraktiven Farbmischung vorausgesetzt. Hagemesters Ausführungen über die "bunte, farbenfrohe Natur", die in den einschlägigen Experimenten zur Farbmischung im Physikunterricht zur Geltung kommen sollte, sind für die Beurteilung des Items völlig irrelevant. Wollte man Kenntnisse in diesem Bereich prüfen, müßte man die Aufgabe etwa durch die Einführung farbigen Lichts abwandeln - und sie würde, wie wir aus den Arbeiten von Gleixner und Wiesner (1995) wissen, für Mittelstufenschüler an die Grenze der Unlösbarkeit geführt. Da die Testaufgabe R2 diese fachliche Eindringtiefe *nicht* erreichen soll, bleibt auch die vorgegebene Lösung bezüglich der Zusammensetzung des Lichts, das die Wahrnehmung Blau hervorruft, vage. Der mit gutem Grund vorsichtig formulierte Text der Antwort B erlaubt folgende Präzisierung: In dem gestreuten Licht sind die Spektralfarbe Blau enthalten sowie weitere nicht benannte Spektralfarben, die im Vergleich zum absorbierten Spektrum den kleineren Teil ausmachen. Damit sind gerade die beiden Möglichkeiten ausgeschlossen, die Hagemester thematisiert bzw. problematisiert: Die Farbpigmente des Hemdes absorbieren weder ausschließlich Orange, so daß alle übrigen Spektralfarben als Komplementärfarbe Blau gesehen werden, noch streuen sie ausschließlich spektralreines Blau. Bei dem wahrgenommenen Blau könnte es sich gut um den realistischen Fall einer Mischfarbe handeln, die grüne, blaue und violette Anteile enthält. Die

Testkonstrukteure haben sich bei der Abfassung der richtigen Lösungsalternative offensichtlich große Mühe gegeben, eine Formulierung zu finden, die es erlaubt, ein Grundverständnis des Konzepts der selektiven Absorption und Streuung ohne Einführung von *termini technici* zu erfassen und den weiterführenden Problembereich der Komplementärfarben und Farbmischung zu vermeiden. Ebenso ist es für das zur Lösung der Aufgabe R2 erforderliche mentale Situationsmodell irrelevant, ob ein Schüler über eine physikalische Erklärung darüber verfügt, weshalb Antwortalternative A nicht richtig sein kann.

Werfen wir einen näheren Blick auf die Distraktoren der Aufgabe. Das Testitem R2 ist nicht zuletzt deshalb ein guter Indikator für die Verfügbarkeit fachlichen Schulwissens, weil die Distraktoren systematisch auf Alltagsvorstellungen zu Licht und Farbe zurückgreifen und dadurch sehr attraktiv sind (Driver u.a., 1994). Wenn das Thema Optik im Physikanfangsunterricht behandelt wird, ist die Lehrkraft bei der Mehrzahl der Schüler mit zwei änderungsresistenten Alltagsvorstellungen zum Licht und Sehen konfrontiert: (1) weißes Licht sei farblos und klar und (2) es illuminiere die Gegenstände, so daß wir sie sehen könnten. Die Vorstellung von gestreutem oder reflektiertem Licht, das auf unsere Netzhaut trifft, ist selten und wenn überhaupt vornehmlich bei spiegelnden Gegenständen anzutreffen (Anderson / Smith, 1983; Andersson / Karrquist, 1983; Feher / Rice, 1985; Gleixner / Wiesner, 1995; Wiesner, 1994). Körperfarben werden von den meisten Schülern als Eigenschaften von Gegenständen aufgefaßt, unabhängig von der Lichtquelle oder dem Rezeptor. Ein roter Gegenstand ist auch im Dunkeln rot. Weißes Licht bringt die Farbe der Körper zum Leuchten (Anderson / Smith, 1983; Guesne, 1985; Wiesner, 1994). Farbige Filter färben weißes Licht ein, indem sie die jeweilige Farbe des Filters hinzufügen (Andersson / Karrquist 1983; Watts, 1985; Wiesner, 1994). Farbige Licht wird als dynamisch aufgefaßt, das sich interagierend mit der Körperfarbe mischt oder den Körper neu einfärbt (Rice / Feher, 1987; Feher / Rice Meyer, 1992). Insgesamt sind dies Alltagsvorstellungen, die der Entwicklung eines fachlichen Verständnisses von Körperfarben als Folge selektiver Absorption und Streuung im Wege stehen.

Distraktor D greift solch eine verbreitete Schülervorstellung von Körperfarben auf. Danach sehen wir die eigene Farbe von Peters Hemd, das dieses durch die Beleuchtung mit

weißem Licht abstrahlt. Die Formulierung des Distraktors ist so gewählt, daß auch Schüler diese Alternative wählen können, die wissen, daß wir Gegenstände durch gestreutes Licht sehen.

