16.10.2006

Georg Lind

Viel ‘Wissen’ kann schädlich sein

In den sechziger Jahren ist der Soziologe Omar K. Moore von der Universität von Pittsburg bei Studien über die Inuits (Eskimos) auf eine merkwürdige Jagdvorbereitung gestoßen, die jeder Rationalität zu widersprechen schien. Statt sich moderner Ortungsmethoden für die Rentier-Jagd to bedienen, hielten sie noch immer an der überlieferten Methode fest: ein Speer wurde in die Höhe geschleudert und die Richtung, in der er zu liegen kam, bestimmte die Richtung der Jagdunternehmung. Moore erklärte die Rationalität dieser Methode so: würde man mit "moderneren" Methoden die Jagdtiere orten, wäre der Jagd sicher kurzfristig ein größerer Erfolg beschieden. Aber bald wären die Tiere so dezimiert, dass die Ernährungsgrundlage der Inuits gefährdet wäre. Durch das Zufallsprinzip mit dem Speerwurf werde sichergestellt, dass sich Jagderfolg und Tierbestand die Waage hielten. Das Prinzip widerspräche vielleicht unserer kapitalistischen, auf größtmögliche Ausbeutung gerichteten Rationalität, aber nicht der Lebenrationalität der Inuits. (Literaturstellen dazu stehen mir nicht zur Verfügung. Moore hatte mir seine Forschungsbefunde privat mitgeteilt.)

Jetzt haben *Gerd Gigerenzer* und seine Kollegen vom Max-Planck-Institut ein neues Beispiel für scheinbare Irrationalität erbracht, nämlich für die Tatsache, dass manchmal weniger Wissen vorteilhafter ist als zuviel Wissen. Es wäre reizvoll diese Erkenntnisse auf die Interpretation von Schulleistungstests anzuwenden. Ich habe manchmal den Eindruck, dass man dort besser abschneidet, wenn man nicht zuviel weiß bzw. nicht versucht, genau zu sein.

10.6.2003

Gerrit Stratmann[1]

Vorteil durch Dummheit

BILDUNGSFORSCHUNGWissen ist Macht, sagt Francis Bacon - doch nun haben Gerd Gigerenzer und sein Team die handfesten Vorzüge der Unwissenheit erforscht

Wenn Sie Ihr Geld in Aktien anlegen wollten, auf welche Kenntnisse würden Sie dabei eher vertrauen: auf die eines Fondsmanagers bei einer Bank, der viele Firmen, ihre Bilanzen und

[1] (c) Freitag, http://www.freitag.de/2003/24/03241801.php

Allianzen kennt, oder auf die eines Laien von der Straße, der kaum in die Feinheiten der Finanzwelt eingeweiht ist?

Wenn Sie verständlicherweise dem Bankangestellten die bessere Beratung zutrauen, könnte Ihnen unter Umständen viel Geld durch die Lappen gehen. Im experimentellen Vergleich haben Depots, die aufgrund von Laienwissen zusammengestellt worden sind, nämlich deutlich besser abgeschnitten als Expertenportfolios. Ein Zufallstreffer? Anfängerglück?

Weniger ist mehr, heißt es oft. Aber niemand nimmt an, dass das auch für Wissen gelten könnte. Kaum jemand würde behaupten, es sei besser, weniger zu wissen, um bei Günther Jauch Millionär zu werden. Es scheint doch so einleuchtend: Je mehr einer weiß, desto verständiger ist sein Blick auf die Dinge, desto treffsicherer sind seine Urteile, desto unfehlbarer seine Entscheidungen. Kann sein. Muss aber nicht.

Smart heuristics - clevere Entscheidungsregeln

"Viele glauben: Je mehr ich weiß, je mehr ich berücksichtige, desto besser wird die Entscheidung. Unsere Simulationen zeigen, das ist nicht immer der Fall", sagt *Ulrich Hoffrage* von der Arbeitsgruppe Adaptives Verhalten und Kognition am Max Planck Institut für Bildungsforschung. Dass die richtige Portion Unwissenheit sehr nützlich sein kann, ist ihm klar geworden, seit er eine Studie mit deutschen und amerikanischen Studenten durchgeführt hat, in der die Teilnehmer ein Urteil über die Größe von Städten abgeben sollten. Welche Stadt hat mehr Einwohner, fragte er. San Diego oder San Antonio? Es überraschte niemanden, dass fast zwei Drittel der Amerikaner die richtige Antwort wussten (San Diego). Überrascht waren Ulrich Hoffrage und seine Kollegen allerdings, als sie feststellten, dass von den deutschen Teilnehmern alle diese Frage richtig beantwortet hatten.

Kannten sich deutsche Studenten tatsächlich besser in Amerika aus als amerikanische? Natürlich nicht. Statt dessen keimte in den Forschern der Verdacht, genau dieser scheinbare Nachteil, das Weniger-Wissen, gereichte den Deutschen zum Vorteil. Die Amerikaner wussten zu viel über San Antonio. Ihr Wissen reichte zwar nicht immer, um die Frage sicher zu beantworten, aber es war genug, um sie zu verunsichern. Die Deutschen wussten wahrscheinlich gar nichts über San Antonio. San Diego aber hatten sie alle schon mal gehört. Um die Frage zu beantworten, verließen sie sich deshalb nicht auf ihr möglicherweise unsicheres Halbwissen, sondern eine ganz einfache Regel: nimm das, was du wiedererkennst.

Solche Regeln haben es Gerd Gigerenzer und seinen Mitarbeitern am Institut angetan. Sie suchen nach Heuristiken. Das sind Verfahren, die einem dabei helfen sollen, ein Problem zu lösen. In diesem Fall lautet das Problem: Wie treffen wir Entscheidungen, wenn wir nicht genau wissen, was die richtige oder beste Entscheidung ist? Die Herausforderung liegt darin, möglichst einfache Entscheidungsregeln zu formulieren, die sie als smart heuristics bezeich-

nen, als clevere Entscheidungsregeln. Was unterscheidet diese Regeln von anderen Vorstellungen darüber, wie wir Entscheidungen treffen?

Klassischerweise erwartet die Entscheidungstheorie, dass der Mensch sich absolut rational verhält. Rational gesehen, würde er alle Alternativen einer Entscheidung einschließlich ihrer möglichen Konsequenzen bedenken, sie mit Erwartungswerten versehen und dann diejenige wählen, die ihm den größtmöglichen Vorteil verspricht.

"Das Problem", so Gigerenzer, "ist, dass wir aus Versuchen wissen, dass die Leute sich nicht so verhalten." Die Vorstellung vom rationalen Handeln hat offensichtlich einen Schönheitsfehler. Sie stimmte nur, wenn der Mensch sämtliche Informationen hätte, sowie genügend Zeit und Ressourcen, um die Daten adäquat zu verarbeiten. Ein solcher Idealfall ist aber eine Illusion. Ein Personalchef, der eine neue Stelle besetzen will, hat niemals alle Informationen über die Bewerber, und selbst wenn er sie hätte, bräuchte er ein halbes Menschenleben oder einen Supercomputer, wenn er sie wirklich alle berücksichtigen wollte. Entscheidungen werden nicht auf diese Weise optimiert. Sie werden aufgrund begrenzter Informationen in begrenzter Zeit getroffen. Das klappt nur, wenn man einige der verfügbaren Informationen ignoriert.

Die Kunst zu wissen, was man nicht wissen muss

Rot oder gelb, Rom oder Prag, Eis oder Keks, Saft oder Sekt, verkaufen oder halten, jetzt oder nie oder vielleicht doch lieber später? Im Alltag treffen wir offensichtlich Entscheidungen nach einfachen Regeln, die schnell sind und sich mit wenigen Informationen begnügen. Fast and frugal, nennt Gigerenzer das. Solch eine einfache Regel ist zum Beispiel die Rekognitionsregel: Wenn du von zwei Objekten nur eins erkennst, dann ziehe den Schluss, dass das wiedererkannte Objekt den höheren Wert hat.

Was sich anhört wie eine Spielerei, kann zu erstaunlich handfesten Ergebnissen führen, nicht nur bei der eher theoretischen Frage nach der größeren von zwei Städten, sondern auch ganz praxisnah wie bei den eingangs erwähnten Investitionen am Aktienmarkt. In einer Umfrage unter Laien ermittelten die Forscher, welche von 800 börsennotierten Firmen namentlich bekannt waren. Aus den bekanntesten zehn stellten sie ein Depot zusammen und verglichen die Entwicklung mit den am wenigsten bekannten Firmen, dem Dax, dem Dow Jones und mit von Profis zusammengestellten Fonds. Im Ergebnis konnten die durch die einfache Rekognitionsregel ermittelten Firmen sehr gut mit den Indices und den Fonds konkurrieren – ganz ohne ausgeklügeltes Expertenwissen und ohne Entscheidungsstrategie. Diese Studie wurde 1997 in Zeiten eines wachsenden Marktes durchgeführt. Eine aktuelle Gegenprobe hat aber gezeigt, dass das Prinzip auch während einer Baisse funktioniert. Intelligentes Verhalten, so die Schlussfolgerung, hängt nicht von der Menge unseres Wissens ab. Die Kunst ist, wie Gigerenzer es formuliert, zu wissen, was man nicht wissen muss.

Am Zentrum für Adaptives Verhalten und Kognition kennt man noch andere Regeln unter klingenden englischen Namen wie Take the Best, QuickEst oder Categorisation by Elimination. Ihnen allen ist gemein, dass sie mit relativ wenig Informationen auskommen, um trotzdem relativ zuverlässige Entscheidungen zu ermöglichen. "Diese Heuristiken sind Annahmen darüber, wie Leute Informationen verarbeiten", erklärt Ulrich Hoffrage. "Und unter bestimmten Bedingungen haben diese einfachen Heuristiken eine höhere Erklärungskraft als komplizierte mathematische Verfahren, wenn es darum geht, zu schauen, was die Leute in manchen Situationen machen."

