

## AKTUELLE BILDUNGSPOLITIK

Volker Hagemeister:

### Was wurde bei TIMSS erhoben?

Eine Analyse der empirischen Basis von TIMSS

#### „Wie ein Gespenst“ geistert TIMSS durch die Lande

„Allmählich geistert die internationale Untersuchung ... wie ein Gespenst durch die Lande. Es verkündet: *Deutschlands Schüler sind ... bestenfalls Mittelmaß.*“<sup>6</sup>. Die Ergebnisse von TIMSS wirken wie ein unabänderliches Urteil, das höhere Mächte über uns gefällt haben. „Statt auf andere Sachverständige zu warten, die vielleicht weniger Schlimmes sagen, sollte sofort einer der Hauptgedanken aufgegriffen werden. Er betrifft die Unterrichtsmethode ... Denn die TIMS-Studie hat deutlich gemacht, dass die Schüler noch viel zu unselbstständig denken und lernen.“<sup>7</sup>

Dass sofort gehandelt werden muss, darin besteht breiter Konsens: Die *Kultusminister wollen* „so schnell wie möglich an allen Schulen *regelmäßig die Leistungen testen lassen.*“<sup>8</sup>

Nun wurde in der Vergangenheit immer wieder die Erfahrung gemacht, dass, wenn ein Netz *überregionaler Leistungstests* über die Schulen gelegt wird, selbstständiges Denken und Kreativität im Unterricht eher vernachlässigt werden:

- „überregionale Leistungsmessungen fördern die Entwicklung *zentraler Curriculumvorgaben* ...
- Schulleistungstests insbesondere bergen die Gefahr, den Unterricht auf die *Vermittlung* von Fakten zu reduzieren“<sup>9</sup>

Demnach wäre es also möglicherweise widersinnig, gleichzeitig

- mehr Selbstständigkeit und Kreativität im Lehren und Lernen zu fordern und andererseits
- überregionale Tests einführen zu wollen.

Wie kommt es, dass aus der TIMS-Studie *bildungsplanerische Maßnahmen* abgeleitet werden, die sich nicht sinnvoll ergänzen, sondern gegenseitig stören? Was bislang meiner Meinung nach fehlt, ist eine *sorgfältige Analyse der TIMS-*

---

<sup>6</sup> Schiller, Joachim 1998: „Deutsche Schulpolitik im Dilemma“, in: Der Tagesspiegel vom 3.7.98, Seite 27

<sup>7</sup> Schiller 1998

<sup>8</sup> Der Tagesspiegel vom 10.6.98, Seite 31

<sup>9</sup> Thomas, Helga 1989: Mögliche pädagogische Gefahren und Nebenwirkungen von Lernerfolgsmessungen. In: Ingenkamp, K. und Schreiber, W. H. (Hg.) 1989: Was wissen unsere Schüler, Überregionale Lernerfolgsmessung aus internationaler Sicht, Weinheim, Seite 242

*Studie.* Da die Studie auf empirisch ermittelten Daten aufbaut, muss die Beschäftigung mit TIMSS bei der empirischen Basis beginnen. - Ich stelle deshalb in den Abschnitten 3.1 bis 3.8 einige Testaufgaben vor, die in exemplarischer Weise zeigen, was bei TIMSS gemessen worden ist.

### **Gibt es anspruchsvolle Multiple-Choice-Aufgaben?**

Es ist „ein Irrglaube, man könne im *Multiple-Choice-Format* keine anspruchsvollen Aufgaben entwickeln, mit denen selbstständiges Denken ... oder Problemverständnis erfasst werden.“<sup>10</sup> Dieser Einschätzung stimme ich zu. Ein Test, der Multiple-Choice-Aufgaben enthält, kann ein sehr leistungsfähiges Diagnose-Instrument sein, sofern die Aufgaben mit entsprechender Sorgfalt entwickelt wurden.

Ein bewährtes Verfahren zur Entwicklung *leistungsfähiger Tests* besteht z. B. darin, dass man einer kleinen, repräsentativen Schülergruppe die geplanten Aufgaben zunächst in offener Form präsentiert. Die verschiedenen freien Antworten der Testgruppe können dann zu Alternativen für die fertigen Multiple-Choice-Aufgaben umformuliert werden.<sup>11</sup> Die falschen Alternativen in einer Multiple-Choice-Aufgabe spiegeln so ganz bestimmte Fehlvorstellungen von Schülern wider.

Vielfältige Informationen über die *Qualität von Multiple-Choice-Aufgaben* erhält man bei Interviews mit potentiellen Adressaten. Ich habe deshalb im September 1997 eine Reihe freigegebener Items aus der TIMS-Studie Schülern aus 2 Gymnasien vorgelegt. Es waren 3 Mädchen und ein Junge aus 8. Klassen und 3 Jungen aus einer 9. Klasse. Meine kleine Testgruppe entsprach also vom Alter her in etwa der Schülerpopulation, die bei TIMSS getestet worden war. Bei TIMSS wurden die Tests vor den Sommerferien, bei mir kurz nach den Sommerferien eingesetzt. Von den Leistungen her war meine Interview-Gruppe allerdings nicht repräsentativ, denn ich hatte mir Mädchen und Jungen aussuchen lassen, die als „sehr gut“ in Mathematik eingestuft wurden.

Die *Mädchen und Jungen meiner Testgruppe* haben jeweils einzeln ein Test-Paket mit 27 Aufgaben, die aus allen Feldern der TIMS-Studie ausgewählt waren, bearbeitet. Bei jeder Aufgabe wurde zunächst der Text still durchgelesen. Erst, wenn eine Antwort aufgeschrieben oder angekreuzt war, haben wir uns über die Aufgabe unterhalten.

---

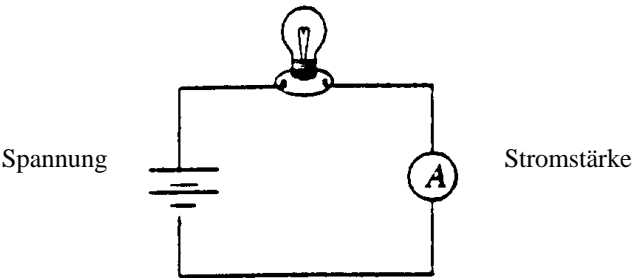
<sup>10</sup> Baumert, Jürgen, Köller, Olaf 1998: Nationale und internationale Schulleistungsstudien. In: Pädagogik, 50. Jg. 1998, Heft 6, Seite 12 bis 18

<sup>11</sup> Preibusch, W., Hagemeyer, V., Schuricht, K., Seyhan, H. 1984: Wer sind „die türkischen Schüler“? In: DDS, 76. Jg., 1984, Heft 3, Seite 224 - 238

**Naturwissenschaftliche Items, die bei TIMSS am Ende der 7. und 8. Klasse eingesetzt wurden**

**Item M12 bei TIMSS<sup>12</sup>**

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	

Meine Testkandidaten aus der 8. und 9. Klasse haben hier - meist nach nur kurzem Nachdenken - die Zahl 40 eingetragen. Schaut man im Internet nach, dann sieht man unter „Item key“, dass tatsächlich die Zahl 40 hier die richtige Lösung sein soll.<sup>13</sup>

Keiner meiner Interviewpartner hatte im Physikunterricht bereits die Größe „Spannung“ kennen gelernt. Keiner hatte ein Experiment - wie im Bild dargestellt - bislang gesehen oder gar selbst durchgeführt. Deshalb konnten meine Testkandidaten hier ganz unbeschwert die Zahl 40 eintragen. - Dass diese Zahlenangabe falsch ist, konnten meine Test-Schüler nicht wissen.

