Was sagen uns PISA & Co, wenn wir uns auf sie einlassen?

Peter Bender

1 Vorbemerkungen

In den letzten fünf Jahren habe ich mich durch viele tausend deutsch- und englisch- sowie einige französischsprachige Seiten mit Berichten von, Kommentaren zu und Folgerungen aus den internationalen Vergleichsuntersuchungen PISA (Programme for International Student Assessment), TIMSS (Third International Mathematics and Science Study; später: Trends in International Mathematics and Science Study) und IGLU (Internationale Grundschul-Lese-Untersuchung; mit einer Erweiterung in Deutschland um Mathematik: IGLU-E) gearbeitet. Dabei habe ich zahlreiche wertvolle Erkenntnisse gewonnen, wurde aber auch mit mehr oder weniger versteckten Ungereimtheiten, unsauberen Argumentationen, gewagten Interpretationen und offensichtlichen Missbräuchen konfrontiert. Um zu einem ausgewogenen Bild von diesen Studien beizutragen, werde ich einen Akzent auf diese Mangelerscheinungen setzen. Da ich hierfür nur einige Seiten zur Verfügung habe, muss ich mich sehr knapp fassen und für detailliertere Untersuchungen auf die Literatur und die Internetseiten verweisen, insbesondere auch auf die Arbeiten (Meyerhöfer 2004a, 2005) mit ihrer fundamentalen erkenntnis- und wissenschaftstheoretischen Kritik sowie Analysen von vielen Aufgaben und sonstigen Einzelheiten, sowie natürlich auf die Beiträge in diesem Buch. Ich habe zwar die grundsätzlichen Probleme immer auch in den Blick genommen, einen Schwerpunkt aber darauf gelegt, *mit den Konstrukten und Daten der Studien selbst zu argumentieren*, weil ich meine, dass dann niemand die Ausrede hat, man ginge nicht wirklich auf PISA & Co ein. Dabei mache ich wohl oder übel eigentlich unzulässige Vergleiche von

JAHNKE/MEYERHÖFER (HG.), PISA & CO, 2. Aufl., 285–341

Punktzahlen u. ä. und Einordnungen in eigentlich ungeeignete bzw. schlecht begründete Kategoriensysteme mit. – Ohne es jedes Mal zu erwähnen, geht es bei mir fast durchweg um den jeweils mathematischen Teil der Studien, und im Mittelpunkt steht PISA.

2 Organisation und Filosofie von PISA & Co

In der „Organisation for Economic Co-Operation and Development" (OECD) in Paris haben sich entwickelte, den USA verbundene Staaten und einige, die auf dem Weg dahin sind, zusammengeschlossen. Eine der Aufgaben der OECD ist die Erhebung und Publikation ökonomischer Daten zur Unterstützung der Entscheidungsträgerinnen & -träger in Politik, Wirtschaft usw. Aus ihrer Perspektive gehört auch der Bildungsstand der 15-Jährigen (am Ende der üblicherweise obligatorischen allgemein bildenden Schule) zu den ökonomischen Daten eines Landes, und er wird in der PISA-Studie in den Jahren 2000, 2003 und 2006 in den OECD- und einigen sog. Partner-Ländern in Form von Leistungstest in den Fächern „Lesen", „Mathematik" und „Naturwissenschaften" partiell erhoben.

Die OECD bedient sich dafür fünf professioneller „Bildungsdienstleister", von denen vier privat sind: ACER aus Australien, ETS und WSTAT aus USA sowie CITO aus Niederlande, „die PISA entwickelt und an bisher achtundfünfzig Staaten verkauft haben" (Flitner 2006). In ihrem Aufsatz macht Flitner die kommerziellen Interessen und die professionelle Öffentlichkeitsarbeit dingfest, die mit PISA, auch und gerade in Deutschland, einhergehen und die man, je nach Standpunkt, als Einstieg oder Meilenstein auf dem Weg in die Privatisierung und Kommerzialisierung auch des Schulsystems sehen kann, u. a. mit der Entwicklung von Testbatterien, die nur gegen Gebühr benutzt werden dürfen.

Von Seiten der Wissenschaft wird PISA von einem internationalen Konsortium und nationalen Konsortien in den Ländern verantwortet. Ein Teil der Arbeit wird von fächerbezogenen internationalen und nationalen Expertinnen- & Experten-Gruppen geleistet, darunter der PISA-Deutschland-Mathematik-Gruppe. In Deutschland gibt es zusätzlich einen Beirat und gehört die Kultusministerkonferenz zu den Auftrag- und Geldgeberinnen & -gebern.

Die Konkurrenten TIMSS und PISA

Mit vielen Grundsätzen und Details hat man an TIMSS 1995 angeknüpft, unter teilweiser Wahrung der personellen Kontinuität, z. B. in der prägenden Person des (mit) federführenden TIMSS-Deutschland-Planers und PISA-Deutschland-Konsortiums-Mitglieds Jürgen Baumert (2000 ebenfalls federführend). Umso mehr wundert man sich, dass auf die späteren Durchgänge von TIMSS (1999, 2003 und, geplant, 2007), jedenfalls in deutschsprachigen PISA-Verlautbarungen, so gar kein Bezug genommen wird. Dass es solche späteren TIMSS-Durchgänge gibt, ist in Deutschland so gut wie unbekannt (ich selbst bin erst im Frühjahr 2005 zufällig auf dieses Faktum aufmerksam geworden). Da Deutschland an ihnen nicht teilgenommen hat, sind sie für Politik und Medien hierzulande nicht sonderlich interessant. Aber ein Unternehmen wie PISA mit seiner Reklamierung von Wissenschaftlichkeit müsste sich mit dieser Konkurrenz auseinandersetzen und deren Ergebnisse in ihrer Gleichartigkeit sowie Unterschiedlichkeit gegenüber den eigenen analysieren. Allerdings müsste man dazu den Glauben an die eigene Einzigkeit ablegen. TIMSS hat immerhin den Vorzug, dass sie schon über einen viel längeren Zeitraum läuft und, zumindest von 1995 bis 1999 sowie von 2003 bis 2007, den Langzeitvergleich von bestimmten Kohorten ermöglicht, weil jedes Mal das 8. Schuljahr und 1995 sowie 2003 außerdem das 4. Schuljahr Zielpopulation war, und 2007 (wie schon 1995) zusätzlich das 12. Schuljahr. Weiterhin sehe ich das Prinzip von TIMSS, sich um Validität bezüglich der nationalen Curricula zu bemühen, als Vorzug, während ich das gegenteilige Vorgehen von PISA für im Ansatz verfehlt (besonders in Anbetracht der von aller Welt inklusive PISA besonders wichtig genommenen Länderrangfolgen) und in der Praxis gescheitert halte, wie ich später noch darlegen werde.

Gewisse Animositäten zwischen PISA und TIMSS rühren nicht nur vom Einzigkeits-Anspruch von PISA (der bei TIMSS allerdings kaum minder ausgeprägt sein dürfte), sondern auf höherer Ebene wohl auch daher, dass TIMSS von der „International Association for the Evaluation of Educational Achievement“ (IEA) in Amsterdam verantwortet wird, die wiederum eine Einrichtung der UNESCO ist, deren politische Ausrichtung sich von der der OECD bekannt-

lich deutlich unterscheidet. Dieser Unterschied zeigt sich nicht zuletzt an der stärkeren Diversität (auch punktemäßig) der bei TIMSS teilnehmenden Länder. Aus praktischen Gründen ist hier PISA mit seiner größeren Homogenität im Vorteil. Die politische Bewertung wiederum muss Jede & Jeder mit der eigenen Überzeugung in Einklang bringen (das gilt auch für so manche pädagogische Romantikerinnen & Romantiker, die trotz deren ihnen eigentlich fernstehender Filosofie auch gerne aus PISA und TIMSS Honig für ihre Überzeugungen saugen). Der ganze Komplex der internationalen Vergleichsuntersuchungen ist jedenfalls auf mehreren Ebenen unmittelbar von der Politik kontaminiert, so dass hier den Angehörigen des Bildungssystems und der wissenschaftlichen Kommunität das viel geübte Ignoranzverhalten gegenüber politischen und gesellschaftlichen Belangen als Ausweg nicht wirklich offen steht.

Den Mangel fehlender Untersuchungen zum Bereich „Lesen“ hat die IEA zwischenzeitlich beseitig und 2001 und 2006 PIRLS (Progress in International Reading Literacy Study) auf den Weg gebracht. Auch IGLU ist eine Veranstaltung der IEA. Im Gegensatz zu PIRLS beteiligte sich Deutschland an IGLU 2001, und hierbei ist eine durchaus fruchtbare Zusammenarbeit mit PISA zustande gekommen; IGLU ist keine Konkurrenz, sondern Ergänzung. Für 2007 ist auch wieder eine Teilnahme Deutschlands an TIMSS für das 4. Schuljahr vorgesehen.

Berechnung der (Länder-) Punktzahlen

Bei PISA werden in jedem Land Stichproben von 15-Jährigen gezogen, (meistens) in der Größenordnung von 5000. Die Jugendlichen werden in den drei Fächern („Inhaltsbereichen“) getestet, und die Ergebnisse werden so geeicht, dass in jedem Bereich für die *OECD*-Jugendlichen der Mittelwert für die Leistungspunktzahlen *500* und die Standardabweichung *100* beträgt (ist a das arithmetische Mittel und s die Standardabweichung aller OECD-Punktzahlen, dann wird jede Punktzahl t in $100 \cdot (t - a)/s + 500$ transformiert). Bei TIMSS und IGLU sieht es *im Prinzip* ähnlich aus.

Die Zahlen sagen jedenfalls nur etwas über den *relativen* Stand beim jeweiligen Test; unmittelbare Punktzahlvergleiche zwischen verschiedenen inhaltlichen Bereichen oder zwischen verschiedenen

Durchgängen lassen meistens keine sinnvollen Aussagen zu und sind unbedingt mit größter Vorsicht zu ziehen. Dies gilt erst recht für die Durchschnittspunktzahlen von Teilpopulationen, z. B. von den Ländern. Wenn ich mich denn trotzdem auf solche Vergleiche einlasse, wie sie von den PISA-Leuten selbst (z. B. zwischen den Punktzahlen in „Problemlösen“ und in Mathematik u.v.a.) und von Anderen gezogen werden, dann geht es mir gemäß meinem in den Vorbemerkungen erläuterten Ansatz jeweils darum aufzuzeigen, dass die daraus gezogenen Schlüsse nicht in Ordnung sind, – und es muss immer dabei mit gedacht werden, dass schon das Vergleichen selbst nicht in Ordnung ist.

Außerdem ist zu beachten, dass alle diese Zahlen auf Stichproben beruhen, d. h. dass der Wert der jeweiligen Population durchaus ein Stück weit abweichen kann. Deshalb ist in den Berichten zu jedem Stichprobenmittelwert der sog. Standard-Error SE angegeben (T95.2, 88 ff, Knoche & Lind 2000, 11 f), ein Intervall, in dem sich der Wert der Population mit hoher Wahrscheinlichkeit befindet. Alle Rangfolgen gelten nur für die gezogenen Stichproben und könnten sich, unbeschadet der Repräsentativität dieser Stichproben, sehr wohl ändern, wenn man die Werte für die jeweiligen kompletten Populationen wüsste, zumindest zwischen Populationen, deren Punktzahlen nahe beieinander liegen. Eigentlich ist die Rede vom soundsovielten Platz eines Landes nicht korrekt; man muss vielmehr ein ganzes Intervall von Plätzen benennen, in dem sich ein Land befindet; und auch diese Angabe trifft dann nur mit einer gewissen (durchaus hohen) Wahrscheinlichkeit zu.

Die aus den Untersuchungen abgeleiteten Länderrangfolgen haben, vor allem in den „schlechten“ Ländern, in den Medien, in der politischen Klasse, in der Gesellschaft und auch bei vielen Angehörigen des Bildungssystems verständlicherweise dennoch breite Resonanz gefunden. Aus ausländischer Sicht besonders hysterisch aufgeladen war die Atmosphäre in Deutschland schon bei TIMSS 1995, wo *509*, und bei PISA 2000, wo gar nur *490* Punkte erreicht wurden. Auf die Publikation Ende 2004 der deutschen Mathematik-Punktzahl bei PISA 2003 (*503*) reagierte man schon gelassener; die Erwartungen waren zwischenzeitlich bescheiden geworden, es war ja eine (leichte) Verbesserung eingetreten, und man lag oberhalb der Mitte von *500*. – Als schließlich im November 2005 das Buch über

den innerdeutschen Vergleich von PISA 2003 der Öffentlichkeit vorgestellt wurde (nachdem schon im Juli 2005 wegen der vorgezogenen Bundestagswahl „auf Drängen der Politik" – wieso eigentlich? – eine Vorab-Information erfolgt war), wärmte man das schon früher hochgespielte Faktum von dem in Deutschland besonders starken Zusammenhang zwischen sozialem Status und PISA-Punktzahl auf, weil man einen Aufmacher brauchte, um in die Schlagzeilen zu kommen, aber nichts großartig Neues zu bieten hatte. „Chancenungleichheit wächst" u. ä. konnte man prompt Ende Oktober 2005 (nach einer gezielten Vorab-Lancierung dieser Schlagzeile um einige Tage) unisono in den führenden deutschen Zeitungen lesen.

Unerwünschte Einflüsse auf die Ergebnisse

Außer vom Leistungsvermögen der Jugendlichen (wie gut haben sie die Unterrichtsinhalte – auswendig – gelernt? wie gut können sie damit umgehen? wie gut kommen sie mit standardisierten, von Fremden gestellten Tests zurecht? wie valide – wofür auch immer – sind die Tests von PISA & Co? usw.) und der Angepasstheit des jeweiligen nationalen Curriculums an die inhaltlichen Vorgaben von PISA & Co sind die Leistungspunktzahlen von vielen weiteren Einflussfaktoren abhängig, z. B.:

– Bei TIMSS waren die schwedischen Jugendlichen im Durchschnitt ein halbes Jahr älter als die anderen und ein Teil ihrer hohen Punktzahlen und des schwedischen Images als TIMSS- und PISA-(!) Musterland geht ganz banal auf diesen Altersvorsprung zurück (T95.2, 90, 98, Knoche & Lind 2000, 12, T03, 34).
– Das stark unterschiedliche Abschneiden von Luxemburg bei PISA 2000 (*446* Punkte) und 2003 (*493*) erklärt man inzwischen (allerdings zu knapp und daher nicht verstehbar) mit „Unterschieden in der … Zuordnung der Testhefte nach Sprachgruppen" (P03.1, 39).
– Wieso sind in Sachsen-Anhalt die Punktzahlen von 2000 bis 2003 so viel stärker gestiegen als in Deutschland, in Mathematik von *477* auf *502* und in Lesen von *455* auf *482*? Bzw.: Wieso waren dort diese Zahlen 2000 so niedrig? (P00.2, 65, 104, P03.2, 60, 88)
– Bekanntlich bestand bei PISA 2000 in Hamburg und in Berlin im Gesamtschulbereich mancherorts eine ausgeprägte Verweige-

rungshaltung (die man ja verstehen kann), so dass in diesen beiden Bundesländern die Repräsentativität verfehlt wurde (P00.3, 32 f).
– Die Niederlande bei PISA 2000 (O00.2, 186 ff) und Großbritannien bei PISA 2003 (P03.1, 26) wurden wegen eines ungenügenden Ausschöpfungsgrads beim Ziehen der Stichprobe nicht in den Ländervergleich einbezogen.

In den Berichten finden sich noch manche Beispiele, wo die Vorgaben unzulänglich erfüllt waren und wie man damit umgegangen ist. Dies alles nährt den Verdacht, dass es neben den bewältigten Abweichungen noch mancherlei unbewältigte gibt, seien sie unabsichtlich oder absichtlich, zur Verbesserung oder zur Verschlechterung von Ergebnissen herbeigeführt worden, seien sie von den Leuten von PISA & Co erkannt worden oder nicht. Vergleicht man einmal die drei Länderrangfolgen bei TIMSS (Mathematik 8. Schuljahr 1995, 1999, 2003) oder die beiden entsprechenden Rangfolgen bei PISA (2000, 2003) oder gar diese fünf alle miteinander, stößt man auf zahlreiche Wellen-, aber auch ausgeprägte unidirektionale Bewegungen (in jeweils kurzer Zeit!), die nicht wirklich mit TIMSS- und PISA-„legitimen“ Einflüssen erklärt werden können (Tab. 1).

Nachdem bei TIMSS 1995 der Bereich „Lesen“ nicht Gegenstand gewesen war, wurde er bei PISA 2000 in den Mittelpunkt gerückt, während der Schwerpunkt bei PISA 2003 auf Mathematik lag und bei PISA 2006 auf den Naturwissenschaften liegt. In Mathematik gab es folglich bei PISA 2000 „nur“ 31 Aufgaben, bei PISA 2003 aber 84. Aufgaben aus den drei Bereichen (2003 zusätzlich aus „Problemlösen“) wurden auf sog. Testhefte verteilt. Jede Probandin & jeder Proband hatte ein Heft in zwei Stunden zu bearbeiten. Hier erwies sich so manche Aufgabe als verschieden schwer (d. h. sie hatte verschieden hohe Lösungsquoten), je nach dem, in welchem Testheft sie enthalten war (O00.2, 157 ff). Diese Auffälligkeit konnte zwar durch Korrekturen bei der Auswertung rein rechnerisch eingeebnet werden; das Problem wurde aber nicht wirklich angegangen, es besteht nach wie vor, und es wirft einen ersten ausgeprägten Schatten von Fragwürdigkeit auf die noch zu diskutierende Punkteskala für die Schwierigkeit der Aufgaben, der ja der Charakter der Messung einer Art von Konstanten zugesprochen wird.

