Peter Bender
Universität Paderborn EIM-Fakultät
bender@upb.de
Erscheint in: Der Mathematikunterricht 2005

Die etwas andere Sicht auf PISA, TIMSS und IGLU

0. Vorbemerkungen

In den letzten 24 Monaten habe ich mich durch viele tausend deutsch- und englisch- sowie einige französischsprachige Seiten von Berichten von, Erläuterungen zu und Folgerungen aus den internationalen Vergleichsuntersuchungen PISA (Programme for International Student Assessment), TIMSS (Third 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 ausgewogeneren Bild von diesen Studien beizutragen, werde ich einen gewissen 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 auf die Arbeiten (Meyerhöfer 2004a, 2005) mit ihrer grundsätzlichen erkenntnis- und wissenschaftstheoretischen Kritik sowie Analysen von vielen Aufgaben und sonstigen Einzelheiten. Ich habe diese Gesichtspunkte 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 usw. ein. Dabei mache ich wohl oder übel eigentlich unzulässige Vergleiche von 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.

1. Organisation und Filosofie von PISA (u.a.)

In der "Organisation for Economic Co-Operation and Development" (OECD) haben sich entwickelte, den USA verbundene Staaten und einige, die auf dem Weg dahin sind, zusammengeschlossen. Einer der Aufträge 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.

Von Seiten der Wissenschaft wird PISA von einem internationalen Konsortium und nationalen Konsortien in den Ländern verantwortet. Ein Großteil der Arbeit wird von fächerbezogenen internationalen und nationalen Expertinnen- & Experten-Gruppen geleistet, darunter die PISA-Deutschland-Mathematik-Gruppe ("PDMG"). In Deutschland gibt es zusätzlich einen Beirat und gehört auch die Kultusministerkonferenz zu den Auftrag- und Geldgeberinnen & -gebern. Die Statistik liegt international und national in den Händen professioneller Institutionen. Mit vielen Grundsätzen und Details hat man an TIMSS aus den 1990-er Jahren 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 federführend). IGLU seinerseits ist stark an PISA angelehnt.

In jedem Land werden jedes Mal 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). Die Zahlen sagen also 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 jedenfalls mit größter Vorsicht zu ziehen. Dies gilt erst recht für die Durchschnittspunktzahlen von Teilpopulationen, z.B. von den Ländern.

Die daraus 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 verständlicher Weise dennoch breite Resonanz gefunden. Aus ausländischer Sicht besonders hysterisch aufgeladen war die Atmosphäre in Deutschland schon bei TIMSS, wo 509, und bei PISA 2000, wo gar nur 490 Punkte erreicht wurden. Auf die 503 Punkte bei PISA 2003 reagierte man dann gelassener; die Erwartungen waren zwischenzeitlich bescheiden geworden.

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 usw.? usw.) und der Angepasstheit des jeweiligen nationalen Curriculums an die inhaltlichen Vorgaben von PISA usw. 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 (T2, 90, 98, Knoche & Lind 2000, 12).
— 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" (P2, 39).
— Bekanntlich bestand bei PISA 2000 in Hamburg und in Berlin im Gesamtschulbereich mancherorts eine ausgeprägte Verweigerungshaltung (die man ja verstehen kann), so dass in diesen beiden Bundesländern die Repräsentativität verfehlt wurde (P1.3, 32f).
— Die Niederlande bei PISA 2000 (O1.2, 186ff) und Großbritannien bei PISA 2003 (P2, 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 usw. erkannt worden oder nicht. Z.B. haben die meisten guten TIMSS-Länder (wie auch Schweden) bei PISA erheblich weniger Punkte erzielt; die osteuropäischen Staaten haben da so viel verloren, dass sie sogar ihre guten Rangplätze eingebüßt haben; extrem war der Rückgang außerdem bei Thailand von 522 (TIMSS) auf 417 (2003), deutlich der Anstieg bei Island von 487 (TIMSS) auf 514 (2000), Luxemburg von 446 (2000) auf 493 (2003), Polen von 470 (2000) auf 490 (2003), auch Deutschland von 490 (2000) auf 503 (2003). Eine Berg- und Talfahrt hat Neuseeland mit 508, 537, 523 gemacht. — Vieles davon ist nicht wirklich erklärt. — Ich erinnere an Griechenland, wo ja kürzlich erst herausgekommen ist, dass es jahrelang seine Staatshaushaltszahlen gefälscht hat, weil es das von der EU erlaubte Defizit von 3% des BIP jedes Mal deutlich überschritten hat, oder an die Sportnation DDR, die einen erheblichen Teil ihrer Erfolge geheimem systematischem Doping zu verdanken hatte.