In etlichen Untersuchungen funktionieren diese einfachen Regeln so gut, dass sie es mit aufwändigen statistischen Modellen zur Vorhersage aufnehmen können. Da liegt die Frage nahe, ob sie diese komplizierten Methoden irgendwann überflüssig machen und ersetzen werden? "Nein. Schon insofern nicht, weil man die komplexen Verfahren braucht, um zu wissen, wie gut die einfachen Heuristiken sind", sagt Ulrich Hoffrage und lächelt.

Damit kein Missverständnis aufkommt: Wirklich sichere Entscheidungen lassen sich mit Hilfe keiner Regel fällen, da wir in einer unsicheren Welt leben, die von keiner noch so komplizierten oder genial einfachen Regel in einen vorherbestimmbaren Kosmos verwandelt werden kann. Es bleibt immer eine Wahrscheinlichkeit des Irrtums. Und wer trotz aller Hilfestellungen und Informationen nicht weiß, was die richtige Entscheidung ist, für den hat Ulrich Hoffrage noch eine letzte Regel parat: "Ich weiß nicht, ob man das als Heuristik bezeichnen würde: Wenn du nix mehr weißt, rate!"

10.6.2003

Heinrich Bauersfeld

Weniger Wissen ist manchmal vorteilhafter

Jetzt haben Gerd Gigerenzer und seine Kollegen vom Max-Planck-Institut ein neues Beispiel für scheinbare Irrationalität erbracht, nämlich für die Tatsache, dass manchmal weniger Wissen vorteilhafter ist als zuviel Wissen. Es wäre reizvoll diese Erkenntnisse auf die Interpretation von Schulleistungstests anzuwenden. Ich habe manchmal den Eindruck, dass man dort besser abschneidet, wenn man nicht zuviel weiß bzw. nicht versucht, genau zu sein. Insbesondere bei Mathematik-Tests kann es sehr schädlich sein, viel zu wissen (oder gar intelligent zu sein).

Wie heißt in der Folge die nächste Zahl? 1; 2; 4; 8; ?
Weltweit wird die Eindeutigkeits-Illusion aufrecht erhalten, es handele sich ganz klar um eine fortlaufende Verdopplung, also müsse dort unausweichlich 160 stehen.

Meine besonders befähigten Grundschüler (die aus dem 2.-4.Schuljahr einer Schule freiwillig an einer Förderstunde teilnehmen) fanden das keineswegs.
Ihre Vermutungen, wie das "Strickmuster" heißen könnte, lauteten u.a.
+1, +2, +4, +1, +2, +4, und immer so weiter. Also nächste Zahl 9.
+1, +2, +4, +2, +3, +5, +3, +4, +6, also immer eins mehr im Strickmuster. Mithin nächste Zahl 10.
+1, +2, +4, +2, +1. und damit Ende der Fahnenstage, Folge bricht nach schlichter 'Spiegelung' des Strickmusters ab. Nächste Zahl 10.
Usw. noch viele andere Angebote.

Genauer gesagt: Es gibt unendlich viele mögliche Fortsetzungen, insbesondere auch einfachere als die fortgesetzte schrittweise Verdoppelung!

Schon im Vergleich mit den wenigen Beispielen ist die Verdopplungshypothese keineswegs einfacher oder sinnfälliger oder gar zwingend zu nennen. Zumal für Grundschüler in der Regel die Multiplikation (als Operationsvermutung) ferner liegt als die Addition.

Na ja, man könnte noch fragen, wer hier eigentlich mit weniger Wissen gearbeitet habe. Sollten es die Testkonstrukteure selbst gewesen sein?

11.6.2003

San Diego oder San Antonio?

Marie Berger-Battran

Heute früh habe ich mich besonders amüsiert, weil mein Mann auf die Testfrage, was größer sei, San Diego oder San Antonio, auf San Antonio (!) tippte, was leider falsch ist. Interessant war seine Begründung, die in etwa so lautete: Ich kenne San Antonio nicht und muss damit rechnen, dass die Fragerin (also ich) damit rechnet, dass ich darauf vertraue, dass ich nur deshalb auf San Diego tippe, weil ich davon schon mal was gehört habe! Wenn ich diese Antwort nun nicht als Zeichen des Misstrauens einer "gestörten "Partnerbeziehung (i.S. von: typisches akademisches Misstrauen bei einem Lehrerehepaar oder so!) nehmen will, sondern als Erklärungsmuster nehme, so werden vielleicht Transferprobleme bezüglich des Schulalltags deutlich: In der Schule, so wie sie im Moment leider oft ist, wird ja meist nicht der produktive Zweifel belohnt oder der Mut zur "Lücke", sondern eher die Reproduktion des angeblich gesicherten lexikalischen Wissens. Nach meiner Erfahrung sind Schüler leider dann eher motivierbar zum stumpfen "Lernen" lexikalischen Wissens, wenn sie damit die vermeintliche Wahrscheinlichkeit erhöhen können, bei Günter Jauch eine Million zu gewinnen, als wenn sie durch Angebote zum kreativen Neuentdecken und Anerkennen des Nicht- oder noch-Nicht-

Wissens beim selbständigen Denken die Zukunftsprobleme der Gesellschaft lösen lernen. Die Frage nämlich, inwiefern ein Lernprozess im Hinblick auf relativ "unsicheres Halbwissen" auch eigentlich reflektieren kann, dass die sogenannte "Sicherheit" beim Wissen eine gelegentlich arrogante und vorschnelle Lösung darstellen kann, wird kaum diskutiert. Die "Standard"-Diskussionen bezüglich der neuen Lehrpläne wird – fürchte ich – in vielen Lehrerzimmern aufgrund der ungenügenden Zeit und der mangelnden qualifizierten Fortbildung eher auf eine Ansammlung von lexikalischen Wissenselementen a la Günter Jauch hinauslaufen . Und die vielen Schüler , die dann dennoch nicht eine Million gewinnen, werden nützliche " Module "dort sein, wo sie in der Produktions- oder Dienstleistungsgesellschaft durch "Reproduzieren" bequemer sind als durch eine eigene Zweifel und deren Folgen. Die Problemlösung wird so den Herrschaftseliten überlassen und denen ist San Antonio sowieso egal und die Millionen haben sie auf dem Konto aus ganz anderen Gründen als die Gewinner bei Günter Jauch.

Im Übrigen lese ich zur Zeit ein interessantes Buch von Donald D. Hoffman:" Visuelle Intelligenz. Wie die Welt im Kopf entsteht."(dtv33088, Jan 2003). Hoffman ist Professor für Kognitionswissenschaft, Philosophie und Computerwissenschaften an der University of California, Irvine. Er beschreibt das schöpferische Genie des Sehens, das jeder Mensch ohne sein Wissen hat und dort wird klar, welche Rolle die Reduktion auf ganz wenige Grundmuster beim Erkennen und KONSTRUIEREN der Welt immer wieder spielen. Dies deckt sich möglicherweise mit den Erkenntnissen der “smart heuristics”, die in dem von dir geschickten Bericht genannt werden . Spannend!

2000

Volker Hagemeister:

Irrwege und Wege zur „Testkultur“ Kann die „empirische Wende“ zur Qualitätssicherung beitragen?

Zusammenfassung:

Bedenken gegen den Einsatz von Tests sind keineswegs nur das Resultat irrationaler Ängste vor modernen Formen der Leistungsmessung. Im Gegenteil, gerade wer Erfahrungen mit standardisierten Tests hat, wird sich von der durch TIMSS ausgelösten Test-Euphorie nicht mitreißen lassen. Denn es ist sehr mühsam, sinnvolle Aufgaben zu konstruieren. Außerdem werden Test-Ergebnisse allzu leicht fehlinterpretiert. Zwar können valide Tests für Diagnosezwecke von großem Wert sein. Andererseits wird jedoch der überregionale Einsatz von

Testbatterien keineswegs zu mehr Gerechtigkeit oder sinnvollerem Lernen führen, wenn mit Hilfe der Testergebnisse Länder oder Schularten oder Schüler in Rangreihen gebracht werden.

Angst vor der empirischen Dampfwalze?

Wer zurzeit in Deutschland Bedenken gegen den überregionalen Einsatz von Tests äußert, dem wird gesagt, dass er

- „rührend naiv“ und „fachlich steril“ sei, „trotzig auf bisherigen Glaubenspositionen“ beharre und sich „mit dem Nichtwahrhabenwollen von Ergebnissen aus empirischen Erhebungen“ in der Sackgasse befinde (Sygusch 1999, S. 184, 185),
- am „Dogma von der Unfehlbarkeit des Pädagogen“ festhalte (Lange 1999, S. 152)
- und sich von „irrationalen Ängsten“ vor dem Einsatz von Tests lösen müsse (Schweitzer 1999, S. 13).

Der Kampf für die überregionalen Tests wird im Stil von Glaubenskriegen geführt: Der Gegner ist dogmatisch, trotzig, irrational, steril und schäbig: Anstatt sich als guter Verlierer dem schlechten Abschneiden der deutschen Schüler bei TIMSS zu stellen und nun ohne weiteres Herumlamentieren zu handeln, damit die Missstände bald behoben sind, gerät mir mit meiner kleinmütigen Kritik an den TIMSS-Items „die Notwendigkeit der Veränderung des Unterrichts ... aus dem Blick.“ (Bethge 1999 S. 178). Außerdem würde durch mein Herumkritisieren an einzelnen Items von TIMSS keineswegs „die Aussagekraft der gesamten Untersuchung ... in Frage“ gestellt (Bethge 1999, S. 179). - Ob ich nun tatsächlich nur unbedeutende Fehler in einigen Aufgaben aufgebauscht habe oder ob meine Kritik an dem Science-Testpaket die ganze Untersuchung in Frage stellt, kann zurzeit nicht in der notwendigen rationalen Weise öffentlich diskutiert werden, weil ein erheblicher Teil der TIMSS-Items noch nicht zur Veröffentlichung freigegeben wurde.