<sup>12</sup> TIMSS, Population 2, Science Content, Grouping G, © IEA, The Hague, 1994

<sup>13</sup> Einen Teil der Items, die bei TIMSS eingesetzt worden sind, findet man im Internet (in englischer Sprache) unter: <http://www.cstep.bc.edu/TIMSS1/Items.html>

Wer *Experimente mit Glühlampen* macht, so wie in Aufgabe M12 dargestellt, der wird feststellen, dass es keinen Glühlampen-Typ gibt, bei dem die Stromstärke linear mit der Spannung im Bereich von 1,5 bis 6 Volt zunimmt. - Für eine 40-Watt-Glühlampe haben wir in einer Schaltung, die Aufgabe M12 entspricht, folgende Ströme gemessen:

40 W/230V-Glühlampe, Schaltung entsprechend M12	
U (Volt) vorgegeben	I (Milliampere) gemessen
1,5	10
3,0	16
6,0	22

Die *Tabelle mit Messwerten* zeigt: Bei der 4-fachen Spannung steigt der Strom nicht auf das Vierfache, sondern es wird nur etwas mehr als das Doppelte für die Strom gemessen. Warum ist das so? Der Metallfaden in der Lampe wird umso wärmer, je mehr Strom fließt. Dadurch nimmt der Widerstand des Metallfadens stark zu. Dies gilt auch für andere Lampentypen.

Mit einem Experiment, wie in Aufgabe M12 dargestellt, kann man zeigen, dass bei Glühlampen *kein linearer Zusammenhang* zwischen Strom und Spannung besteht. Die wichtige Erkenntnis, dass in der Natur Linearität nur sehr eingeschränkt gültig ist, ließe sich also gerade anhand einer solchen Anordnung mit Glühlampe vermitteln. - Ein Schüler, der dies im Physikunterricht erfahren hat, hat mit Aufgabe M12 erhebliche Probleme.

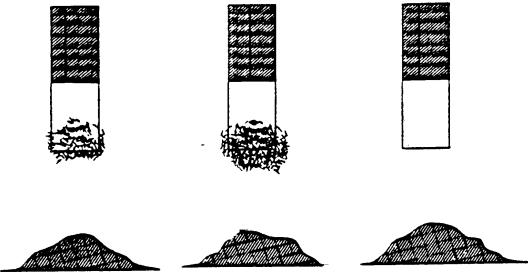
Die fachliche Fehlerhaftigkeit dieser Aufgabe zeigt, dass die Items der *TIMS-Studie nicht mit der notwendigen Sorgfalt entwickelt* worden sind. Bei der Entwicklung und Erprobung eines solchen Items hätte man irgendwann das Experiment vorführen müssen, das zu diesem Item gehört, um dann die Mädchen und Jungen, die zugeschaut haben, einzeln zu interviewen. Man hätte also in einer frühen Erprobungsphase der Tests bemerken müssen, dass diese Aufgabe so nicht gestellt werden darf.

Weshalb wurde hier in den Schaltkreis kein technischer Widerstand (z. B. von 150  $\Omega$ ), sondern eine Glühlampe eingezeichnet? - Ich vermute, man hat die *Glühlampen gewählt, weil sie eher Bezug zur Lebenswelt von Kindern* hat. Mädchen und Jungen der 8. Klasse sind Glühlampen sicherlich vertrauter als technische Widerstände. Bei ihrem Bemühen um Nähe zur Lebenswelt haben die TIMSS-Testkonstrukteure jedoch den Fehler gemacht, der Realität einen simplen, linearisierenden Ansatz überzustülpen.

Auch wenn Item M12 korrigiert würde (in dem in der Zeichnung die Glühlampe durch einen technischen Widerstand von  $150\ \Omega$  ersetzt wird) entsteht damit kein sinnvolles Item für einen Physiktest, denn durch die in der Tabelle vorgegebenen Werte wird ja schon suggeriert, dass es hier nur um eine lineare Fortsetzung geht. Bei Aufgabe M12 wird die Fertigkeit getestet, eine Tabelle linear fortzusetzen. Das Bild mit dem Stromkreis und die ersten beiden Sätze könnte man einfach weglassen. Deshalb konnten meine Interviews-Partner diese Aufgabe lösen, obwohl „elektrische Spannung“ im Unterricht noch nicht behandelt worden war.

### Item D2 bei TIMSS<sup>12</sup>

Jeder der abgebildeten Magnete 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

Stoff A      Stoff B      Stoff C

Sicherlich kann man sich leicht darauf verständigen, dass „Stoff C“ Kaffee sein könnte. - *Aber was für Stoffe liegen bei A und bei B?* Um reines Eisenpulver handelt es nicht, denn ein solches Häufchen Eisenpulver würde vollständig an einem Magneten hängen bleiben, wie jeder weiß, der schon einmal mit Eisenpulver und Magneten experimentiert hat.

Zwei der Neunt-Klässler, die ich interviewt hatte, haben mir erzählt, dass sie im Chemieunterricht *mit einem Magneten Schwefelpulver und Eisenpulver getrennt* hätten. Beide meinten dann, dass im Bild von Item D2 bei A und B eigentlich auch Kaffee liegen könnte. Kaffee, der mit Eisenpulver vermischt war. Das Eisenpulver ist durch den Magneten aus dem Kaffee herausgezogen worden.

Durch das Item D2 wird wiederum bestätigt, dass *zur Entwicklung der TIMSS-Tests keine physikalischen Experimente durchgeführt* worden sind. Man hätte beim Experimentieren nicht nur bemerkt, dass ein Häufchen Eisenspäne (wie im Bild bei D2 dargestellt) an den heute üblichen Dauermagneten vollständig hängenbleibt. Man hätte außerdem z. B. bemerkt, dass die Magnete in dem Stoff, in den sie „eingetaucht worden“ sind, Abdrücke hinterlassen hätten.

Bei einer *Einführung in wissenschaftliche Arbeitsmethoden* muss man meiner Meinung nach zunächst vor allem lernen, Phänomene genau zu beobachten.

Durch dieses sorgfältige Beobachten werden nicht nur wichtige Erkenntnisse gewonnen, sondern es wird auch Achtung vor der Natur vermittelt. Ferner können Bilder, die Magnete in Eisenspäne zeichnen, sehr schön aussehen. - Doch davon ist in dem TIMSS-Item D2 nichts präsent.

Bei dieser Aufgabe ist es wieder wie bei der Stromkreis-Aufgabe: Konkrete *experimentelle Erfahrungen stören*. Wer Experimente mit Magneten und Eisenspäne kennt, weiß, dass die kleinen Bilder mit den Magneten die Realität nicht befriedigend wiedergeben.

Einer der Neunt-Klässler, die ich zu den TIMSS-Items interviewt habe, meinte schließlich - nach dem er einige Zeit über diese Aufgaben nachgedacht hatte: „*Naja, wenn man so fragt, dann kann eigentlich - Nur C@ richtig sein.*“ - Wiederum wird bei dieser Aufgabe vor allem Textverständnis getestet. Elaborierte Physikkenntnisse hindern auch bei Item D2 daran, die Lösung schnell zu finden.

Dass fehlende Multiple-Choice-Routine auch für leistungsstarke Schüler bei TIMSS ein Handikap gewesen sein dürfte, zeigt die Beschäftigung mit Item D2: Mit dieser Aufgabe haben sich alle von mir interviewten Mädchen und Jungen unverhältnismäßig lange aufgehalten. Sie haben nach einem tieferen Sinn gesucht, weil sie es nicht gewohnt sind, in einer Physikarbeit mit fehlerhaft Texten oder Zeichnungen konfrontiert zu werden.