Tabelle 1. Eine Auswahl der Länder-Punktzahlen mit auffälligen Differenzen bei den 5 Untersuchungen TIMSS 1995, 1999, 2003 (T03, 42 ff), PISA 2000 (P00.1, 173 f), 2003 (P03.1, 70) in Mathematik

Test (Anzahl)	*T 95*	(39)	*T 99*	(38)	*T 03*	(45)	*P 00*	(31)	*P 03*	(40)
Land	Pkte.	Pl.	Pkte.	Pl.	Pkte.	Pl.	Pkte.	Pl.	Pkte.	Pl.
Maximum	*609*		*604*		*605*				*550*	
Bulgarien	*527*	13	*511*	17	*476*	25				
Deutschland	*(502)*	23					*490*	20	*503*	19
Finnland			*520*	14			*536*	4	*544*	2
Israel			*466*	28	*496*	19				
Lettland	*488*	28	*505*	18	*505*	11	*463*	25	*483*	27
Litauen	*472*	33	*482*	22	*502*	16				
Neuseeland	*501*	24	*491*	21	*494*	20	*537*	3	*523*	12
Norwegen	*498*	(25)			*461*	27	*499*	17	*495*	22
Russland	*524*	17	*526*	12	*508*	12	*478*	22	*468*	29
Schweden	*540*	7			*499*	17	*510*	15	*509*	17
Slowakei	*534*	8	*534*	8	*508*	13			*498*	21
Thailand	*(510)*	22							*417*	36
Tschechien	*(550)*	6					*498*	18	*516*	13
Tunesien			*448*	29	*410*	35			*359*	39
Ungarn	*527*	16	*532*	9	*529*	9	*488*	21	*490*	24
USA	*492*	27	*502*	19	*504*	15	*493*	19	*483*	27

Anmerkungen: Pkte. = Punkte, Pl. = Platz.
Die Punkte bei TIMSS 1995 waren zum Zwecke des Vergleichs mit TIMSS 1999 und 2003 unter Weglassen der Länder, die weder 1999, noch 2003 teilgenommen hatten, neu berechnet worden (T99, 334, T03, 371). Von den Ländern, die nur 1995 teilnahmen, sind mir lediglich die Zahlen aus T95.2 (S. 90 f) bekannt. Diese lauten für Deutschland *509*, Thailand *522* und Tschechien *564*. Zwecks Harmonisierung habe ich sie überschlägig herabgesetzt und mich dabei an den Punktzahlen der anderen Länder orientiert. Norwegen taucht in T95.2 (S. 90 f) nicht auf, wohl aber in T03 (S. 42 ff). Bei TIMSS wurden in Lettland nur die lettisch-sprachigen Jugendlichen getestet (und nicht die russisch-sprachigen, die etwa ein Drittel ausmachen).

Zusätzlich zum Bearbeiten der Aufgaben mussten die Probandinnen & Probanden Fragebögen ausfüllen, mit denen zahlreiche Merkmale erhoben wurden, wie Migrationshintergrund (ist gegeben, wenn wenigstens ein Elternteil im Ausland geboren ist), sozialer Status, Bildungsnähe des Elternhauses (Anzahl der Bücher zu Hause; I01.1, 50) usw., insgesamt das sog. „soziale und kulturel-

le Kapital“ (I01.1, 270, P00.1, 326). Darüber hinaus wurden durch Auswertung von allerlei zusätzlichen Quellen, etwa Befragung der Schulleitungen, viele weitere Daten ermittelt. Durch PISA & Co hat man z. B. Informationen über die Migrationsquoten in verschiedenen (geografisch definierten) Teilpopulationen der deutschen Gesellschaft erhalten, die einem bis dahin nicht geläufig waren, sei es, dass sie der Öffentlichkeit vorenthalten worden waren, sei es, dass sie den Behörden nicht wirklich bekannt waren, weil diese sich nach der häufig nichtssagenden Staatsangehörigkeit richten (müssen). Allerdings ist auch der so definierte Begriff des Migrationshintergrunds unzureichend, weil solche Jugendliche nicht unter ihm subsumiert werden, deren Eltern zwar beide im Gastland geboren sind, aber ihrerseits Migrationshintergrund haben und unzulänglich integriert sind. In Deutschland z. B. handelt es sich hierbei um eine nicht zu vernachlässigende Population, deren Behandlung als Einheimische durch PISA zu Verfälschungen führt, die mit der Ersetzung des Merkmals „Staatsangehörigkeit“ durch „Migrationsstatus“ ja vermieden werden sollten.

Der Eiertanz um den innerdeutschen Ländervergleich

Bei PISA 2000 wurde der Mathematik-Test Deutschland um einen nationalen Test auf insgesamt 117 Aufgaben erweitert, die etwas stärker an „das“ deutsche Curriculum angepasst waren. Dadurch konnten innerdeutsche Vergleiche u. a. zwischen den Bundesländern und zwischen den Schulformen angestellt werden. Da bei TIMSS 1995 in den kleineren Bundesländern die Repräsentativität nicht erreicht worden war (die Stichproben waren ja nur für den internationalen Vergleich vorgesehen), diskutierte man lediglich Tendenzen bei (anonymisierten „Typen“ von) Ländern A und B. Allerdings pfiffen schon damals die Spatzen von den Dächern, dass bei den gemessenen Leistungen ein ausgeprägtes Südost-Nordwest-Gefälle herrscht, wie es dann bei beiden PISA-Durchgängen bestätigt wurde. Auch das überraschend schlechte Abschneiden der Gesamtschule (weit unter der Realschule knapp über der Hauptschule) ergab sich so bereits bei TIMSS 1995.

Im *internationalen* Vergleich belegte Bayern bei PISA-2003-Mathematik (unter Einschluss der Nicht-OECD-„Partnerländer“) den

Tabelle 2. Vergleich der Punktzahlen der Bundesländer (P00.2, 104, P03.3, 370; sortiert nach den Ergebnissen bei PISA 2003) und der Schulformen (T95.2, 90, P00.3, 273) in Deutschland. Die Zahl für die Gesamtschule bei TIMSS 1995 habe ich aus (T95.2, 136) abgelesen und für alle Schulformen bei PISA 2003 aus den Angaben in (P03.1, 68) erschlossen. Die Bedeutung von „adjustiert" ist weiter unten im Text erläutert.

(a)	PISA 00	PISA 03 beobachtet	PISA 03 adjustiert
Bayern	*512*	*533*	*531*
Sachsen	*501*	*523*	*516*
Baden-Württemberg	*512*	*512*	*512*
Thüringen	*493*	*510*	*501*
Sachsen-Anhalt	*477*	*502*	*498*
Saarland	*487*	*498*	*499*
Hessen	*486*	*497*	*498*
Schleswig-Holstein	*490*	*497*	*490*
Niedersachsen	*478*	*494*	*497*
Mecklenburg-Vorp.	*484*	*493*	*492*
Rheinland-Pfalz	*488*	*493*	*494*
Brandenburg	*472*	*492*	*485*
Berlin		*488*	*487*
Nordrhein-Westfalen	*480*	*486*	*491*
Hamburg		*481*	*488*
Bremen	*452*	*471*	*481*

(b)	TIMSS 95	PISA 00	PISA 03
Gymnasium	*573*	*574*	*(585)*
Realschule	*504*	*501*	*(500)*
Gesamtschule	*(465)*	*455*	*(458)*
Hauptschule	*446*	*439*	*(412)*

7. Rang, und würde man beim besten Dutzend der Länder der Welt (inklusive Bayern) einmal nur die Leistungen der Jugendlichen ohne Migrationshintergrund betrachten, würde Bayern noch um einige Plätze nach vorne rücken.

Immer wieder wurde in Diskussionsbeiträgen versucht, diesen bayrischen Erfolg madig zu machen, i.W. mit folgendem Argumentationsmuster: In Bayern seien soziales Gefälle und Auslese in der schulischen Laufbahn besonders ausgeprägt; zu Wenige würden

das Abitur erwerben, und zu Viele „nur“ den Hauptschulabschluss; auch die Gesamtschule und überhaupt die Errungenschaften der modernen Pädagogik kämen in Bayern zu kurz; schlagwortartig zusammengefasst: das bayrische Schulsystem sei zu konservativ.

Dazu hat man auch mit Zahlen aufgewartet. Für den innerdeutschen Vergleich 2003 hat man sich als ein Maß für den Zusammenhang zwischen sozialer Herkunft und Bildungsbeteiligung die „relative Wahrscheinlichkeit des Gymnasialbesuchs“ (P03.2, 261 ff) ausgedacht. Obwohl schon in früheren Berichten (zu PISA 2000 und 2003) hervorgehoben worden war, dass dieser Zusammenhang in Deutschland stärker als in fast allen anderen Ländern ist, wurde die Verwendung dieses neuen Maßes zum Anlass genommen, dieses Faktum mit der o.a. Schlagzeile und der Ergänzung „Chancenungleichheit in Bayern am größten“ noch einmal hochzuspielen.

Man hatte die Bevölkerung nach ihrem „ökonomisch-sozial-kulturellen Status“ (ESCS) in vier Quartile eingeteilt und dann in Deutschland und in jedem Bundesland in jedem Quartil die Anteile der Jugendlichen, die ein Gymnasium besuchen, und der Jugendlichen, die eine Schule einer anderen Schulart besuchen, miteinander verglichen und so die o. g. „relativen Wahrscheinlichkeiten“ erhalten. Dann wurden für Deutschland und für die 16 Bundesländer die „relative Wahrscheinlichkeit“ des ersten durch die des dritten (nicht des vierten!) Quartils dividiert, mit dem reißerisch aufgemachten Ergebnis, dass in Deutschland die Wahrscheinlichkeit, das Gymnasium zu besuchen, für die „Reichen“ 4,01-mal so groß ist wie für die „Armen“ und in Bayern dieses Verhältnis sogar 6,65 beträgt.

Diese Rechnung hat man mit zwei verschiedenen Modellen durchgeführt, und zwar einmal mit und einmal ohne Kontrolle der Mathematik- und der Lese-Kompetenzen (wie immer man diese Kontrolle bewerkstelligt hat). Die beiden genannten Werte stammen aus dem Modell „mit Kontrolle“, und in dem Modell „ohne Kontrolle“ sind alle Werte naturgemäß höher (weil ja statistisch in höheren sozialen Schichten höhere Kompetenzen vorliegen). Da liegen dann allerdings auf einmal Sachsen-Anhalt mit 10,44 und Bremen mit 9,06 vorne, während Bayern mit 7,77 etwa in der Mitte und nahe bei Deutschland mit 6,87 platziert ist. Der PISA-Bericht stellt selbst fest, dass für jedes Bundesland der „wahre“ Wert (was immer man damit meint) im Intervall zwischen den Werten der beiden Modelle

liegt (P03.2, 262), und da steht Bayern auf einmal gar nicht mehr so „schlecht“ da. Aber in die breite Öffentlichkeit wurden nur die für Bayern negativen Zahlen gebracht.

Außerdem wird im PISA-Bericht zutreffend darauf hingewiesen, dass viele Bundesländer (darunter häufig solche mit niedriger Verhältniszahl) in nennenswertem Umfang Gesamtschulen besitzen, die zum Abitur führen, die aber bei diesen Rechnungen nicht berücksichtigt sind (S. 262, Fußnote 5), und dass es zusätzlich auf den Expansionsgrad der Gymnasien ankommt (S. 263), der ebenfalls nicht einbezogen wurde.

Die Willkür aller dieser Setzungen (zu den von mir bereits genannten z. B. noch: die ESCS-Skala überhaupt; die Einteilung in Quartile; der Anteil des Gymnasialbesuchs wurde nicht zu *allen* Jugendlichen, sondern zu denen, die Schulen anderer Schularten besuchen, in Beziehung gesetzt, wodurch die Verhältniszahlen vergrößert werden; usw.) und die Unmöglichkeit eines Vergleichs mit dem Ausland werfen ein merkwürdiges Zwielicht auf die PISA-Schlagzeile vom Oktober 2005.

Zurück zu Bayern: Auch mir missfallen z. B. das Zentral-Abitur und die Verkürzung der gymnasialen Schulzeit um ein Jahr (es ist allerdings zu konstatieren, dass die anderen Bundesländer sich derzeit diesen „Bildungs“-Maßnahmen mehr oder weniger eifrig anschließen). Aber den Erfolg des bayrischen Schulsystems muss man anerkennen, gerade auch in Anbetracht des vergleichsweise hohen Hauptschulniveaus dort (*462* Punkte; P03.2, 178). Und dass in den anderen Bundesländern höherwertige Zertifikate für schwächere Schulleistungen vergeben werden, ist doch kein Zeichen von Qualität; – eher im Gegenteil!

Für eine Geringschätzung der bayrischen Punktzahlen stünden ja durchaus logische Argumente zur Verfügung: Vielleicht hält man von Untersuchungen wie PISA nichts. Oder man akzeptiert sie mit ihren Ergebnissen, findet aber andere in der Schule zu erwerbende Tugenden wichtiger als die (mit PISA & Co gemessenen) kognitiven Leistungen, usw.

Wenn man sich jedoch auf PISA & Co einlässt und zugleich einem fortschrittlichen – in Bayern ja als weniger ausgeprägt vorausgesetzten (zur Herausarbeitung der Argumentationsstruktur muss ich hier manichäisch vereinfachen) – Schulsystem (inklusive Päd-

agogik) frönt und einem solchen unterstellt, dass es auch kognitiven Leistungen (wie sie mit PISA & Co gemessen werden) förderlich ist, – dann müsste man sich über das schwache Abschneiden (nicht zuletzt auch beim schwachen Viertel der Jugendlichen) vieler Bundesländer mit fortschrittlicherem Schulsystem (inklusive Pädagogik) im Vergleich zu Bayern wundern. Dann müssten Zweifel wach werden, ob die fortschrittlicheren Schulsysteme (inklusive Pädagogik) wirklich so überlegen sind. – Dieser Konflikt wird aber bei der Bayern-Schelte durch die o.a. Argumentations-Volte aufgelöst: „Das bayrische Schulsystem (inklusive Pädagogik) ist schlecht, weil es nicht fortschrittlich ist, und daran ändern auch die hohen PISA-&-Co-Punktzahlen nichts".

Wenn allerdings Länder mit einem „fortschrittlichen" Schulsystem (das sich nicht zuletzt in der Einheitsschule bis etwa zum 9. Schuljahr manifestiert; bis vor kurzem wurde dafür besonders gerne Schweden als prototypisch herangezogen) bei PISA & Co gut abschneiden, dann sind diese Punktzahlen auf einmal doch ein Beleg für dessen Überlegenheit. Dass, im konkreten Fall, Bayern schon immer mit Schweden gut mithalten konnte und in PISA 2003 deutlich besser ist, wird hier geflissentlich ignoriert. Leider habe ich zu spät angefangen, Zeitschriften- und Zeitungsartikel mit solchen Argumentationsmustern zu sammeln, so dass ich nur über wenige verfüge. Als prominenten Vertreter kann ich aber den (zufällig) deutschen OECD-Bildungsstudien-Koordinator Andreas Schleicher mit seiner folgenden Aussage gegenüber dem Fernsehsender 3sat (AP-Meldung vom 15.07.05, zitiert nach GDM 2005, 88) anführen:

„Auch in den stärksten Bundesländern gibt es riesige Leistungsunterschiede zwischen den einzelnen Schülern, die ein Vielfaches größer sind als die Unterschiede zwischen den Ländern." Damit wollte er anlässlich der Vorabveröffentlichung des innerdeutschen Vergleichs am 15.07.05 begründen, dass man innerhalb Deutschlands aus höheren PISA-Zahlen einzelner Bundesländer nicht auf die Überlegenheit von deren Schulsystem schließen darf. Hier begegnet uns ein Muster an gespaltener Argumentation. In Schleichers verbissenen Feldzug gegen die Dreigliedrigkeit des deutschen Schulsystems passt die bayrische Überlegenheit so ganz und gar nicht. Dagegen führt er selbstverständlich u. a. die Punktevorsprünge der (internationalen) Spitzenländer als „Beweis" für die Über-

legenheit des Einheitsschulsystems an, obwohl auf internationaler Ebene seine zitierte Aussage ja genauso gilt. – Auf diesen ganzen Komplex gehe ich in Kapitel 3 ausführlicher ein.

Die Argumentation hat mich aus einem weiteren Grund erschüttert: In der Statistik ist es fast immer so, dass die Unterschiede zwischen den beiden Mittelwerten zweier Populationen viel kleiner als deren beiden Spannweiten sind, und trotzdem werden, je nach dem, weitreichende, tiefgehende und handfeste Schlüsse aus solchen Mittelwertdifferenzen gezogen, etwa auch über die Überlegenheit einer Methode, der die eine Population ausgesetzt ist und die andere nicht. Solche Schlüsse zu ziehen und sie möglichst weit abzusichern ist gerade das Wesen der Beschreibenden und der Beurteilenden Statistik! Mit welchen Absichten stellt ein PISA-Koordinator und studierter Mathematiker dieses Fundament jeglicher empirischer Forschung in Abrede, obwohl er seine ganze berufliche Tätigkeit doch genau darauf aufbaut? – Konkret war es um die Frage gegangen, ob die innerdeutschen PISA-Zahlen nicht dafür sprechen, „dass die Bildungssysteme einer bestimmten Regierung besser sind als die einer anderen“. In der Tat ist ja die positive Korrelation zwischen PISA-Leistung und längerfristig wirksamem Grad an Konservatismus der Landesregierung unübersehbar. Aber da gibt es ein ganzes Bündel an Einflussgrößen, nämlich der ökonomische, soziale, kulturelle und vor allem der Migrations-Status der Jugendlichen (die auch besser definier- und messbar sind als der längerfristig wirksame Grad an Konservatismus und in der damals aktuellen Vorwahlperiode im Juli 2005 nicht ganz so polarisierend wirken wie dieser) und hochkomplexe Zusammenhänge zwischen allen. Man kann also guten *statistischen* Gewissens in Abrede stellen, dass die politische Prägung einer Landesregierung als solche bessere Schulsysteme mit sich bringe. Aber das von Schleicher benutzte Argument ist völlig sachwidrig.

Die statistische Kontrolle der gerade genannten Einflussgrößen führt auf die sog. „adjustierten“ Werte in Tab. 2a. Man tut dabei so, als ob in allen Ländern die entsprechenden Merkmale gleichstark ausgeprägt wären und erhält eine rechnerische Korrektur der tatsächlich aufgetretenen Werte (P03.2, 370). Während sich bei den meisten alten Bundesländern kaum Veränderungen ergeben, geht die Punktzahl bei den neuen Bundesländern mehr oder weniger

deutlich zurück und steigt sie bei Bremen, Hamburg und Nordrhein-Westfalen mehr oder weniger deutlich an. Hier spiegelt sich unmittelbar die Migrationsquote unter den 15-Jährigen in diesen Bundesländern wider. Während sie bei den neuen Bundesländern insgesamt bei etwa 5% liegt, lautet sie in Nordrhein-Westfalen 32%, in Bremen 41% und in Bayern 21% (P00.2, 190; von Hamburg ist dort die Quote nicht angegeben; 2003 beträgt sie 35%; P03.2, 272). Natürlich besteht hier kein monokausaler Zusammenhang, aber vermutlich wären die Unterschiede zwischen tatsächlichem und adjustiertem Wert teilweise noch größer, wenn zur Adjustierung *allein die Migrationsquote* verwendet würde. Auch den Einfluss der Migrationsquote diskutiere ich in Kapitel 3 noch ausführlicher.

Jedenfalls liegt Bayern auch nach der Adjustierung und ebenfalls nach alleiniger Berücksichtigung der Migrationsquote mit weitem Vorsprung vorne und hat den Löwenanteil daran, dass Deutschland überhaupt über *500* Punkte gekommen ist.