Die empirische Wende beginnt also in Deutschland mit einem Fehlstart: Anstatt dass wir uns zunächst mit der Basis der Empirie, mit den Testaufgaben ernsthaft beschäftigen, wird die Diskussion über die Validität der TIMSS-Items als nicht relevant oder gar „rührend naiv“ abgetan.[2] – Weil die Ergebnisse von TIMSS längst mit der bekannten großen Publizität

[2] Bethge erweckt zwar den Anschein einer inhaltlichen Diskussion, aber es ist unübersehbar, dass er sich mit meiner Itemkritik nur flüchtig beschäftigt hat. So reißt er Argumente aus dem Zusammenhang und ordnet sie falsch zu. Weil ich die TIMSS-Items analysiert hätte, „ohne einen Bezug zum Antwortverhalten der Schülerinnen und Schüler herzustellen“, hätte ich das Item I10 als besonders US-typisch bezeichnet, obwohl die deutschen Schüler dieses Item häufiger richtig gelöst hätten als die US-Schüler. Tatsächlich habe ich nur bei Item M4 den Inhalt als US-typisch bezeichnet, woran auch nichts zurückzunehmen ist. Über das Item I10 habe ich dagegen gesagt, dass sich hier ein überholtes Konzept von Ernährungslehre widerspiegelt. Diese Aussage wird nicht dadurch falsch, dass die deutschen Schüler bei Item I10 besser abgeschnitten haben als die US-amerikanischen. Vielmehr könnte man hier z. B. folgern, dass entweder viele deutsche Kinder einen überholten Ernährungslehreunterricht hatten oder dass ihre Informationen über „Mineralien und Vitamine“ aus außerschulischen Quellen stammen. - Ich habe zwar das ganze TIMSS-Testpaket als US-typisch charakterisiert. Dies bezog sich jedoch vor allem auf die Item-Formate und darauf, dass komplexe Fragen als „understanding

vorgestellt wurden, ist es nun nicht mehr statthaft, die empirische Basis von TIMSS in Frage zu stellen, denn ...

> … weit reichende Beschlüsse sind von wichtigen Gremien wegen TIMSS gefasst worden – da kann doch nicht plötzlich falsch sein, was jetzt gar nicht mehr falsch sein darf (andernfalls müsste sich ja die KMK geirrt haben).
> … von meinen Ko-Gutachtern kann man nicht erwarten, dass sie nun einräumen, sich seinerzeit zu hastig und oberflächlich die TIMSS-Items angesehen zu haben.

Meine Item-Diskussion mit den Hinweisen auf „länderspezifische Besonderheiten" sei ein sich „Einlassen auf ein Nationen-Ranking mit dem Versuch, den Rangplatz zu verschieben." (Bethge 1999, S. 179).

Hierzu muss ich richtig stellen, dass ich gerade wegen der „länderspezifischen Besonderheiten" ein „Nationen-Ranking" sehr fragwürdig finde. „Und wenn wir anstelle einer eng verstandenen 'Leistung' tatsächlich »Bildung« wollten, dann wäre das mäßige Abschneiden bei solchen Leistungs-Vergleichen geradezu als Erfolg zu werten!" (Schlömerkemper 1998, S. 264)[3]

Bedenken gegen einen allzu leichtfertigen Einsatz irgendwelcher Tests sind keineswegs nur das Resultat irrationaler Ängste vor modernen Formen der Leistungsmessung. Im Gegenteil, gerade wegen eigener Erfahrungen im Entwickeln und im Einsatz solcher Tests, ist mir die schlechte Qualität der Science-Items von TIMSS aufgefallen. Einerseits wird zurzeit Test-Euphorie verbreitet, aber andererseits wird nicht thematisiert, was genau mit den überregional gewonnenen Daten gemacht werden soll. Lediglich bei Arnold (1999) wird konkret dargestellt, in welch unterschiedlicher Weise Testergebnisse benutzt werden können. In dem Artikel von Arnold wird u.a. das von einer englischen Universität entwickelte Testprojekt „CEM" beschrieben. Auch bei „CEM" kommen Tests überregional zum Einsatz, jedoch erhält nur der einzelne Lehrer oder die jeweilige Schule einen Diagnosebericht von der Universität. Ein solches Projekt darf man nicht mit dem in England ansonsten praktizierten Schul-Ranking in einen Topf werfen, in dem man pauschal sagt, in vielen Ländern sei die externe Beurteilung von Lehrern selbstverständlich, „ohne dass negative Folgen für die Qualität des Unterrichts zu beobachten wären, eher im Gegenteil." (Schweitzer 1999, S. 141)

Weil angeblich im Ausland überall gute Erfahrungen mit dem Testeinsatz gesammelt werden, wird jetzt bei uns z. B. die Meinung vertreten, dass durch überregionale Tests keineswegs „Paukschulen" begünstigt würden, denn in solchen „Paukschulen" wären, wegen des „schlechten Lernklimas ... die fachlichen Leistungen schlechter" (Sygusch 1999, S. 184). Gerade weil das ungünstige Lernklima in einer Paukschule langfristig zu schlechteren fach-

---

simple information" eingestuft werden.

[3] Ich hätte mir bei etlichen TIMSS-Items gewünscht, die deutschen Schüler hätten schlecher abgeschnitten. Z. B. bei Item M12 könnte man ein schlechtes Abschneiden als Folge guten Physikunterrichts interpretieren (siehe hinten, Abschnitt 4).

lichen Leistungen führt, muss man vor den negativen Rückwirkungen warnen, die der überregionale Einsatz von Tests haben kann. Dies zeigen Erfahrungsberichte aus den USA und aus Japan, die ich in den Abschnitten 2 und 3 zur Diskussion stellen möchte.

Die Aussage, dass „Leistungsmessung sehr wohl einen fruchtbaren Beitrag zu Entwicklung von Schulen leisten kann“ (Bethge 1999, S. 181), ist ohne Zweifel richtig. Im Vertrauen auf die Wirksamkeit moderner Testverfahren habe ich selbst solche Verfahren zur Analyse von Lernprozessen eingesetzt (siehe z. B. Preibusch/Hagemeister 1984). Trotzdem habe ich Angst vor der „empirischen Dampfwalze“, die, ausgelöst durch TIMSS, zurzeit durch unser Land rollt, weil die „empirische Wende“ von Verwaltungen vorangetrieben wird, wo es kaum jemanden gibt, der Zeit und Muße hätte, sich über Risiken und Nebenwirkungen, die der überregionale Testeinsatz[4] mit sich bringt, hinreichend zu informieren. Dies wäre jedoch dringend nötig, denn es kann kein Zweifel daran bestehen, dass auch für Deutschland gilt, was in einer OECD-Studie aus dem Jahre 1994 für die USA festgestellt wird:

> „Bildungspolitiker müssen eine realistische Vorstellung davon entwickeln, was standardisierte Tests leisten können. Sie müssen zur Kenntnis nehmen, dass auch die besten Testverfahren mit Messungenauigkeiten behaftet sind und dass sie von begrenzter Gültigkeit und Nützlichkeit sind.“ (OECD/CERI 1994, S. 38)

Der Sorge, dass „Meinungsbildner“ in der „Bildungsgewerkschaft mit dem Nichtwahrhabenwollen von Ergebnissen aus empirischen Erhebungen“ die GEW in die „Sackgasse manövrieren könnten“ (Sygusch 1999, S. 185), muss man eine sehr gravierende Befürchtung entgegenhalten: Weil viele Bildungspolitiker und Bildungsplaner in Deutschland keine realistischen Vorstellungen von der begrenzten Gültigkeit und Nützlichkeit standardisierter Tests haben, besteht wieder einmal die Gefahr, dass in unseren Schulen mit schlecht geplanten Reformen Schaden angerichtet wird. Denn die zentrale Prüfungen, die heute überall gefordert werden, führen weder zu mehr Gerechtigkeit, noch hat der Einsatz von Papier-und-Bleistift-Tests zur Folge, dass sinnvoller gelernt wird, wie die Erfahrungen aus anderen Ländern zeigen (siehe die nachfolgenden Abschnitte 2, 3 und 4).

Überregionaler Testeinsatz und Schulaufsicht

Angestoßen durch die TIMS-Studie will man heute in Deutschland mit überregionalen Leistungsvergleichen gegen das Mittelmaß ankämpfen:

[4] Im Übrigen habe ich mich sehr wohl mit dem „Antwortverhalten der Schülerinnen und Schüler“ beschäftigt. Meine Aussagen zur „inneren Konsistenz“ des TIMSS-Testpakets basierten natürlich auf der Betrachtung des Antwortverhaltens (siehe S. 173, 174 in Heft 2 von DDS). Aber solche, aus dem Antwortverhalten abgeleiteten Reliabilitäts-Betrachtungen ersetzen eben nicht die Diskussion über die äußere Validität von Items.

- „Die schlechten Leistungen deutscher Schüler in nationalen und internationalen Tests habe die Bildungspolitiker alarmiert.“ (Süddeutsche Zeitung vom 5.2.99, S. 24 „Lehren ohne Leistungsnachweis“)
- „Deutschland gehört zu den wenigen Industriestaaten, die bisher nicht auf die Erträge der Schule geachtet haben ... Unser Bildungshaushalt ist der zweitgrößte öffentliche Haushalt. Da darf man schon fragen, ob die Mittel gut eingesetzt werden.“ (Baumert 1999, S. 24)

Der Vorwurf, bei uns würde „nicht auf die Erträge der Schule geachtet“, ist ganz eindeutig für die Bundesländer mit „Zentralabitur“ unzutreffend. Die Prüfungsaufgaben sind für alle gleich und sie werden nach ihrem Einsatz publiziert. Die Standards sind also bekannt. Sie werden öffentlich diskutiert und bei der Arbeit in der Schule sehr ernst genommen. - Herrscht dagegen in Ländern ohne Zentralabitur Schlendrian und Chaos?