Eines der von mir interviewten Mädchen sagte, nachdem sie die Bilder bei Item D2 ein Weilchen betrachtet hatte: „*Leider weiß ich nicht, ob Kaffee Eisen enthält.*“ Experimente zum Trennen von Schwefel- und Eisenpulver waren bei ihr zu Beginn der 8. Klasse noch nicht vorgeführt worden, aber sie wusste, dass ein solches Häufchen Eisenspäne vollständig an einem Magneten hängen bleiben würde. Deshalb kam sie auf die Idee, dass Stoff A und Stoff B in Aufgabe D2 vielleicht stark eisenhaltiger Kaffee sein könnte, weil ja immerhin etwas Pulver von dem Magneten mitgenommen worden war. - Sie hat dann lieber keine der Alternativen angekreuzt.

Auch in meiner kleinen, nicht-repräsentativen Stichprobe haben die *Mädchen*, so wie bei TIMSS weltweit registriert, *schlechter abgeschnitten als die Jungen*. In meiner Interview-Gruppe lag dies weder an Mängeln im Physikwissen noch an Defiziten im logischen Denken und auch nicht an geringerem Textverständnis. Die Mädchen in meiner Interview-Gruppe sind jedoch häufiger als die Jungen an Ungenauigkeiten in der Aufgabenstellung hängen geblieben und die Mädchen waren weniger bereit, Halbwissen einzubringen, um Aufgaben zu lösen.

Warum sind die *vielen fachlichen und sprachlichen Ungenauigkeiten* beim Gelesen der deutschen Testaufgaben nicht bemerkt worden? Meine eigenen Erfahrungen bei der Begutachtung der TIMSS-Tests liefern hier eine plausible Erklärung: Ich bekam ein dickes Paket mit Testaufgaben zugeschickt und wurde gebeten, in einem Begleitbogen bei jeder Aufgaben-Nummer anzukreuzen, in

welchem Schuljahr entsprechende Inhalte in Berlin durchgenommen werden. Für diese Form der Begutachtung reichte meist ein flüchtiger Blick auf die einzelnen Aufgaben. So haben die meisten meiner Ko-Gutachter sicherlich viele Aufgaben nur überflogen und deshalb auch die vielen Fehler gar nicht bemerkt.

Mit Sicherheit sind auch zu der deutschen Variante der TIMSS-Tests nicht *die notwendigen Vorstudien* durchgeführt worden, indem man zu einzelnen Items Experimente vorgeführt hat, um anschließend Kommentare und Fragen der Mädchen und Jungen, die zugeschaut haben, sammeln zu können.

### **Item R2 bei TIMSS<sup>12</sup>**

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.<sup>14</sup>
- C. Es nimmt nur den blauen Teil des Lichts in sich auf.
- D. Es gibt sein eigenes blaues Licht von sich.

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% angestiegen.

Ob nun A oder B richtig ist, können deutsche Acht-Klässler nicht entscheiden. Dass Variante A mit dem Energiesatz nicht vereinbar ist, weil die Quanten des blauen Lichts zu den energiereichsten des sichtbaren Spektrums gehören, wird bei uns in der 8. Klasse in der Regel nicht mitgeteilt.<sup>15</sup>

Nach Lösungsschlüssel soll Variante B richtig sein. Allerdings ist das, was bei B in Item R2 steht, für die Farben in unserer Umwelt nicht relevant. - Warum sieht Peters Hemd oder *warum sehen Farben*, die im Kunstunterricht zum Malen benutzt werden, *blau aus*? Wir nehmen blaue Farbtöne wahr, wenn orange Anteile

---

14 In der Englischen Variante dieses Items steht bei B. „It reflects the blue part of the light and absorbs most of the rest.“ In der deutschen Fassung des Items wurde „absorb“ mit „in sich aufnehmen“ umschrieben. Dass der Übersetzer diese etwas umständliche Formulierung gewählt hat, kann man ihm nicht vorwerfen. Unbefriedigend ist jedoch, dass beim Gegenlesen der deutschen Fassung die unpassende Formulierung „in sich aufnehmen“ nicht durch „absorbieren“ ersetzt worden ist.

15 Weil von Rot nach Blau im sichtbaren Spektrum die Frequenz  $f$  zunimmt, sind (wegen  $E = hf$ ) die „Quanten“ des blauen Lichts energiereicher als die des grünen, gelben oder roten Lichts. Die von Max Planck gefundene Beziehung  $E = hf$  wird in der Sek. II eventuell behandelt. Um mit dieser Gleichung arbeiten und argumentieren zu können, muss man z. B. Experimente kennen, mit denen die Frequenz  $f$  von Spektralfarben gemessen wird.

aus dem sichtbaren Spektrum stärker absorbiert werden als die anderen Regenbogenfarben.

Die Regeln über „*Komplementärfarben*“ sind ein wichtiger Schlüssel zum Verständnis unserer Farbwahrnehmung: Addiert man alle Regenbogenfarben, außer Orange, so macht unser Gehirn aus diesem Farbgemisch den Eindruck „Blau“. Umgekehrt sehen wir einen orangen Farbton, wenn alle Regenbogenfarben bis auf Blau addiert werden. Blau und Orange nennen wir deshalb Komplementärfarben.

Das Blau in unserem Tuschkasten ist nicht nur für blaue und grüne, sondern oft auch für rote, unter Umständen sogar für gelbe Spektralfarben durchlässig. *Deshalb entsteht beim Mischen von blauer und gelber Tuschfarbe der Eindruck Grün*, weil sowohl das Tusch-Blau wie auch das Tusch-Gelb grüne Spektralfarben nicht absorbieren. Wäre das Blau aus unserem Tuschkasten nur für blaue und Tusch-Gelb andererseits nur für gelbe Spektralfarben durchlässig, dann würde Schwarz entstehen, wenn man Blau und Gelb übereinander malt. - Wäre Lösung „B.“ aus Item R2 für die Farben in unserer Umwelt relevant, dann würden wir beim Mischen von Pigmenten unterschiedlicher Farben immer nur Schwarz oder dunkle Braun-Töne erhalten<sup>16</sup>.

Ich möchte meine Ausführungen zum Mischen von Farben hier nicht weiter vertiefen, weil alles graue Theorie bleibt, wenn keine Experimente gezeigt werden. Zur „additiven“ und „subtraktiven“ Farbmischung gibt es viele eindrucksvolle und einige überraschende Experimente. Eine ganze Reihe solcher Experimenten muss man in Ruhe betrachtet und besprochen haben, ehe sich das Verständnis dafür entwickelt, warum Peters Hemd blau oder gelb oder grün aussieht und warum beim Mischen von blauer und gelber Tuschfarbe der Eindruck „Grün“ entsteht.

Sehr aufschlussreich ist bei Item R2 die Kommentierung im Internet: Zum Schwierigkeitsgrad steht bei R2 „Understanding Simple Information“. Die Testkonstrukteure haben also tatsächlich angenommen, dass es hier lediglich um das „*Verstehen simpler Informationen*“ geht.