Weitere externe und interne Spekulationen zu PISA & Co

Wohl haben PISA & Co, wie gesagt, Informationen über gewisse gesellschaftliche Strukturen unseres Landes geliefert, die man sonst nicht hätte; an anderer Stelle strotzen sie dagegen von Geheimnissen: Schon bei TIMSS 1995 werden „die Leistungstests insgesamt … nicht veröffentlicht, da (?) sie für weitere Forschungszwecke nutzbar sein sollen“ (T95.2, 68, Fragezeichen von mir), wenigstens werden sie dort noch als „ausschließlich für Wissenschaftler unter Beachtung üblicher Professionsregeln zugänglich“ erklärt. Bei PISA und IGLU-E werden zahlreiche Aufgaben und konkrete Ergebnisse dezidiert geheim gehalten. Dies wird damit gerechtfertigt, dass man einen Teil der Aufgaben identisch in verschiedenen Durchgängen einsetzen möchte, um Vergleiche ziehen zu können. Dieses Ziel ist zunächst einmal OECD-typisch, aber auch aus wissenschaftlicher Sicht durchaus ehrenwert. Angesichts

– des Flickenteppichs von Länderpunktzahlen-Entwicklungen zwischen PISA 2000 und 2003 (noch schlimmer, wenn man auch noch TIMSS einbezieht), für den *vor* irgendwelchen Leistungsentwicklungen viele exogene Einflussgrößen verantwortlich sind;

– der im Standard-Error zum Ausdruck kommenden Unsicherheit dieser Punktzahlen;
– der Abhängigkeit der Lösungshäufigkeit vom Ort der jeweiligen Aufgabe im jeweiligen Testheft und im Testhefte-Ensemble (O00.2, 157 ff);
– der geringen kurzfristigen (3 Jahre bzw. eigentlich von der Veröffentlichung zu einem Durchgang bis zur Durchführung des nächsten Durchgangs knapp 1 1/2 Jahre) Fortschrittsmöglichkeiten eines Landes (jedenfalls wenn es etwas größer als Liechtenstein ist);
– der Probleme mit der Beibehaltung von Aufgaben – seien da (u. a. fachdidaktische) Mängel erkannt, mögen sie weiter entwickelten (u. a. fachdidaktischen) Paradigmen nicht mehr entsprechen, würden sie in bestimmten Regionen doch bekannt werden usw. – bei wirklich längerfristigen Vergleichen

muss man sich als PISA-Mensch mit wissenschaftlichen Ansprüchen fragen, ob es sich wirklich lohnt, diesen Ruch von Geheimwissenschaft in Kauf zu nehmen, zumal PISA ja mit öffentlichen Geldern finanziert wird. Makaber kommt mir jedenfalls vor, wenn (m. W. nur mündlich) einem Kritiker die Berechtigung zum Kritisieren abgesprochen wird, weil er nicht alle Aufgaben kennt. Die IGLU-Deutschland-Mathematik-Gruppe, deren Mitglied ich war, hatte z. B. auch Aufgaben in den Test eingebracht, bei denen wir später mathematikdidaktischen Forschungsfragen auf der Basis der Bearbeitungen durch die Probandinnen & Probanden nachgehen wollten. Noch nicht einmal zu den publizierten Aufgaben sind diese Bearbeitungen der (wissenschaftlichen) Öffentlichkeit zugänglich gemacht worden. Da häufen sich seit Jahren (auch bei TIMSS und PISA) Schätze von Forschungsmaterial; – wegen des Primats der Statistik dürfen sie von der mathematikdidaktischen Kommunität nicht gehoben werden, und sie veralten mittelfristig.

Auch gelang es mir trotz intensiven Studiums der Berichte, insbesondere auch des Technical Reports (O00.2) und des Database-Manuals (O00.1) nicht, die Rechnungen zu erschließen. Es hätte einmal konkret vorgeführt werden müssen, wie man für eine Probandin bzw. einen Probanden von den Aufgabenbearbeitungen zu den Roh- und schließlich zu den PISA-Punktzahlen gelangt. Da muss man nicht jedes statistische Detail breit treten, und da kann man so-

gar die Geheimnisse wahren, indem man fiktive Probandinnen & Probanden nur publizierte Aufgaben lösen lässt.

Es ist allzu menschlich und zur Rechtfertigung von eingesetzten sowie Akquirierung von neuen Finanzmitteln unabdingbar, dass man an die Bedeutsamkeit der eigenen Arbeit glaubt und sie dem Publikum vermittelt. In den Berichten von PISA & Co wird zwar ein objektiv erscheinender Stil gepflegt, auf willkürliche Setzungen hingewiesen und zur Vorsicht bei Vergleichen und bei Interpretationen gemahnt (z. B. T95.2, 88), aber der Öffentlichkeit werden die Ergebnisse in einer Weise serviert, dass sie hektisch darauf reagiert. Mit verantwortlich ist die statistik- bzw. testbezogene Sprache, deren Begriffe i. d. R. stochastisch gemeint sind, aber leicht kausal oder gar wertend verstanden werden können, z. B. „benachteiligen“ oder „Chance“ oder die Rede von „Kompetenzen“, etwa beim „Problemlösen“, wenn die Lösungsquote bei einem bestimmten Sortiment von Aufgaben gemeint ist, die von PISA einem bestimmten Bereich zugeordnet sind, der eben als „Problemlösen“ bezeichnet wird. Man müsste einmal prüfen, wie weit PISA-Leute selbst solchen Bedeutungsvermischungen unterliegen. Vor allem aber wird zu wenig gegen absichtliche oder unabsichtliche Fehlverständnisse und -interpretationen unternommen, mit denen einschlägig Interessierte die PISA-Statistiken mit z.T. haarsträubenden Argumenten zur Unterstützung ihrer (bildungs-) (politischen, pädagogischen) (Vor-) Urteile (Entscheidungen) missbrauchen, z. B.: *Schließung* von Stadtteilbüchereien in Frankfurt (Frankfurter Rundschau, FR, April 03), Anschluss aller Schulen ans Internet, *Verkürzung* der gymnasialen Schulzeit in Nordrhein-Westfalen um ein Zeitjahr und faktisch ein halbes Unterrichtsjahr, Abschaffung der Dreigliedrigkeit des deutschen Schulsystems usw. (Hiermit ist noch nichts gegen diese Maßnahmen gesagt, sondern nur gegen ihre Begründung mit PISA.)

Aber auch im System von PISA & Co selbst finden sich auf allen Ebenen spekulative Elemente:

– Für den schon erwähnten OECD-Bildungsstudien-Koordinator Andreas Schleicher ist das „international nicht mehr vermittelbare“ deutsche Schulsystem mit seiner Dreigliedrigkeit eine wichtige Ursache für das langsame Wachstum des deutschen Bruttoinlandprodukts (BIP) in den letzten Jahren (FR, 15.09.04). – Schneller wuchs das BIP in vielen weniger entwickelten EU-Ländern

(mit allerdings durchweg weniger PISA-Punkten als Deutschland), u. a. in Spanien (*485*). Ein anderes „positives" Beispiel ist Mexiko mit einem einsam hohen Anteil von 24,3% Bildungsausgaben an allen öffentlichen Ausgaben (mit sogar nur *385* PISA-Punkten) gegenüber 9,7% in Deutschland. – Für das langsame BIP-Wachstum hierzulande gibt es doch ökonomische Einflussgrößen von ganz anderem Kaliber!

– In PISA 2003 wird Problemlösen als eigenständiger Inhaltsbereich überbewertet.
– Im Bericht für Deutschland wird der Faktor „Migrationshintergrund" systematisch zugunsten des Faktors „soziale Stellung" unterschätzt.
– Die PISA-Deutschland-Mathematik-Gruppe bemüht sich, die willkürliche Stufen-Einteilung der Leistungspunkteskala, die das internationale Konsortium vorgenommen hatte, zu einem quasi naturgegebenen universellen mathematikdidaktischen Analyse-Instrument hochzustilisieren.

Ich war vorübergehend von der Publizität angetan, die Bildung, insbesondere mathematische Bildung, im Gefolge von PISA & Co in Deutschland und anderswo erfahren hat. Ich werde mir aber in dieser positiven Einschätzung zunehmend unsicher, weil ich immer wieder erlebe, wie es bei den Schlüssen und Konsequenzen (z. B. „Bildungsstandards", Leistungsvergleiche, Schulzeitverkürzungen usw.), die „die" Politik unter tätiger Mithilfe von (meistens durchaus wohlmeinenden) Kolleginnen & Kollegen aus Schule, Seminar, Hochschule, Behörden usw. zieht, nicht um Bildung, sondern bestenfalls um Leistung, oft um Wahlkampfparolen, Haushaltsumschichtungen, Bedienung von Ideologien usw. geht.

3 Auf der Suche nach Ursachen für die PISA-&-TIMSS-Mittelmäßigkeit Deutschlands

Jedenfalls nicht die Dreigliedrigkeit

In den Berichten zu TIMSS und PISA wurde von Anfang an betont, dass sich aus diesen Studien keine Schlüsse auf die Überlegenheit eines Schulsystems ziehen lassen (T95.2, 18 f, 89, u. a.), nicht

zuletzt wohl auch ein wenig zum Schutz der deutschen Gesamtschule, die ja zur großen Verblüffung von mir und vielen Anderen sehr schlecht abgeschnitten hatte. Inzwischen werden die Zahlen für sie noch nicht einmal mehr in dem Bericht über den innerdeutschen Vergleich separat ausgeworfen. Stattdessen wird die Gesamtheit aller Schulen typisiert nach den merkwürdigen weichen Kategorien „belastet/unbelastet“ und „aktiv/passiv“ (P03.2, 301 ff). So musste ich die Angaben für meine o.a. Tabelle 2b aus anderen Zahlen erschließen.

Eine entscheidende Ursache für die Schwäche der deutschen Gesamtschule ist, in der Tat unbestreitbar, die bloße Existenz des Gymnasiums in Deutschland (allerdings wird auch die Hauptschule durch die bloße Existenz der Gesamtschule geschwächt). Trotzdem sind interessierte Kreise nicht müde geworden, aus den Zahlen von PISA & Co Honig für die Gesamtschule saugen zu wollen. Ein, zugegeben, zurückhaltendes Beispiel liefert etwa Herrlitz (2003). In der Tat gibt es im internationalen Vergleich zahlreiche Länder mit Einheitsschule und mehr Punkten als Deutschland, und schon meint man, die Überlegenheit der Gesamtschule mit PISA-Zahlen belegt zu haben. – Dass auch fast alle Länder mit weniger Punkten als Deutschland über die Einheitsschule verfügen, wird da geflissentlich übersehen.

Immer wieder wird das – gemäß PISA & Co „schlechte“ – deutsche Schulsystem insgesamt für allerlei Mängel verantwortlich gemacht und eine umwälzende Veränderung desselben gefordert, wobei die Gegliedertheit als ein ausschlaggebender Faktor unterstellt wird, der mit verändert werden muss. Das langsame deutsche BIP-Wachstum habe ich bereits als Beispiel erwähnt.

Der Hamburger Erziehungswissenschaftler Peter Struck hat einen weiteren Mangel identifiziert (FR, 05.01.05): Bei PISA 2003 haben die deutschen Jugendlichen im Bereich „Problemlösen“ mit *513* Punkten ja deutlich besser abgeschnitten als in Mathematik mit *503*. Da Problemlösen nicht in der Schule gelernt wird, zeigt sich angeblich hier, wie auch z. B. bei den Straßenkindern in Mittelamerika oder in Rumänien, die Überlegenheit des Lebens als Lehrmeister in manchen Bereichen gegenüber der Schule. – Diese, in Deutschland insbesondere ihre Dreigliedrigkeit, müsse also verändert werden. Dieses Argumentationsgrundmuster wird im Artikel außerdem auf

viele verfälschte Daten gestützt, und ich führe es deswegen hier aus, weil mir der Missbrauch von PISA & Co in der Geballtheit noch nirgends sonst begegnet ist.

Auch das relativ gute Abschneiden der deutschen Grundschulkinder bei IGLU-E wurde mancherorts als Pluspunkt für die Gesamtschule verbucht: Nachdem Deutschland bei TIMSS 1995 nur mit den Sekundarstufen teilgenommen hatte, wurde die Überprüfung der Primarstufe in Mathematik (und Naturwissenschaften) durch eine entsprechende Erweiterung von IGLU nun 2001 nachgeholt und mit Hilfe von Ankeraufgaben mit TIMSS-1995-Primarstufe vergleichbar gemacht (s. a. T95.1). In Mathematik erzielte Deutschland hierbei *545* Punkte (das Spitzenland Singapur *625*; T95.1, 24). – Solange also alle Kinder gemeinsam unterrichtet werden, erzielen sie hohe Punktzahlen, und sobald sie nach Schularten sortiert werden, geht es mit den Punkten bergab. Wenn das nicht für die Überlegenheit der Einheitsschule spricht! – Wenigstens zwei Umstände dämpfen jedoch die Euphorie: Zum einen hatten die Deutschen 2001 deutlich bearbeitungsfreundlichere Aufgabenformulierungen zur Verfügung als z. B. die Kinder aus Österreich 1995, zum anderen und hauptsächlich war beim Test der Deutschen nur das vierte Schuljahr beteiligt, während bei TIMSS 1995 auch das dritte dabei war. Schränkt man den TIMSS-1995-Datensatz auf das vierte Schuljahr ein, beträgt der Durchschnitt nicht mehr *500*, sondern *529* (T95.1, 24 ff, I01.1, 207), und der deutsche Vorsprung ist nicht mehr ganz so fulminant.

Außerdem stehen mit Variablen wie „Leistungsbereitschaft“, „Pubertätsprobleme“, „Stoffschwierigkeit“, „Stoffcharakter“ usw. (viele ihrerseits vom Lebensalter abhängig) gewichtige Einflussgrößen zur Erklärung der Unterschiede zwischen Primar- und Sekundarstufe I bei PISA & Co zur Verfügung. Komplementär zu den Argumenten der Gesamtschul-Befürworterinnen & -Befürworter stellt sich die Frage, ob in Deutschland eine gegliederte Grundschule nicht noch mehr IGLU-Punkte (aufgrund besserer individueller Förderung) und eine ungegliederte Sekundarstufe I nicht noch weniger PISA-Punkte (aufgrund eines Rückgangs der hohen Punktzahlen des Gymnasiums) erzielt hätte. – Aber selbst wenn es so wäre: Das (Nicht-) Erreichen wesentlicher Bildungs- und Erziehungsziele wird doch durch die Zahlen von PISA & Co überhaupt nicht erfasst,

und hohe Zahlen bei PISA & Co könnten gerade Ausweis eines mangelhaften Bildungssystems sein (Leistungsdruck, Engführung beim Lernen, Pauken für Tests usw.)!

Geringere Leistungsbereitschaft als in den Spitzenländern

Während die skandinavischen Länder (bis auf Finnland *544*; s. dazu jedoch u. a. Freymann 2004) sich inzwischen bei PISA 2003 auf Augenhöhe mit Deutschland befinden: Island *515*, Dänemark *514*, Schweden *509* (bei TIMSS 2003 nur *499*, obwohl die schwedischen Probandinnen & Probanden erneut ein halbes Jahr älter als der Weltdurchschnitt waren), Norwegen *495* (bei TIMSS 2003 nur *461*); – ist Ostasien die wahre Spitzenregion.

Diese Staaten haben zwar alle das Einheitsschulsystem; aber sie würden (nach meinem Dafürhalten: mit Sicherheit) diese guten Platzierungen auch mit einem gegliederten Schulsystem erreichen. Denn diese Leistungen beruhen auf weit wichtigeren Faktoren, nämlich der Leistungsorientiertheit der dortigen Gesellschaften sowie der Hochschätzung von Schulbildung, zumal in Mathematik. – Man müsste auch einmal intensiver prüfen, wie ausgeprägt jeweils

Tabelle 3. Die Länder-Punktzahlen der ostasiatischen Spitzenländer und Finnlands bei den 5 Untersuchungen TIMSS 1995, 1999, 2003 (T99, 32, T03, 42 ff), PISA 2000 (P00.1, 173 f), 2003 (P03.1, 70) sowie bei den Viertklässlerinnen & Viertklässlern bei TIMSS 1995 (V 95; T95.1, 24), 2003 (V 03; T03, 35) in Mathematik

Test (Anz.)	*T 95*	(39)	*T 99*	(38)	*T 03*	(45)	*P 00*	(31)	*P 03*	(40)	*V95*		*V03*	(25)
Land	Pte.	Pl.	Pte.	Pl.	Pte.	Pl.	Pte.	Pl.	Pte.	Pl.	Pte.	Pl.	Pte.	Pl.
Singapur	*609*	1	*604*	1	*605*	1					*625*	1	*594*	1
Südkorea	*581*	2	*587*	2	*589*	2	*547*	2	*542*	3	*611*	2		
Hongkong	*569*	4	*582*	4	*586*	3			*550*	1	*587*	4	*575*	2
Taiwan			*585*	3	*585*	4							*564*	4
Japan	*581*	3	*579*	5	*570*	5	*557*	1	*534*	6	*597*	3	*565*	3
Macau									*527*	9				
Finnland			*520*	14			*536*	3	*544*	2				

Anmerkung: Die Punkte der Achtklässlerinnen & Achtklässler bei TIMSS 1995 waren zum Zwecke des Vergleichs mit TIMSS 1999 und 2003 unter Weglassen der Länder, die weder 1999, noch 2003 teilgenommen hatten, neu berechnet worden (T99, 334, T03, 371) und unterscheiden sich deutlich von den ursprünglich in (T95.2) veröffentlichten.

das System von Privatschulen ist, mit dem dort doch wieder differenziert wird (auch in USA und anderen Ländern), und zwar viel stärker nach Reichtum als in Deutschland mit seinem nach wie vor in weiten Bereichen funktionierenden öffentlichen Schulsystem.

Der mit der Leistungsorientierung verbundene Leistungsdruck entspricht natürlich nicht den romantischen Vorstellungen westlicher Pädagogik, und deswegen wird der Vergleich nicht so gern mit den ostasiatischen Tigerstaaten gezogen. Hinzu kommt, dass dort sowie in Kanada, Australien und Neuseeland die Einwanderungsstruktur hohe Punktzahlen bei PISA & Co eher zulässt als in den älteren EU-Ländern oder in den USA: Entweder ist die Eingewandertenquote fast 0 (Südkorea, Japan, auch Finnland), oder die eingewanderten Familien haben im Durchschnitt ein relativ hohes soziales und kulturelles Niveau, so dass in einigen dieser Länder einige der Ergebnisse durch die Probandinnen & Probanden mit Migrationshintergrund sogar verbessert werden (z. B. Singapur, Neuseeland bei den Viertklässlerinnen & -klässlern in „Lesen“; I01.1, 296).