> „Die Ergebnisse der TIMS-Studie haben ... das Bewusstsein geschärft, dass das Gesamtsystem (Schule) so gut wie keine objektivierbaren Daten über seine Leistungsfähigkeit besitzt, obwohl es eine teure Schulaufsicht hat, die von der entsprechenden Fiktion lebt.“ (Konzept ... 1998)

Ist es eine Fiktion, Schulaufsicht ohne zentrale Prüfungen betreiben zu wollen? Eine solche Erkenntnis lässt sich aus der TIMS-Studie keineswegs ableiten. Die Ergebnisse, die bei TIMSS-III in 12. und 13. Klassen erzielt wurden, zeigen: Staaten mit Zentralabitur schneiden im Mittel etwas schlechter ab als Staaten ohne Zentralabitur (von Saldern 1999) Ähnliche Resultate wurden auch für Deutschland ermittelt: Zwischen Bundesländern mit und ohne Zentralabitur bestehen keine bedeutsamen Leistungsunterschiede (Baumert/Bos/Watermann 1998, S. 118, 119)

In Bundesländern, die kein Zentralabitur haben, werden die von den einzelnen Lehrern entworfenen Abituraufgaben vor deren Einsatz von der Schulaufsicht genehmigt. Also auch ohne Zentralabitur wird zumindest durch die Schulaufsicht auf die „Erträge der Schule geachtet“. Dass diese Schulaufsicht bei uns versagt habe und dass zentral gestellte Aufgaben zu besseren Resultaten führen, lässt sich aus den Tests, die bei TIMSS-III zum Einsatz gekommen sind, nicht ableiten.

Hätte der überregionale Einsatz von Tests effektiveren Unterricht zur Folge, dann hätten die Schüler aus den USA bei TIMSS ganz brillante Ergebnisse erzielen müssen, denn „in den amerikanischen Schulen (werden) Tests etwa 30 bis 50-mal häufiger eingesetzt“ als an „bundesdeutschen Schulen...“ (Ingenkamp/Schreiber 1989, S. 9) Hinzuzurechnen wäre obendrein die Wirkung der US-weit eingesetzten Hochschuleingangstests. Trotzdem haben bei TIMSS-II die 7und 8-Klässler aus den USA nicht besser abgeschnitten als die deutschen Schüler, obwohl die Items bei TIMSS-II fast ausschließlich auf Unterricht in Nordamerika zugeschnitten waren (siehe Hagemeister 1999). Bei TIMSS-III schließlich haben die US-

Schüler deutlich schlechtere Resultate erzielt als die deutschen Schüler aus 12. und 13. Klassen (Baumert/Bos/Watermann 1998, S. 75 bis 79 und 84 bis 87).

Weder aus TIMSS-II noch aus TIMSS-III lässt sich die Erkenntnis ableiten, dass der überregionale Einsatz von Tests eine Verbesserung des Schulsystems zur Folge hat. Im Einklang hiermit stehen Aussagen in der OECD-Studie über das Bildungswesen in den USA:

> „Die permanent unbefriedigenden Resultate des US-Schulsystems haben zur Entzauberung der Methode ... der Messung von Leistung durch standardisierte Tests geführt." (OECD/CERI 1994, S. 130)
>
> „Standardisierte Tests messen in der Regel isolierte Fertigkeiten, ohne Bezug zu dem Unterricht, den die Schüler hatten ..." (OECD/CERI 1994, S. 128)
>
> „Es kann kaum bezweifelt werden, dass sorgfältige Inspektionen ... eine wichtige Rolle bei der Weiterentwicklung einer Schule spielen ... Viele Aspekte, die sich nicht direkt in messbarer Leistung widerspiegeln, können sehr wichtig sein, um sich ein Bild von einer Schule zu machen ... die Wahrnehmung eines breiten Bandes von Indikatoren kann sehr wertvoll bei einer Schulbesichtigung sein." (OECD/CERI 1994, S. 38)

Während es zurzeit in Deutschland als ein Zeichen von Rückständigkeit angesehen wird, dass es bei uns noch keine überregional einsetzbaren Testbatterien gibt, steht man in den USA nach Jahrzehnten intensiven Testeinsatzes vor der Erkenntnis, dass die Schulen nicht besser geworden sind und es wird auf die Vorteile der individuellen Schulaufsicht und der informellen Schul-Inspektion hingewiesen.

Nun könnte man ja sagen, die Schulen in den USA sind in den vergangenen Jahrzehnten deshalb nicht besser geworden, weil die üblicherweise eingesetzten Testbatterien ein zu niedriges Niveau hatten. Es müsste sich doch gerade durch den Einsatz anspruchsvoller Testverfahren das Niveau der Schulen systematisch anheben lassen. Entsprechend wird in Deutschland zurzeit die Hoffnung geäußert, mit modernen Tests könnte wieder eine „Kultur der Anstrengung" in unseren Schulen einziehen. Ehe wir uns jedoch daran machen, möglichst schwierige Testaufgaben zur Niveausteigerung einzusetzen, ist es wiederum hilfreich, Entwicklungen in test-geprüften Ländern zu betrachten: In etlichen Staaten werden sehr anspruchsvolle Tests bei Aufnahmeprüfungen zu angesehenen Gymnasien oder Universitäten eingesetzt. Dies hat zur Folge, dass Kinder wohlhabender Eltern neben der staatlichen Schule private Lehranstalten besuchen, wo das Ausfüllen von Testbögen trainiert wird. In Japan spielen die privaten Paukschulen, die Jukus, eine derart dominierende Rolle im Leben von immer mehr Schulkindern, dass wiederholt durch bildungspolitische Maßnahmen versucht wurde, den Einfluss der Jukus und der Test-Industrie zurückzudrängen: So „verbot die japanische Erziehungsbehörde im April 1993 ... die Abhaltung der von der Erziehungsindustrie" entwickelten Tests „in den öffentlichen Mittelschulen." (Ito 1997, S. 458) Was ist die Folge? – Die Mittelschüler

bereiten sich nun zunehmend „auch am Wochenende ... auf die Aufnahmeprüfungen derjenigen Gymnasien“ vor, „die ihrem Rang entsprechen.“ (Ito 1997, S. 458) Verschiedene Maßnahmen, die den Leistungsdruck reduzieren sollten, der auf Japans Schülern lasten, brachten nicht den gewünschten Effekt. Nicht nur am Wochenende, auch in den Sommerferien besuchen viele japanische Schüler inzwischen private Paukschulen (Schümer 1998).

> *In den privaten Jukus wird „geübt, und wiederholt und auswendiggelernt, und zwar mit einer Ausdauer und Intensität, die ... deutsche Kinder außerordentlich befremden würde.“ (Schümer 1998, S. 215)*
>
> *„Das japanische Bildungssystem besteht aus der »Fassade« eines ministerial verordneten Harmonieprinzips und der »wirklichen Absicht« des der selektiven Gesellschaft dienenden Konkurrenzprinzips.“ (Ito 1997, S. 449)*
>
> *„Japan leidet deshalb seit langem unter der Unmöglichkeit einer tiefergehenden Reform.“ (Ito 1997, S. 455)*

Es ist nahezu unvermeidbar, dass mit dem Einsatz standardisierter Tests Faktenwissen und andere isolierte Fertigkeiten in den Vordergrund rücken. Je größer die Bedeutung der Tests ist (etwa für den weiteren Bildungsweg), umso eher werden geistige Selbständigkeit und Kreativität an Gewicht verlieren, weil sie sich in den Testergebnissen in der Regel nicht widerspiegeln. Nun könnte man erwidern, dass man auch im Multiple-Choice-Format „anspruchsvolle Aufgaben entwickeln (kann), mit denen selbstständiges Denken ... oder Problemverständnis erfasst werden.“ (Baumert/Köller 1998, S. 15) Dies setzt allerdings voraus, dass die Aufgaben sorgfältig entwickelt und erprobt werden. Die Rahmenbedingungen für die Testentwicklung sind jedoch denkbar ungünstig, wenn die Tests überregional eingesetzt werden und wenn ihre Ergebnisse für Schulen oder Schüler Rechtsfolgen haben:

Aufgaben zu entwickeln, die allen gerecht werden, ist, wenn nicht unmöglich, so doch sehr schwierig, denn die Unterrichtsinhalte und -methoden unterscheiden sich von Lehrer zu Lehrer, von Land zu Land, von Stadt zu Stadt. Der Item-Entwicklung müsste eine breite Analyse von Methoden und Inhalten vorausgehen. Kommunikation mit allen betroffenen Schulen wäre notwendig. Genau dies ist aber nicht möglich, wenn die überregional eingesetzten Tests Rechtsfolgen haben: Weil Jahr für Jahr neue Tests produziert werden müssen, die möglichst niemand kennen darf, fehlt es an Zeit und an Kommunikation für eine sorgfältige Item-Entwicklung.[5]

[5] Dieser Mangel an Kommunikation wird nicht dadurch aus der Welt geschafft, dass die allgemeinen Ziele, die der Item-Konstruktion zu Grunde gelegen haben, vor dem Testeinsatz veröffentlicht werden (wie es in DDS, Heft 2, 1999 auf den Seiten 136 und 137 bei Schweitzer beschrieben wird). Erst wenn man alle Items kennt, kann man beurteilen, was die Test-Autoren unter „grundlegenden Kompetenzen“ oder unter „fundamentalen Prinzipien der Physik“ verstanden haben. Wenn die Testautoren z. B. meinen, mit ihrem Test „die Konzepte Spannung und Stromstärke“ zu erfassen, dazu dann aber lediglich das TIMSS-Item M12 anbieten, dann wird deutlich, dass es unverzichtbar gewesen wäre, über das gesamte Science-Testpaket eine breite öffentliche Diskussion zu führen, bevor es bei TIMSS zum Einsatz kommt (siehe auch weiter hinten, den Abschnitt 4).