Bei der Bewertung der Items der TIMS-Studie geht es also nicht nur darum, ob z. B. Farbenlehre in der 8. Klasse behandelt wurde oder nicht. Ganz entscheidend ist auch die Frage, *wie die Farbenlehre behandelt wurde*: Macht man Experimente, die sorgfältig beobachtet werden und über deren Interpretation in Ruhe nachgedacht wird, oder präsentiert man ein paar Faustregeln über Farben als „simple

---

16 Es gibt Filter, die entsprechend Lösung „B.“ von Item R2 alle Farben bis auf Blau absorbieren. Solche „monochromatischen“ Filter kosten einige Hundert DM, weil es sehr aufwändig ist, Pigmente so zusammenzustellen, dass alle Regenbogenfarben bis auf eine absorbiert werden.

Information“, wobei nach Übereinstimmung mit Experimenten gar nicht erst gefragt wird.

Auch bei Item R2 ist erneut das Bemühen der Science-Testkonstrukteure um Nähe zur Lebenswelt von Kindern erkennbar: Schließlich geht es hier um die Farbe von Peters Hemd. *Dabei stülpt man jedoch wiederum der bunten, farbenfrohen Natur ein simples, monokausales Schema über.*

### **Item L7 bei TIMSS**

Die Besatzungen zweier Schiffe auf dem Meer können sich durch lautes Rufen verständigen. Weshalb ist dies den Besatzungen zweier Raumschiffe bei gleichem Abstand 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.

Ein Junge aus meiner Interview-Gruppe meinte, die richtige Antwort sei „Die Raumschiffe bewegen sich schneller als der Schall.“ Die drei anderen Jungen haben

„Es gibt keine Luft im Weltraum, in der sich der Schall fortbewegen kann“, angekreuzt. Die Begründung klang jedes Mal sinngemäß gleich: „*Im Weltraum gibt es keine Luft, das weiß ich*“. Keiner meiner Interview-Partner aus 8. und 9. Klassen wusste dagegen, ob Schall zu seiner Ausbreitung Materie benötigt. - Akustik wird nach Berliner Rahmenplan nur als Wahlthema in Klasse 10 angeboten.

Ziemlich sicher waren sich die drei Jungen, die „D.“ angekreuzt hatten, dass sich Raumschiffe nicht schneller als der Schall bewegen können. Offenbar wurde von ihnen „Schallgeschwindigkeit“ mit „Lichtgeschwindigkeit“ gleichgesetzt. Dass nichts schneller als Licht sein kann, hatten sie schon irgendwo gehört. - Hier zeigt sich wiederum, wie wichtig es gewesen wäre, vor dem Einsatz der TIMSS-Tests Vorstudien mit Interviews durchzuführen. Denn dass drei meiner Interview-Partner bei der Raumschiff-Aufgabe die „richtige“ Lösung „D.“ angekreuzt hatten, war *durch Kenntnisse über Schallausbreitung überhaupt nicht beeinflusst.*

*Die drei interviewten Mädchen haben jeweils lange über die Raumschiffaufgabe nachgedacht.* Zwei von ihnen haben dann schließlich ebenfalls Lösung „D.“ angekreuzt. Die Dritte konnte sich zu keiner Alternative durchringen, weil sie in Sorge war, sie könne etwas Falsches ankreuzen: „Also, ich hab ja keine Ahnung, wie schnell die fliegen.“

Auch wenn Akustik und Schallausbreitung in Klasse 7 oder 8 behandelt worden sind, wäre die „Raumschiffaufgabe“ der TIMS-Studie kein sinnvolles Item für einen Physiktest:

1. Bei dieser Aufgabe sind auch die Antworten, die bei B und C stehen, richtig. So ist z. B. eine *Verständigung durch Rufen bei Überschall nie möglich*.
2. In 500 bis 1000 Kilometer Höhe, wo heute Raumschiffe mit Astronauten um die Erde kreisen, sind immerhin noch *einige Millionen Moleküle und Atome pro Kubikmeter vorhanden*. - Falls also im Unterricht die wichtige Erkenntnis vermittelt worden ist, dass „Vakuum“ eine Fiktion ist, so könnten Testteilnehmer aus dieser Klasse zu der (zutreffenden) Einsicht gelangen, dass Antwort „D. es gibt keine Luft im Weltraum . .“ bei der Raumschiffaufgabe nicht korrekt formuliert ist.

Sehr bemerkenswert ist, was in den „Deskriptiven Befunden“ über die Raumschiffaufgabe gesagt wird:

Zu dem „*erfahrungsnahen Alltagswissen* ... gehört die Vorstellung von einer an Stoffe gebundenen Schallausbreitung, wie sie die Raumschiffaufgabe<sup>17</sup> erfasst.“<sup>17</sup>

Hierzu muss man erwidern, dass „die Vorstellung von einer an Stoffe gebundenen Schallausbreitung“ keineswegs „erfahrungsnahes Alltagswissen“ ist. *Es gibt keine Situationen im Alltag, bei denen man die Erfahrung machen könnte, dass Schall im luftleeren Raum nicht übertragen wird*. Falls eine Schulklasse tatsächlich in einem Raumschiff mitfliegen könnte, würden die beteiligten Mädchen und Jungen dort sicherlich vielfältige Eindrücke mitnehmen. Aber die Erfahrung, dass sich Schall im Weltall nicht ausbreitet, könnten sie auch dort nicht machen. Wenn sie einem draußen vorbeifliegenden Mitschüler etwas zurufen wollten, müssten sie ja die Luken am Raumschiff oder am Raumanzug öffnen, was sofort zum Tode führen würde.

Dass Schall im luftleeren Raum nicht übertragen wird, kann man aus Alltagserfahrungen nicht ableiten. Im Gegenteil: Viele Schüler kennen *Science-Fiction-Filme mit heftigen Kämpfen und lauten Explosionen im intergalaktischen Raum*. Solche Kämpfe müssten eigentlich völlig lautlos im Film dargestellt werden (womit dieser Film natürlich nur geringe Resonanz beim Publikum finden würde).

Im Physikunterricht wird, sofern die entsprechenden Geräte zur Verfügung stehen, demonstriert, dass man eine Klingel, die unter einer Glasglocke liegt, irgendwann nicht mehr hören kann, wenn Luft unter der Glasglocke herausge-

---

17 Baumert, Lehmann u.a. 1997, Seite 83

pumpt wird. Dieses Experiment löst immer lebhaftere Reaktionen aus, gerade *weil hier kein erfahrungsnahes Alltagswissen vermittelt wird*. Die Ergebnisse dieses Labor-Experiments sind von erheblicher Bedeutung auf dem Weg zu der Einsicht, dass Schallausbreitung an Materie gebunden ist.

Mit der Raumschiffaufgabe wird der „Lebensbezug“, um den die Science-Test-Konstrukteure laufend bemüht sind, *ad absurdum* geführt. Lebensbezug ist hier nur ganz vordergründig vorhanden, weil jeder heute Raumschiffe kennt. Die Physik, um die es in der Raumschiffaufgabe geht, ist jedoch weit entfernt von „erfahrungsnahem Alltagswissen“.

Indem die Raumschiffaufgabe mit „erfahrungsnahem Alltagswissen“ verknüpft wird, wird hier unfreiwillig dokumentiert, dass bei der TIMS-Studie die *fächerübergreifende Zusammenarbeit zwischen Sozialwissenschaften und Physik-Fachdidaktik in vielfältiger Weise versagt hat*.

#### **Item N4 bei TIMSS<sup>12</sup>**

Auch das Item N4 wurde unter die Rubrik „Verstehen simpler Informationen“ eingeordnet:

Vor Jahren haben Landwirte herausgefunden, dass Maispflanzen besser gedeihen, wenn man daneben verwesenden Fisch vergräbt. Was hat der verwesende Fisch wahrscheinlich an die Pflanzen abgegeben, um ihr Wachstum zu steigern?