M.E. sind die Leistungsorientierung der Gesellschaft (die nicht ohne sichtbare Autoritätsstrukturen auskommt) und eine entwickelte Wirtschaft wesentliche Faktoren für die Punktzahlen eines Landes bei PISA & Co. Diese Einflussgrößen können von PISA & Co gar nicht erfasst werden, während die meisten der untersuchten Faktoren dagegen zweitrangig sind. Dies gilt z. B. auch für die TIMSS-1995-Videostudie, in der „die“ Unterrichtsstile in Mathematik in Japan, USA und Deutschland verglichen wurden (s. T95.2, 215 ff). Fachdidaktisch ist diese Studie hoch-interessant. Aber abgesehen davon, dass dabei von Repräsentativität keine Rede sein kann, ist eine Auswirkung der identifizierten Stile auf Punktzahlen von PISA & Co nicht nachgewiesen. Sie ist noch nicht einmal plausibel, weil ja das inhaltliche Paradigma der „Mathematical Literacy“, zu dem die Unterrichtsstile mehr oder weniger gut passen, prinzipiell und praktisch nicht à la PISA & Co getestet werden kann (s. dazu Kap. 4).

Ein großes „Experiment“, das nach meinem Dafürhalten die Wirkungsmächtigkeit des Faktors „Leistungsorientierung“ auf die Zahlen von PISA & Co eindrucksvoll belegt, ist der Zerfall der Sowjetunion verbunden mit der Auflösung autoritärer Strukturen dort und überhaupt in Osteuropa (inklusive DDR) (s. Tab. 1). Dieser

Umschwung hat im Laufe der 1990-er Jahren die Schulen in dieser Region voll erfasst und, pauschal gesprochen, zu einem Straffheitsabbau an Schulorganisation und im Unterricht, verbunden mit einer Lässigkeitszunahme bei allen Beteiligten geführt. Die PISA-Leute machen es sich arg einfach, wenn sie diesen Trend wirklichkeitsfremd gerade umgekehrt mit der Perpetuierung der „traditionell dort vorherrschenden Methode eines stark lehrergesteuerten, auf Kenntnis mathematischer Fakten ausgerichteten Unterrichts“ (P00.1, 177) erklären, als ob die PISA-Fragen sich so stark von den TIMSS-Fragen unterscheiden würden, dass Faktenwissen nichts mehr nützt!

Die ausgeprägte Schwäche des schwachen Viertels der deutschen Jugendlichen, verbunden mit der mangelhaften Integration vieler Jugendlichen mit Migrationshintergrund

Die „nur“ mittleren Ränge Deutschlands bei PISA & Co sind die Folge vor allem des ausgeprägt schwachen Abschneidens des schwachen Viertels unserer Jugendlichen (P00.1, 176 u.v.a.). Eine Hauptursache hierfür sehe ich in einer Distanz zur Leistung in relevanten Gruppen unserer Gesellschaft. Zum einen wirkt die 68-er-Bewegung mit ihrem antiautoritären Prinzip (gegenüber Bildungsinstitutionen, deren Vertreterinnen & Vertretern sowie den Fächern) in Teilen der Pädagogik, kurz gesagt, in einer ungesunden leistungsfernen Grundatmosphäre nach. Zum zweiten haben die Medien bei uns im Großen und Ganzen nicht gerade einen den Leistungsgedanken fördernden Einfluss auf unsere Heranwachsenden. Zum dritten sehen unsere Jugendlichen, besonders die mit schlechten Schulleistungen, für sich wenig Zukunftsperspektiven. Diese Stimmung kann man nicht mit Schulleistungstests erfassen, aber z. B. mit den Schulschwänzerinnen- & -schwänzerzahlen: Die Bertelsmann-Stiftung geht hier von einer halben Million „Schulmüder“ vor allem in Haupt- und Sonderschulen aus, die wöchentlich mehrere Stunden unentschuldigt fehlen (FR, 17.05.03).

In den Berichten wird hervorgehoben, dass in Deutschland der Einfluss des sozialen Status auf die Leistungen bei PISA & Co besonders ausgeprägt ist (P00.1, 319 ff u. v. a.). In einer Regressionsanalyse zur Abhängigkeit der Mathematik-Punktzahlen von acht

verschiedenen Einflussfaktoren (die schrittweise nacheinander einbezogen wurden) wurde deren Anteil an der aufgeklärten Varianz zu 67% für den sozialen Status, zu 12% für den Migrationshintergrund und zu 21% für die anderen (Kindergartenbesuch, Vater-Erwerbstätigkeit, Familien-Umgangssprache usw.) bestimmt (P03.1, 274). Dass in Deutschland und in den älteren EU-Ländern alle diese Faktoren wiederum stark vom Faktor „Migrationshintergrund" abhängen (der als einziger wirklich unabhängig ist und mit dem eigentlich anzufangen wäre) und ihr Aufklärungsanteil wieder großenteils diesem zugeschlagen werden müsste, wird hierbei nicht deutlich. Diese systematische Verharmlosung des Einflusses der Migrationsquote auf die deutschen PISA-Zahlen zieht sich durch alle deutschen PISA-Berichte. Sie wurde schon früh von Borsche (2002) zu Recht scharf kritisiert; die Quantifizierung dieses Einflusses, die Borsche vornimmt, ist jedoch leider untauglich.

Möglicherweise wird aber z. Z., wohl in Reaktion auf die aktuelle politische Diskussion in Deutschland, von der OECD ein Paradigmenwechsel vollzogen. So ist man Mitte Mai 2006 mit dem aus PISA 2000 und 2003 (P03.1, 262 ff) eigentlich zur Genüge bekannten schwachen Abschneiden der Jugendlichen mit Migrationshintergrund als scheinbar neue Sensation an die breite Öffentlichkeit getreten und hat nun das (dreigliedrige) Schulsystem für dieses schwache Abschneiden verantwortlich gemacht (FR, 16.05.06).

Insgesamt haben 21% (P03.1, 271) aller 15-Jährigen in Deutschland Migrationshintergrund, in Westdeutschland (ohne Stadtstaaten!) 27%, in westdeutschen Großstädten über 300 000 Einwohner im Durchschnitt 36%, in der alten DDR nur etwa 5% (P00.2, 190). Probandinnen & Probanden mit Migrationshintergrund erbringen im Mittel erheblich schlechtere Leistungen, z. B. die Viertklässlerinnen & -klässler in Lesen *55* IGLU-Punkte weniger als die anderen (I01.1, 296), ein Unterschied, mit dem Deutschland in der Spitzengruppe liegt. Die deutsche PISA-Leseleistung liegt mit *484* bzw. *491* besonders niedrig. Wo nur ein Elternteil im Ausland geboren ist, fallen die Leistungen lange nicht so stark ab (P00.1, 378, P03.1, 257). Besonders drastisch sind die Auswirkungen in den Bundesländern Nordrhein-Westfalen und Bremen mit Quoten von 32% bzw. 41% Jugendlichen mit Migrationshintergrund (P00.3, 247), wo dann in bestimmten Regionen in vielen Hauptschulklassen deren Anteil so hoch ist, dass

sowohl die mit, als auch die ohne Migrationshintergrund schlecht gefördert werden. (S. hierzu auch den Vergleich der tatsächlichen und der „adjustierten" Werte in Tab. 2a.)

Zur Stützung der These von der Überlegenheit der Gesamtschule wurde (jedenfalls bis zur Veröffentlichung von PISA 2003) Deutschland immer gern mit Schweden verglichen, u. a. wegen der ähnlichen Bevölkerungsstruktur inklusive Eingewandertenquote. Einen deutlichen Unterschied gibt es aber doch. Während in Deutschland unter allen Jugendlichen mit Migrationshintergrund 75% sogar *doppelten* Migrationshintergrund haben, beträgt in Schweden dieser Anteil nur gut die Hälfte (P03.1, 257); und diese sind es ja, die besondere Schwierigkeiten haben (z. B. P00.3, 249). Rechnet man außerdem einmal sämtliche Jugendliche mit Migrationshintergrund (also auch die leistungsstärkeren) heraus, liegt Deutschland mit *527* Punkten plötzlich deutlich *vor* Schweden mit *518* Punkten (P03.1, 257) und fällt nicht mehr so stark gegen Finnland mit seinen *544* Punkten und fast 0 Einwanderung ab.

Wer jetzt hier herauslesen möchte, dass das schwedische Schulsystem besser für die Bewältigung der Einwanderungsproblematik geeignet ist, muss zugleich mitlesen, dass es die autochthone Bevölkerung benachteiligt. Außerdem ist die Punktzahl von Schweden bei TIMSS inzwischen auf *499* (mit deutlich älteren Probandinnen & Probanden!) gesunken, so dass von einem besonders erfolgreichen Bildungssystem (Prinzip der Einheitsschule und Integration der Jugendlichen mit Migrationshintergrund) aus Sicht von PISA & Co nicht mehr die Rede sein kann. Inzwischen ist man in Schweden so weit, dass man diese beiden Prinzipien aufweicht und z. B. in den größeren Städten Jugendliche mit Migrationshintergrund in Mathematik (neben Schwedisch und Englisch das Hauptprüfungsfach) in ihrer Herkunftssprache, z. B. Arabisch oder Serbo-Kroatisch, unterrichtet (Engström, 2005, in seinem Vortrag).

Die genannte Differenz (Punktzahl ohne und mit Migrationshintergrund-Jugendliche) von *24* Punkten in Deutschland (bei PISA 2000 betrug sie *23*; P00.1, 245), die auf der Welt sonst nirgends so hoch ist, ist unabhängig von dem Vergleich mit Schweden ein Indikator für die bei uns besonders ausgeprägte Einwanderungsproblematik. – Um nicht falsch verstanden zu werden: Wenn man sich überhaupt auf das Zahlenwerk von PISA & Co einlässt, dann ist *503*

genau die richtige Punktzahl; denn diese steht für die Leistungen, die die 15-jährigen Jugendlichen 2003 in Deutschland erbracht haben.

Natürlich bilden die Eingewanderten keine homogene Gruppe, und man muss genauer hinschauen. So verdanken wir PISA 2003 die Erkenntnis, dass Jugendliche mit doppeltem Migrationshintergrund, die im Ausland geboren sind, im Mittel deutlich höhere Leistungen erbringen als solche, die hier geboren sind. Dieser Umstand findet seine einfache Erklärung darin, dass die im Ausland Geborenen vorwiegend aus den ehemals sozialistischen Ländern in Osteuropa mit ihren guten Schulsystemen und häufig aus Familien mit durchaus nennenswertem sozialen und kulturellen Standard stammen (P03.1, 271 f).

Im Zuge der erwähnten OECD-Kampagne Mitte Mai 2006 wurde diese PISA-Erkenntnis in den Zeitungen schon arg tendenziös dargestellt, und drei Wochen später wurde die bis dahin lediglich unterschwellig suggerierte Schuldzuweisung an das deutsche Schulsystem in einer dpa-Meldung über den gerade veröffentlichten jüngsten deutschen Mikrozensus auf ganz üble Art explizit gemacht. In der Meldung spielte die Migrationsproblematik eine große Rolle, und m. W. hatte man bei diesem Mikrozensus (von 2005) erstmals das untaugliche Kriterium der Staatsbürgerschaft durch das des Migrationshintergrunds ersetzt (PISA sei Dank). Der letzte Absatz lautete schließlich:

> Der Rückgang der Bevölkerung vollzieht sich ausschließlich bei den Deutschen ohne Migrationshintergrund. Was das für die Integrationspolitik heißt, liegt für Experten auf der Hand. Beispielsweise versagt das deutsche Schulsystem nach der jüngsten OECD-Studie wie kein anderes vergleichbarer Industrienationen bei der Förderung von Migrantenkindern. [Hier haben wir wieder den unseriösen Vergleich mit Australien, Kanada, Neuseeland, der Schweiz und Luxemburg.] Die Schulleistungen von Zuwandererkindern werden mit Dauer des Aufenthaltes ihrer Familien in Deutschland sogar deutlich schlechter. (Westfälisches Volksblatt, WV, 07.06.06)

Dümmer geht's nimmer. – Von einer protestierenden oder wenigstens aufklärenden Reaktion der OECD auf diese bösartige Fehlformulierung hat man nichts gehört. Angesichts ihrer permanen-

ten Propaganda gegen das deutsche Schulsystem muss man sogar befürchten, dass diese Formulierung von ihr selbst lanciert wurde.

Der o. a. Leistungspunkteunterschied (bei den Auswertungen mit und ohne Migrationshintergrund-Jugendliche) in Verbindung mit der großen Leistungsbandbreite in Deutschland konfrontiert uns schmerzlich mit den Fehlern und Versäumnissen unserer globalen Einwanderungspolitik seit langem, die inzwischen in eine völlig unzulängliche Integration weiter Teile der eingewanderten Familien gemündet ist. Es ist sachwidrig und unfair, die Verantwortung hierfür der deutschen Schule zuzuschieben. Vielmehr wurde eine gewaltige Aufgabe für die ganze Gesellschaft aufgetürmt, die nicht vom Bildungssystem allein zu schultern ist. Ein Verdienst von PISA & Co ist es, sie uns vor Augen geführt zu haben. Wie weit wir bei der Bewältigung dieser Jahrhundert-Aufgabe Erfolg haben werden, sei dahin gestellt, zumal angesichts des überlasteten Sozialsystems. Wenigstens sollten die Kritikasterinnen & Kritikaster sich jetzt aber nicht wieder für die „Freiheit", das Deutsch-Lernen zu verweigern, stark machen, – auch wenn der UN-Inspektor Vernor Muñoz am Ende seines, von interessierten Kreise als Schulsystem-Kontrolle instrumentalisierten, Deutschland-Besuchs im Februar 2006 ohne weitere Begründung die ausschlaggebende Bedeutung des Spracherwerbs für die Integration in Abrede stellte (FR, 22.02.06).

4 Was soll bei PISA & Co mit welchen Aufgaben getestet werden?

Welche mathematikbezogenen Fähigkeiten, Fertigkeiten, Wissensbestände, Einstellungen usw. hält man für wichtig, so dass man mit dem Grad ihres Vorhandenseins mathematische Leistungsfähigkeit eines Individuums oder einer ganzen Population bestimmt? Wie misst man diese Tugenden? – Üblicherweise lässt man übliche Aufgaben lösen, und zwar bei einem voluminösen Unternehmen wie PISA & Co überwiegend solche, bei denen die Antwort entweder richtig oder falsch ist, also 1 oder 0 Punkte ergibt (ob im Multiple-Choice- oder in einem anderen Format). Dabei unterstellt man Validität, d. h. dass tatsächlich die interessierenden Tugenden relevant

sind. Allerdings gibt es dazu keine robusten mathematikdidaktischen Forschungsergebnisse.

Es ist klar, dass in einem solchen Test viele durchaus wichtige Tugenden nicht berücksichtigt werden können: Die Fähigkeit, komplexe Probleme anzugehen, überhaupt Mathematisierbarkeit zu prüfen, ein Problem längerfristig und mit Ausdauer zu bearbeiten, Ansätze zu verwerfen oder weiter zu verfolgen, das Problem einmal eine Zeit lang liegen zu lassen, Anderen es verständlich darzustellen, von Gesprächen mit Anderen zu profitieren, Medien inklusive Internet zu nutzen, usw. Wenn z. B. jemand – außerhalb der Testsituation – den Flächeninhalt der Antarktis (PISA-2000-international; Neubrand 2004, 269) durch geometrische Aktivitäten auf einer Landkarte bestimmt, statt in einem gedruckten oder elektronischen Lexikon nachzuschauen, muss man schon an ihrer oder seiner Problemlösefähigkeit zweifeln.

Problematische Aufgabenformulierungen und -übersetzungen

Wie bei allen Tests wird auch bei PISA & Co ganz wesentlich eine extrinsische Fähigkeit der Probandinnen & Probanden abgeprüft, nämlich herausfinden zu können, was die Aufgabenautorinnen & -autoren wohl gemeint haben. Diese Herausforderung ist bei innermathematischen Aufgaben naturgemäß geringer, aber bei den Aufgaben mit einem irgendwie gearteten außermathematischen Kontext beliebig schwierig (und wird durch misslungene Formulierungen bzw. Übersetzungen noch verschärft); denn die aufgeworfenen Probleme sind ja nie die der Probandinnen & Probanden. Da sind natürlich wieder Diejenigen im Vorteil, die an solche (standardisierten, insbesondere von Fremden gestellten) Tests gewöhnt sind und in deren Sprache und Kultur die Aufgaben ursprünglich angesiedelt sind. Bei der angelsächsischen Dominanz haben es die Probandinnen & Probanden aus Deutschland da auf verschiedenen Ebenen allerdings leichter als etwa die aus Mexiko (*385*), Türkei (*423*), Brasilien (*356*) u. a. (P03.1, 70).

Aus Platzgründen kann ich in diesem Aufsatz nur wenige Aufgaben, und diese nur knapp analysieren. Aber man wird bei vielen leicht selbst feststellen können, dass man sie, wenn man nicht durch eine englische (oder gar französische; s. dazu O00.2, 57 ff) Vorlage

gebunden und beeinflusst wäre, anders formulieren würde (vgl. zu dieser Problematik auch Rindermann 2006). Bei meiner Arbeit an den IGLU-Mathematik-Aufgaben stellte ich an über der Hälfte der 102 auf deutsch gegebenen TIMSS-1995-Aufgaben, die als Vorlagen zur Verfügung standen, entsprechende Mängel fest, z. B.:

> Dies ist ein Rechteck mit einer Länge von 6 cm und einer Breite von 4 cm. Die Strecke rund um seine Form nennt man Umfang. *(Zeichnung)*
> Was gibt den Umfang des Rechtecks in Zentimetern an?
> A. $6+4$ B. $6 \bullet 4$ C. $6 \bullet 4 \bullet 2$ D. $6+4+6+4$

Das Wort „Strecke“ ist hier schlicht falsch; die Flecken sind nicht ohne Weiteres als Multiplikationszeichen zu verstehen (was aber vor Fehlern bewahrt); deutsche Kinder sind an weniger lakonische Fragen, d. h. mit Substantiv gewöhnt, etwa: „Welcher Ausdruck gibt … an?“, „Welche Formel gibt … an?“

Oder:

> Dies ist ein Zahlenmuster: 100, 1, 99, 2, 98, ☐, ☐, ☐.
> Welche drei Zahlen passen in die Kästchen?
> A. 3, 97, 4 B. 4, 97, 5 C. 97, 3, 96 D. 97, 4, 96

Natürlich alle vier Tripel.