Diese kritischen Betrachtungen betreffen vor allem den überregionalen Einsatz von Testbatterien, durch die Schüler, Lehrer, Schulen oder Staaten in irgendwelche Rangfolgen gebracht werden sollen. Ganz anders ist der Einsatz standardisierter Testverfahren zu bewerten, wenn es um Begleituntersuchungen zu einzelnen Reformvorhaben geht. Hier können Testverfahren auf ein bestimmtes Curriculum, auf Unterrichtsmethoden und auf die Schülerpopulation zugeschnitten werden (siehe z. B Preibusch 1984). Solche Tests können sehr wertvolle Teilinformationen liefern, wobei Leistungsdaten nie isoliert betrachtet und zum absoluten Standard gemacht werden dürfen. Dass Schüler mehr Interesse und Freude an einem Fach haben, kann langfristig gesehen wichtiger sein, als ein kurzfristig messbarer Zuwachs an kognitivem Wissen. – Wissenschaftliche Begleituntersuchungen, die sorgfältig geplant und durchgeführt werden, gibt es in Deutschland kaum noch. Mit Hinweis auf die Kosten wird auf wissenschaftliche Begleitung meist verzichtet. Stattdessen werden Schulreformen an modischen Trends ausgerichtet und nach Gutdünken bewertet. Hier besteht in der Tat ein großer Mangel an „objektivierbaren Daten". Am Anfang einer „empirischen Wende" sollten in Deutschland kleine, überschaubare wissenschaftliche Begleituntersuchungen stehen. Hier könnten wir lernen, „Testkultur" zu entwickeln.

Überregionaler Testeinsatz und Chancengleichheit

Durch eine „zentrale Prüfung in den Fächern Deutsch, Mathematik und l. Fremdsprache zum Abschluss der 10. Jahrgangstufe" soll nun „erreicht werden, dass

- der mittlere Bildungsabschluss unabhängig von der Schulart ... in Kernbereichen vergleichbare Aussagen über den Leistungsstand vermittelt und
- der Übergang in die gymnasiale Oberstufe unter Berücksichtigung der Leistungsaussagen ... geregelt wird, was zwischen Gesamtschule, Realschule und Gymnasium eine fairere Konkurrenzsituation bzw. Chancengleichheit herstellt." (Konzept ... 1998)

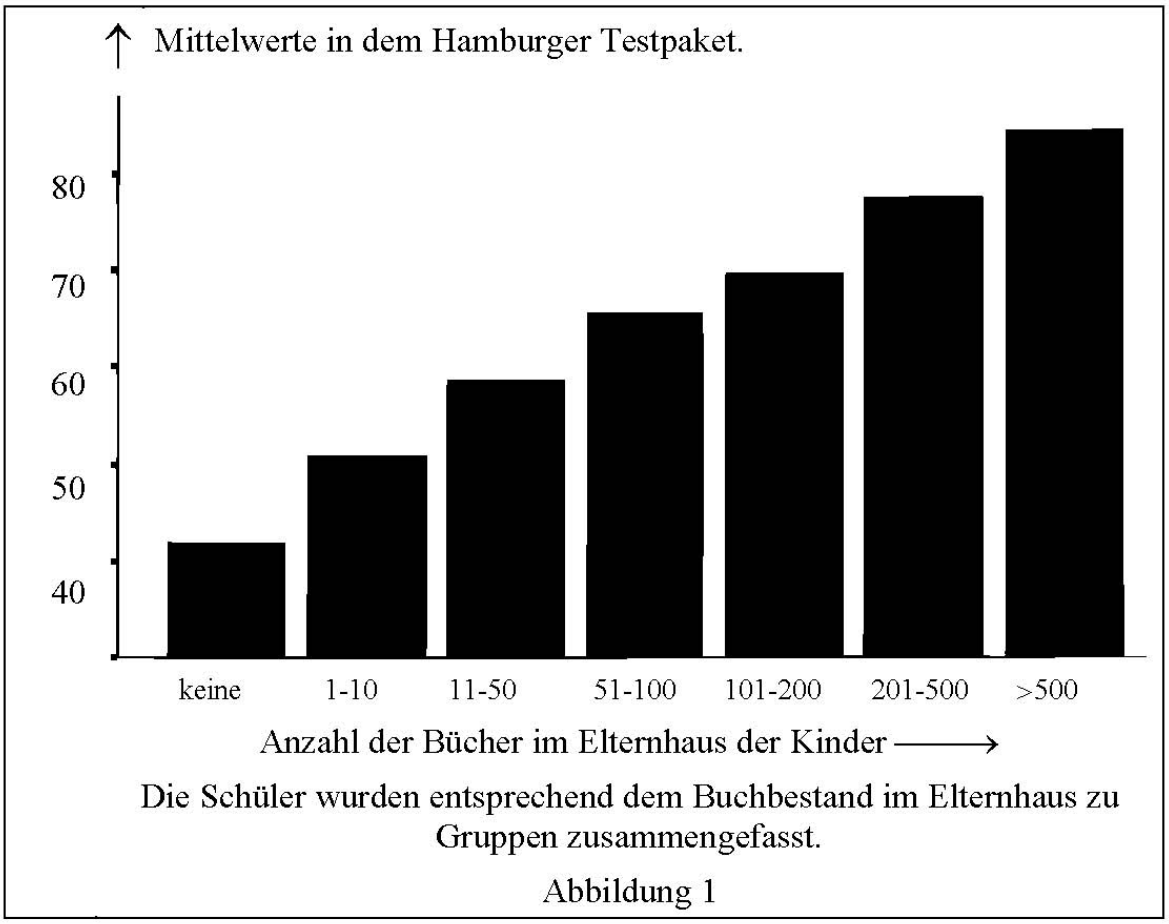
Die Erfahrungen aus Ländern, wo seit Jahrzehnten Selektion mit Hilfe standardisierter Tests betrieben wird, zeigen, dass Kinder wohlhabender Eltern in immer größerer Zahl am Nachmittag private Paukschulen besuchen, wo das Ausfüllen von Testbögen trainiert wird (siehe hier Abschnitt 2). Durch den überregionalen Einsatz von Tests wird also nicht Chancengleichheit realisiert, sondern es werden vorgegebene soziale Ungleichheiten stabilisiert.

Wo immer in den vergangenen Jahrzehnten Leistungen von Schülern mit sozialen Daten zusammen erhoben worden sind, hat sich gezeigt, dass das soziale Umfeld einer Schule großen Einfluss auf die Testergebnisse von Schülern hat. So hat die Untersuchung der „Lernausgangslage ... an Hamburger Schulen", die im Jahre 1996 in allen fünften Klassen zu Beginn des Schuljahres durchgeführt wurde, gezeigt, ...

„dass sich der Buchbestand im Elternhaus ... als einer der besten Indikatoren für den erreichten Lernstand erweist ... Schülerinnen und Schüler, in deren Elternhäusern (praktisch) keine Bücher vorhanden sind, erreichen im Durchschnitt weniger als 42 Rohpunkte“ in dem Hamburger Testpaket,

„solche, deren Eltern 500 Bücher oder mehr besitzen, erzielen mit durchschnittlichen 84,3 Rohpunkten mehr als doppelt so viele richtige Lösungen“ (Lehmann u.a. 1997, S. 68)

Abbildung 1: Testleitung und Buchbestand im Elternhaus; Mittelwerte; die Schüler wurden



entsprechend dem Buchbestand im Elternhaus zu Gruppen zusammengefasst (aus: Lehmann u.a. 1997, S. 59):

Mit Ranking-Listen werden Schulen, die in sozialen Brennpunkten liegen, systematisch benachteiligt. Die Lehrer, die an solchen Schulen möglicherweise sehr viel Kraft und Zeit in die intellektuelle und emotionale Entwicklung ihrer Schüler investieren, werden als untalentiert oder unengagiert abqualifiziert, weil die von ihnen unterrichteten Schüler im Mittel in einem zentralen Test nicht so gut abschneiden wie Kinder, die in einer gutbürgerlichen Gegend wohnen.

Durch zentrale Prüfungen wird also keine fairere Konkurrenzsituation zwischen Gesamtschule, Realschule und Gymnasium herstellt, weil die Kinder, die Gesamtschulen besuchen, in der Regel schlechtere Voraussetzungen von zu Hause her mitbringen als die Mehrzahl der Gymnasiasten. Wie bedeutsam die unterschiedlichen Startbedingungen je nach Elternhaus sind, zeigt Abbildung 1 in eindrucksvoller Weise.

Nun könnte man ja anstelle absoluter Testergebnisse die „Lernfortschritte", die von Jahr zu Jahr erzielt werden, als Maß für die Qualität einer Schule einführen. Dem stehen jedoch kaum lösbare methodische Schwierigkeiten entgegen, wie in der OECD-Studie über die Test-Realität in den USA festgestellt wird: Die hohe Fluktuation in der Schülerschaft macht es gerade in sozialen Brennpunkten unmöglich, Lernfortschritte von Jahr zu Jahr mit der benötigten Genauigkeit zu messen (OECD/CERI 1994, S. 59). Außerdem gibt es für Lernfortschritte kein gesichertes Maß. Wie stark sich Testergebnisse von einem Schuljahr zum anderen unterscheiden, hängt entscheidend von ab, wie gut die Items dem tatsächlich behandelten Unterrichtsstoff entsprechen.

Auch wenn es gelingen sollte, Lernfortschritte in allen Schulen hinreichend genau zu messen, so ist trotzdem nicht damit zu rechnen, dass Schulvergleiche „fairer" werden, wenn man anstelle des Leistungs-Endstandes die Lernfortschritte für Ranking-Listen benutzt. Denn auch die Lernfortschritte sind bei Kindern, die aus stabilen und behütenden Verhältnissen kommen, größer als bei unterpriviligierten Kindern. Die „Schülerinnen und Schüler in den Gymnasialklassen" haben von der 5. bis zur 7. Klasse „ihren Vorsprung erheblich ausbauen können", wie sich bei der Messung „der Lernausgangslagen und der Lernentwicklung" in Hamburg gezeigt hat (Lehmann u.a. 1998, S. 53). Dies ist nicht überraschend, wenn man bedenkt, dass man die fürsorgliche Zuwendung und Hilfe, die Kinder in bildungsorientierten Elternhäusern in der Regel erfahren, durch schulische Maßnahmen nur begrenzt ersetzen kann.

Es ist nicht zu erwarten, dass durch den überregionalen Testeinsatz Mechanismen ausgelöst werden, die die Leistungsunterschiede zwischen Kindern aus bildungsfernen und bildungsnahen Elternhäusern kleiner werden lassen. Im Gegenteil: Wenn erst einmal regelmäßig getestet wird, dann werden die Kinder aus besser gestellten Familien gezielt auf die Tests vorbereitet (wie in Japan, wo immer mehr Kinder neben der Schule private „Jukus" besuchen, um den Aufnahme-Test an einem angesehenen Gymnasien zu bestehen, siehe z. B. Schümer 1998, Ito 1997).