Distraktor C nimmt die Vorstellung des dynamischen, den Gegenstand einfärbenden Lichts auf und verbindet sie mit dem Konzept der Spektralfarben. Der blaue Anteil des weißen Lichts färbt Peters Hemd blau ein. Diese Alternative ist gerade für Gymnasiasten besonders attraktiv, wenn die einschlägigen Themen in der Optik bereits durchgenommen sind, aber das Konzept der Körperfarben nicht richtig verstanden wurde. Die Ankreuzwahrscheinlichkeit der Antwort C ist für Gymnasiasten mit .24 signifikant höher als für Schüler anderer Schulformen.

Distraktor A schließlich gibt eine Erklärung für Körperfarben, die zur Zeit Newtons wissenschaftlich diskutiert wurde (Feher / Rice Meyer, 1992). Dieser Distraktor überträgt die vorherrschende Schülervorstellung über die Wirkung von Farbfiltern auf undurchsichtige Gegenstände. Körperfarben werden durch eine Interaktion von weißem Licht mit Eigenschaften des Gegenstandes erklärt.

Ein Wort zur Übersetzung von Testaufgaben sei hinzugefügt, da Hagemeyer bei dem Testitem R2 die Übersetzung des englischen Begriffs "*absorb*" mit "in sich aufnehmen" als Musterbeispiel schlechter Übersetzung bemängelt. Die Herstellung äquivalenter Übersetzungen bei internationalen Schulleistungsvergleichen ist ein dorniges Problem und in vielen Fällen wird man sich trotz aller Bemühungen mit zweitbesten Lösungen zufrieden geben müssen. Von kultureller Äquivalenz von Testitems spricht man, wenn Personen unterschiedlicher kultureller Herkunft, aber gleicher latenter Fähigkeit bei derselben - möglicherweise übersetzten - Testaufgabe identische Lösungswahrscheinlichkeiten besitzen. Im Rahmen der Konstruktion der TIMSS-Tests sind alle Items mit Hilfe der Analyse differentieller Itemfunktionen (DIF) statistisch auf Äquivalenz überprüft worden (Garden / Orpwood, 1996). Die meisten auffälligen Items wiesen in der Tat Übersetzungsprobleme auf, die entweder korrigiert werden konnten oder zum Ausschluß der Testaufgabe führten. Dennoch gibt es in der deutschen Übersetzung eine Reihe von Testaufgaben, bei denen man sich beim dritten und vierten Durchlesen bessere Lösungen vorstellen könnte - auch wenn die Items in ihren Meßeigenschaften dadurch nicht tangiert werden. Die Übersetzung von Aufgabe R2 ist nun allerdings ein

ausgesprochen ungeeignetes Beispiel, um den Übersetzern - durchweg Fachlehrern - Nachlässigkeit vorzuwerfen. Die naheliegende Übersetzung von *absorb* mit "absorbieren", die sich Hagemeyer wünscht, führt gerade zu keiner äquivalenten Lösung. Da der Begriff *absorb* im Englischen den Charakter eines Fremdwortes praktisch verloren hat, dies aber für "absorbieren" im Deutschen nicht gilt, wird das Item bei einer Wort-zu-Wort-Übersetzung im Deutschen für Schüler mit geringerem Wortschatz schwieriger als dies im Englischen der Fall ist. Die Übersetzer haben deshalb nach einer Alternative gesucht. In deutschen Physik-Lehrbüchern wird "absorbieren" gelegentlich mit dem umgangssprachlichen Ausdruck "verschlucken" beschrieben und erläutert. Diese Sprachebene zu wählen, wäre aber im Vergleich zum Englischen unkorrekt, denn in den englischsprachigen Lehrbüchern ist der äquivalente Ausdruck *soak up*. Die Übersetzer haben sich schließlich für die Fassung "in sich aufnehmen" entschieden, um schwächeren Schülern gerecht zu werden - und dies hat sich bewährt, wie die Kennwerte der Aufgabe und eine Analyse der differentiellen Itemfunktion zeigen.

Viertes Beispiel: TIMSS-Aufgabe L7

Abbildung 8: TIMSS-Aufgabe L7

- L7. Die Besatzungen zweier Schiffe auf dem Meer können sich durch lautes Rufen verständigen. Weshalb ist dies den Besatzungen zweier Raumschiffe bei gleichem Abstand voneinander im Weltraum nicht möglich?
- A. Der Schall wird im Weltraum stärker reflektiert.
 - B. Der Druck im Inneren der Raumschiffe ist zu groß.
 - C. Die Raumschiffe bewegen sich schneller als der Schall.
 - D. Es gibt keine Luft im Weltraum, in der sich der Schall fortbewegen kann.

