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

Konzeption und Aussagekraft der TIMSS Leistungstests. Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik [Teil 2]

Die Deutsche Schule 92 (2000) 2, S. 196-217

Jürgen Baumert, Eckhard Klieme, Manfred Lehrke und Elwin Savelsbergh

Konzeption und Aussagekraft der TIMSS-Leistungstests

Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik

Teil 1 dieses Beitrags wurde in Heft 1/00 publiziert!

1. Analyse der Kritik an den TIMSS-Testaufgaben

1.1 Situationsmodelle als Grundlage der Lösung von Testaufgaben

Will man einen Schulleistungstest beurteilen, ist es, wie in Abschnitt 2.1 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 Hagemeister 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/Sibberns 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 nahe legen. Die Schwierigkeit von *multiple choice*-Aufgaben wird nicht zuletzt dadurch bestimmt, inwieweit die Distraktoren die Entwicklung attraktiver, aber nicht adäquater Situationsmodelle nahe legen.

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 $3x = y\text{-}7/x = -2$ 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 schluntypische 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 anschließenden 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 Anhieb, 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.

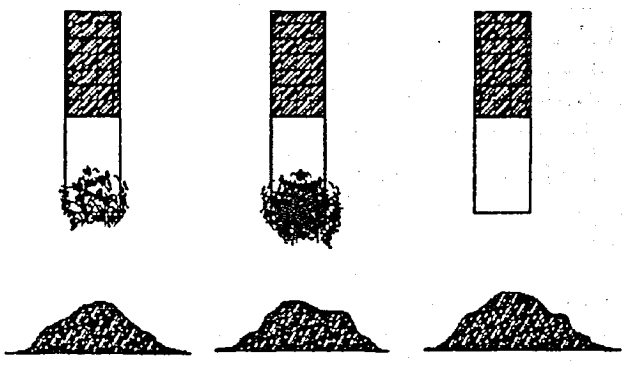
1.2 Überprüfung der Aufgabenkritik

Erstes Beispiel: Testaufgabe D2

Abb. 2

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

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

Hagemeister 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 hängen bleibe. 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. Hagemeisters 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 14-jährigen 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 ($r_{bis} = .36$) 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?" (wohlgemerkt: *nicht* „enthalten") vorgezeichnet. Die Aufgabe verlangt die Aktivierung einer einzigen Wissenseinheit, 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 hängen bleiben 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 adäqua-

ten 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.

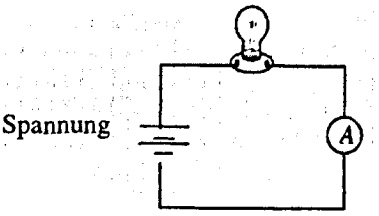
Wenn Hagemeister schließlich von Modellskizzen in Testaufgaben die Wiedergabe ästhetisch faszinierender Naturerscheinungen oder gar eine Erziehung zur Achtung vor der Natur erwartet, werden seine Einwände abwegig. Man kann diese Argumentation überhaupt nur verstehen, wenn nicht zwischen den Funktionen, die Aufgaben in einen 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. Hagemeister 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 Gesamttestwert 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 eine Benachteiligung von Schülern mit experimentell orientiertem Unterricht: der differentielle Schwierigkeitsindex ist nicht signifikant und im Vorzeichen überdies negativ (DIF = .07; SE = .09).

Zweites Beispiel: Test-Aufgabe M12

Abb. 3

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 Physikschüler bevorzugt werden, wenn sie die Vervollständigung der Messwerttabelle als einfache Mathematik- oder Denkaufgabe behandelten. 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 $r_{bis} = .38$ 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 Wissenseinheit – 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 Hagemeister 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 verliefe, 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 Gesamttestleistung 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.