Nachdem bei TIMSS der Bereich "Lesen" nicht Gegenstand gewesen war, wurde er bei PISA 2000 in den Mittelpunkt gestellt, während der Schwerpunkt bei PISA 2003 auf Mathematik lag und bei PISA 2006 auf den Naturwissenschaften liegen wird. 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 (P&P) 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 (O1.2, 157ff). Dieses Phänomen konnte zwar durch Korrekturen bei der Auswertung aufgefangen werden, es wirft aber einen ersten Schatten von Fragwürdigkeit auf die noch zu diskutierende Punkteskala für die Schwierigkeit der Aufgaben.

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. Da-

durch wurden innerdeutsche Vergleiche u.a. zwischen den Bundesländern und zwischen den Schulformen möglich: Dabei schnitt Bayern mit 516 Punkten am besten, Bremen mit 452 Punkten am schlechtesten ab (Hamburg und Berlin waren, wie gesagt, nicht in die Auswertung aufgenommen worden) (P1.2, 104). Insgesamt bildete sich ein deutliches Südost-Nordwest-Gefälle aus. Die vier Schulformen erreichten in Mathematik (P1.3, 273, in Klammer die von mir aus den Angaben in P2, 68, erschlossenen Werte für 2003): Gymnasium 574 (ca. 585), Realschule 501 (ca. 500), integrierte Gesamtschule 455 (ca. 458), Hauptschule 439 (ca. 412) Punkte.

Zusätzlich zum Bearbeiten der Aufgaben mussten die P&P Fragebögen ausfüllen, mit denen zahlreiche Merkmale erhoben wurden, wie Migrationshintergrund (MH; ist gegeben, wenn wenigstens ein Elternteil im Ausland geboren ist), sozialer Status, Bildungsnähe des Elternhauses (Anzahl der Bücher zu Hause; I1, 50) usw., insgesamt das sog. "soziale und kulturelle Kapital" (I1, 270, P1.1, 326). Darüber hinaus wurden durch Auswertung von allerlei zusätzlichen Quellen, z.B. Befragungen der Schulleitungen, viele weitere Daten ermittelt. Durch PISA usw. hat die Öffentlichkeit einen tiefen Einblick in die Struktur der deutschen Gesellschaft gewonnen (in vielen anderen Staaten vermutlich ähnlich), der so vorher nicht gegeben war.

An anderer Stelle dagegen strotzt PISA usw. von Geheimnissen: Schon bei TIMSS werden "die Leistungstests insgesamt ... nicht veröffentlicht, da (?) sie für weitere Forschungszwecke nutzbar sein sollen" (T2, 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 Leitungsentwicklungen viele exogene Einflussgrößen verantwortlich sind;
- der in dem sog. Mittelwerte-Standardfehler (T2, 88, Knoche & Lind 2000, 11f) zum Ausdruck kommenden Unsicherheit dieser Punktzahlen;
- der Abhängigkeit der Lösungshäufigkeit von Aufgaben von deren Ort im jeweiligen Testheft und im Testhefte-Ensemble (O1.2, 157ff);
- der geringen kurzfristigen (3 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 (zugegeben: 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 P&P nachgehen wollten. Noch nicht einmal zu den veröffentlichten 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 (O1.2) und des Database-Manuals (O1.1) nicht, die Rechnungen zu erschließen. Es hätte einmal konkret vorgeführt werden müssen, wie man für die P&P 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 sogar die Geheimnisse wahren, indem man fiktive P&P nur solche Aufgaben lösen lässt, die öffentlich gemacht sind.

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 usw. wird zwar ein objektiv erscheinender Stil gepflegt, auf willkürliche Setzungen hingewiesen und zur Vorsicht bei Vergleichen und bei Interpretationen gemahnt (z.B. T2, 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 verstanden werden können, z.B. "benachteiligen", 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, verbunden mit Stresserhöhung für alle Betroffenen, 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 usw. selbst finden sich auf allen Ebenen spekulative Elemente:

— Für den (zufällig deutschen) 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 "MH" systematisch zugunsten des Faktors "soziale Stellung" unterschätzt.
— Die PDMG bemüht sich, die willkürliche Einteilung der Leistungspunkteskala in Stufen durch das internationale Konsortium zu einem quasi naturgegebenen universellen mathematikdidaktischen Anaylseinstrument hochzustilisieren.