Bei diesem Item meint Hagemeyer, fachliche und technische Mängel erkennen zu können. Auf Grund seiner Schülergespräche kommt er zum Schluß, daß ein Ankreuzen der als richtig bezeichneten Lösung D durch (in der Schule erworbene) physikalische Kenntnisse über Schallausbreitung unbeeinflusst sei. Was Sinn macht, weil „Akustik ... nach dem Berliner Rahmenplan nur als Wahlthema in Klasse 10 angeboten [wird]“ (Hagemeyer, S. 167). Er bestreitet auch daß die Vorstellung von einer an Stoffe gebundenen Schallausbreitung zum erfahrungsnahen Alltagswissen von Jugendlichen ge-

höre, wie es die Autoren des TIMSS-Berichts dargestellt haben (Baumert u.a., 1997, S. 83). Außerdem sei die als richtige bezeichnet Antwort D falsch weil, „in 500 bis 1000 Kilometer Höhe, wo heute Raumschiffe mit Astronauten um die Erde kreisen ... immerhin noch einige Millionen Moleküle und Atome pro Kubikmeter mitvorhanden [sind]“. Schließlich seien auch die Distraktoren B und C fachlich richtige Antworten.

Die Raumschiffaufgabe weist einen internationalen Schwierigkeitsindex von 473 aus. In unserer Analyse der Fähigkeitsniveaus konnten wir die Aufgabe als Markier-Item für die zweite Kompetenzstufe (d.h., die Anwendung alltagsbezogener vorwissenschaftliche Konzepte) identifizieren. Bei der Lösung von Aufgaben auf diesem Schwierigkeitsniveau reichen lebenspraktische Erfahrungen alleine nicht aus, sondern man muß - wenn auch auf Alltagsniveau - erste qualitative physikalische Konzepte einbringen. Die Lösungswahrscheinlichkeit für die 8. Jahrgangsstufe beträgt international 70 Prozent, für die deutsche Stichprobe 74 Prozent. Das Item weist mit $r_{bis} = .34$ eine gute Diskriminationsfähigkeit auf, die Trennschärfeindizes aller Distraktoren haben negative Vorzeichen.

Der Itemstamm beschreibt zwei Situationen, in denen unterschiedliche Bedingungen für Verständigung durch Rufen gelten. Der Proband wird zu einem vergleichenden Gedankenexperiment aufgefordert, in dem er - ausgehend von einer leicht vorstellbaren Situation auf der Erde - eine Erklärung für die Unmöglichkeit der Verständigung im Weltraum finden soll. Die Beschreibung beschränkt sich auf zentrale Elemente des Situationsmodells. Es ist weder davon die Rede, daß die Raumschiffe sich in einer bestimmten Höhe um die Erde bewegen - sie könnten sich auch weit entfernt von Himmelskörpern und ihren Atmosphären befinden -, noch spielen zu öffnende Luken oder Raumanzüge eine Rolle. Das für die Lösung erforderliche mentale Situationsmodell verlangt a) die Aktivierung der *Alltagsvorstellung*, daß die Ausbreitung von Schall an ein Trägermedium wie Luft gebunden ist und b) den Umkehrschluß, daß beim Fehlen des Trägermediums eine Ausbreitung des Schalls nicht möglich ist. Ähnliche Fragestellungen sind in nahezu jedem Physiklehrbuch der Mittelstufe zu finden:

- Warum könnten die Menschen auf der Erde niemals eine Explosion hören, die sich

im Weltraum ereignet? (Walz, 1993, S. 47)

- Zwei Astronauten stehen mit Schutzanzügen bekleidet auf dem Mond nebeneinander. Obwohl der eine sehr laut spricht, hört der andere nichts davon. Warum? (Feuerlein / Näpfel, 1992, S. 26)
- Warum hören wir im Weltraum nicht das Klingeln eines Weckers? (Dorn / Bader, 1992).

Drei Viertel der befragten deutschen Achtkläßler lösen das Gedankenexperiment richtig. Ähnlich erfolgreich sind auch Hagemesters Interviewpartner, obwohl diese seiner Ansicht nach nicht wissen konnten, ob Schall zu seiner Ausbreitung Materie benötige, da es "keine Situationen im Alltag gebe, bei denen man die Erfahrung machen könnte, daß Schall im luftleeren Raum nicht übertragen wird" (Hagemeister, S. 168). Wie ist dies zu erklären? Prüfen wir anhand der verfügbaren empirischen Forschungsliteratur, wie sich die Schülervorstellungen zur Schallausbreitung entwickeln.

Die für die Primarstufe vorliegenden Untersuchungen zeigen, daß für Schülerinnen und Schüler in diesem Alter die Schallausbreitung offenbar noch nicht an ein Trägermedium gebunden ist: Der Ton fliegt „wie ein Ball durch die Luft“ (Kircher / Engel, 1994; vgl. Wulf / Euler 1995). Ab der 4. Jahrgangsstufe wird dieses Modell allmählich durch Strahlenvorstellungen ersetzt. Ein Trägermedium scheint für Schallausbreitung allerdings nicht notwendig zu sein (Watt / Russel, 1990). Die Töne reiben sich eher an der Luft (Wulf / Euler, 1995). Diese Vorstellung ändert sich in der nachfolgenden Entwicklungsphase. Ab 13 Jahren gehört die Vorstellung, daß Schall zur Ausbreitung eines Mediums - natürlicherweise der Luft - bedürfe, zum Alltagswissen der großen Mehrheit von Jugendlichen, wie Bar, Zinn und Rubin (1997) in mehreren Untersuchungen mit israelischen Jugendlichen zeigen konnten (vgl. auch Driver u.a., 1994). Bemerkenswerterweise wird in diesem Alter die Vorstellung, daß zur Wirkung über Distanz ein Träger- oder Interaktionsmedium erforderlich sei, über unterschiedliche Phänomene hinweg generalisiert. Dazu gehören die Gravitation, Wärmeausbreitung, elektrostatische Anziehung oder der Magnetismus (Bar / Zinn / Rubin, 1997).