Abb. 4

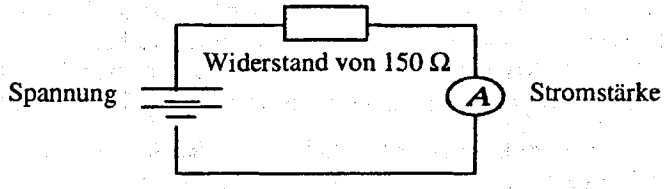
Bitte vervollständige die Wertetabelle:

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.

Abb. 5

Einige Schüler benutzen ein Amperemeter *A*, um die Stromstärke 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

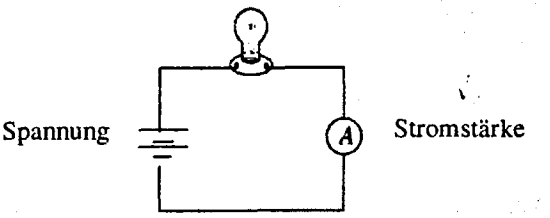
Spannung (Volt)	Stromstärke (Milliampere)
1,5	10
3,0	
6,0	

Begründe Deine Wahlen.

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

Abb. 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 Wissenseinheit 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 Wissenseinheiten 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 Konstantandraht-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 = 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 Hagemeisters 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 Gesetzt 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 Hagemeisters 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 eine Weiterentwicklung des TIMSS-Tests sollte die Aufgabe M12 ausgeschlossen werden. Ob eine Aufgabengestaltung, die auf den Standardversuch mit einem Konstantandraht zurückgreift, zu einem besseren internationalen Testitem führt, ist offen.

Drittes Beispiel: TIMSS-Aufgabe R2

Abb. 7

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 ($r_{bis} = .27$). 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. Hagemeisters 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 Hagemeister 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 Problemkreis 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). Farbiges 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 Hagemeister 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 identisch 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 nahe liegende Übersetzung von *absorb* mit „absorbieren", die sich Hagemeister 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

Abb. 8

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 Hagemeister, 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]" (Hagemeister, S. 167). Er bestreitet auch, dass die Vorstellung von einer an Stoffe gebundenen Schallausbreitung zum erfahrungsnahen Alltagswissen von Jugendlichen gehöre, wie es die Autoren des TIMSS-Berichts dargestellt haben (Baumert u.a. 1997, S. 83). Außerdem sei die als richtig bezeichnete 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 vorwissenschaftlicher 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 *Alltags*vorstellung, das 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 Hagemeisters 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.

2. 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.

Abb. 9

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 Hagemeister 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.

Abb. 10

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 Hagemeister: „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 auseinander gehalten.

3. Zusammenfassung

Die Hagemeistersche 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, welch 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 Fehlurteilen 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 Fähigkeitsmodel. 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 Fehlurteile 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 Hagemeister 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.

Literatur

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) Cestnut 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 ist 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 Schulwisssen – Eine kulturvergleichende Studie. In: A. Leschinsky (Hg.): Die Institutionslisierung 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 Schulleistungssgtudien: 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) Cestnut 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-Komission 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

Feuerlein, R. / Näpfel, H. / Schäflein, H.: Physik 3. München: Bayerischer Schulbuch-Verlag. 1994, S. 29-60
Fischer, H.E. (im Druck): Schlußfolgerungen aus der TIMS-Studie. Naturwissenschaften im Unterricht, Sonderheft TIMSS
Garden, R.A. / Orpwood, G.: Development of the TIMSS achievement tests. In M.O. Martin / D.L. Kelly (Hg.), Third international mathematic 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 wor(l)d 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 Problemösen 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-Operataion 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
Prenzel, M. / Duit, R.: Elf Ansatzpunkte für einen besseren Unterricht. Der BLK-Modellversuch „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts". In: Naturwissenschaften im Unterricht Physik, 47, 1999, 54.
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 1. Hannover: Schroedel, 1993, S. 1-62
Walz, A.: Blickpunkt Physik. Hannover: Schroedel. 1997, S. 72-82 und S. 188-211
Watt, D. / Russell, T.: Sound. Liverpool: Liverpool University Press Sience 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

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