Um eine faire Konkurrenzsituation herzustellen, könnte man nun daran denken, Testmittelwerten immer nur zusammen mit der Bewertung des sozialen Umfeldes einer Schule zu veröffentlichen. Mit einem solchen sozialen Bonus werden jedoch Schulen, die ein ungünstiges Umfeld haben, im Konkurrenzkampf um Eltern keineswegs besser gestellt:

„Berücksichtigt man den sozialen Hintergrund der Kinder, dann wird damit ein ungünstiges Signal an ... die Eltern gegeben“, die für ihre Kinder eine Schule suchen, wo nicht relativ gute, sondern absolut hohe Testwerte erzielt werden (OECD/CERI 1994, S. 25)

Neben familiären Einflüssen haben natürlich auch die Fähigkeiten der einzelnen Lehrer erhebliche Bedeutung für die Testergebnisse von Schülern. Ob ein Schüler zu Beginn der

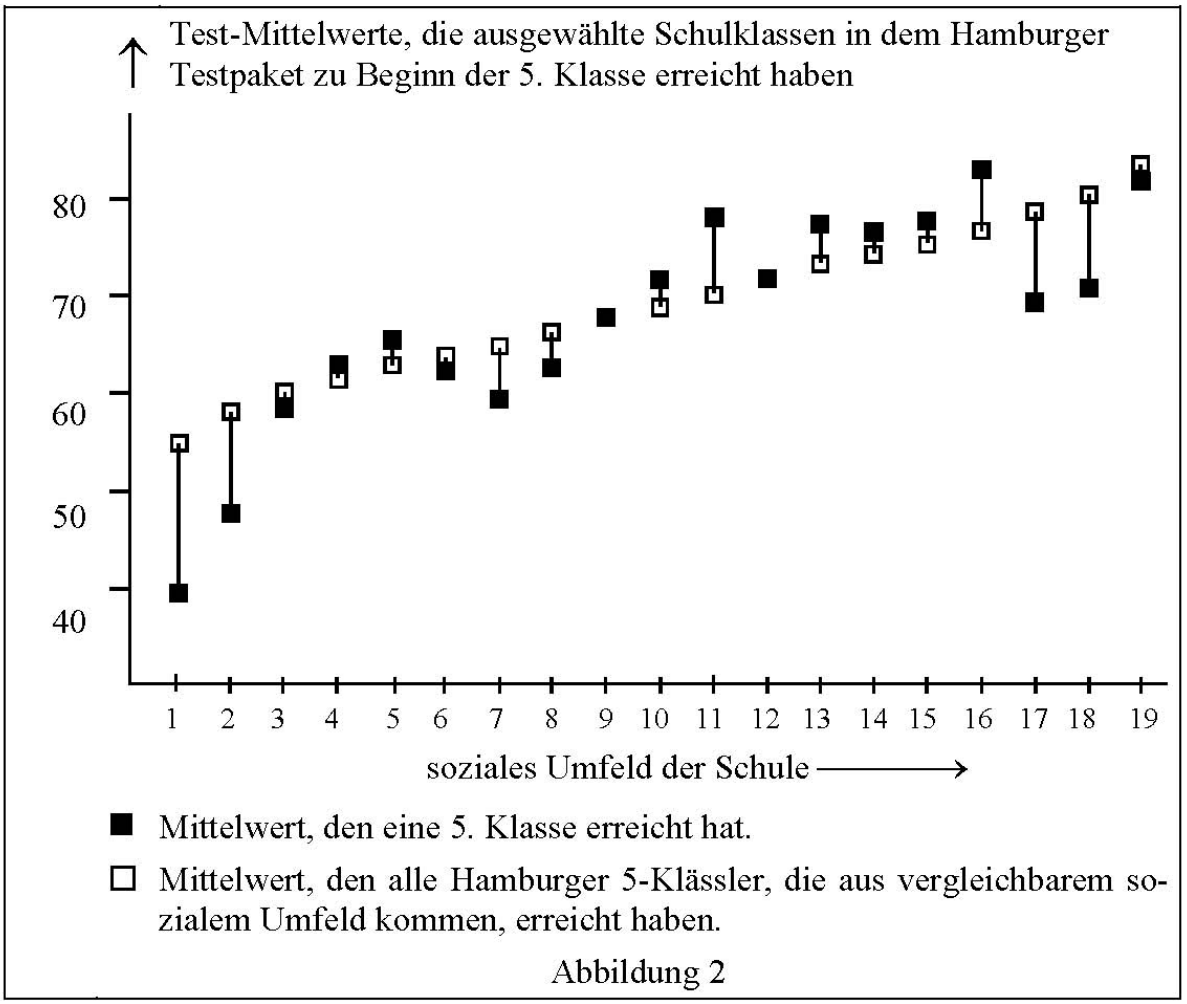


Abbildung 2

fünften Klasse im Hamburger Testpaket 40 oder 55 Punkte erreicht, hängt unter Umständen nur davon ab, welche Lehrer er in der Grundschule hatte (siehe Abbildung 2):

Abbildung 2: Test-Mittelwerte, die ausgewählte Grundschulklassen in dem Hamburger Testpaket zu Beginn der 5. Klasse erreicht haben (aus Lehmann u.a. 1997, S. 59):

Wer alle Schüler eines Bundeslandes die gleichen Testaufgaben bearbeiten lässt, benachteiligt diejenigen Schüler, die in einem ungünstigen sozialen Umfeld aufgewachsen sind und die nicht so talentierte Lehrer hatten. Zentrale Prüfungen bringen nicht mehr, sondern weniger Gerechtigkeit.

Nun könnte man sagen, nicht wegen eines Mehr an Gerechtigkeit, sondern aus Gründen der Staatsräson brauchen wir zentrale Prüfungen: Man müsse den Leistungsstand aller Abiturienten objektiv ermitteln, damit nicht Menschen mit objektiv schlechten Leistungen unsere Hochschulen bevölkern. Für diese Position liefert Abbildung 2 eine wichtige Erkenntnis, aus der scheinbar zwingend folgt, dass auf überregionale, objektive Leistungsstandards gar nicht verzichtet werden kann. Das Problem ist nämlich, dass sich die Verteilung der Schulnoten in allen bei der Hamburger Studie erfassten Klassen kaum unterscheidet. In Schulklassen, die im Hamburger Test sehr schlecht abgeschnitten haben, gibt es die Noten Zwei, Drei oder Vier genauso häufig wie in den test-besten Klassen. Das heißt aber auch, für Leistungen, die in der einen Klasse mit „sehr gut" bewertet werden, erhält man in einer anderen Klasse lediglich ein „Mangelhaft". Folgt daraus nicht unwiderlegbar, dass die Noten, die Lehrer geben, willkürlich und unbrauchbar sind? Erstaunlicherweise haben jedoch die von Lehrern am Ende der Schulzeit erteilten Noten in der Regel einen besseren Prognosewert für den Studienerfolg als Studieneingangstests.

Schulnoten erweisen sich immer wieder als recht zuverlässiges Mittel, um den späteren Erfolg in akademischen Prüfungen vorherzusagen. „Weltweit gilt die Durchschnittsnote im Abschlusszeugnis der Sekundarstufe als das beste einzelne Merkmal zur Vorhersage des Studienerfolgs" (Trost 1989, S. 79, siehe hierzu auch Schumann/Claus 1970, S. 17). Der mit Hilfe eines Tests gemessene momentane Kenntnisstand eines Menschen ist für den Erfolg im nachfolgenden Studium von relativ geringer Bedeutung. Sehr viel wichtiger sind „nichtintellektuelle" Persönlichkeitsmerkmale, wie der „Arbeitsstil" oder das „Studienengagement" (Fisseni, 1993). Deshalb sind die Noten, die relativ zu den Arbeitsbedingungen in einer Schulklasse erteilt werden, für ein Hochschulstudium relevanter als überregionale, objektive Standards.[6] Hier bestätigt sich, was auch in der OECD-Studie von 1994 festgestellt wird: Es darf nicht übersehen werden, dass standardisierte und unter normierten Bedingungen durchgeführte Tests mit Messungenauigkeiten behaftet sind und dass solche Tests stets nur einen kleinen Ausschnitt der Fähigkeiten eines Menschen erfassen. (OECD/CERI 1994, S. 38)

„Forschungsergebnisse deuten daraufhin, dass die Papier-und-Bleistift-Methode", die in Tests aus Kostengründen überall angewendet wird, mit der „direkten Beobachtung" von Schülern beim Experimentieren höchstens 10% an gemeinsamer Varianz erfasst (vgl. Shavelson/Ruiz-Rrimo 1999). In den Ergebnissen von Papier-und-Bleistift-Tests spiegeln sich experimentelle Fertigkeiten nur in geringem Maße wider. - Daraus folgt, dass mit zentral vorgegebenen Prüfungs-Standards das Experimentieren im naturwissenschaftlichen Unterricht an Bedeutung

[6] Es ist allgemein üblich, Testbatterien durch Korrelation mit Schulnoten zu validieren. Ein international anerkanntes Gütekriterium für einen Test ist, wenn er die Schüler nahezu in die gleiche Rangfolge bringt, wie es die Noten der Lehrer tun. - Ist umgekehrt die Korrelation zwischen Test und Lehrernoten klein, dann wird gerne behauptet, die Lehrernoten seien mangelhaft. Tatsächlich ist es jedoch wahrscheinlicher, dass das Testinstrument Mängel hat. Damit soll jedoch keineswegs gesagt werden, dass Lehrernoten stets gerecht und angemessen sind. Die Beurteilung der Fertigkeiten und Fähigkeiten eines Menschen ist immer mit Fehlern behaftet. Die exakte und gerechte Leistungsmessung gibt es nicht.

verliert, wohingegen Fertigkeiten, die sich leicht testen lassen, immer mehr Gewicht im Schulalltag erhalten. Auch wenn Zeit und Mühe dafür aufgewendet würden, anspruchsvolle Items zu entwickeln, so wird man doch mit der Papier-und-Bleistift-Methode experimentelle oder künstlerischkreative Fähigkeiten kaum erfassen.