Wie kommt es zu dieser verbreiteten Vorstellung von einer an einen Träger gebundenen Schallausbreitung? Der Alltag von Jugendlichen ist reich an direkten und indirekten einschlägigen Erfahrungen: Das Fadentelefon, die Verständigung unter Wasser, der Rohrtelegraph, der Lauscher an der Tür oder das Horchen an der Eisenbahnschiene im Western. Der hypothetische Umkehrschluß, daß beim Fehlen eines Trägermediums die Schallausbreitung unterbunden werde, fällt Jugendlichen offensichtlich leicht - wie die Ergebnisse von Bar / Zinn / Rubin (1997) und die Item-Kennwerte der TIMSS-Aufgabe L7 zeigen.

Die gleichmäßig gewählten Distraktoren sind bei dieser Aufgabe besonders interessant, da sie unter anderen als den idealisierten Weltraumbedingungen plausible Erklärungen für die Unmöglichkeit der Verständigung durch Rufen abgeben könnten. Alternative C wäre eine gute Antwort, wenn es im Weltraum eine definierte Schallgeschwindigkeit gäbe. Antwort B macht auch Sinn, insoweit Schall beim Übergang zwischen Medien unterschiedlicher Dichte schlecht übertragen wird. Im Weltraum greift diese Erklärung jedoch zu kurz. Daß die fachwissenschaftlich richtige Abhandlung eines solchen Themas auch Lehrbuchautoren und Lehrerfortbildnern Schwierigkeiten bereitet, zeigt die Behandlung des Klingel-Experiments, das nach Hagemeister "von erheblicher Bedeutung auf dem Wege zu der Einsicht [sei], daß Schallausbreitung an Materie gebunden ist" (Hagemeister, S. 168). So heißt es in einem kürzlich veröffentlichten Survey über amerikanische Physiklehrbücher:

"It's hard to wipe out the old-physicists' tale about sound not being able to travel in a partial vacuum. The experiment with the ringing alarm clock in a bell jar being evacuated seems so convincing. However, it isn't a matter of sound not traveling in a low-pressure region. The effect is due to poor impedance match between the bell and low-density air, and between the air and the jar." (The Physics Teacher 1999, S. 299).

Das gleiche Mißverständnis findet man in deutschen Physikbüchern, wie z.B. Dorn / Bader (1992, S.7). Solche physikalischen Feinheiten gehen aber weit über die Kompetenzebene hinaus, die das Item anzeigen soll, nämlich die Ebene alltagsbezogener vorwissenschaftlicher Konzepte, zu denen tatsächlich die Vorstellung von einer an ein Trägermedium gebundenen Schallausbreitung gehört, wie Baumert u.a. (1997) geschrieben haben.

Die übrigen Items im Überblick

Die übrigen vorgeführten Items wollen wir nur kurz streifen, da die Kritik nach demselben Muster gestrickt ist und auf den gleichen Mißverständnissen beruht.

Abbildung 9: TIMSS-Aufgabe I10

- I10. Was ist der BESTE Grund dafür, daß eine gesunde Ernährung auch Obst und Gemüse enthalten soll?
- A. Sie haben einen hohen Wassergehalt.
 - B. Sie sind die besten Eiweißspender.
 - C. Sie haben viele Mineralien und Vitamine.
 - D. Sie sind die besten Kohlenhydratspender.

In Aufgabe I10 legt Hagemester viel komplizierende Interpretation hinein. Dazu gehört eine lange Ausführung zum Unterschied zwischen "Mineralien" (die in einer Lösungsalternative umgangssprachlich erwähnt werden) und "Mineralstoffen" (die eigentlich gemeint sind). Er hat recht, auch wenn in der deutschen Umgangssprache die Differenz eingeebnet ist. Wenn Nahrungsmittelpackungen lebenswichtige Mineralien anbieten, erwartet kein Verbraucher eine Wundertüte mit Bergkristallen. Ausschlaggebend ist aber, daß man zum Finden der korrekten Lösung nur eine Assoziation zwischen Obst, Gemüse, Gesundheit und Vitaminen herstellen muß. Dies ist bereits aufgrund lebenspraktischer Erfahrungen möglich. Dementsprechend liegt der Schwierigkeitsindex dieser Aufgabe noch unter der zweiten Stufe unserer Einteilung. Nicht mehr und nicht weniger als die unterste Stufe des naturwissenschaftlichen Alltagswissens soll mit dieser Aufgabe indiziert werden. Wer sie mit Erziehungsideen zum selbständigen Denken oder Theorien der Ernährungslehre verknüpft, verkennt den Meßzweck der Aufgabe.