Wenn man sonst nichts aus der Testtheorie weiß: die Lösungswahrscheinlichkeit für eine Aufgabe kann auf kleinste Veränderungen in der Formulierung empfindlich reagieren. Bei PISA & Co sind aber eben nicht alle Probandinnen & Probanden von unglücklichen Formulierungen gleichermaßen betroffen, sondern die nicht-englisch- (und eventuell nicht-französisch-) sprechenden verstärkt. So weit ich das überblicken kann, hat man aber bei PISA gegenüber TIMSS 1995 Fortschritte bei der sprachlichen und inhaltlichen Qualität der Aufgaben gemacht. – Noch ein Beispiel auf Oberstufenniveau (T95.4, 93):

> Eine Schnur ist symmetrisch um einen zylindrischen Stab gewickelt. Die Schnur windet sich genau viermal um den Stab. Der Umfang des Stabes beträgt 4 cm und seine Länge 12 cm. *(Zeichnung)* Bestimmen Sie die Länge der Schnur.

Bei dieser Aufgabe ist im deutschen Sprachgebrauch das Wort „symmetrisch“ inkorrekt verwendet (allerdings rührt die Schwierigkeit

nicht allein davon; s. die vorzügliche Analyse von Kießwetter, 2002).

Wenn in Schweden die Lösungshäufigkeit hier mit 24% sechsmal so hoch ist wie in Frankreich, so ist das ein Indiz, dass im schwedischen Curriculum raumgeometrische Aufgaben, womöglich von diesem speziellen Typ, intensiver behandelt werden als im Geometrie-Mutterland Frankreich, wo allerdings bekanntlich der Geometrieunterricht sowieso stark algebraisiert ist und sich hauptsächlich auf die affine Ebene beschränkt.

Das problematische Testkonstrukt „Mathematical Literacy"

Hier drängt sich die Frage nach der Unterrichts- und Curriculumsvalidität auf. Während bei TIMSS die Validität bezüglich der Ländercurricula noch ein erklärtes Ziel war und ist (T95.2, 21, 47 usw.), findet bei den PISA-Tests ein „Verzicht auf transnationale curriculare Validität" statt, stattdessen „führen sie ein didaktisches und bildungstheoretisches Konzept mit sich, das normativ ist" (P00.1, 19). Als hierfür „in mancher Hinsicht vorbildlich" (P00.1, 25) werden die NCTM-Standards für den Mathematikunterricht zitiert (NCTM 1989, überarbeitet 2000). Das Konzept wird verkörpert durch die sog. „Mathematical Literacy", die von der OECD so definiert ist:

> Die Rolle zu erkennen und zu verstehen, die die Mathematik in der Welt spielt, fundierte mathematische Urteile abzugeben und sich auf eine Weise mit der Mathematik zu befassen, die den Anforderungen des gegenwärtigen und künftigen Lebens einer Person als konstruktivem, engagiertem und reflektierendem Bürger entspricht. (P00.1, 23)

Wie den meisten Kolleginnen & Kollegen aus der deutschsprachigen mathematikdidaktischen Kommunität sagt mir das Konzept der Mathematical Literacy von PISA durchaus zu, und ich teile die schon von den TIMSS-Leuten vertretene Auffassung, dass der deutsche Mathematikunterricht allzu sehr auf Faktenwissenserwerb und die Beherrschung von Verfahren zielt (T95.2, 31). Es ist klar, dass Länder, die ihr geschriebenes und reales Curriculum stärker daran ausgerichtet haben, ob durch explizite Übernahme oder durch eigene Entwicklung, von PISA „bevorzugt" werden. Die PISA-Deutschland-Mathematik-Gruppe hat diesen Grundmangel des PISA-An-

satzes erkannt, bei den nationalen Ergänzungsaufgaben von PISA 2000 „das“ deutsche Curriculum stärker berücksichtigt und den Mathematical-Literacy-Begriff hin zu „mathematische Grundbildung“ leicht abgewandelt (Neubrand 2004, 15 ff).

Der Mathematical-Literacy-Begriff passt sehr wohl zur Tradition der deutschen bildungstheoretischen Didaktik, wie sie z. B. vom *alten* Wolfgang Klafki (1958) oder speziell zum Mathematikunterricht von Heinrich Winter (1975) mit seinem Begriff der „Umwelterschließung“ verkörpert wird. Allerdings enthält die Konzeption von PISA einen stärker pragmatischen Zug (P00.1, 19). Ich persönlich vermisse dabei z. B. die Rolle der Mathematik als Kulturgut.

Der dominierende Bestandteil von Mathematical Literacy im Sinne von PISA ist die Kompetenz zum Modellieren (offenbar i. W. mit Problemlösen und sogar erklärtermaßen mit Aufgabenlösen zu identifizieren; I01.2, 118 f). Besonders wichtig unter den NCTM-Standards scheint der folgende zu sein: „Vorbereitung auf offene Aufgabenstellungen, da realistische Probleme und Aufgaben in der Regel nicht gut definiert sind“ (P00.1, 25).

Insgesamt kann ein Test wie PISA oder IGLU, zumal mit vielen Multiple-Choice-Aufgaben, der Mathematical-Literacy-Definition natürlich nicht gerecht werden (so schon Kießwetter, 2002; s. dazu auch die nicht überzeugende Replik von Reiss & Törner, 2003). Kein einziger Aspekt dieser Definition kann sich in solchen Aufgaben wiederfinden: Es ist nirgends nötig, eine vorgelegte Situation überhaupt auf Mathematisierbarkeit zu prüfen; denn es ist immer klar, dass zu mathematisieren ist. Es kann nirgends das Erkennen und Verstehen der Rolle der Mathematik in der Welt wirklich aufgezeigt werden. Usw. Keine einzige dieser Häppchenaufgaben, sei sie noch so komplex aufgebaut, stellt ein authentisches Sachproblem dar, gar ein Problem der Probandinnen & Probanden selbst. Natürlich ist keine Aufgabe wirklich offen; es ist lediglich immer wieder der Versuch erkennbar, ein direktes Anwenden von Faktenwissen und Fertigkeiten durch häufig textlastige Einkleidung des mathematischen Gehalts in allerlei inner- und außermathematische Kontexte zu verhindern, wobei die Autorinnen & Autoren immer wieder über ihre eigenen Füße stolpern.

Alle diese Mangelerscheinungen sind prototypisch in der folgenden Aufgabe versammelt. Es sei an dieser Stelle betont, dass ich

noch nie Zeit darauf verwendet habe, solche Aufgaben bei PISA & Co zu suchen, sondern dass sie alle von Irgendjemand, und zwar ironischerweise meistens von PISA & Co selbst, neben den vielen geheim gehaltenen, als musterhaft der Öffentlichkeit vorgestellt wurden. Die Aufgaben „Gehen“ sollte im WV vom 15.07.05 den (ja vorgezogenen) Teilbericht über den innerdeutschen Vergleich bei PISA 2003 illustrieren.

> Gehen: *(Zeichnung)* Das Bild zeigt die Fußabdrücke eines gehenden Mannes. Die Schrittlänge P entspricht dem Abstand zwischen den hintersten Punkten von zwei aufeinander folgenden Fußabdrücken. – Für Männer drückt die Formel *n/P=140* die ungefähre Beziehung zwischen *n* und *P* aus, wobei *n = Anzahl der Schritte pro Minute* und *P = Schrittlänge in Meter*
> Frage 1: Wenn die Formel auf Daniels Gangart zutrifft und er 70 Schritte pro Minute macht, wie viel beträgt dann seine Schrittlänge?
> Frage 2: Bernhard weiß, dass seine Schrittlänge 0,80 Meter beträgt. Die Formel trifft auf Bernhards Gangart zu. Berechne Bernhards Geschwindigkeit in Metern pro Minute und in Kilometern pro Stunde.

Im Original hat die Aufgabe ein übersichtlicheres Layout. Sie ist sehr schlecht formuliert (vielleicht unvermeidbar wegen der erforderlichen Knappheit). Dies trägt mit dazu bei, dass man sogar als Experte für anwendungsorientierten Mathematikunterricht Schwierigkeiten hat, den behaupteten Zusammenhang überhaupt zu verstehen. Was ist das für eine Konstante mit der Dimension 1/(Minute mal Meter)? Alle Erfahrungen besagen doch, dass n und P nicht proportional, sondern eher umgekehrt proportional sind. Ihr *Produkt* ist eine sinnvolle physikalische Größe, die Geschwindigkeit (Meter/Minute). Normalerweise übt man die Tätigkeit des Gehens zu einem bestimmten Zweck aus, vor allem um von einem Ort zu einem anderen zu gelangen, aber auch um zu schlendern oder sich sportlich zu betätigen. Besonders auch wenn man gemeinsam mit anderen Menschen geht, legt man quasi die Geschwindigkeit fest, und auch wenn man sie in verschiedenen Zeitintervallen ändert, so ist sie doch die *primäre* Größe, und die Schrittfrequenz stellt sich entsprechend der Schrittlänge ein (wobei diese Abhängigkeit nicht unidirektional ist).

Nun hatten die PISA-Leute eine sportmedizinische Untersuchung ausgegraben, nach der der o. a. Zusammenhang gilt, jedenfalls im Bereich von 0,5 bis 0,9 Meter Schrittlänge. Natürlich müsste man hier genauer wissen, wie in dieser Untersuchung erreicht wurde, dass die Versuchspersonen gehen, ohne dass sie sich eine Geschwindigkeit vorgeben. Mit realistischen Kontexten dürften diese Versuche wenig zu tun haben. Jedenfalls hängt unter diesen Bedingungen in dem genannten Bereich die Geschwindigkeit quadratisch von der Schrittlänge ab. Und dann könnte dieser Zusammenhang sogar plausibel sein, wenn man bedenkt, dass die Muskelmasse ja prinzipiell kubisch mit der Beinlänge wächst.

Im „üblichen" Mathematikunterricht der Sekundarstufe I würde man eher den mathematischen Zusammenhang n·P=v (Geschwindigkeit) analysieren: man nimmt jeweils eine der drei Größen als Parameter und betrachtet den funktionalen Zusammenhang der beiden anderen. In der Aufgabe hat man dagegen die Menge aller Menschen als Definitionsbereich, und jedem Menschen ist die konstante Zahl n/P zugeordnet, die sich als solche ergibt, wenn man unter ganz bestimmten, Nicht-Alltags-, Bedingungen bei ihm n und P misst. Das wesentliche Ergebnis dieser sportmedizinischen Untersuchung lautet also, dass sich (bei Männern im Bereich $0,5 < P < 0,9$) überhaupt eine konstante Zahl ergibt, dass n und P dabei proportional sind und dass diese Zahl (unabhängig von Alter, Körperbau und Konstitution) etwa 140 beträgt. Das wird jedenfalls in der Aufgabe „Gehen" so behauptet. Dass der Gültigkeitsbereich fehlt, ist nicht korrekt, aber verständlich, weil seine Angabe das Ganze für die Probandinnen & Probanden noch unübersichtlicher machen würde. Eine uralte mathematikdidaktische Weisheit besagt, dass die Nennung einer Größe wie 140 in einem Anwendungskontext nur sinnvoll ist, wenn man sie mit anderen Größen konfrontiert, also z. B. wenigstens mit der Konstanten von Frauen (wenn es sie denn dort gibt), was hier allerdings versäumt ist.

Man könnte auf den Gedanken kommen, diesen Sachkomplex wenigstens für den Unterricht fruchtbar zu machen, um vielleicht herauszuarbeiten, wie wichtig es ist, sich klar zu machen, welche Größen man konstant hält, welche man variiert, und überhaupt, was Definitions- und Wertebereiche sind. Vom Inhaltlichen her ist mir das Thema zu weit hergeholt und irrelevant, und ich stelle in Abre-

de, dass man seine knappe Unterrichtszeit darauf verwenden sollte. Vom mathematischen Gehalt her sind „gewöhnliche“ Schülerinnen und Schüler bis in die Oberstufe m.E. überfordert (man hat doch selbst seine anfängliche Mühe, diese funktionale Begrifflichkeit auf die Reihe zu kriegen).

Aus fachdidaktischer Sicht ist diese Aufgabe als *Test*-Aufgabe völlig ungeeignet. Der Kontext, der da vor den Probandinnen & Probanden entworfen wird, ist nicht nur ungewohnt, sondern seine zentrale Aussage ist unplausibel bis hin zur Unverständlichkeit. Den meisten Probandinnen & Probanden dürfte es weder innerhalb der stressigen Testsituation, noch gar aus irgendeiner Erinnerung heraus gelingen, sich den Sachverhalt sinnvoll zu machen. Sie verhalten sich falsch, wenn sie lange über den Sinn nachdenken, weil sie da Zeit verlieren, und sie tun gut daran, die Werte in die Formel einzusetzen und das Ergebnis rechnerisch (und sinnlos) zu produzieren. Lediglich die Umrechnung am Schluss von m/min in km/h erscheint mir sinnvoll.

Da haben wir die vielbeklagte Pervertierung der Anwendungsorientierung prototypisch vor uns: Für die Probandinnen & Probanden reduziert sich die Aufgabe auf das Einsetzen von Werten in die Formel, und der Kontext erweist sich als letztlich irrelevante Einkleidung, die man wieder beseitigen muss, was bei dieser Aufgabe relativ einfach möglich ist. Probandinnen & Probanden mit einer gesunden Mathematical Literacy werden durch diese Aufgabe benachteiligt, wenn sie (gemäß dem, was sie in gutem Mathematikunterricht gelernt haben sollten) versuchen, den Anwendungsgehalt zu erschließen, dabei viel Zeit verlieren und vielleicht sogar vor lauter Verwirrung die Proportionalität durch die scheinbar plausiblere Antiproportionalität ersetzen, usw. Ein besonders hohes Maß an Mathematical Literacy läge vor, wenn sie sich weigern würden, diese Aufgabe überhaupt zu bearbeiten.

Bei der PISA-Deutschland-Mathematik-Gruppe ist ja einige stoffdidaktische Kompetenz versammelt, und so hat man auch dort die Mangelhaftigkeit dieser Aufgabe bemerkt und im Internet eine erläuternde Sachanalyse hinzugefügt: bei pisa.ipn.uni-kiel.de im Text auf „Beispielaufgaben PISA-Testung“ klicken und dann unter der Überschrift „Beispielaufgaben aus PISA 2003“ die Datei „Mathematische Grundbildung (Beispielaufgaben)“ (nicht: „Lösungen“)

downloaden. – Diese Mangelhaftigkeit spielt aber für PISA keine Rolle, da ja die Aufgabe „empirisch gut gelaufen“ ist, so dass man sie nach 2000 nun schon zum zweiten Mal eingesetzt hat. Hier fragt sich natürlich, wie schlecht eine Aufgabe noch sein muss (ich kann mir das kaum vorstellen), damit sie trotz empirisch guten Laufens eliminiert wird. Die Grenze ist erklärtermaßen erst erreicht, wenn fachliche Fehler enthalten sind, wie z. B. bei der weiter unten zu besprechenden Aufgabe „Tageslicht 1“, die ebenfalls empirisch gut gelaufen sein muss, wurde sie doch gleich zu mehreren Gelegenheiten in der Presse veröffentlicht. – Mit „empirisch gut gelaufen“ wiederum ist, so fürchte ich, lediglich gemeint, dass zum tatsächlichen Lösungsverhalten der Probandinnen & Probanden gut eine Aufgabencharakteristik wie in Abb. 1a passt.

Ist folgende Aufgabe nur schlecht oder schon falsch? Sie stammt aus TIMSS 1995 für die Sekundarstufen I und II, wo sie u. a. für „Social Utility“ steht (T95.2, 73, *675* Punkte, T95.3, 164, *554* Punkte), und ob sie bei PISA wieder verwendet wurde, weiß ich natürlich nicht; sie wurde aber noch Ende 2004 in der PISA-Satelliten-Veröffentlichung (Neubrand 2004, 246) als Exempel für eine bestimmte „Kompetenzklasse“ angeführt:

> Diese beiden Anzeigen sind in einer Zeitung in einem Land erschienen, in dem die Währungseinheit *zeds* ist: GEBÄUDE A: Büroräume zu vermieten: 85–95 qm 475 *zeds* pro Monat; 100–120 qm 800 *zeds* pro Monat; GEBÄUDE B: Büroräume zu vermieten: 35–260 qm 90 *zeds* pro Quadratmeter pro Jahr. – Eine Firma ist daran interessiert, ein 110 qm großes Büro in diesem Land für ein Jahr zu mieten. In welchem Bürogebäude, A oder B, sollte sie das Büro mieten, um den niedrigeren Preis zu bekommen? Wie rechnest du? bzw. Wie rechnen Sie?
> (Im Original ist der Text übersichtlicher gestaltet.)

Vermutlich wird erwartet: Bei A muss man $12 \cdot 800 = 9600$ *zeds*, bei B $110 \cdot 90 = 9900$ *zeds* zahlen. Die Probandinnen & Probanden haben natürlich, schon aus Zeitgründen, jedes Nachdenken über den Realgehalt dieser Situation auszuschalten, und das wissen sie auch. – Problematisieren müssten sie eigentlich: Ist überhaupt ein genau 110 qm großes Büro vorhanden? Vor allem aber: Wo gibt es das, dass man 475 *zeds* pro Monat für 95 qm und 800 *zeds* pro Monat (fast das Doppelte) für 100 qm (fast die gleiche Größe) zahlen muss? Da wä-

re doch eine Firma mit dem Klammersack gepudert, wenn sie sich für das eine Jahr nicht mit 95 qm bescheiden würde (und zur Not noch 35 qm im anderen Gebäude hinzu mieten würde mit einer Gesamtsumme dann von nur $475 \cdot 12 + 35 \cdot 90 = 8850$ *zeds*!). – Solche Überlegungen wären Teil der Mathematical Literacy. Genau diese werden hier nicht erwartet und können nicht erwartet werden; sie waren auch von den Aufgabenautorinnen & -autoren offensichtlich nicht angestellt worden.

Durchweg gilt: ist eine Aufgabe den Probandinnen & Probanden schon einmal so oder in ähnlicher Form begegnet und erkennen sie, dass sie Gelerntes einsetzen können (was PISA & Co natürlich nicht messen können), steigt die Lösungsquote. – Dies alles liegt am Schonraum-Vermitteltheits-Pädagogik-Charakter von Schule, worin die Testsituation ja integriert ist; – und dies ist auch gut so.

Es sei konzediert, dass der Anspruch von PISA & Co, weltweit sehr viele Jugendliche (schriftlich, mit begrenzter Zeit, allgemein vergleichbar, ökonomisch auswertbar) zu testen, wohl nur mit relativ einfachen Aufgabenformaten zu realisieren ist. Aber mit diesen Formaten (und mit vielen Inhalten und Fragestellungen) ist man vom Testen der Mathematical Literacy viel weiter entfernt, als man glaubt und glauben machen möchte.