- A. Energie
- B. Mineralien
- C. Eiweiß
- D. Sauerstoff
- E. Wasser

Für Item N4 wird Faktenwissen benötigt. Denn Schüler der 8. Klasse kennen weder die Biochemie der Verfalls- und Stoffwechselprozesse, um die es bei Item N4 geht, noch wissen sie, was hier mit „Eiweiß“ oder „Mineralien“ gemeint ist. Dass man solche Themen routinemäßig als „simple Information“ abhakt, ist zum Glück bislang bei uns nicht üblich.

Hinzu kommt bei dieser Aufgabe noch, dass *für deutsche Kinder das Gedeihen von Maispflanzen nicht so alltäglich* ist wie für Kinder, die in den USA leben. Diese Aufgabe hätte also einer grundlegenden Transformation bedurft. Mit einer schlechten wortwörtlichen Übersetzung war es auch bei diesem Item nicht getan.

Mit Item N4 wird Vokabelwissen abgefragt, wobei in der deutschen Fassung an der entscheidenden Stelle eine falsche Vokabel steht, denn anstelle von „Mineralien“ hätte man bei B. „Mineralstoffe“ oder „Mineralsalze“ schreiben müssen.

Mit meiner Kritik an der Vokabel „Mineralien“ möchte ich keineswegs für einen wortklaubenden Unterricht plädieren. Wenn jede Schüleräußerung korrigiert

wird, weil immer irgend ein Wort falsch gewählt wurde, dann gehen die Schüler schließlich dazu über, sich die Formulierungen des Lehrers wortwörtlich einzuprägen, wodurch selbständiges, sinnvolles Lernen behindert wird.<sup>18</sup> - Aber auch wenn man ein Gegner von Wortklaubereien im Fachunterricht ist, so wird man doch einräumen müssen, dass in einem überregionalen Test die verwendeten Fachtermini stimmen sollten. Es kann ja nicht der Sinn eines Tests sein, dass Schüler dann schlechter abschneiden, wenn ihr Lehrer im Unterricht korrekte Fachtermini benutzt hat. Außerdem wird in Chemiebüchern zwischen „Mineralien“ und „Mineralstoffen“ unterschieden<sup>19</sup>. Falls dies im Unterricht thematisiert wurde, könnten die Mädchen und Jungen aus einer solchen Klasse bei Item N4 die „richtige“ Lösung „Mineralien“ nicht ankreuzen.

**Item I10 bei TIMSS**

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 Kohlehydratspender.

Erstaunlicher Weise wurde dieses Item ebenfalls unter „Verstehen simpler Information“ eingestuft. Die richtige Lösung ist nach „Item key“ „C. Mineralien und Vitamine“

Für die Achtklässler, die die Testaufgaben bearbeitet haben, war wohl kaum von Bedeutung, dass auch bei diesem Item anstelle von „Mineralien“ die Vokabel „Mineralstoffe“ hätte stehen müssen. Viel wichtiger ist wiederum, dass diese Aufgabe nicht mit einem Unterricht vereinbar ist, dessen Ziel es ist, *selbständiges Denken* zu fördern. Wenn die Inhalte aus Item I10 als „simple Information“ abgehandelt werden, dann müssen Merksätze mit Vokabeln, die nicht bekannt sind, auswendig gelernt werden. Außerdem sollte man Fragen der Ernährungslehre nicht als simple Botschaft abhandeln, nach dem Motto: Vor allem Vitamine, Spurenelemente oder Proteine sind wichtig. Damit wird dann z. B. die Fehlvor-

---

18 Die Nachteile des wortwörtlichen Lernens werden ausführlich diskutiert in:  
Ausubel, David P. u.a. 1980: „Psychologie des Unterrichts“, Weinheim, Basel, 1980

19 „Sofern es sich um feste Stoffe aus der unbelebten Natur handelt, bezeichnet man die Reinstoffe als Mineralien... Bekannte Mineralien sind z. B. Bergkristall, Saphir, Kalkspat ...“ (zitiert nach „Kemper-Fladt, Chemie“, Klett Schulbuchverlag, Stuttgart 1976, Seite 6). Mineralstoffe oder Mineralsalze nennt man dagegen z. B. Calcium- und Magnesium-Verbindungen, die Pflanzen oder Tieren (etwa aus wässriger Lösung) aufnehmen und verarbeiten können.

stellung erzeugt, man müsse nur in großer Menge synthetische Vitamine konsumieren, dann würde man nie wieder krank werden.<sup>20</sup>

**Item P7 TIMSS**

- Wenn Wissenschaftler irgendeine Größe mehrere Male sorgfältig messen, erwarten sie, dass
- A. alle Messwerte genau übereinstimmen.
  - B. nur zwei der Messwerte genau übereinstimmen.
  - C. alle Messwerte bis auf einen genau übereinstimmen.
  - D. die meisten Messwerte nahe beieinander liegen, jedoch nicht genau übereinstimmen.

Diese Aufgabe ist für alle Mädchen und Jungen leicht zu lösen, die „Messfehler“ nur aus Erzählungen des Lehrers kennen. Sie werden sicherlich ohne langes Zögern Lösung „D.“ ankreuzen. - Auch Multiple-Choice-Routine führt hier zur richtigen Lösung: „D.“ muss wohl richtig sein, denn mit der Aussage bei „D.“ haben sich die Aufgabensteller die meiste Mühe gegeben. - Dass man die richtige Lösung an ihrer Länge oder an der anspruchsvolleren grammatischen Struktur erkennen kann, ist ein Fehler, den Item-Konstrukteure häufig machen. - Dieser Fehler wurde z. B. auch bei der „Raumschiffaufgabe“ gemacht (Item L7 in Abschnitt 3.4).

Die Aussage, dass beim sorgfältigen Messen die Messwerte „nahe beieinander liegen, jedoch nicht genau übereinstimmen“ trifft sowohl in der Forschung, wie auch bei Experimenten, die in der Schule möglich sind, oft nicht zu. Werden z. B. „Nullraten“ mit einem Geiger-Müller-Zählrohr mehrere Male sorgfältig gemessen, dann werden immer wieder Messwerte „genau übereinstimmen“.

Wenn Wissenschaftler, die in der aktuellen Forschung tätig sind, eine „Größe mehrere Male sorgfältig messen“ kann es sehr wohl vorkommen, dass „alle Messwerte genau übereinstimmen“ (z. B. wenn es darum geht, zu bestimmen, ob die Schadstoffbelastung in einer Probe oberhalb eines Grenzwertes liegt).

Nach Meinung der TIMSS-Testkonstrukteure geht es auch bei Item P7 wiederum nur um das „Verstehen simpler Information“.

Es kann kein Zweifel daran bestehen, dass das Thema *Messungenauigkeiten* in der Schule immer wieder aufgegriffen werden muss. Beim eigenen Experimentieren müssen Mädchen und Jungen erfahren, dass Messungenauigkeiten unvermeidbar sind, auch wenn sehr sorgfältig experimentiert wird.

---

20 Aus den USA wird immer wieder von Menschen berichtet, die erkranken, weil sie synthetische Vitamine über Jahre hinweg in unsinnig hohen Dosen zu sich genommen haben. Die Items N2 und I10 aus der TIMS-Studie lassen die Vermutung zu, dass eine veraltete Ernährungslehre in den Schulen der USA solche Fehlorientierungen bestärkt.