Abbildung 10: TIMSS-Aufgabe N3

- N3. Eine Tasse Wasser und eine gleich große Tasse Benzin werden an einem heißen, sonnigen Tag auf einen Tisch ans Fenster gestellt. Ein paar Stunden später ist festzustellen, daß es in beiden Tassen weniger Flüssigkeit hat, aber vom Benzin noch weniger übrig ist als vom Wasser. Was zeigt dieses Experiment?
- A. Alle Flüssigkeiten verdunsten.
 - B. Benzin wird heißer als Wasser.
 - C. Einige Flüssigkeiten verdunsten schneller als andere.
 - D. Flüssigkeiten verdunsten nur bei Sonnenschein.
 - E. Wasser wird heißer als Benzin.

An N3 kritisiert Hagemester: "Textverständnis und ein bißchen Logik führen wieder einmal zu der Lösung". Er diskutiert die Aufgabe im Hinblick auf unterschiedliche physikalische Anforderungen, zum Beispiel die Gefahren, die auftreten, wenn das Gasgemisch explosiv werden könnte, oder was Verdunstung mit der Absorption von Infrarotstrahlung zu tun habe. Aber auch diese Aufgabe hat keinen didaktischen Anspruch: es wird auch nicht empfohlen, den Versuch zu Hause nachzuvollziehen. Die Aufgabe ist ein Gedankenexperiment, das überhaupt kein Fachwissen erfassen soll, sondern alltagsbezogenes vorwissenschaftliches Denken.

In der Kritik zu den Aufgaben N4 und P7 kehren im Grunde dieselben Argumente wieder. Es werden unbelegte und falsche Behauptungen über mangelnde Trennschärfen vorgetragen, es wird ein kultureller *bias* unterstellt, der - wie die empirische Prüfung zeigt - unzutreffend ist, es taucht wiederum die Verwechslung von theoretischen Testdimensionen und Aufgabenschwierigkeiten auf, und abermals werden normative Ansprüche an Unterricht und Meßzwecke von Items nicht auseinandergehalten.

5. Zusammenfassung

Die Hagemestersche Kritik des TIMSS-Tests setzt darauf, daß der Leser der Deutschen Schule seine Mittelstufenphysik vergessen hat, mit der einschlägigen fachdidaktischen Forschungsliteratur nicht vertraut ist und unbewiesene Behauptungen für bare Münze nimmt. Es war leicht zu zeigen, daß seine Behauptungen, die technische Mängel von

Items betreffen, mit einer einzigen Ausnahme (Item M12) nicht nur unbelegt, sondern falsch sind. Ebenso war es nicht schwierig, seine Validitätskritik zurückzuweisen. Verwundert hat uns in diesem Zusammenhang nur, welche geringe Begründungspflichten dem Kritiker seitens der Herausgeber der Deutschen Schule auferlegt wurden. Hauptanliegen unserer ausführlichen Reanalyse der Hagemeisterschen Kritik war es jedoch, auf basale Mißverständnisse hinzuweisen, die auch bei anderen Testkritikern anzutreffen sind und zu grundsätzlichen Fehltrüben über Leistungstests führen. Wir wollen die wichtigsten Kritikpunkte am Hagemeisterschen Vorgehen noch einmal aufführen:

- Hagemeister unterscheidet nicht zwischen den komplexen Funktionen von Aufgaben im Unterricht und der spezifischen Indikatorfunktion einer Testaufgabe in einem Fähigkeitsmodell. Er vermengt *normative* didaktische Vorstellungen über Unterricht mit Fragen der Diagnostik.
- Es ist absolut unzulässig und im schlimmsten Fall geradezu irreführend, Einzelitems ohne Berücksichtigung ihres Schwierigkeitsniveaus und ihrer Funktion im Gesamttest herauszugreifen, um sie vor dem Hintergrund willkürlich festgelegter fachlicher Eindringtiefe zu beurteilen. In der Regel steht hinter einem solchen Vorgehen eine bildungspolitische Absicht.
- Es fehlt eine klare Vorstellung vom Zusammenhang zwischen Meßeigenschaften einer Testaufgabe und den zur Lösung notwendigen mentalen Situationsmodellen bei Schülern. Hagemeister bringt Fachvorstellungen eines Physikers in Anschlag, wo Alltagsvorstellungen von Schülern erfaßt werden sollen. Dies führt dazu, daß er fast immer, wenn er fachliche Mängel eines Items meint entdecken zu können, die falsche Referenz wählt, da er nicht die jeweilige Indikatorfunktion des Items im Meßmodell berücksichtigt. Die fachliche Eindringtiefe einer Aufgabe ist nur vor dem Hintergrund des zu erfassenden Kenntnisniveaus zu beurteilen.
- Diese Fehltrüben sind unter anderem darauf zurückzuführen, daß Hagemeister mit der einschlägigen fachdidaktischen Literatur zu Schülervorstellungen nicht

vertraut ist und unrealistische Vorstellungen von der physikalischen Kompetenz von Mittelstufenschülern hat.