Mit dem Bereich „Lesen“ habe ich mich nicht befasst; aber ich schätze, dass die diskutierten Probleme dort eher größer sind. Auch der Bereich „Naturwissenschaften“ widersetzt sich einem Abtesten à la PISA & Co wohl stärker als „Mathematik“, weil in den Naturwissenschaften das Curriculum weltweit viel uneinheitlicher ist und die eingekleideten Aufgaben immer an der „wirklichen“ Realität gemessen werden können. Schon Hagemeister (1999) hat da sehr verdienstvoll bei mehreren TIMSS-1995-Aufgaben entsprechende Mängel herausgearbeitet (worauf Baumert u. a., 2000, wenig überzeugend reagiert haben). Ähnliche Analysen zu PISA 2003 u. a. findet man bei (Braams 2002). Abgesehen davon, dass die Aufgabenautorinnen & -autoren von PISA & Co anscheinend hin und wieder selbst Lücken bei ihrer „Scientific Literacy“ haben, muss man ihnen zugute halten, dass sie zum Zwecke der Genießbarkeit durch die Probandinnen & Probanden oft zu Vereinfachungen gezwungen sind, die leicht in Verfälschungen umschlagen können (was ihr Vorgehen dann doch in Frage stellt).

Beispiel (Hinweis von Anselm Lambert): Anlässlich der Veröffentlichung der Ergebnisse von PISA-2003-international im Dezember 2004 wurde in der „Zeit“ (09.12.04) und dann noch einmal anlässlich der Vorabveröffentlichung einiger Ergebnisse von PISA-2003-Deutschland im Juli 2005 in der FR (15.07.05) folgende Aufgabe als beispielhaft vorgestellt (P03.1, 394, *591* Punkte):

> Tageslicht 1: Welche Aussage erklärt, warum es auf der Erde Tageslicht und Dunkelheit gibt? A. Die Erde rotiert um ihre Achse. B. Die Sonne rotiert um ihre Achse. C. Die Erdachse ist geneigt. D. Die Erde dreht sich um die Sonne.

Alle Antworten sind falsch, insbesondere auch A. Die Erklärung lautet vielmehr: „Die Erde rotiert mit einer anderen Winkelgeschwindigkeit um die eigene Achse, als sie sich um die Sonne dreht.“ Wären die beiden Winkelgeschwindigkeiten gleich, würde die Erde der Sonne immer dieselbe Seite zuwenden, und es gäbe keinen Tag-Nacht-Wechsel. Bei der Drehung des Mondes um die Erde z. B. besteht genau dieser Zustand. (Wie in der Aufgabe ist auch bei meinen Erläuterungen das übliche einfache geometrische Modell des Systems „Sonne-Erde“ mit zunächst einmal konstanten Winkelgeschwindigkeiten zugrunde gelegt.)

Die Frage ist außerdem für das Gemeinte schlampig gestellt. Die genaue Antwort auf sie lautet nämlich: „Weil das Sonnenlicht nur aus einer Richtung kommt, liegt immer eine Hälfte der Erde im Tageslicht und die andere in der Dunkelheit.“ Man hätte die Frage deshalb vielleicht so formulieren sollen: „. . . , warum an jedem Ort der Erde sich Tageslicht und Dunkelheit regelmäßig abwechseln.“

Allerdings spielt hier sogar doch noch die Neigung der Erdachse eine Rolle: Wäre sie nämlich nicht geneigt, dann hielte sich die Sonne am Nord- und am Südpol immer am Horizont auf, und an diesen beiden Orten fände nie ein Wechsel zwischen Tageslicht und Dunkelheit statt, während an allen anderen Orten Tageslicht und Dunkelheit immer genau 12 Stunden dauern würden. Diesen Zustand gibt es auf der Erde übrigens tatsächlich jährlich zweimal, nämlich am Frühlings- und am Herbstanfang. – Man müsste also die von mir vorgeschlagene Fragestellung noch modifizieren, etwa: „. . . , warum *in unseren Breiten* sich Tageslicht und Dunkelheit regelmäßig abwechseln.“ Dadurch käme allerdings die erschwerende Re-

de von „in unseren Breiten“ ins Spiel, die ja von Vielen nicht verstanden würde.

Die IGLU-Stromkreis-Aufgabe (Viertklässlerinnen & Viertklässler sollen bei einigen Stromkreisen in der „üblichen“ schematischen Darstellung feststellen, wo Strom fließt; I01.1, 158, *546* Punkte) kann man, zumal als Grundschulkind, bestenfalls lösen, wenn das Thema und insbesondere die Art der grafischen Darstellung schon behandelt wurden. Bei dieser Aufgabe kommt erschwerend hinzu, dass das naturwissenschaftliche Problem durch die Antwortvorgaben von einem kombinatorischen überlagert wird: „Welche der Glühbirnen werden leuchten: 1 und 2? 1, 2 und 3? 2, 3 und 4? 2 und 3? 3 und 4?“

Der problematische Testbereich „Problemlösen“

Vergleichbare Schwierigkeiten wie bei der Mathematical Literacy hat man sich bei PISA 2003 mit dem Abtesten eines eigenen Bereichs „Problemlösen“ eingehandelt. Ich gehe nicht auf das läppische sog. *dynamische* Problemlösen ein (das sowieso nur Teil des deutschen Ergänzungstests war; P03.1, 162) und konzentriere mich auf das *analytische*. Fraglich ist für mich immer, was ein Problem im Sinne des Problemlöse-Paradigmas von einem Nicht-Problem unterscheidet. Im angelsächsischen Raum z. B., wo das Konstrukt „problem solving“ besonders gepflegt wird, werden dieselben Wörter auf das Bearbeiten irgendwelcher (z. B. Mathematik-) Aufgaben gemünzt. Die PISA-Definition, nach der analytisches Problemlösen hauptsächlich „in der Analyse gegebener oder erschließbarer Informationen und dem Entwickeln einer Lösung“ einer sich aus einer „verbal, oft auch unter Nutzung von Graphiken, beschriebenen Ausgangslage … ergebenden Problemstellung“ (P03.1, 148 f) besteht, hilft nicht weiter, weil der Begriff „Problemlösen“ da i. W. mit sich selbst erklärt wird. Der Bedingung, „dass der Lösungsweg nicht unmittelbar erkennbar ist“ (P03.1, 148), sollten schließlich alle Testaufgaben (englisch: „problems“) und nicht nur solche in einem abgetrennten Teilbereich „Problemlösen“ unterworfen sein! – Das ganze Definitionsproblem liegt aber in der Natur der Sache, und ich mache nicht den PISA-Bericht dafür verantwortlich. Nach meinem Verständnis ist Problemlösen bei jeglicher geistigen Arbeit allgegenwärtig wie

das Atmen beim Leben. Im Folgenden verwende ich daher in klassischer Weise die Arbeitsdefinition „Problemlösen ist, was der PISA-Problemlöse-Test misst".

Bei den oft notgedrungen textlastigen Aufgaben handelt es sich zumeist um „Logeleien", grob gesprochen, vom Typ des alten Stundenplanproblems, z. B. (kurz gefasst):

> „Kinobesuch" (P03.1, 152): Drei 15-Jährige mit bestimmten Zeitwünschen und -restriktionen und weiteren Bedingungen wollen in den Ferien gemeinsam ins Kino gehen. Wann klappt es?

Oder:

> „Bewässerung" (P03.1, 401), wo statt eines Stromkreises mit Schaltern ein System von Wasserkanälen mit Schleusen zu analysieren ist (damit es sich nicht um eine Aufgabe aus dem herkömmlichen Physikunterricht – eigentlich: Mathematikunterricht – handelt, aber mit derselben logischen Situation und denselben kognitiven Ansprüchen).

Viele der vorgestellten Aufgaben könnten sich, mit denselben oder anderen Kontexten, als Knobelaufgaben in Zeitungswochenendbeilagen finden.

Es besteht, wen wundert's?, eine besonders starke Korrelation zwischen der Problemlöse- und der mathematischen Kompetenz (P03.1, 167). Deutschland ist mit *513* Punkten um *10* Punkte besser, die Niederlande dagegen z. B. mit *520* Punkten um *18* Punkte schlechter als in Mathematik (P03.1, 157 f). „An deutschen Schulen" zeigt sich „im Hinblick auf die Entwicklung mathematischer Kompetenz eine mangelnde Ausschöpfung des kognitiven Potentials zum analytischen Problemlösen", dagegen wird „in den Niederlanden … dieses Potential … optimal genutzt beziehungsweise sogar überkompensiert" (P03.1, 170 f). – Was sagt uns das?

Zunächst muss darauf hingewiesen werden, dass die deutschen Jugendlichen im mathematischen Teilbereich „Quantität", der mit drei anderen bei PISA 2003 separat ausgewertet wurde, *514* Punkte erzielten (P03.1, 75), also dort das Potenzial doch ausgeschöpft, sogar leicht „überkompensiert" haben. Die Mathematikdidaktik weiß schon lange, dass in den anderen drei Bereichen, zumal in Geometrie und Stochastik, Intensivierungsbedarf besteht. – Mit der Überlegenheit des Lebens als Lehrmeister und der Schwäche des deut-

schen dreigliedrigen Schulsystems im Sinne Strucks hat das allerdings nichts zu tun.

Der *unmittelbare* Vergleich der PISA-Punktzahlen in Problemlösen und in Mathematik ist, wie schon in Kapitel 0 bemerkt, eigentlich unzulässig. Es könnte doch sein, dass, wenn man die Leistungen in den beiden Bereichen irgendwie „objektiv“ in Bezug zueinander setzen könnte, die Deutschen in Problemlösen sogar „schlechter“ als in Mathematik sind. – Außerdem könnte dieses Problemlösen doch sehr wohl ein Ergebnis schulischen Unterrichts sein, insbesondere im Fach „Mathematik“ mit seinem Anspruch einer allgemeinen Denkförderung. Ich will nun nicht behaupten, dass man dabei erfolgreich ist; aber ebenso wenig ist bewiesen, dass man keinen Erfolg hat.

Zweifelhaft ist natürlich, ob das Lösen dieser Knobelaufgaben überhaupt eine eigene, hervorhebenswerte Kompetenz anzeigt. Müssten dazu nicht noch ganz andere Aufgaben herangezogen werden? Und noch weiter gehend: so wenig wie Mathematical Literacy lässt sich m.E. Problemlösen letztlich in Tests wie PISA & Co prüfen. Meine diesbezügliche Argumentation zu Mathematical Literacy lässt sich wörtlich auf das Problemlösen übertragen.

5 Das didaktischen Belangen fern stehende Testmodell von PISA & Co

Auf internationaler Ebene wurden die Punkteskalen für die Testleistungen der Probandinnen & Probanden (*Leistungsskalen*) in den verschiedenen Inhaltsbereichen in Stufen eingeteilt, und zwar im Prinzip willkürlich („dividing … into levels, though useful for communication …, is essentially arbitrary“, O00.2, 197). Bei PISA 2000 wurde dabei (angeblich wegen der geringen Anzahl von Mathematikaufgaben) auf eine inhaltliche Beschreibung der Stufen verzichtet und eine solche den nationalen Gruppierungen überlassen (P00.1, 159 f). Es wäre einmal interessant zu verfolgen, was die etwa 30 bzw. 40 PISA-Länder aus dieser Möglichkeit gemacht haben, insbesondere ob die Anderen zum selben Kategoriensystem mit derselben inhaltlichen Füllung gekommen sind wie die Deutschen, wo man jedenfalls Großes damit vorhatte.

Eine gemeinsame Skala für Probandinnen-&-Probanden-Leistungen und für Aufgaben-Schwierigkeiten

Aus den Testleistungen der Probandinnen & Probanden lassen sich leicht *Schwierigkeitsskalen für die Aufgaben* ermitteln (nicht vergessen: wie viele Begriffe bei PISA & Co ist auch „Schwierigkeit" nicht inhaltlich, sondern rein statistisch zu verstehen). Für jede Aufgabe wird die relative Häufigkeit derjenigen, die sie nicht gelöst haben, auf dem Intervall [0;1] notiert. Je höher darauf eine Aufgabe angesiedelt ist, desto schwerer ist sie (zumindest bei dieser Stichprobe). Bei hoher Repräsentativität (wie sie bei PISA & Co i. A. wohl gegeben ist) kann man sogar (statistisch!) von *der* (testunabhängigen) Schwierigkeit einer Aufgabe bei einer bestimmten Population reden (bei PISA 2003 die 15-Jährigen in den teilnehmenden Ländern).

Die Dualität zwischen der Probandinnen-&-Probanden-Leistungsskala und der Aufgaben-Schwierigkeitsskala lässt sich aber nicht ohne Weiteres auf diesen Schluss von der Stichprobe auf die Grundgesamtheit übertragen. Während die Probandinnen-&-Probanden-Grundgesamtheit bei PISA & Co jeweils feststeht, gilt das Entsprechende für die Aufgaben-Grundgesamtheit nicht. Das obige Kapitel 4 behandelt genau diese Problematik. Über Mathematical Literacy (bzw. „mathematische Grundbildung") gibt es wohl einen gewissen Grundkonsens, aber schon nicht mehr darüber, was vielleicht über die o. a. Definition von Mathematical Literacy hinaus noch dazu gehört, und erst recht nicht darüber, *wie* bzw. *ob* überhaupt Mathematical Literacy in einem Test wie PISA abgeprüft werden kann. – Man beziehe sich also tunlichst nur auf die jeweilige konkrete Aufgabenkollektion, auch wenn wir aus Erfahrung wissen, dass die Ergebnisse bei anderen Kollektionen ähnlich ausfallen würden; – aber eben nicht hinreichend sicher hinreichend ähnlich, um angesichts von Ländervergleichen und weiteren subtilen Aussagen von *der* (testunabhängigen) mathematischen Leistung von Probandinnen & Probanden reden zu können. In der Filosofie von PISA & Co wird das anders gesehen, und es werden (unter dem Einfluss der test-orientierten Psychologie) auch andere Wörter verwendet, z. B. „Kompetenz" statt „Leistung". Dieser Konflikt zwischen den fachdidaktischen und statistischen Belangen und Schlussweisen ist in

verschiedenen Facetten Thema dieses Kapitels 5 (vgl. auch Meyerhöfer, 2005, und Jahnke, 2005, in seinem Vortrag).

„Kompetenz" als Disposition einer Person kann auch auf Aufgaben bezogen gesehen werden: Zur Lösung einer Aufgabe sind bestimmte „Kompetenzen" erforderlich, z. B. stochastisches Denken, Umgang mit dem Taschenrechner, mathematisches Modellieren einer außermathematischen Situation usw. Diese Sichtweise legt nahe, die Probandinnen-&-Probanden-Leistungs- und die Aufgaben-Schwierigkeits-Skala zu vereinigen. Dies wird bei PISA & Co in der Tat gemacht, und zwar nach folgendem Prinzip: Man zerlegt die Stichprobe S in nichtleere Schichten S_t, jeweils bestehend aus den Probandinnen & Probanden mit genau t Leistungspunkten. Für jede Aufgabe A wird nun für jede solche Schicht S_t der Anteil $p_A(t)$ der Probandinnen & Probanden in dieser Schicht ermittelt, die A lösen. Man unterstellt noch, dass für jede Aufgabe A die Funktion p_A (i. W. streng) monoton wächst, d. h. dass gilt: Ist $u>t$, dann löst in der Schicht S_u ein größerer Anteil der Probandinnen & Probanden die Aufgabe A als in der Schicht S_t. Man legt nun einen Schwellenwert p_0 fest (bei PISA 62%) und ermittelt für jede Aufgabe A die kleinste Punktzahl t, für die $p_A(t)\geq p_0$ ist, so dass man also sagen kann: Diese Aufgabe wird von 62% derjenigen Probandinnen & Probanden gelöst, die t Punkte erreicht haben (und von denen mit mehr als t Punkten mit einem höheren Anteil). Dann erhält diese Aufgabe die Punktzahl t zugewiesen.

Aus mathematischer Sicht muss man bei dieser Definition noch Sorge tragen für mögliche Randfälle, die allerdings praktisch irrelevant sein dürften. Der Schwellenwert von 62% geht auf einen Wert von 65% zurück, der „nach internationaler Absprache … Lösen … mit ‚einiger Sicherheit'" bedeutet und deswegen z. B. in TIMSS 1995 verwendet wurde (T95.2, 67). Wegen der geringen Anzahl von 31 Mathematikaufgaben bei PISA 2000 wurde er dort auf 62% vermindert und bei der deutschen Ergänzung trotz deren Umfang von 117 Aufgaben nicht erhöht (persönliche Mitteilung von Detlef Lind).

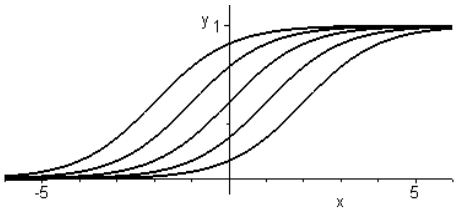
Die (für Testleute) ideale Testbatterie besteht aus Aufgaben, für die eine schärfere Monotoniebedingung als die o.a. existiert: Die Aufgaben lassen sich in einer Folge A_1, A_2, …, A_m anordnen, so dass für jede Probandin, jeden Probanden P gilt: Löst P die Aufgabe A_k, dann löst P auch sämtliche Aufgaben A_j mit $j\leq k$. Aus dieser

Bedingung folgt direkt die o.a. Monotonie, und man kann die Aufgabenpunktzahlen genau wie oben definieren.

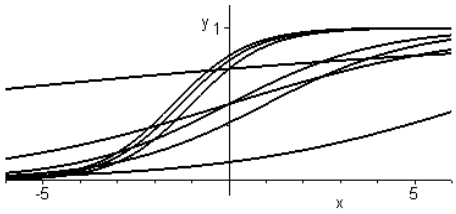
Um diese Ideen deutlich zu machen, habe ich mit Lösungshäufigkeiten usw. bei konkreten Tests argumentiert. Tatsächlich will man aber bei PISA & Co (nicht nur dort) Aufgaben unabhängig von realisierten Tests charakterisieren und verwendet dazu ein probabilistisches Testmodell, wo zu jeder Aufgabe eine bestimmte Funktion gehört, nämlich die, die jeder Probandinnen-&-Probanden-Punktzahl die Wahrscheinlichkeit zuordnet, dass Probandinnen & Probanden mit dieser Punktzahl die Aufgabe lösen (sog. Aufgabencharakteristik): Auch hier wird Monotonie unterstellt, d. h. zu höherer Probandinnen-&-Probanden-Punktzahl gehört eine höhere Lösungswahrscheinlichkeit.