Wenn deutsche Schüler am Ende der 8. Klasse beim TIMSS-Test nicht gewusst haben, was sie bei Item P7 ankreuzen sollen, dann deshalb, weil es bislang bei uns nicht üblich ist, Kenntnisse über Messungenauigkeiten als „simple Information“ in der 7. oder 8. Klasse abzuhandeln, in dem erzählt wird, was Wissenschaftler erwarten, wenn sie „sorgfältig messen“. Werden *Erkenntnisse über Messungenauigkeiten* schrittweise anhand eigener Experimente gewonnen, dann kann man nicht damit rechnen, dass am Ende der 8. Klasse fertige Faustregeln über Messungenauigkeiten parat sind.

**Item N3 bei TIMSS<sup>12</sup>**

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.

Wer über hinreichendes *Textverständnis* verfügt, wird sich bei dieser Aufgabe sagen können, dass die Aussagen, die bei A., B., E. und bei D. stehen, im Zusammenhang mit dem Vorspann wohl als irrelevant eingestuft werden können. Unter der Zusatz-Annahme, dass bei jedem Item eine Antwort richtig sein soll, folgt schließlich, dass er hier wohl

„C. Einige Flüssigkeiten verdunsten schneller als andere“

angekreuzt werden muss. *Textverständnis und ein bisschen Logik* führen wieder einmal zu der Lösung, die die Testkonstrukteure für die richtige halten. Tatsächlich stellt aus Sicht der Physik auch die Formulierung bei „C.“ eine unzulässige Verallgemeinerung dar. Bei „C.“ hätte z. B. stehen müssen:

„C. Benzin verdunstet unter Einwirkung von Sonnenlicht schneller als Wasser.“

Außerdem könnte die *Variante bei „B.“* auch eine richtige Lösung sein, denn es ist ja möglich, dass Benzin stärker als Wasser die IR-Strahlung der Sonne absorbiert. Oder absorbiert Wasser womöglich die IR-Strahlung der Sonne stärker als Benzin? Das hieße, Wasser würde unter Wirkung von Sonnenlicht heißer als Benzin werden (womit dann auch „E.“ richtig wäre). - Hier wird deutlich: Bei dieser Aufgabe hätte man die Sonne nicht ins Spiel bringen dürfen. Diese Aufgabe ist mit Nebenbedingungen und mit schlecht formuliertem Text überladen.

Vermutlich ist auch zu Item N3 nie ein Experiment durchgeführt worden. Von einem solchen Experiment müsste man auch abraten, denn wenn man eine „Tasse Benzin ... an einem heißen, sonnigen Tag auf einen Tisch ans Fenster“ stellt, könnte sich ein *explosives Luft-Gas-Gemisch* bilden. Und falls eine Explosion nicht stattfindet, hätte man es immerhin mit ziemlich giftigen Gasen zu tun, wenn Benzin in der Sonne verdunstet.<sup>21</sup>

## **Folgerungen aus den Tests der TIMS-Studie**

### *Charakteristische Merkmale der Science-Items bei TIMSS*

Das Science-Testpaket der TIMS-Studie lässt sich in folgender Weise charakterisieren:

- Experimente werden falsch dargestellt.
- Komplexe Sachverhalte werden simplifiziert und linearisiert.
- Das Science-Testpaket von TIMSS enthält sehr viel Text, der außerdem oft unpräzise formuliert und schlecht übersetzt wurde.
- Multiple-Choice-Items sind fehlerhaft, weil mehrere Alternativen richtig sind (wie bei der Raumschiffaufgabe, siehe Abschnitt 3.4)
- Bei einigen Items ist die richtige Lösung für einen multiple-choice-test-erfahrenen Schüler leicht zu finden, weil die richtige Variante sprachlich anspruchsvoller gestaltet ist als die falschen Alternativen.
- Hinderlich beim Lösen vieler TIMSS-Aufgaben sind Erfahrungen aus eigenen Experimenten und Reflexionen über die begrenzte Gültigkeit von Gesetzen.

Die fachlichen und sprachlichen Fehler, die viele TIMSS-Items enthalten, zeigen, dass die *Testbatterien der TIMS-Studie nicht mit der notwendigen Sorgfalt konstruiert und erprobt worden sind*. Ebenso hat es an Sorgfalt bei der Herstellung der deutschen Testvariante gemangelt, wie allein schon die vielen Vokabelfehler zeigen (siehe vorne z. B. 3.5 und 3.6, siehe auch die ausführlicher Darstellung in [22]).

### *Betrachtungen zur Validität der TIMSS-Tests*

So wie die *fächerübergreifende Kooperation* zwischen Fachlehrern und Sozialwissenschaftlern bei der Bewertung der deutschen Test-Fassung *versagt* hat, so

---

21 Wer ein Experiment zu dem TIMSS-Item N3 durchzuführen will, muss außerdem berücksichtigen, dass man Benzin nicht in einer „Tasse“ irgendwo stehen lassen darf. In einen Behälter, der auch zum Trinken benutzt wird, darf man nach allgemein anerkannten Sicherheitsregeln keine giftigen Flüssigkeiten füllen.

22 Hagemeister, Volker 1999: Analyse der Tests, die bei TIMSS eingesetzt worden ist, Berliner Institut für Lehrerfort- und -weiterbildung und Schulentwicklung, Berlin 1999 (in Druck)

unzureichend war die Beteiligung fachdidaktischer Kompetenz bei der Interpretation der Ergebnisse von TIMSS. Dem entsprechend enthalten die „Deskriptiven Befunde“<sup>23</sup> gravierende Fehlinterpretationen, wie z. B. die Aussage, durch die Raumschiffaufgabe würde „erfahrungsnahes Alltagswissen ... erfasst.“ (siehe vorne 3.4). Nicht zutreffend sind auch folgende Aussagen:

„... für den naturwissenschaftlichen Unterricht belegt die TIMSS-Curriculum-Studie die Existenz eines internationalen *Kerncurriculums* für die Mittelstufe.“<sup>24</sup>

„95% der Mathematikaufgaben und 88% der naturwissenschaftlichen Aufgabenstellungen repräsentieren Lehrplanstoff, der bis zum Ende der 8. Jahrgangsstufe in den Schulen der Bundesrepublik durchgenommen worden sein sollte.“<sup>25</sup>

Bezogen auf den Berliner Physikunterricht sind *über 50% der Aufgaben nicht rahmenplankonform*<sup>26</sup>. Aber selbst wenn z. B. im Rahmenplan für die 8. Klasse die Stichworte „Farbaddition“ und „Farbabsorption“ genannt würden, wäre das Item R2 der TIMS-Studie nicht valide, weil in diesem TIMSS-Item ein komplexes Phänomen linearisiert und auf nicht verstandenes Faktenwissen reduziert wird (siehe vorne 3.3).

Da die Testbatterien von TIMSS einerseits sehr viel Text enthalten und da andererseits Experimente falsch dargestellt werden, dürften die Tests von TIMSS insbesondere den Mädchen und Jungen Schwierigkeiten bereitet haben, die *gut in Physik und Mathematik aber schlecht in Deutsch und Sprachen* sind. Solche Schüler sind in Haupt- und Gesamtschulen überrepräsentiert. Dies ist nach meiner Einschätzung der entscheidende Grund dafür, warum Testteilnehmer aus Haupt- und Gesamtschulen bei TIMSS sehr viel schlechter als die getesteten Gymnasialschüler abgeschnitten haben.