- Schließlich hat Hagemeyer den Grundgedanken des Klassifikationssystems der TIMSS-Testaufgaben nicht verstanden. Infolge eines Übersetzungsfehlers verwechselt er theoretische Testdimensionen mit Aufgabenschwierigkeiten.

Bei einem so großen und komplexen internationalen Forschungsvorhaben wie TIMSS gelingt vieles weniger gut, als man es sich gewünscht hätte. Dennoch ist das Projekt insgesamt ein gutes Beispiel für internationale Forschungskooperation, bei der es gelungen ist, Vertreter der pädagogischen und psychologischen Forschung und der Mathematik- und Naturwissenschaftsdidaktik einzubinden. Dies gilt in ähnlicher Weise für die deutsche Testadaptation, die nationale Durchführung und Auswertung der Studie und nicht zuletzt für die anschließenden Entwicklungsvorhaben, die im Rahmen eines Modellversuchsprogramms der Bund-Länder-Kommission stattfinden. Lehrkräfte und Fachdidaktiker waren nicht nur als Berater beteiligt, sondern sie haben die Testadaptation und die Validitätsprüfungen inhaltlich getragen. Die Auswertungen sind - glücklicherweise - schon längst in die Fachdidaktiken gegangen (Kaiser u.a., 1999, Blum / Neubrand, 1998; Fischer, im Druck) und die langfristig wichtigen Entwicklungsvorhaben liegen in den Händen von Fachlehrkräften an Schulen, die von der institutionalisierten Fachdidaktik unterstützt werden (Bund-Länder-Kommission 1997). Und selbst unsere Kritik der Kritik ist ein Beispiel dieser Kooperation.

Literaturverzeichnis

- Adams, R.J. / Wu, M.L. / Macaskill, G.: Scaling methodology and procedures for the mathematics and science scales. In: M. O. Martin / D. L. Kelly (Hg.): Third International Mathematic and Science Study. Technical report. Vol.II: Implementation and analysis. Primary and middle school years. (Chap. 7) Chestnut Hill, MA: Boston College 1997, S. 111-146.
- Anderson, C.W. / Smith, E.L.: Children's conceptions of light and colour: Understanding the concept of unseen rays. East Lansing: Michigan State University 1983
- Andersson, B. / Karrqvist, C.: How Swedish pupils, aged 12-15 years, understand light and its properties. In: European Journal of Science Education. 5, 1983, S. 387-402
- Arnold, K.-H.: Fairneß bei Schulsystemvergleichen. Münster: Waxmann, 1999,
- Bar, V. / Zinn, B. / Rubin, E.: Children's ideas about action at a distance. In: International Journal of Science Education, 19, 1997, 10, 1137-1157.
- Bassok, M.: Transfer of Domain-specific Problem-Solving procedures. In: Journal of Experimental Psychology: Learning, memory and Cognition 16, 1990, 3, S. 522-533.
- Baumert, J.: Technisches Problemlösen im Grundschulalter: Zum Verhältnis von Alltags- und Schulwissen - Eine kulturvergleichende Studie. In: A. Leschinsky (Hg.): Die Institutionalisierung von Lehren und Lernen. Weinheim: Beltz, 1996 (Zeitschrift für Pädagogik, 34. Beiheft). S. 187-209.
- Baumert, J. / Evans, R.H. / Geiser, H.: Technical problem solving among 10 year-old students as related to science achievement, out of school experience, domain-specific control beliefs, and attribution patterns. Journal of Research in Science Teaching, 35, 1998, 9, S. 987-1013.
- 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: Herber, H.-J. / Hofmann, F. (Hg.): Schulpädagogik und Lehrerbildung. Innsbruck: Studien Verlag 1998, S. 301-324.
- Baumert, J. / Köller, O.: Nationale und internationale Schulleistungsstudien: Was können sie leisten, wo sind ihre Grenzen? In: Pädagogik, 50, 1998, 6, S. 12-18.
- Baumert, J. / Lehmann, R., u.a.: TIMSS-Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen: Leske und Budrich, 1997.
- Baumert, J. u.a. (Hg.): Testaufgaben Naturwissenschaften. TIMSS 7./8. Klasse (Population 2). (Materialien aus der Bildungsforschung Nr.61). Berlin: Max-Planck-Institut für Bildungsforschung, 1998.
- Beaton, A.E. / Allen, N.L.: Interpreting scales through scale anchoring. In: Journal of Educational Statistics, 17, 1992, 2, S. 191-204.
- Beaton, A.E. / Martin, M.O. / Mullis, I.V.S. / Gonzalez, E.J. / Smith, T.A. / Kelly, D.L.: TIMSS. Science achievement in the middle school years. Chestnut Hill, MA: Boston College, 1996.
- Beaton, A.E. / Gonzalez, E.J.: TIMSS test-curriculum matching analysis. In: M. O. Martin / D. L. Kelly (Hg.): Third International Mathematic and Science Study.