Ich war vorübergehend von der Publizität angetan, die Bildung, insbesondere mathematische Bildung, im Gefolge von PISA usw. 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.

2. Auf der Suche nach Ursachen für die Mittelmäßigkeit Deutschlands

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 (T2, 18f, 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. Eine entscheidende Ursache für diese Schwäche 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 usw. 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 Ein-

heitsschule 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 usw. "schlechte" — deutsche Schulsystem insgesamt für allerlei Mängel verantwortlich gemacht und eine umwälzende Veränderung desselben gefordert, wobei die Gegliedertheit als ein ausschlaggebende 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 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, muss also verändert werden. (Dieses Argumentationsgrundmuster wird im Artikel u.a. auf viele verfälschte Daten gestützt, und ich führe es deswegen hier aus, weil mir der Missbrauch von PISA usw. 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 Mitte der 90-er Jahre bei TIMSS 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-Primarstufe vergleichbar gemacht (s.a. T1). In Mathematik erzielte Deutschland hierbei 545 Punkte (gegenüber dem Spitzenland Singapur mit 625; I1, 209). — 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 1994, zum anderen und hauptsächlich waren beim Test der Deutschen nur Viertklässlerinnen & -klässler (V&V) beteiligt, während bei TIMSS auch die Drittklässlerinnen & -klässler dabei waren. Schränkt man den TIMSS-Datensatz auf V&V ein, beträgt der TIMSS-Durchschnitt nicht mehr 500, sondern 529 (T1, 24ff, I1, 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 usw. 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 usw. überhaupt nicht erfasst, und hohe Zahlen bei PISA usw. könnten gerade Ausweis eines mangelhaften Bildungssystems sein (Leistungsdruck, Engführung beim Lernen, Pauken für Tests usw.)!

Während die skandinavischen Länder (bis auf Finnland 544) sich (inzwischen) auf Augenhöhe mit Deutschland befinden (Island 515, Dänemark 514, Schweden 509, Norwegen 495), ist Ostasien die wahre Spitzenregion (Hongkong 550, Südkorea 542, Japan 534, Macau 527 und Singapur, das bei PISA nicht beteiligt, aber bei TIMSS und IGLU Spitze war; T2, 90, I1, 210). Diese Staaten haben zwar alle das Einheitsschulsystem; aber sie würden selbstverständlich diese guten Platzierungen auch mit einem gegliederten Schulsystem erreichen. Denn diese Leistungen beruhen doch 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 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 Tiger-Staaten gezogen. Hinzu kommt, dass dort sowie in Kanada, Australien und Neuseeland die Einwanderungsstruktur hohe Punktzahlen bei PISA usw. 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 P&P mit MH sogar verbessert werden (z.B. Singapur, Neuseeland bei den V&V in "Lesen"; I1, 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 usw. Diese können von PISA usw. gar nicht erfasst werden, während die meisten der untersuchten Faktoren dagegen zweitrangig sind. Dies gilt z.B. auch für die TIMSS-Videostudie, in der "die" Unterrichtsstile in Mathematik in Japan, USA und Deutschland verglichen wurden. Fachdidaktisch ist diese Studie hoch-interessant (s. T2, 215ff). Aber abgesehen davon, dass dabei von Repräsentativität keine Rede sein kann, ist eine Auswirkung der identifizierten Stile auf Punktzahlen von PISA usw. nicht nachgewiesen. Sie ist noch nicht einmal plausibel, weil ja das inhaltliche Paradigma der "Mathematical Literacy" (ML), zu dem die Unterrichtsstile mehr oder weniger gut passen, prinzipiell und praktisch nicht à la PISA usw. getestet werden kann (s. dazu Kap. 3).

Ein großes "Experiment", das nach meinem Dafürhalten die Wirkungsmächtigkeit des Faktors "Leistungsorientierung" auf die Zahlen von PISA usw. eindrucksvoll belegt, ist der Zerfall der Sowjetunion verbunden mit der Auflösung autoritärer Strukturen dort und überhaupt in Osteuropa. Dieser Umschwung hat im Laufe der 1990-er Jahren die Schulen dort voll erfasst und, pauschal gesprochen, zu einem Straffheitsabbau an Schulorganisation und im Unterricht, verbunden mit einer Lässigkeitszunahme bei allen Beteiligten geführt. Tschechien (564, 498, 516), Ungarn (537, 488, 490) und Russland (535, 478, 468) haben von TIMSS 1994 bis PISA 2000 erheblich verloren, und zwar, wohlgemerkt, an den Rängen. Die PISA-Leute erklären diesen Trend übrigens wirklichkeitsfremd gerade umgekehrt mit der Perpetuierung der "traditionell dort vorherrschenden Methode eines stark lehrergesteuerten, auf Kenntnis mathematischer Fakten ausgerichteten Unterrichts" (P1.1, 177), als ob die PISA-Fragen sich so stark von den TIMSS-Fragen unterscheiden würden, dass Faktenwissen nichts mehr nützt!