Über die Schwierigkeiten, gute Testaufgaben zu konstruieren

Es wird vielfach unterschätzt, wie schwierig es ist, valide Aufgaben für einen überregionalen Testeinsatz zu entwickeln. Wenn durch solche Tests Schüler, Lehrer, Schulen oder Länder in irgendwelche Rangfolgen gebracht werden sollen, dann ist die Testentwicklung mit schweren Handikaps belastet: Einerseits wäre, um die unterschiedlichen Lehrmethoden und Unterrichtsinhalte hinreichend zu berücksichtigen, für die Item-Entwicklung ein vielfältiger Informationsaustauch, der sich kurzfristig nicht realisieren lässt, unverzichtbar, andererseits stehen überregionale Test-Projekte immer aus irgendwelchen Gründen unter Zeitdruck. Der Informationsaustausch über das Wesentliche, über die Items, muss sogar verhindert werden, damit es beim Test gerecht zugeht (siehe vorne, Abschnitt 2). Auch die Testentwicklung zur TIMS-Studie hat offenbar unter einem erheblichem Mangel an Zeit und an Kommunikation gelitten. In Heft 2/99 dieser Zeitschrift hatte ich gezeigt (Hagemeister 1999), dass es um die Validität von TIMSS-II-Items sehr schlecht bestellt ist. Die Mängel, die ich hier an einzelnen Items dargestellt habe, sind symptomatisch für das Science-Testpaket. Wie ein roter Faden ziehen sich folgende Konstruktionsmängel durch das Science-Testpaket bei TIMSS:

- Man ist laufend um Lebensbezug bemüht, aber dem Leben werden simplifizierende und linearisierende Schemata übergestülpt.
- Reflexionen über die begrenzte Gültigkeit von Gesetzen haben sich in den TIMSS-Items nicht niedergeschlagen.
- Sofern physikalische Experimente in Items angesprochen werden, so ist unübersehbar, dass die Testkonstrukteure diese Experimente nicht selbst durchgeführt haben können.[7] Die Physik tritt uns hier in einer Papier-und-Bleistift-Welt entgegen.

Der Unterricht ist zum Spiegelbild der Testrealität geworden. Weil in den massenhaft eingesetzten Multiple-Choice-Tests reale Experimente keinen Platz haben, sind sie auch im Physikunterricht bedeutungslos geworden.

Wer im naturwissenschaftlichen Unterricht nicht experimentiert, läuft Gefahr, Regeln und Gesetze unzulässig zu verallgemeinern. Dies lässt sich an dem TIMSS-Item M12 ablesen: Das Item sollte als richtig gelöst gelten, wenn in eine Tabelle der Wert „40“ eingetragen

[7] Fachlich korrekt sind einige TIMSS-Items zur Reflexion an Spiegeln, was aber nicht heißen muss, dass hierzu experimentiert wurde, denn die Reflexion an Spiegeln lässt sich ja sehr leicht rein zeichnerisch abhandeln (siehe hierzu auch Hagemeister 1999).

worden war (Hagemeister 1999, Seite 161). Die Testautoren waren offenbar der Meinung, dass bei Glühlampen das „Ohmsche Gesetz“ gilt, d.h. zwischen Strom und Spannung ein linearer Zusammenhang besteht. Tatsächlich gibt es jedoch keine Glühlampen, für die man Messwerte erhält, wie sie in der Tabelle bei TIMSS-Item M12 stehen. Das TIMSS-Item M12 ist also aus fachlicher Sicht eindeutig falsch, denn anstelle von 40 werden hier im Experiment allenfalls 20 bis 25 mA ermittelt.[8] Diese Feststellung muss nicht deshalb korrigiert werden, weil bei Item M12 die „Lösungshäufigkeit ... in Deutschland deutlich über ... dem internationalen Mittelwert“ gelegen hat (Bethge 1999, S. 179). Ob für Glühlampen das „Ohmsche Gesetz“ gilt, wird ja nicht durch Umfragen unter Schülern aus 8. Klassen geklärt, sondern durch geeignete Experimente. – Das Item M12 führt uns vor Augen, dass man bei der Test-Entwicklung nicht auf das mühsame Ringen um fachliche Richtigkeit und curriculare Validität verzichten kann, weil die spätere Analyse der Lösungshäufigkeiten schon zeigen wird, ob ein Item sinnvoll oder unbrauchbar war.

Es soll hier nicht etwa die mathematische Analyse der Lösungshäufigkeiten als unwichtig oder uninteressant hingestellt werden. Z. B. werden mit der Parallel- oder Retest-Reliabilität wichtige Erkenntnisse über die Qualität einer Testbatterie gewonnen. Ebenso kann man zur Bewertung einzelner Aufgaben nicht auf die Berechnung von Schwierigkeitsindizes und Trennschärfe-Koeffizienten verzichten. Wenn die ansonsten besseren Testteilnehmer ein bestimmtes Item schlechter lösen als die schwächeren Testteilnehmer, dann muss ein solches Item umformuliert oder herausgenommen werden. Ein negativer Trennschärfewert zeigt, dass ein Item nicht in Ordnung ist. Umgekehrt ist jedoch ein positiver Trennschärfewert kein sicheres Indiz dafür, dass ein Item gut ist. Beispiele dafür sind das Stromkreis-Items M12 oder das Raumschiff-Item L7. (Hagemeister 1999, S. 161-168) Obwohl beide Items aus Sicht der Physik eindeutig falsch sind, werden hier, wie bei den meisten anderen Items des Science-Testpaketes, Lesefertigkeit und vor allem Textverständnis benötigt, um die Lösung zu finden, die die Aufgabenkonstrukteure für die richtige halten. Ohne Zweifel kann man das Science-Testpaket als homogen bezeichnen, weil in erster Linie Textverständnis getestet wird (Hagemeister 1999, Abschnitt 3.2). Insofern ist es nicht überraschend, dass die mathematische Analyse der Testergebnisse bei TIMSS zu guten Werten für Schwierigkeit, Trennschärfe oder Reliabilität geführt hat. Solche homogenen Testpakete besitzen allerdings in der Regel „eine relativ geringe praktische Validität.“ (Lienert 1994 S. 255)

Warum das Science-Testpaket von TIMSS, trotz guter Trennschärfewerte, von geringer curricularer Validität ist, zeigt wiederum das Item M12: Bei diesem Item soll sich zeigen, ob „die Konzepte Spannung und Stromstärke verfügbar (sind) und ... richtig zueinander in Beziehung gesetzt werden“ können (Baumert/Lehmann u.a. 1997, S. 84). Meine Testschüler, die ich

[8] Bei einer Spannung von 6 Volt haben wir Werte von 20 bis 25 mA für 25- und 40-Watt-Haushaltslampen gemessen (Hagemeister 1999, Seite 162). - Diese Abweichungen von einer linearen Strom-Spannungs-Kennlinie sind bei Haushaltslampen (für 230 Volt) noch relativ gering. Sehr viel stärker werden die Abweichungen von der Linearität, wenn man Strom und Spannung bei Halogenlampen oder bei Lämpchen für die Fahrradbeleuchtung misst.

nach jeder Aufgabe kurz interviewt hatte, konnten alle das Item M12 lösen, obwohl ihnen die physikalische Größe Spannung noch unbekannt war. (Hagemeister 1999, S. 160-162) Meine Testkandidaten haben sich einfach nur mit der Tabelle befasst und nach kurzem Nachdenken an der freien Stelle eine „40“ eingetragen. Ein erheblicher Mangel bei Item M12 besteht also darin, dass man die erwartete „40“ in die Tabelle eintragen kann, ohne irgendetwas über Strom und Spannung zu wissen.

Aus Sicht eines 8-Klässlers ist es ja durchaus sinnvoll, in die Tabelle bei Item M12 die „40“ einzutragen. Die wenigsten 8-Klässler haben schon so viel im Physikunterricht experimentieren können, dass sie aus eigener Anschauung genau wüssten, dass „40“ hier nicht stehen darf.[9]

Diese Wenigen dürften jedoch im statistischen Rauschen der Testanalyse untergegangen sein. Insofern wird also die Item-Statistik ergeben, dass die test-besten Schüler mit größerer Wahrscheinlichkeit die „40“ wählen als schwache Schüler (die im Lesen oder im logischen Schlussfolgern weniger gut sind). Der Eintrag „40“ ist jedoch kein Indiz für guten Physikunterricht. Im Gegenteil: Wenn relativ viele deutsche Schüler „40“ in die Tabelle bei Item M12 eingetragen haben, so zeigt dies, dass bei uns im Physikunterricht zu wenig experimentiert wird.

Das TIMSS-Item M12 ist nicht nur aus fachlicher Sicht und aus test-technischen Gründen abzulehnen. Vor allem sprechen pädagogische Argumente gegen dieses Item: Hier wird ein linearisierendes Schema unzulässig verallgemeinert, obwohl gerade am Beispiel der Glühlampe die wichtige Erkenntnis vermittelt werden könnte, dass physikalische Gesetze in der Regel nicht unbegrenzt gültig sind. - Das „Konzept von Strom und Spannung“, das der Konstruktion von Item M12 zugrunde gelegen hat, ist ein ungeeignetes, nicht tragfähiges Konzept. Hier wird Linearität zum Prinzip erhoben, obwohl es in der Natur nur selten linear zugeht.