In einer Kurzinformation des MPI für Bildungsforschung zum Thema TIMSS steht:

„Wenn *mathematisch-naturwissenschaftliche Bildung* überhaupt ... substantiierbar<sup>27</sup> ist, *deckt der TIMSS-Test diese in bislang nicht erreichter Qualität ab*“ Die Tests von TIMSS sind „bereits übermäßig breit angelegt. Durch das Hinzufügen weiterer Aufgaben wird praktisch kein Informationsgewinn erzielt.“

---

23 Baumert, Lehmann u.a. 1997

24 Baumert, Lehmann u.a. 1997, Seite 62

25 Baumert, Lehmann u.a. 1997, Seite 29

26 Hagemeister 1999

27 substantiierbar: durch Tatsachen belegbar, begründbar (nach Duden, Fremdwörterlexikon)

Diese Aussagen scheinen auf den ersten Blick in krassem Gegensatz zu dem zu stehen, was hier in den Abschnitten 3.1 bis 3.8 über einzelne Items der TIMSS-Studie gesagt worden ist. Tatsächlich passt jedoch insbesondere der letzte Satz des Zitats sehr gut zu meiner Kritik an den TIMSS-Tests: Die meisten Science-Items bei TIMSS messen offenbar gleiche oder eng verwandte Fertigkeiten. Dies sind vor allem *Lese- und Textverständnis*. Deshalb führt das Hinzufügen oder Wegnehmen einzelner Items nicht zu veränderten Ergebnissen.

Die Aussage, dass mit den TIMSS-Tests mathematisch-naturwissenschaftliche Bildung „in bislang nicht erreichter Qualität“ gemessen wird, ist offenbar das Resultat rein test-interner Überprüfungen. Das Science-Testpaket hat sicherlich eine hohe Konsistenz (in Bezug auf Leseverständnis). - Dass jedoch die *innere Konsistenz eines Testpakets nicht mit Validität gleichgesetzt* werden darf, dafür ist die TIMSS-Studie ein exzellentes Beispiel.

TIMSS bestätigt: „*Tests test tests*“<sup>28</sup>

Zusammenfassend muss man meines Erachtens sagen, dass es nicht sinnvoll ist, aus den Ergebnissen der TIMSS-Studie Aussagen über den naturwissenschaftlichen Unterricht in Deutschland abzuleiten. *Vergleiche etwa zwischen Schularten oder Aussagen über Lernfortschritte von einem Schuljahr zum anderen, lassen die Ergebnisse von TIMSS nicht zu*, weil nur wenige Items dem bei uns üblichen Unterricht entsprechen.

Wenn hier die Ergebnisse der TIMSS-Studie überwiegend als irrelevant oder auch irreführend eingestuft werden, so soll damit nicht etwa gesagt werden, dass bei uns im Physik- oder Mathematikunterricht alles zum Besten steht. Z. B. werden im Physikunterricht der Sek. I Übungen und Wiederholungen stark vernachlässigt. Damit fehlen im Physikunterricht wichtige Voraussetzungen für sinnvolles Lernen, denn neue Unterrichtsinhalte werden nur dann mit früher Gelerntem sinnvoll vernetzt, wenn beim Üben und Wiederholen die Bedingungen variiert werden und wenn gleichzeitig gefordert wird, Beobachtungen und Gedanken in eigenen Worten zu beschreiben.<sup>29</sup> Für einen solchen Physikunterricht braucht man viel Zeit und kleine Klasse. Gerade dies aber wird, ausgelöst durch TIMSS, neuerdings in Frage gestellt. So wird in Presse- und Zeitschriftenartikeln immer wieder hervorgehoben, dass in Japan angeblich bei gleicher Unterrichtszeit in Klassen mit 40 Schülern viel mehr erreicht wird als bei uns. Hierbei wird jedoch

---

28 Zitiert nach „Reinhard Kahl's Kolumne“. In: Pädagogik, 50. Jg., 1998, Heft 10, S. 64

29 Zum Thema „Variieren der Bedingungen beim Üben“ siehe z. B.:

- Bleichroth, Wolfgang 1998: „Mehr Üben!“ In: NiU Physik, 9. Jg., 1998, Heft 48, Seite 4 bis 8
- Steiner, Gerhard 1995: „Übung macht den Meister - unter welchen Bedingungen?“ In: Computer + Unterricht, 1995, Heft 9, Seite 7 und 8
- Hagemeister, Volker 1991: „Argumente für die Fortsetzung der Koedukation.“ In: DDS, 83. Jg., 1991, Heft 4, Seite 474-492
- Ausubel u.a. 1980

übersehen, dass Japanische Schüler in privaten Nachhilfeschoolen gezielt im Ausfüllen von Testbögen trainiert werden. Deshalb waren die Japanischen Schüler bei TIMSS vor allem im Lösen der Multiple-Choice-Items so erfolgreich.<sup>30</sup>

### *Die Wirkung überregionaler Tests*

Obwohl es nicht sinnvoll ist, aus TIMSS Aussagen über den naturwissenschaftlichen Unterricht in Deutschland abzuleiten, so liefert doch das Science-Testpaket sehr wertvolle Erkenntnisse für die aktuelle bildungspolitische Diskussion:

Das Science-Testpaket zeigt, wo man hinkommt, wenn die Leistungen von Schülern und Schulen über Jahrzehnte hinweg mit *überregionalen Tests* kontrolliert werden:

- Weil die Test-Industrie laufend neue Tests produzieren muss, die möglichst niemand kennen darf, fehlt es an Zeit und an Kommunikation für eine sorgfältige Testentwicklung.
- Wer Testaufgaben kritisiert, kommt immer zu spät, denn die Aufgaben müssen vor ihrem Einsatz schließlich geheim gehalten werden.
- Die Ziele der Schule werden auf Faktenwissen und auf vordergründige, multiple-choice-gerechte Logeleien reduziert.
- Experimente und Lebensbezug sind nur noch Staffage. Dieses Beiwerk wird linearisiert und simplifiziert, damit es in schlicht konstruierte Multiple-Choice-Aufgaben passt. Auf diese Weise kann man problemlos auch solche Experimente, die man nie selber durchgeführt hat, in Multiple-Choice-Form bringen.
- Kinder wohlhabender Eltern besuchen neben der Schule private Einrichtungen, wo trainiert wird, Testbögen rasch und richtig auszufüllen. - In Japan versucht man seit Jahren vergeblich, durch schulreformerische Maßnahmen die Macht und den Einfluss der privaten Paukschulen und der Verlage, die Testbögen herstellen, zurückzudrängen.<sup>31</sup>
- Wenn wir in Deutschland dazu übergehen, Testbatterien regelmäßig überregional einzusetzen, dann werden wir damit einen neuen Industriezweig ins Leben rufen, wo viel Geld verdient werden kann (mit Testentwicklung, Verkauf von Testbögen und privaten Testtrainings-Schulen). Diese Geister, die die KMK jetzt gerufen hat, werden wir auch dann nicht mehr loswerden,

---

30 Hoffmann, Andrea 1996: „Japanische Mittelschüler - schwach bei inhaltlichen Aufgabenstellungen“ ASD No 109 vom 1.12.1996, im Internet unter:  
<http://www.japonet.de/asd/109/inh109.htm>

31 Ito, Toshiko 1997: Zwischen Fassade und wirklicher Absicht, Zeitschrift für Pädagogik, 43. Jg., 1997, Heft 3., Seite 449-466

wenn wir längst bemerkt haben, dass unser Bildungssystem nicht besser aber ungerechter geworden ist.