Die deutsche Mittelmäßigkeit bei TIMSS und IGLU ist die Folge vor allem des ausgeprägt schwachen Abschneidens des schwachen Viertels unserer Jugendlichen (P1.1, 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 (von der ich aktives Mitglied war und deren Ideale ich großenteils heute noch hochhalte) 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 usw. besonders ausgeprägt ist (P1.1, 319ff 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 MH und zu 21% für die anderen (Kindergartenbesuch, Vater-Erwerbstätigkeit, Familien-Umgangssprache usw.) bestimmt (P2, 274). Dass in Deutschland und in den älteren EU-Ländern alle diese Faktoren wiederum stark vom Faktor "MH" 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.

Insgesamt haben 20,7% (P2, 271) aller 15-Jährigen in Deutschland MH, in Westdeutschland über 25% der V&V (I1, 277), in westdeutschen Großstädten über 300 000 Einwohner im Durchschnitt 36%, in der alten DDR nur knapp 4% (P1.2, 190). P&P mit MH erbringen im Mittel erheblich schlechtere Leistungen, z.B. die V&V in Lesen 55 IGLU-Punkte weniger als die anderen (I1, 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 (P1.1, 378, P2, 257). Besonders drastisch sind die Auswirkungen in den Bundesländern Nordrhein-Westfalen und Bremen mit Quoten von 32% bzw. 41% Jugendlichen mit MH (P1.3, 247), wo dann in bestimmten Regionen in vielen Hauptschulklassen deren Anteil so hoch ist, dass sowohl die mit, als auch die ohne MH schlecht gefördert werden.

Zur Stützung der These von der Überlegenheit der Gesamtschule wurde (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 MH 75% *doppelten* MH haben, beträgt in Schweden dieser Anteil nur gut die Hälfte (P2, 257); und diese sind es ja, die besondere Schwierigkeiten haben (z.B. P1.3, 249). Rechnet man außerdem einmal sämtliche Jugendliche mit MH (also auch die leistungsstärkeren) heraus, liegt Deutschland mit 527 Punkten plötzlich deutlich *vor* Schweden mit 518 Punkten (P2, 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 ist, muss zugleich mitlesen, dass es die autochthone Bevölkerung benachteiligt. — Ich meine, dass diese Differenz von 24 Punkten in Deutschland (bei PISA 2000 betrug sie 23; P1.1, 245), die auf der Welt sonst nirgends so hoch ist, unabhängig von dem Vergleich mit Schweden ein Indikator für die bei uns besonders ausgeprägte Einwanderungsproblematik ist.

Um nicht falsch verstanden zu werden: Wenn man sich überhaupt auf das Zahlenwerk von PISA usw. einlässt, dann ist 503 genau die richtige Punktzahl; denn diese steht für die Leistungen, die die 15-jährigen Jugendlichen 2003 in Deutschland erbracht haben. Diese Zahl in Verbindung mit der großen Leistungsbandbreite konfrontiert uns schmerzlich mit den Fehlern und Versäumnissen unserer Einwanderungspolitik seit langem, die inzwischen in eine völlig unzulängliche Integration weiter Teile unter den eingewanderten Familien gemündet ist. Hier wurde eine gewaltige Aufgabe für die ganze Gesellschaft aufgetürmt, die keineswegs vom Bildungssystem allein zu schultern ist, und ein Verdienst von PISA usw. ist es, sie uns vor Augen geführt zu haben.

Während man aber in Schweden schon so weit resginiert hat, dass man in den größeren Städten Jugendliche mit MH in Mathematik (neben Schwedisch und Englisch das Hauptprüfungsfach) in ihrer Herkunftssprache, z.B. Arabisch, unterrichtet (Engström, 2005, in seinem Vortrag), hat man in Deutschland doch noch größere Integrierungs-Ambitionen. Wie weit diese 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.