Diese Monotonie ist nicht denknotwendig (noch nicht einmal die o. a. Monotonie bezüglich der Punkteschichten), und in realen Tests wird sie immer wieder verletzt (mündlicher Bericht von Norbert Knoche), wo also bestimmte Aufgaben von insgesamt schwächeren Probandinnen & Probanden häufiger gelöst werden als von stärkeren, möglicherweise, weil sie unbefangener zu Werke gehen und diese Unbefangenheit bei solchen Aufgaben hilfreich ist (wieder ein Beispiel für Meyerhöfers, 2003 ff, Konstrukt der inhaltsunabhängigen Testfähigkeit). (S. dazu auch die gut verständliche Darstellung in Kleine, 2004, 87 ff.) Man wüsste gerne, ob solche Phänomene bei PISA & Co auch aufgetreten sind; Monotonie bei allen Aufgaben wäre angesichts der großen Probandinnen-&-Probanden-Zahlen aber durchaus plausibel.

In starker Idealisierung werden bei PISA & Co (auf Rasch zurückgehend) für alle Aufgaben logistische Charakteristiken mit einheitlichen Parametern, also der Form $1/(1+\exp(c-t))$, angesetzt (Lind 1994, 279 ff), so dass sie (bis auf waagrechte Verschiebung) alle denselben Graf haben (s. Abb. 1a). Das Schaubild eines Bruchrechentests (Lind 1994, 317) zeigt, wie eine solche Kollektion realistischerweise auch aussehen kann (s. Abb. 1b) (es könnten, wie gesagt, auch noch Berge und Täler vorkommen). Natürlich kann man seine Testaufgaben so auswählen und gestalten, dass man dem idealen Bild näher kommt. Wie weit das ohne Verlust von fachdidaktischer Substanz möglich ist, steht jedoch dahin; die oben analysierte Aufgabe „Gehen“ zeigt das ganze Dilemma auf. Das Rasch-Modell je-



(a)



(b)

Abbildung 1. Die beiden Abbildungen sind nicht Kopien der Originale aus (Lind 1994), sondern jenen nur nachempfunden. Auch jene stellen ja starke Idealisierungen real aufgetretener Lösungshäufigkeiten dar.

denfalls war ursprünglich für Tests mit primitiv strukturierten Aufgaben und nicht für so etwas Komplexes wie die Untersuchung von Mathematical Literacy gedacht. In Lind (1994, 283, 303 ff, 316 ff), Knoche & Lind (2000), Knoche u. a. (2002) wird die beschränkte Eignung des Rasch-Modells wiederholt angedeutet.

Konsterniert hat mich an der Kalibrierung der Aufgabenschwierigkeit die Tatsache, dass nicht i. W. sämtliche Probandinnen-&-Probanden-Punktzahlen herangezogen wurden, sondern aus jedem OECD-Land gleichgewichtig je 500 Probandinnen & Probanden, d. h. etwa ein Zehntel (P00.1, 51, 520 f, O00.2, 105). Hierbei haben also die Jugendlichen aus Island dasselbe Gewicht wie die aus den USA, obwohl diese für fast 1000-mal so viele stehen (O00.2, 135 f), und z. B. die aus dem Nicht-OECD-Staat Brasilien haben das Gewicht 0. Ich konzediere, dass sich die resultierenden Über- und Unterschätzungen, jedenfalls innerhalb der OECD, wohl ungefähr ausgleichen. Womöglich trifft diese Ausgleichsannahme auch auf

die Verfälschung der „wahren“ Aufgabencharakteristiken durch die Verwendung des Rasch-Modells zu.

Das Kompetenzstufenmodell der PISA-Deutschland-Mathematik-Gruppe

Zweck dieser Identifizierung der Probandinnen-&-Probanden-Leistungsskala mit der Aufgaben-Schwierigkeitsskala war nicht zuletzt die simultane Verwendung der Stufung, das sog. Kompetenzstufenmodell, das die PISA-Deutschland-Mathematik-Gruppe ja gerade für die Analyse der Aufgaben ausbeuten wollte. Die Stufen waren im Prinzip folgendermaßen festgesetzt worden: Zunächst wurde ein bestimmter Punktewert als Untergrenze von Stufe I ausgewählt, und von da aus wurden nach oben vier Stufen in folgender Weise festgelegt: die Obergrenze einer Stufe ist dadurch bestimmt, dass die Probandinnen & Probanden, die auf der Untergrenze dieser Stufe angesiedelt sind, 50% aller Aufgaben auf dieser Stufe lösen. Diese Obergrenze ist zugleich Untergrenze der nächsten Stufe. Außer den vier begrenzten gibt es noch je eine nach oben und nach unten offene Stufe (letztere von mir „Stufe 0“ genannt), insgesamt also 4+2. Das Rasch-Modell für die Aufgabencharakteristiken impliziert nun zusammen mit dieser Definition, dass alle vier begrenzten Stufen gleichbreit sind, bei PISA-2000-Mathematik mit der Breite *91,75* und den Grenzen *329*, *421*, *512*, *604*, *696* (P00.1, 160).

Solche Schwierigkeits-Leistungs-Stufungen wurden für sämtliche TIMSS-, PISA- und IGLU-Studien aufgestellt, und es wurde versucht, sie inhaltlich zu füllen. Aber sobald mehr ausgesagt wird, als dass die Aufgaben mit zunehmender Punktzahl schwerer werden, tauchen immer wieder „widerspenstige“ Aufgaben auf, die mit ihrer tatsächlichen Schwierigkeitspunktzahl in einer offensichtlich falschen Stufe landen. Beispiel:

Die IGLU-Naturwissenchaften-Aufgabe „Welches Tier säugt seine Jungen? Huhn, Frosch, Affe, Schlange?“ gerät mit *474* Punkten in die Stufe III „Anwenden naturwissenschaftsnaher Begriffe“. Sie gehört weder hier hinein, noch in II „Anwenden alltagsnaher Begriffe“ (*401–468*), sondern in I „Einfache Wissensreproduktion“ (*323–400*) oder gar in 0 „Vorschulisches Alltagswissen“ (s. I01.1, 156 ff) (je nach den Fernseherfahrungen der Kinder).

Die inhaltlichen Beschreibungen bleiben notgedrungen, nicht nur hier, trivial. In der internationalen Mathematikskala bei PISA 2000 etwa ist die Stufenabfolge durch zunehmende Leistungen beim Modellieren (inkonsistent von „nicht“ über „elementar“, „SI-Niveau“, „umfangreich“ bis „komplex“) mit nicht-konsistenten und nichtssagenden Zusätzen („begriffliches Verknüpfen“, „anspruchsvolle Begriffe“, „innermathematisches Argumentieren“) geprägt (P00.1, 168). Diese von mir wörtlich zitierten Schlagwörter stellen Kurzfassungen aus dem PISA-Bericht selbst dar. Die ausführlicheren Beschreibungen (P00.1, 160) wiederum sind nicht genügend operational. Verbindet man die Eigenschaften einer Stufe mit einem logischen „und“, werden viele Aufgaben nicht erfasst, verbindet man sie mit „oder“, verliert die Stufe ihre Identität. Bei PISA 2000 ist viel die Rede von „Modellbildung“, und genau diese wird in unserer Vergleichsuntersuchungen-Kommunität insgesamt so inflationär gebraucht, dass sie banal wird. Wie von IGLU (in Deutschland ja sehr affin zu PISA) explizit eingeräumt (I01.2, 118 f), wird sie nämlich mit jeglichem Bearbeiten von Aufgaben identifiziert.

Die PISA-Deutschland-Mathematik-Gruppe hat die Problematik früh erkannt. Für sie ist sie deswegen besonders relevant, weil sie mit dem Schwierigkeits-Leistungs-Modell ehrgeizigere Pläne hat: Es soll ermöglichen, dass man für jede Aufgabe aufgrund einer fachdidaktischen Analyse der erforderlichen Kompetenzen und weiterer Faktoren voraussagen kann, auf welcher Stufe sie landet. Damit sollen die Basis für wissenschaftliche, klare Vorgaben für die Konstruktion von schulmathematischen Tests gelegt werden, dieser Zweig der Mathematikdidaktik vom Ruch der Beliebigkeit und des Laientums befreit werden und nicht zuletzt Aufgabenkollektionen entstehen, die von vorneherein vielleicht sogar dem Idealbild mit breit gestreuten einheitlichen Aufgabencharakteristiken besser entsprechen.

Weitere Kategorien: „Arten“ bzw. „Typen mathematischen Arbeitens“ und „inhaltliche Teilbereiche“ bzw. „Big Ideas“

So hat die PISA-Deutschland-Mathematik-Gruppe als zusätzliche Kategorie die „Art“ bzw. den „Typ mathematischen Arbeitens“ mit

„den“ drei Ausprägungen „technische Aufgaben“, „rechnerische“ und „begriffliche Modellierungs- und Problemlöseaufgaben“ eingeführt (z. B. Neubrand 2004, 88 ff). Im Gegensatz zur Kategorie „PISA-Punktestufen“ ist diese nun weich mit unklarer Begrifflichkeit, fließenden Übergängen, weiten Überschneidungsbereichen und vermutlich großen Lücken.

Versucht man dennoch, Aufgaben einzusortieren, so wird man oft feststellen, dass verschiedene Aufgabenteile zu verschiedenen Ausprägungen gehören. – Oder: Für den Einen handelt es sich bei der Anwendung einer Formel vielleicht um eine rein „technische Aufgabe“, für die Andere bedarf es dazu möglicherweise einer aufwändigen Herleitung, also eines ganz anderen „Typs mathematischen Arbeitens“. – Und: Auch aufgrund unterschiedlicher Lösungswege können mit ein- und derselben Aufgabe (sogar ein- und demselben Aufgabenteil) unterschiedliche „Typen mathematischen Arbeitens“ angesprochen sein.

Einen Ausschnitt aus dieser Problematik liefert exemplarisch die 31-Pfennig-Aufgabe (Neubrand 2004, 89, *797* Punkte):

> Wie kannst du einen Geldbetrag von genau 31 Pfennig hinlegen, wenn du nur 10-Pfennig-, 5-Pfennig- und 2-Pfennig-Münzen zur Verfügung hast? Gib *alle* Möglichkeiten an.

Sie wird unter „begriffliches Modellieren und Problemlösen“ eingeordnet. Ich kann diese Einordnung nicht nachvollziehen. Es handelt sich doch um eine begrifflich völlig anspruchslose Abzählaufgabe an vorgestellten oder aufgezeichneten konkreten Objekten, und die Schwierigkeit liegt in der Erfassung aller Fälle.

Analog die Pyramiden-Aufgabe (P00.1, 151 ff, *810* Punkte):

> Die Grundfläche einer Pyramide ist ein Quadrat. Jede Kante der skizzierten Pyramide misst 12 cm. *(Zeichnung)* Bestimme den Flächeninhalt einer der dreieckigen Seitenflächen. Erkläre, wie du eine Antwort gefunden hast.

Man muss nur den Flächeninhalt eines gleichseitigen Dreiecks bei bekannter Seitenlänge a = 12 cm ermitteln, und der beträgt $\sqrt{3}/4 \cdot a^2$ = 62 cm^2. So gesehen, ist das eine reine Wissensaufgabe, und eine „komplexe Modellierung und innermathematisches Argumentieren“ vermag ich nicht zu erkennen. – Eine Schwierigkeit liegt

darin, dass i.d.R. keine Formeln auswendig gewusst werden. Die andere Schwierigkeit resultiert daraus, dass die Dreiecke in einer dreidimensionalen Situation gegeben sind und weltweit zu wenig Raumgeometrie getrieben wird, so dass bei den meisten Probandinnen & Probanden schon deswegen die Klappe fällt, obwohl es sich um ganz gewöhnliche (notwendig ebene) gleichseitige Dreiecke handelt. Wenn man in der Schule intensiv solche dreidimensionalen Situationen behandelt hat, fällt einem diese Sichtweise leicht; und wenn nicht, dann eben nicht.

Offensichtlich ist bei vielen Aufgaben die Kategorie „inhaltliche Teilbereiche“ (≈ „Big Ideas“) hoch-relevant und die Schwierigkeit rührt oft daher, dass es sich etwa um ungewohnte Kombinatorik, unbewältigte Raumgeometrie oder unverstandene Stochastik handelt.

In PISA 2003 war ja Mathematik der Schwerpunkt und mit 84 (statt 31) Aufgaben vertreten (P03.1, 51). Deswegen wurde dieses Fach noch einmal zerlegt, allerdings nicht etwa in die Typen mathematischen Arbeitens (die tauchen im internationalen Bericht P03.1 gar nicht auf), sondern in die vier „Big Ideas“: „Quantität“, „Veränderung und Beziehungen“, „Raum und Form“ und „Unsicherheit“ mit vier eigenen Länderrangfolgen, wenn auch (ohne inhaltliche Begründung) mit gemeinsamer Stufung (komplett in O03 veröffentlicht). Es gibt jetzt eine Stufe mehr, also 5+2 statt 4+2, was natürlich die Willkür dieser Stufung unterstreicht und die Bemühungen um eine Kanonisierung auch PISA-endogen konterkariert.
So relevant diese Unterteilung in „Big Ideas“ zu sein scheint, so unscharf ist auch sie, und in der Mathematikdidaktik weiß man dies schon lange. „Raum und Form“ spielt fast überall eine Rolle, nämlich bereits, sobald es um das Anfertigen oder Interpretieren eines Funktionsgrafs geht. Vergleichbares gilt sowohl für „Quantität“, als auch für „Veränderung und Beziehungen“.

Bei „Unsicherheit“ dagegen besteht das Problem, dass es bei den gern gestellten Aufgaben aus der Beschreibenden Statistik oft gar nicht um Unsicherheit geht, und aus diesem Grunde wurde ja (in P03.1, 49) die Überschrift „Daten und Zufall“ vorgeschlagen (aber im Bericht nicht realisiert). Diese beiden Ideen wiederum werden zwar aus mathematiksystematischen Gründen üblicherweise gemeinsam behandelt; der epistemologische und psychologische

Umgang mit ihnen ist jedoch völlig unterschiedlich, und das Thema „Daten“ gehört m. E. in die Bereiche „Quantität“ bzw. „Veränderung und Beziehungen“. Die PISA-Leute scheinen das auch zu ahnen, und sie haben bei den beiden strukturgleichen Statistikaufgaben *Größer werden 2* und *Raubüberfälle* den Kompromiss geschlossen, die erste bei „Veränderung und Beziehungen“ und die zweite bei „Unsicherheit“ einzusortieren. Bei beiden sind Daten in Grafen repräsentiert, bei der ersten „muss eine Graphik interpretiert werden“, bei der zweiten „muss eine Graphik verständig interpretiert werden“ (P03.1, 54 f).

Unter den kognitionsbezogenen Ansätzen scheint der von Cohors-Fresenborg, Sjuts & Sommer (in Neubrand 2004, 109 ff) am besten fundiert zu sein. Wie weit er zum „Kompetenzstufenmodell“ passt, steht dahin; einen direkten Niederschlag konnte ich nicht beobachten (P03.1, 61).

Die unerwünschte, aber unerlässliche Kategorie der Aufgaben-Lösungswege

Mit Händen und Füßen (z. B. auf der Tagung des Arbeitskreises „Vergleichsuntersuchungen“ in der GDM am 26.11.04 oder in Lind u. a., 2005) wehrt sich die PISA-Deutschland-Mathematik-Gruppe aber gegen eine wirklich erforderliche Differenzierung, wie sie Meyerhöfer (u. a. 2004b) ins Spiel gebracht hat: Je nach Lösungsweg kann eine Aufgabe unterschiedliche Kompetenzen erfordern, vielleicht zu verschiedenen Typen mathematischen Arbeitens (bzw. „Kompetenzklassen“) gehören und bei entsprechend differenzierter Auswertung auf verschiedenen Kompetenzstufen landen, und zwar unabhängig davon, ob verschiedene Typen mathematischen Arbeitens involviert sind oder nicht. Dabei spielt es keine Rolle, in welchem Umfang bei einem bestimmten Testdurchgang die diversen Lösungswege überhaupt benutzt wurden (wobei aus mathematikdidaktischer Sicht deren tatsächliche Verteilungen jeweils hochinteressant wären). Meyerhöfer hat für alle Aufgaben aus PISA 2000 analysiert, wie stark Aufgabenanforderungen infolge unterschiedlicher Lösungswege differieren, und die Analysen der nicht geheimen Aufgaben publiziert (2004a, 2005). Sein Befund stellt die Zuordnung einer jeden *Aufgabe* zu einer bestimmten Punktzahl und

damit das Paradigma einer eindimensionalen Skala, die es erlauben würde, mehr als die Schwierigkeit (= Lösungshäufigkeit) abzulesen, absolut in Frage.

Meyerhöfer versucht nun gerade (von Lind u. a., 2005, völlig verkannt; s. dazu Bender, 2005d), das Paradigma der eindimensionalen Aufgaben-Schwierigkeitsskala zu retten, indem er es erweitert und sinngemäß feststellt, dass man statt Aufgaben: *Paare von Aufgaben und Lösungswegen* betrachten müsste. Durch die Arbeit mit den Testheften wäre dieses Vorgehen prinzipiell praktikabel, ohne dass die Probandinnen & Probanden unnötig verwirrt würden. Es soll jedoch nicht verschwiegen werden, dass bei der Differenzierung nach Lösungswegen ähnliche Probleme auftreten wie beim Konstrukt der Typen des mathematischen Arbeitens. Zusätzlich würden Vorbereitung und Auswertung des Tests erheblich aufgebläht und erschwert. Der Vorschlag wird daher von niemandem ernsthaft gemacht, sondern soll lediglich die Problematik des „Kompetenzstufenmodells" von PISA & Co auf den Punkt bringen.

Insgesamt sind die Gesichtspunkte, die von der PISA-Deutschland-Mathematik-Gruppe im Zuge ihres Schwierigkeits-Leistungs-Stufen-Modells berücksichtigt werden, alle interessant und haben eine Rolle bei jeglicher Testkonstruktion zu spielen. Ich habe den Eindruck, dass von der PISA-Deutschland-Mathematik-Gruppe der ganze PISA-Komplex gründlicher durchdacht ist als von anderen Gruppierungen und dass sie von manchen Anfangssetzungen und fortwährenden Restriktionen gehemmt wird. Aber ihr Schwierigkeits-Leistungs-Stufen-Modell hat trotzdem nicht die Aussagekraft, die sie ihm zuspricht, da diesem die willkürliche Stufensetzung von Beginn an anhaftet und eine stringente wissenschaftliche Begründung nicht erkennbar ist.