Niemand kann im Alleingang fachlich korrekte und valide Tests herstellen. *Für die Entwicklung guter Multiple-Choice-Tests braucht man Teams aus Fachlehrern und Sozialwissenschaftlern.* Die Mitglieder solcher Teams müssen Erfahrungen in der Konstruktion, der Anwendung und der Auswertung von Tests erwerben. Dazu gehört z. B. auch, dass alle Teammitglieder an Interviews zur Itemprobung beteiligt sind, damit im Team kompetent über Validitätsprobleme diskutiert werden kann. Dies zeigt, dass der Anspruch, valide Multiple-Choice-Tests für eine überregionale Leistungskontrolle zu entwickeln, nicht kompatibel ist mit der Fließbandproduktion von Tests, wie sie in den USA üblich ist.

### ***Der Leistungsvergleich zwischen Staaten***

Eine Studie hat Ende der 60-er Jahre ergeben, „ *dass in den amerikanischen Schulen Tests etwa 30 bis 50mal häufiger eingesetzt werden*“ als an „ *bundesdeutschen Schulen... Heute ist die Häufigkeit des Testens in deutschen Schulen eher noch geringer.*“<sup>32</sup> Nun ist bekannt, dass „Test-Coaching“ erheblichen Einfluss auf Testergebnisse hat. Wer Übung im Ausfüllen von Multiple-Choice-Tests hat, wird auch in einem inhaltlich neuen Multiple-Choice-Test in der Regel besser abschneiden als ein multiple-choice-unerfahrener Mensch. Da aber die deutschen Schüler vergleichsweise selten mit Multiple-Choice-Tests konfrontiert werden, hätte man eine kleine, repräsentative Stichprobe zunächst intensiv im Ausfüllen von Testbögen trainieren müssen. Die Resultate, die diese vortrainierte Gruppe in den TIMSS-Tests erzielt hätte, hätte dann eine Abschätzung dafür geliefert, in welchem Ausmaß die Resultate von TIMSS durch Test-Coaching in Deutschland verändert worden wären.

Aus meinen Interviews kann man die Hypothese ableiten, dass die fehlende Test-Erfahrung erheblichen Einfluss auf die Ergebnisse bei TIMSS in Deutschland gehabt hat. Alle meine Interviewpartner haben lange über einzelne Aufgaben nachgedacht, weil sie immer wieder mit ungewohnten Formulierungen und Problemstellungen konfrontiert wurden. Wenn man nun bedenkt, dass 70 Aufgaben in 90 Minuten bearbeitet werden mussten, dann wird deutlich, wie wichtig Test-Routine bei TIMSS gewesen ist.<sup>33</sup>

---

32 Ingenkamp, K. und Schreiber, W. H. (Hg.) 1989: Was wissen unsere Schüler, Überregionale Lernerfolgsmessung aus internationaler Sicht, Weinheim, 1989, Seite 9

33 Besonders lange haben sich meine Interviewpartner z. B. bei dem unter „Understanding simple Information“ eingestuftem Item mit den 3 Magneten (D2) und bei der Raumschiffaufgabe (L7) aufgehalten. Weil die von mir interviewten Mädchen und Jungen es nicht gewohnt sind, dass Texte und Zeichnungen in Physikarbeiten fehlerhaft sind, haben sie bei den Items D2 und L7 lange nach einem tieferen Sinn gesucht.

Bei den Tests, die bei TIMSS in 7. und 8. Klassen eingesetzt worden sind, ist unübersehbar, dass sie die *Schulrealität in Nordamerika widerspiegeln*.<sup>34</sup> Für eine internationale Studie hätten die verwendeten Testbatterien aus internationaler Kooperation hervorgehen müssen. Dazu hätte man auch Test-Aufgaben entwickeln und bei TIMSS einsetzen müssen, die eng an die in Deutschland üblichen Rahmenpläne angelehnt sind.

Außerdem hätte, *bevor bildungspolitische Beschlüsse wegen TIMSS gefasst werden, eine breite öffentliche Diskussion über die Validität der bei TIMSS verwendeten Tests geführt werden müssen*. Diese unverzichtbare Diskussion ist bis heute gar nicht möglich, da ein erheblicher Teil der Testaufgaben bislang noch nicht veröffentlicht worden ist. So reden alle über irgendwelche Ergebnisse, obwohl die Messverfahren, mit denen diese Ergebnisse gewonnen wurden, weitgehend unbekannt sind. Dass dies so akzeptiert wird, zeigt, wie schwer es werden wird, die Arbeit der geplanten supranationalen Testinstitute der notwendigen fachlichen und pädagogischen Kontrolle zu unterziehen.

Insbesondere darf man nicht damit rechnen, dass es gelingen wird, den *Missbrauch der Testergebnisse* zu verhindern. Da wird dann ein Lehrer an den Pranger gestellt, wenn die Testergebnisse seiner Klasse unter dem bundesweiten Mittelwert liegen,

- weil die Kinder in seiner Klasse aus ungünstigen sozialen Verhältnissen kommen

und

- weil er sich bemüht hat, die Mädchen und Jungen in seiner Klasse zu selbstständigem Arbeiten und Denken zu ermuntern, anstatt rechtzeitig mit dem Test-Coaching zu beginnen.

### **Zusammenfassung**

Viele *Science-Items* bei TIMSS *enthalten fachliche Fehler*, insbesondere wenn es um die Darstellung von Experimenten geht. Dies zeigt, dass die Testbatterien der TIMS-Studie nicht mit der notwendigen Sorgfalt konstruiert und erprobt worden sind. Bei der Übertragung der Items ins Deutsche sind dann noch *Übersetzungsfehler* hinzugekommen.

So wie die fachlichen und sprachlichen Mängel, die viele TIMSS-Items enthalten, nicht bemerkt worden sind, so ist auch übersehen worden, dass die Testbatte-

---

<sup>34</sup> Die Mittelwerte, die ein Land in den Tests der TIMS-Studie erreicht hat, sind deshalb auch ein Maß dafür, wie klein oder wie groß die Differenzen zum Schulsystem der USA sind. Außerdem wurden die Ergebnisse bei TIMSS sicherlich durch die Lesegeschwindigkeit und das Textverständnis der Testteilnehmer beeinflusst. Die deutschen Schüler haben also das Handikap, z. B. im Ausfüllen von Tests ungeübt zu sein, durch hohe Lesegeschwindigkeit und gutes Textverständnis teilweise kompensieren können.

Volker Hagemeister:

---

rien, die bei TIMSS-II und -III eingesetzt worden sind, schlecht zu dem in Deutschland üblichen Unterricht passen. Deshalb ist es unzulässig, aus den Ergebnissen von TIMSS Aussagen darüber abzuleiten, wie groß (bzw. klein) Lernfortschritte von einem Schuljahr zum anderen sind oder welche Schulart oder welcher Staat am erfolgreichsten ist.

*Bevor bildungspolitische Beschlüsse wegen TIMSS gefasst werden, hätte eine breite öffentliche Diskussion über die Validität der bei TIMSS verwendeten Tests geführt werden müssen. Diese unverzichtbare Diskussion ist bis heute gar nicht möglich, da ein erheblicher Teil der Testaufgaben bislang noch nicht veröffentlicht worden ist. So reden alle über irgendwelche Ergebnisse, obwohl die Messverfahren, mit denen diese Ergebnisse gewonnen wurden, weitgehend unbekannt sind.*

*Der vorstehende Text ist ein Vorabdruck eines Aufsatzes in Heft 2/99 der Zeitschrift „Die Deutsche Schule“, das weitere Artikel zum gleichen Thema enthält.*

*Dr. Volker Hagemeister arbeitet im Berliner Institut für Lehrerfort- und -weiterbildung und Schulentwicklung (BIL).*

*Er ist zuständig für „Fächerübergreifende Projekte und außerschulische Lernorte, Schwerpunkt: Physik“.*