- Technical report. Vol.II: Implementation and analysis. Primary and middle school years. (Chap. 10) Chestnut Hill, MA: Boston College 1997, S. 187-193.
- Bloom, B.S.: Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive Domain. New York: Longman, 1956.
- Blum, W. / Wiegand, B.: Wie kommen die deutschen TIMSS-Ergebnisse zustande? In: Blum, W. / Neubrand, M. (Hg.): TIMSS und der Mathematikunterricht. Informationen, Analysen, Konsequenzen. Hannover: Schroedel, 1998, S. 28-34.
- Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung: Gutachten zur Vorbereitung des Programms "Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts." Bonn: BLK, 1997.
- Camilli, G. / Shepard, L.A.: Methods for identifying biased test items. Vol. 4. Thousand Oaks, London, New Delhi: Sage, 1994.
- Dorn, F. / Bader, F.: Physik Mittelstufe. Hannover: Schroedel, 1992, S. 158-187 und S. 236-249.
- Driver, R. / Squires, A. / Rushworth, P. / Wood-Robinson, V.: Making sense of secondary science. Research into children's ideas. London, New York: Routledge, 1994.
- Feher, E. / Rice Meyer, K.: Development of scientific concepts through the use of interactive exhibits in a museum. In: Curator, 1985, 28, S. 35-46.
- Feher, E. / Rice Meyer, K.: Children's conceptions of color. In: Journal of Research Science Teaching, 29, 1992, No.5, S. 505-520.
- Feuerlein, R. / Näpfel, H.: Physik 1. München: Bayerischer Schulbuch-Verlag, 1992, S. 23-26.
- Feuerlein, R., Näpfel, H. / Schäflein, H.: Physik 2. München: Bayerischer Schulbuch-Verlag. 1996, S. 102-108.
- Fischer, H.E. (im Druck): Schlußfolgerungen aus der TIMS-Studie. Naturwissenschaften im Unterricht, Sonderheft TIMSS.
- Feuerlein, R. / Näpfel, H. / Schäflein, H.: Physik 3. München: Bayerischer Schulbuch-Verlag. 1994, S. 29-60.
- Garden, R.A. / Orpwood, G.: Development of the TIMSS achievement tests. In M.O. Martin / D.L. Kelly (Hg.), Third international mathematics and science study. Technical report. Vol. I.: Design and development (Chap.2). Chestnut Hill, MA: Boston College. 1996.
- Gleixner, C. / Wiesner, H.: Licht und Farbe: Akzeptieren Mittelstufenschüler eine elementarisierte Erklärung für das Sehen farbiger Oberflächen? In: Behrendt, H. (Hg.). Zur Didaktik der Physik und Chemie. Probleme und Perspektiven. Alsbach: Leuchtturm-Verlag, 1995, S. 207-209.
- Greer, B.: The modeling perspectives on word problems. In: Journal of Mathematical Behaviour, 12, 1993, S. 239-250.
- Guesne, E.: Light. In: Driver, R., Guesne, E. / Tiberghien, A. (Hg.): Children's ideas in science. Philadelphia: Open University Press, 1985, S. 10-32.
- Hagemeister, V.: Was wurde bei TIMSS erhoben? Rückfragen an eine standardisierte Form der Leistungsmessung. In: Die Deutsche Schule, 91, 1999, 2, S. 160-177.
- IEA: TIMSS Science items. Released set for population 2. IEA's Third International Mathematics and Science Study, 1998.
- Kaiser, G. / Luna, E. / Huntley, I. (Hg.): International comparisons in mathematics education. In: Ernest, P. (series ed.): Studies in mathematics education series. Vol. 11. Philadelphia, London: Falmer Press, 1999.