6 Ausblick

Zukünftige PISA-Ergebnisse

Eine der zentralen Parolen des PISA-2000-Berichts, nämlich dass nur 44% der 15-Jährigen in Deutschland über den mathematischen Grundbildungsstandard verfügen (P00.1, 161), bedeutet nicht mehr

und nicht weniger, als dass diese 44% bei *jenem* Test mit *seinem* Auswertungs- und Berechnungsverfahren *512* oder mehr Punkte erreicht haben. – Diese „Erfolgs"-Quote ist weit von den Ansprüchen entfernt, die die Lehrerinnen & Lehrer, die Gesellschaft und nicht zuletzt die mathematikdidaktische Kommunität an den deutschen Mathematikunterricht stellt. Allerdings ist dieses schlechte Ergebnis keine Überraschung. Wer wissen wollte, konnte auch schon vorher wissen. Nach meiner Einschätzung werden die deutschen PISA-2006-Mathematik-Zahlen noch etwas höher liegen, da sich Teile unserer Lehrkräfte und in deren Gefolge Teile unserer Jugendlichen besser auf solche Tests einstellen. Dies nützt zwar den guten und den schlechten Probandinnen & Probanden nicht viel, aber in der Mitte, und zwar im unteren Gymnasialbereich, scheint da noch Verbesserungspotenzial zu existieren (so Klaus Klemm vom deutschen PISA-Beirat, FR, 08.12.04, bezogen auf den Vergleich von PISA 2000 mit 2003, wie er in P03.1, 86 ff, dargestellt ist).

Beim schwachen Viertel dürfte das Ende der Talfahrt noch nicht erreicht sein, und eine merkliche Leistungssteigerung wird sich dort bestenfalls langfristig einstellen. Eine Voraussetzung dafür ist, dass die schlecht integrierten Familien mit Migrationshintergrund sowohl durch aktive Maßnahmen, als auch durch die assimilierende Wirkung der Zeitläufte vor allem über die Jugendlichen spürbar besser in die Gesellschaft eingegliedert werden.

PISA und die Mathematikdidaktik

In der deutschsprachigen Mathematikdidaktik hatte die empirische Forschung schon immer einen geringeren Anteil als etwa im angelsächsischen Kulturkreis, und schon immer war die Sehnsucht virulent, unsere Ergebnisse nicht nur durch stoffdidaktische Analysen, (nicht-normierte) Erfahrungen sowie Erfahrungsberichte zu gewinnen, sondern empirisch, oder sie wenigstens empirisch abzusichern. Entsprechende Bemühungen waren von wenig Erfolg gekrönt: Bei den meisten quantitativen Untersuchungen fehlte die Repräsentativität, wurde das statistische Instrumentarium mangelhaft eingesetzt, vermisste man bei den Fragestellungen die Relevanz usw. Die in den 1970-er Jahren populär gewordenen qualitativen Verfahren waren durchaus vielversprechend. Aber ihre Verallgemeinerungs-

probleme in Verbindung mit einer Distanziertheit zum Stoff erwiesen sich als nicht förderlich. So setzte man schließlich einige Hoffnung in sehr breite Untersuchungen mit saubersten Methoden wie PISA & Co. Allerdings entgehen auch diese nicht einem grundsätzlichen Dilemma jeglicher empirischen Forschung in der Fachdidaktik: Das Lehren und Lernen etwa von Mathematik und die von der Fachdidaktik dazu entwickelte Begrifflichkeit sind viel komplexer als die Entsprechungen in den empirischen Wissenschaften wie Medizin, Psychologie, Ökonomie usw., für die die statistischen Methoden entwickelt wurden. Dort ist eine Aussage wie „Intelligenz ist, was der Intelligenz-Test misst“ durchaus eine gute Arbeitsgrundlage.

Genauso muss man auch PISA sehen: „Mathematische Kompetenz ist, was der PISA-Test misst“ oder „Problemlöse-Kompetenz ist, was der Problemlöse-Test misst“ usw. Man würde dann von „mathematischer Kompetenz im Sinne von PISA“ usw. reden, und Alle wüssten, was damit gemeint ist, jedenfalls ein stark reduzierter Begriff im Vergleich zu dem, was man sich als gewöhnliche Mathematikdidaktikerin oder gewöhnlicher Mathematikdidaktiker unter „mathematischer Kompetenz“, „Problemlösen“ oder „Mathematical Literacy“ vorstellt und wie man diese Begriffe auch für die tiefliegenden und komplexen Fragen des Lehrens und Lernens von Mathematik benötigt. Hier tut sich nun der fundamentale Konflikt auf: Obwohl zumindest die PISA-Deutschland-Mathematik-Aktiven ihre Wurzeln in der Stoffdidaktik haben und sich nach wie vor der mathematikdidaktischen Kommunität angehörig fühlen, benutzen sie die Begriffe ohne den Zusatz „im Sinne des PISA-Testmodells“ o.ä. und rufen damit bei ihren Adressatinnen & Adressaten eine Überschätzung der Begriffe hervor, weil diese ja mit dieser unangekündigten Einschränkung nicht rechnen, da mit ihnen ja so wie immer geredet wird. Oder aber die PISA-Deutschland-Mathematik-Aktiven überschätzen die mathematikdidaktische Aussagekraft von PISA wirklich und verstehen die Begriffe doch im tiefliegenden und komplexen Sinn; dann überschätzen sie das, was PISA misst, erheblich und verleiten ihre Adressatinnen & Adressaten, sich dieser Überschätzung anzuschließen. Beide Alternativen suggerieren jedenfalls eine Überschätzung der Reichweite der Aussagen von PISA & Co und sind geeignet, Kritik zu provozieren.

Obwohl nach meinen Erfahrungen der letzten fünf Jahre die große Mehrzahl der Kolleginnen & Kollegen unserer Kommunität dem ganzen Komplex von PISA & Co eher skeptisch gegenübersteht, hat dieser doch einen unübersehbaren Einfluss auch auf die Wissenschaft „Mathematikdidaktik“ gewonnen, nicht zuletzt wegen der gewaltigen Geld- und Personalmittel, die da eingesetzt werden, und der Medienwirksamkeit, die geschickt erzeugt wurde und immer wieder erzeugt wird. Sowohl diese Zuflüsse, als auch die Publizität gereichen der Mathematikdidaktik insgesamt natürlich auch zum Vorteil, aber es muss Sorge dafür getragen werden, dass nicht Paradigmen ein zu großes Gewicht erhalten, die eher in andere Wissenschaften gehören, aus denen aber die den Charakter bestimmenden Leiterinnen & Leiter größerer Projekte wie PISA und die Gutachterinnen & Gutachter über unsere DFG- und sonstigen Anträge kommen (s. dazu ausführlicher Wittmann, 2002).

7 Nachtrag anlässlich der zweiten Auflage

Für den Zeitraum Dezember 2005 bis Juni 2007 habe ich in (Bender 2006 und 2007) vor allem aufgrund von Zeitungsartikeln u.ä. einmal einige Beiträge, die sich auf PISA berufen oder auf die Kritik daran reagieren, zusammengestellt und glossiert. Es ist eine reichhaltige Sammlung von unfundierten Behauptungen, missbräuchlichen Interpretationen, gezielten Falschzitierungen bis hin zu persönlichen Diffamierungen entstanden.

Hinweisen möchte ich außerdem auf einige Schriften, die mir erst nach Beendigung des Textes für die 1. Auflage im Spätsommer 2005 bekannt wurden, die aber mit ihren grundlegenden (bildungs-) politischen Analysen bzw. ihrer breiten Aufarbeitung der Mängel von PISA und von dessen Rezeption ganz wesentliche Beiträge zu unserem Thema leisten: Huisken 2005, Karg 2005, Kraus 2005, Krautz 2007.

Literatur

Berichte

(I01.1, IGLU) Wilfried Bos, Eva-Maria Lankes, Manfred Prenzel, Knut Schwippert, Gerd Walther & Renate Valtin (Hrsg.) (2003): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster u. a.: Waxmann

(I01.2, IGLU) Wilfried Bos, Eva-Maria Lankes, Manfred Prenzel, Knut Schwippert, Renate Valtin & Gerd Walther (Hrsg.) (2004): IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. Münster u. a.: Waxmann

(O00.1, PISA 2000) OECD (Hrsg.) (2002): Manual for the PISA 2000 Database. Paris: OECD

(O00.2, PISA 2000) Ray Adams & Margaret Wu (Hrsg.) (2002): PISA 2000 Technical Report. Paris: OECD

(O03, PISA 2003) OECD (Hrsg.) (2005): Lernen für die Welt von morgen – erste Ergebnisse von PISA 2003. Heidelberg u. a.: Spektrum

(P00.1, PISA 2000) Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Petra Stanat, Klaus-Jürgen Tillmann & Manfred Weiß (= Deutsches PISA-Konsortium) (Hrsg.) (2001): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich

(P00.2, PISA 2000) Jürgen Baumert, Cordula Artelt, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann & Manfred Weiß (= Deutsches PISA-Konsortium) (Hrsg.) (2002): PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich. Opladen: Leske + Budrich

(P00.3, PISA 2000) Jürgen Baumert, Cordula Artelt, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann & Manfred Weiß (= Deutsches PISA-Konsortium) (Hrsg.) (2003): PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland im Vergleich. Opladen: Leske + Budrich

(P03.1, PISA 2003) Manfred Prenzel, Jürgen Baumert, Werner Blum, Rainer Lehmann, Detlev Leutner, Michael Neubrand, Reinhard Pekrun, Hans-Günter Rolff, Jürgen Rost & Ulrich Schiefele (PISA-Konsortium Deutschland) (Hrsg.) (2004): PISA 2003 – der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster u. a.: Waxmann

(P03.2, PISA 2003) Manfred Prenzel, Jürgen Baumert, Werner Blum, Rainer Lehmann, Detlev Leutner, Michael Neubrand, Reinhard Pekrun, Jürgen Rost & Ulrich Schiefele (PISA-Konsortium Deutschland) (Hrsg.) (2005): Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche? Münster u. a.: Waxmann

(T95.1; TIMSS 1995) Ina V.S. Mullis, Michael O. Martin, Albert E. Beaton, Eugenio J. Gonzales, Dana L. Kelly & Teresa A. Smith (1997): Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study. Chestnut Hill, MA, USA: Boston College

(T95.2, TIMSS 1995) Jürgen Baumert, Rainer Lehmann u. a. (1997): TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen: Leske + Budrich

(T95.3, TIMSS 1995) Jürgen Baumert, Wilfried Bos & Rainer Lehmann (Hrsg.) (2000): TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen: Leske + Budrich

(T95.4, TIMSS 1995) Jürgen Baumert, Wilfried Bos & Rainer Lehmann (Hrsg.) (2000): TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. Opladen: Leske + Budrich

(T99, TIMSS 1999) Ina V.S. Mullis, Michael O. Martin, Eugenio J. Gonzales, Kelvin D. Gregory, Robert A. Garden, Kathleen M. O'Connor, Steven J. Chrostowski & Teresa A. Smith (2000): TIMSS 1999 International Mathematics Report. Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade. Chestnut Hill, MA, USA: Boston College

(T03, TIMSS 2003) Ina V.S. Mullis, Michael O. Martin, Eugenio J. Gonzales & Steven J. Chrostowski (2004): TIMSS 2003 International Mathematics Report. Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. Chestnut Hill, MA, USA: Boston College

Weitere Literatur

Baumert, Jürgen, Eckhard Klieme, Manfred Lehrke & Elwin Savelsbergh (2000): Konzeption und Aussagekraft der TIMSS-Leistungstests. In: Die Deutsche Schule 92, 103–115 & 196–217

Bender, Peter (2003): Die etwas andere Sicht auf die internationalen Vergleichs-Untersuchungen TIMSS, PISA und IGLU. In: Paderborner Universitätsreden 89, 35–59

Bender, Peter (2004): Die etwas andere Sicht auf den mathematischen Teil der internationalen Vergleichs-Untersuchungen PISA sowie TIMSS und IGLU. In: Mitteilungen der DMV 12, Heft 2/2004, 101–108, zugleich Mitteilungen der GDM 78, ISSN 0722-7817

Bender, Peter (2005a): Die etwas andere Sicht auf PISA sowie TIMSS und IGLU. In: Beiträge zum Mathematikunterricht 2004. Hildesheim & Berlin: Franzbecker, 81–84

Bender, Peter (2005b): Neue Anmerkungen zu alten und neuen PISA-Ergebnissen und -Interpretationen. In: Beiträge zum Mathematikunterricht 2005. Hildesheim & Berlin: Franzbecker, 73–76

Bender, Peter (2005c): Die etwas andere Sicht auf PISA, TIMSS und IGLU. In: Der Mathematikunterricht 51, Heft 2/3, 36–57

Bender, Peter (2005d): PISA, Kompetenzstufen und Mathematik-Didaktik. In: Journal für Mathematik-Didaktik 26, 274–281

Bender, Peter (2006): Einige Anmerkungen zu PISA, PISA-Reaktionen und Reaktionen auf PISA-Reaktionen. In: Mitteilungen der GDM 82, 39–49, ISSN 0722-7817, http://math-www.uni-paderborn.de/bender/index.html (03.02.07 gesehen)

Bender, Peter (2007): Weitere Anmerkungen zu PISA, PISA-Reaktionen und Reaktionen auf PISA-Reaktionen. In: Mitteilungen der GDM 83, 22–30, ISSN 0722-7817

Borsche, Lorenz (2002): Das Fiasko der Forscher. www.borsche.de (03.02.07 gesehen)

Braams, Bas (2002): math.nyu.edu/mfdd/braams (03.02.07 gesehen)

Engström, Arne (2005): Mittelstadt 1977–1986–2002. Untersuchung mathematischer Fähigkeiten in Kl. 1–9. In: Beiträge zum Mathematikunterricht 2005. Hildesheim & Berlin: Franzbecker, 187–190

Flitner, Elisabeth (2006): Rationalisierung von Schulsystemen durch ‚public-private-partnership' am Besipiel von PISA. In: Jürgen Oelkers, Rita Casale, Rebekka Horlacher & Sabina Larcher Klee (Hrsg.): Rationalisierung und Bildung bei Max Weber. Beiträge zur historischen Bildungsforschung. Bad Heibrunn: Klinkhardt, 245–266

Freymann, Thelma von (2004): Bemerkungen zum finnischen Schulwesen. www.km.bayern.de/km/lehrerinfo/positionen/2004/01219/index.shtml (03.02.07 gesehen)

GDM (2005): Mitteilungen der GDM 80, ISSN 0722-7817

Hagemeister, Volker (1999): Was wurde bei TIMSS erhoben? Über die empirische Basis einer aufregenden Studie. In: Die Deutsche Schule 91, 160–177

Herrlitz, Hans-Georg (2003): Das große Tabu. PISA, IGLU und die Gesamtschulfrage. In: Die Deutsche Schule 95, 262–266

Huisken, Freerk (2005): Der „PISA-Schock" und seine Bewältigung. Wieviel Dummheit braucht/verträgt die Republik. Hamburg: VSA

Jahnke, Thomas (2005): Ideologiekritisches und Versöhnliches zu PISA & Co. In: Beiträge zum Mathematikunterricht 2005. Hildesheim & Berlin: Franzbecker, 267–270

Kaiser, Gabriele (2000): Internationale Vergleichsuntersuchungen im Mathematikunterricht – eine Auseinandersetzung mit ihren Möglichkeiten und Grenzen. In: Journal für Mathematik-Didaktik 21, 171–192

Karg, Ina (2005): Mythos PISA. Göttingen: V&R unipress

Kießwetter, Karl (2002): Unzulänglich vermessen und vermessen unzulänglich: PISA & Co. In: Mitteilungen der Deutschen Mathematiker-Vereinigung 10, Heft 4/2002, 49–58

Klafki, Wolfgang (1958): Didaktische Analyse als Kern der Unterrichtsvorbereitung. In: Die Deutsche Schule 50, 450–471

Kleine, Michael (2004): Quantitative Erfassung von mathematischen Leistungsverläufen in der Sekundarstufe I. Hildesheim & Berlin: Franzbecker

Knoche, Norbert & Detlef Lind (2000): Eine Analyse der Aussagen und Interpretationen von TIMSS unter Betonung methodologischer Aspekte. In: Journal für Mathematik-Didaktik 21, 3–27

Knoche, Norbert, Detlef Lind, Werner Blum, Elmar Cohors-Fresenborg, Lothar Flade, Wolfgang Löding, Gerd Möller, Michael Neubrand & Alexander Wynands (Deutsche PISA-Expertengruppe Mathematik, PISA-2000) (2002): Die PISA-2000-Studie, einige Ergebnisse und Analysen. In: Journal für Mathematik-Didaktik 23, 159–202

Kraus, Josef (2005): Der PISA Schwindel. Wien: Signum

Krautz, Jochen (2007): Ware Bildung. Schule und Universität unter dem Diktat der Ökonomie. München: Diederichs

Lind, Detlef (1994): Probabilistische Testmodelle. Mannheim u. a.: BI Wissenschaftsverlag

Lind, Detlef, Norbert Knoche, Werner Blum & Michael Neubrand (2005): Kompetenzstufen in PISA. In: Journal für Mathematik-Didaktik 26, 80–87

Meyerhöfer, Wolfram (2003): Testfähigkeit: Was ist das? In: Beiträge zum Mathematikunterricht 2003. Hildesheim & Berlin: Franzbecker, 441–444

Meyerhöfer, Wolfram (2004a): Was testen Tests? Objektiv-hermeneutische Analysen am Beispiel von TIMSS und PISA. Potsdam: Dissertation

Meyerhöfer, Wolfram (2004b): Zum Kompetenzstufenmodell von PISA. In: Journal für Mathematik-Didaktik 25, 294–305

Meyerhöfer, Wolfram (2005): Tests im Test: Das Beispiel PISA. Leverkusen: Barbara Budrich

NCTM (Hrsg.) (1989): Curriculum and evaluation standards for school mathematics. Reston, VA: NCTM 1989

NCTM (Hrsg.) (2000): Principles and standards for school mathematics. Reston, VA: NCTM 2000

Neubrand, Michael (Hrsg.) (2004): Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland – Vertiefende Analysen im Rahmen von PISA 2000. Wiesbaden: VS Verlag für Sozialwissenschaften

Reiss, Kristina & Günter Törner (2003): PISA 2000: Eine Klärung von Missverständnissen. In: Mitteilungen der Deutschen Mathematiker-Vereinigung 11, Heft 1/2003, 46–48

Rindermann, Heiner (2006): Was messen internationale Schulleistungsstudien? In: Psychologische Rundschau 57, Heft 2, 69–86

Winter, Heinrich (1975): Allgemeine Lernziele für den Mathematikunterricht? In: Zentralblatt für Didaktik der Mathematik 7, 106–116

Wittmann, Erich C. (2002): Falsch programmiert: Forschungsförderung im Bereich der Mathematikdidaktik in Deutschland. In: Mitteilungen der GDM 75, 62–65, ISSN 0722-7817