Wenn bei uns in Zukunft Tests überregional zum Einsatz kommen, dann werden sich früher oder später viele Lehrer im Unterricht am Inhalt dieser Tests ausrichten, weil sie ja wollen, dass ihre Schüler im Test erfolgreich sind. Falls dann Physik-Items von TIMSS-Typ eingesetzt werden, s3o wird dies zur Folge haben, dass bei uns noch seltener experimentiert wird und dass Regeln und Gesetze nicht gemeinsam erarbeitet, sondern überwiegend mitgeteilt

[9] Ein Schüler, der im Physikunterricht mit Glühlampen experimentieren konnte, könnte z. B. darüber nachdenken, ob in das freie Feld bei Item M12 eventuell eine „Null“ eingetragen werden soll, weil es sich vielleicht um ein 4,5-Volt-Lämpchen handelt, das bei 6-Volt durchbrennt. Auch wenn er schließlich die Zahl 40 in das freie Feld einträgt, weil es sich nach einigem Nachdenken gesagt hat, dass bei Item M12 wohl irgendein exotischer Lampentyp eingesetzt wurde, für den im Bereich 1,5 bis 6 Volt das Ohmsche Gesetz gilt, so war er doch wegen seiner konkreten experimentellen Erfahrungen im Nachteil, denn er hat sich unnötig lange mit dem Item M12 beschäftigt.

werden. Auch Reflexionen über die begrenzte Gültigkeit von Regeln und Gesetzen finden dann nicht mehr statt, weil dafür die experimentelle Basis fehlt. - Hier wird deutlich, wie wichtig es ist, dass die Items, die in überregionalen Studien eingesetzt werden, aus fachlicher und aus pädagogischer Sicht von hoher Qualität sind. Mit meiner Kritik an den TIMSS-Items habe ich also keineswegs „die Notwendigkeit der Veränderung des Unterrichts ... aus dem Blick." (Bethge 1999, S. 178) verloren. Es geht vielmehr darum, zu verhindern, dass wieder einmal Veränderungen in Gang gesetzt werden, die in die falsche Richtung führen.

Die Items, die bei TIMSS-II zum Einsatz gekommen sind, sind in Nordamerika entwickelt worden. Die Mangelhaftigkeit der naturwissenschaftlichen Items führt uns auch vor Augen, dass der massenhafte Einsatz von Tests nicht notwendig zu „Testkultur" führt: Angesichts der vielen Tests, die laufend entwickelt werden, fehlt es an Zeit und an Kraft, um auch noch über die Qualität von Items zu diskutieren. - Falls trotzdem einmal festgestellt wird, dass ein bestimmter Test nicht gut ist, dann können Bildungspolitiker nur zwischen mehreren Übeln wählen, wie in dem OECD-Bericht über den Testeinsatz in den USA festgestellt wird:

> „Man prüft die Schulen weiterhin mit dem ... diskreditierten Testinstrument"
>
> oder
>
> „man benutzt neue, nicht erprobte Messverfahren." (OECD/CERI, Paris 1994, S. 129)

Resümee: Mögliche Wege zur „Testkultur"

Es besteht zurzeit in Deutschland ein großer Mangel an wissenschaftlichen Begleituntersuchungen zu Schulreformprojekten. Solche Begleitforschung könnte eine gute Basis für die Entwicklung von „Testkultur" werden. Dazu müssten sich Teams aus Sozial- und Fachwissenschaftlern bilden, die über Jahre hinweg Erfahrungen in der Itementwicklung und -erprobung sammeln. Erst wenn uns Wissenschaftler-Teams zur Verfügung stehen, die bei überschaubaren Reformprojekten Erfahrungen mit begleitenden Tests gewonnen haben, sollte man sich an die schwierige Aufgabe heranmachen, Items für den überregionalen Testeinsatz zu entwickeln. Überregional einsetzbare Tests, die von hinreichender curricularer Validität sind, können als Diagnose-Instrument von erheblichem Nutzen sein. Mit Hilfe solcher Tests können einzelne Lehrer Rückmeldungen darüber erhalten, wo die von ihnen unterrichteten Schüler Defiziten haben. Die Resultate einer solchen Diagnose sollten dem Datenschutz unterliegen und nur mit Zustimmung des betroffenen Lehrers weitergeben werden dürfen. Das von einer englische Universität entwickelte Testprojekt „CEM" könnte hier als Vorbild dienen. Obwohl bei CEM weder Lehrer noch Schulen an den Pranger gestellt werden, weil keine Ranking-Listen in der Presse erscheinen, hat man festgestellt, dass sich die Diagnoseberichte, die dem einzelnen Lehrer vom CEM-Team geliefert werden, positiv auf Lernergebnisse und Unterrichtsklima auswirken. (Arnold 1999, S. 230)

Zentrale Prüfungen oder Tests sollten nicht dazu verwendet werden, Schüler, Lehrer oder Schulen in Rangreihen zu bringen. Durch ein solches Schüler- oder Schulenranking werden soziale Benachteiligungen verstärkt. Auch wenn Lernfortschritte gemessen werden, werden Schüler in sozial schwachen Regionen in der Regel benachteiligt, denn die Lernfortschritte sind bei Kindern aus gutem Hause im Mittel größer als bei Kindern, die aus desolaten Verhältnissen kommen (siehe vorne, Abschnitt 3).

Durch zentrale Prüfungen werden Schüler benachteiligt, die aus Schulen mit einem ungünstigen sozialen Umfeld kommen und die die weniger talentierten Lehrer hatten. Auch wenn Schulverwaltungen mitteilen, dass ihr Zentral-Abitur gut funktioniert, so muss man doch hinzufügen, dass zentrale Prüfungen weder zu mehr Gerechtigkeit noch zu besseren Leistungen führen (siehe vorne, die Abschnitte 2 und 3). Hinzu kommt noch, dass zentrale Standards, die in Prüfungen eingesetzt werden, negative Rückwirkungen haben: Die kurzzeitig prüfbare Leistung wird zum faktisch bedeutsamste Ziel der Schule. Und je mehr Eltern und Schülern „signalisiert wird, dass es in der Schule in erster Linie ... auf »Leistung« ankomme, desto weniger wird »Bildung« möglich, desto mehr steht das Abfragbare und Abgeschlossene wieder im Vordergrund und um so weniger geht es um die prozesshafte, offene, aktive Auseinandersetzung mit Kultur.“ (Schlömerkemper, 1998, Seite 265)

Literatur

Arnold, Karl-Heinz: „Schulen im Vergleich. Probleme des Ranking und Chancen eines Monitoring.“ In: Die Deutsche Schule. 91, 1999, 2, S. 218-231

Baumert, Jürgen: „Lehren ohne Leistungsnachweis.“ MPI für Bildungsforschung. Zitiert nach Süddeutsche Zeitung vom 5.2.99, S. 24

Baumert, Jürgen; Olaf Köller: „Nationale und internationale Schulleistungsstudien.“ In: Pädagogik. 50, 1998, 6, S. 12-18

Baumert, Jürgen; Rainer Lehmann u.a.: „TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich –Deskriptive Befunde.“ Opladen 1997.

Baumert, Jürgen; Wilfried Bos, Rainer Watermann: „TIMSS/III – Schülerleistungen in Mathematik und den Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich.“ Berlin 1998

Bethge, Thomas: „Zum Umgang mit den Ergebnissen von TIMSS.“ In: Die Deutsche Schule. 91, 1999, 2, 178-181.

Fisseni, H.-J.: „Auswahlgespräche mit Medizinstudenten: Modelle - Erfahrungen - Vorschläge.“ Göttingen 1993

Hagemeister, Volker 1999: „Was wurde bei TIMSS erhoben? Über die empirische Basis einer aufregenden Studie“ In: Die Deutsche Schule. 91, 1999, 2, 160-177.

Ingenkamp, Karlheinz und Walter H. Schreiber (Hg.) 1989: „Was wissen unsere Schüler? Überregionale Lernerfolgsmessung aus internationaler Sicht.“ Weinheim.

Ito, Toshiko: „Zwischen »Fassade«und »wirklicher Absicht«.“ In: Zeitschrift für Pädagogik. 43, 1997, 3, 449.

Konzept für eine Schulgesetzänderung in einem deutschen Bundesland, Entwurffassung aus dem Jahre 1998.

Lange, Hermann: „Qualitätssicherung in Schulen.“ In: Die Deutsche Schule. 91, 1999, 2, 144-159.

Lehmann, Rainer H.; Rainer Peek; Rüdiger Gänsfuß: „Aspekte der Lernausgangslage und der Lernentwicklung - Jahrgangsstufe 7.“ Hamburg 1998.

Lehmann, Rainer H.; Rainer Peek; Rüdiger Gänsfuß: „Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen.“ Hamburg 1997

Lienert, Gustav A; U. Raatz.: „Testaufbau und Testanalyse.“ 5. überarb. u. erw. Aufl., Weinheim 1994

OECD/CERI: „What works in Innovation: The Assessment of School Performance.“ CERI/CD (1994, 14) Paris 1994.

Preibusch, Wolfgang; Volker Hagemeister, K. Schuricht, H. Seyhan: „Wer sind die türkischen Schüler?“ In: Die Deutsche Schule. 76, 1984, 3, 224-238.

Saldern, Matthias von: „TIMSS – kulturell interpretiert.“ In: Die Deutsche Schule. 91, 1999, 2, 186-201.

Schlömerkemper, Jörg 1998: „Bildung bleibt wichtiger als Leistung! TIMSS darf die Bildungsreform nicht in Frage stellen.“ In: Die Deutsche Schule. 90, 1998, 3, 262-265.

Schumann, Karl Ferdinand; Hans J. Claus: „Prognosen des Studienerfolgs.“ In: Blickpunkt Hochschuldidaktik. Nr. 10, 1970, S. 17 Schümer, Gundel: „Mathematikunterricht in Japan – ein Überblick über den Unterricht in öffentlichen Grund- und Mittelschulen und privaten Ergänzungsschulen.“ In: Unterrichtswissenschaft. 26, 1998, 3, S. 195-228 .

Schweitzer, Jochen: „Keine Angst vor PISA!“ In: Die Deutsche Schule. 91, 1999, 134-143

Shavelson, Richard J.; Maria A. Ruiz-Rrimo: „Leistungsbewertung im naturwissenschaftlichen Unterricht.“ In: Unterrichtswissenschaft. 27, 1999, 2, S. 102-127.

Sygusch, Hajo: „Bildung und Leistung gehören zusammen!“ In: Die Deutsche Schule. 91, 1999, 2, 182-185.

Trost, Günter 1989: „Entwertung des Abiturs durch Hochschuleingangstests.“ 90, 1989, 3, 76-79 .