- Kircher, E. / Engel, C.: Schülervorstellungen über Schall. In: Sachunterricht und Mathematik in der Primarstufe. 22, 1994, 2, S. 53-57.
- Klieme, E. / Maichle, U.: Ergebnisse eines Trainings zum Textverstehen und zum Problemlösen in Naturwissenschaften und Medizin. In: G. Trost (Hg.): Test für medizinische Studiengänge. 14. Arbeitsbericht, 1990, S. 258-309. Bonn: Institut für Test- und Begabungsforschung.
- Klieme, E. (im Druck): Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische und methodische Grundlagen. In: Baumert, J. u.a., TIMSS - Mathematisch-naturwissenschaftliche Bildung am Ende der Sekundarstufe II. Opladen: Leske + Budrich.
- Knoche, N. / Lind, D.: Eine Analyse der aussagen und Interpretationen von TIMSS unter Betonung methodologischer Aspekte. Erscheint in: Journal für Mathematikdidaktik, 21, 2000, 1.
- Neubrand, J. / Neubrand, M. / Sibberns, H.: Die TIMSS-Aufgaben aus mathematikdidaktischer Sicht: Stärken und Defizite deutscher Schülerinnen und Schüler. In: Blum, W. / Neubrand, M. (Hg.): TIMSS und der Mathematikunterricht. Informationen, Analysen, Konsequenzen. Hannover: Schroedel, 1998, S. 17-27.
- Nieswandt, M.: Schreiben als Mittel zum verstehenden Lernen und Konsolidierung des Gelernten im Chemieanfangsunterricht des Gymnasiums. Dissertation. Christian-Albrechts-Universität zu Kiel, 1996.
- OECD: Measuring student knowledge and skills. A new framework for assessment. Paris, Organisation for Economic Co-Operation and Development, 1999.
- Orpwood, G./ Garden, R.A.: Assessing mathematics and science literacy. (TIMSS Monograph No. 4), Vancouver, Pacific Educational Press, 1998.
- Physics Teacher, The: Quibbles, Misunderstandings, and egregious mistakes. 37, 1999, S. 299.
- Ramseier, E.: Naturwissenschaftliche Leistungen in der Schweiz. Vertiefende Analyse der nationalen Ergebnisse in TIMSS. Bern: Amt für Bildungsforschung, 1997.
- Ramseier, E.: Leistungsprofil und Unterricht. Eine Analyse der schweizerischen Leistungen im naturwissenschaftlichen Test von TIMSS. In: Bildungsforschung und Bildungspraxis, 20, 1998, 1, S. 8-27.
- Ramseier, E.: TIMSS-Differenzen. Die Leistungen in den Naturwissenschaften und der Mathematik in Deutschland und der Schweiz. In: Die Deutsche Schule, 91, 1999, 2, S. 202-209.
- Reusser, K.: From Cognitive Modeling to the Design of Pedagogical Tools. In: S. Vosniadou, E. / De Corte, R. / Glaser / H. Mandl, (Hg.), International Perspectives on the Design of Technology-Supported Learning Environments (S. 81-103). New Jersey: Lawrence Erlbaum Associates, Publishers, 1996.
- Reusser, K. / Stebler, R.: Every word problem has a solution - the social rationality of mathematical modeling in schools. In: Learning and Instruction, Vol. 7, 1997, No. 4, S. 309-327.
- Rice, K. / Feher, E.: Pinholes and images: Children's conceptions of light and vision I. In: Science Education, 71, 1987, S. 629-639.
- Robitaille, D.F. / Garden, R. (Hg.): Research Questions and Study Design. Vancouver, Pacific Educational Press, 1996.
- Robitaille, D.F. / Schmidt, W.H. / Raizen, S. / Mc Knight, C. / Britton, E. / Nicol, C.: Curriculum frameworks for mathematics and science. TIMSS Monograph, No.1 Vancouver, Canada: Pacific Educational Press, 1993.

- Rymniak, M.J. / Kurlandski, G. / Smith, K.A. (Hg.): TOEFL-Test. New York, Kaplan Books, 1997
- Schmidt, W.H. / Jakwerth, P.M. / McKnight, C.C.: Curriculum sensitive assessment: Content does make a difference. In: International Journal of Educational Research, Chap. 2, 29, 1998, S. 503-527.
- Schmidt, W.H. / McKnight, C.C. / Valverde, G.A. / Houang, R.T. / Wiley, D.E.: Many visions, many aims. A cross-national investigation of curricular intentions in school mathematics. Dordrecht: Kluwer, 1997.
- Verschaffel, L. / De Corte, E. / Lasure, S. :Realistic considerations in mathematical modeling of school arithmetic word problems. In: Learning and Instruction, Vol. 7, 1994, No. 4, S. 273-294.
- Vijver, F. van de / Hambleton, R.K.: Translating tests: Some practical guidelines. In: European Psychologist, 1, 1996, 2, S. 89-99.
- Vijver, F. van de / Tanzer, N.K.: Bias and equivalence in cross-cultural assessment: An overview. In: European Review of Applied Psychology, 47, 1998, 4, S. 263-279.
- Walz, A.: Blickpunkt Physik. Hannover: Schroedel. 1997, S. 72-82 und S. 188-211.
- Walz, A.: Blickpunkt Physik 1. Hannover: Schroedel, 1993, S. 1-62.
- Watt, D. / Russell, T.: Sound. Liverpool: Liverpool University Press Science Process And Concept Exploration, 1990.
- Watts, M.: Student conceptions of light: A case study. In: Physics Education, 20, 1985, S. 183-187.
- Wiegand, B.: (1998). Stoffdidaktische Analysen von TIMSS-Aufgaben. In: mathematik lehren, 1998, Heft 90, S. 18-22.
- Wiesner, H. (1994). Verbesserung des Lernerfolgs im Unterricht über Optik (XIV). Farben. In: Physik in der Schule. Vol.32, Heft 2.
- Wilson, J. W. (1971). Evaluation of learning in secondary school mathematics. In: B. S. Bloom / J. T. Hasting / G. F. Madaus (Hg.) Handbook on formative and summative evaluation of student learning (S. 643-696). New York: McGraw-Hill, 1971.
- Wulf, P. / Euler, M. (1995). Ein Ton fliegt durch die Luft - Vorstellungen von Primarstufenkindern zum Phänomen Schall. In: Physik in der Schule, 33, 7-8, 254-260.