3. Was soll bei PISA usw. 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 usw. überwiegend solche, bei denen die Antwort entweder richtig oder falsch ist, also 1 oder 0 Punkte ergibt. 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 gar nicht berücksichtigt werden: 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 jemand — außerhalb der Testsituation — den Flächeninhalt der Antarktis (PISA-2000-internatio-

nal; Neubrand 2004, 269) durch geometrische Operationen auf einer Landkarte bestimmt, statt im Lexikon nachzuschauen, kann man ihr oder ihm gewiss nicht Problemlösefähigkeit attestieren.

Wie bei allen Tests wird auch bei PISA usw. ganz wesentlich eine extrinsische Fähigkeit der P&P abgeprüft, nämlich herausfinden zu können, was die Aufgabenautorinnen & -autoren (A&A) 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 P&P. 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 P&P aus Deutschland da auf verschiedenen Ebenen allerdings leichter als etwa die aus Mexiko (385), Türkei (423), Brasilien (356) u.a.

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 O1.2, 57ff) Vorlage gebunden und beeinflusst wäre, sie anders formulieren würde. Bei meiner Arbeit an den IGLU-Mathematik-Aufgaben stellte ich an über der Hälfte der 102 auf deutsch gegebenen TIMSS-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 • 4 C. 6 • 4 • 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?"

"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 usw. sind aber eben nicht alle P&P 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 Fortschritte bei der sprachlichen und inhaltlichen Qualität der Aufgaben gemacht.

Noch ein Beispiel auf Oberstufenniveau (T3.2, 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." — Hier ist im deutschen Sprachgebrauch das Wort "symmetrisch" inkorrekt verwendet (allerdings rührt die Schwierigkeit dieser Aufgabe nicht allein davon; s. die vorzügliche Analyse von Kießwetter, 2002). Wenn in Schweden die Lösungshäufigkeit 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.

Hier drängt sich die Frage nach der Unterrichts- und Curriculumsvalidität auf. Während bei TIMSS die Validität bezüglich der Ländercurricula noch ein erklärtes Ziel war (T2, 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" (P1.1, 19). Als hierfür "in mancher Hinsicht vorbildlich" (P1.1, 25) werden die NCTM-Standards für den Mathematikunterricht (MU) zitiert (NCTM 1989, überarbeitet 2000). Das Konzept wird verkörpert durch die sog. "Mathematical Literacy" (ML), die von der OECD so definiert ist (P1.1, 23): "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."

Wie den meisten Kolleginnen & Kollegen aus der deutschsprachigen mathematikdidaktischen Kommunität sagt mir das Konzept der ML von PISA durchaus zu, und ich teile die schon von den TIMSS-Leuten vertretene Auffassung, dass der deutsche MU allzu sehr auf Faktenwissenserwerb und die Beherrschung von Verfahren zielt (T2, 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 PDMG hat diesen Grundmangel des PISA-Ansatzes erkannt, bei den nationalen Ergänzungsaufgaben von PISA 2000 das deutsche Curriculum stärker berücksichtigt und den ML-Begriff hin zu "mathematische Grundbildung" leicht abgewandelt (Neubrand 2004, 15ff).

Der ML-Begriff passt sehr wohl zur Tradition der deutschen bildungstheoretischen Didaktik, wie sie z.B. vom alten Wolfgang Klafki (1958) oder speziell zum MU von Heinrich Winter (1975) mit seinem Begriff der "Umwelterschließung" verkörpert wird. Allerdings enthält die Konzeption von PISA einen stärker pragmatischen Zug (P1.1, 19). Ich persönlich vermisse dabei z.B. die Rolle der Mathematik als Kulturgut.

Der dominierende Bestandteil von ML im Sinne von PISA ist die Kompetenz zum Modellieren (offenbar i.W. mit Problemlösen und sogar erklärtermaßen mit Aufgabenlösen zu identifizieren; I2, 118f). 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" (P1.1, 25).

Insgesamt kann ein Test wie PISA oder IGLU, zumal mit überwiegend Multiple-Choice-Aufgaben, der ML-Definition natürlich nicht gerecht werden (so schon Kießwetter). Kein einziger Aspekt 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 P&P 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 A&A immer wieder über ihre eigenen Füße stolpern. Durchweg gilt aber: ist eine Aufgabe den P&P schon einmal so oder in ähnlicher Form begegnet und erkennen sie, dass sie Gelerntes einsetzen können (was PISA usw. 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.

Meine Kritik richtet sich nicht gegen diese Art von Tests. Man kann so viele P&P weltweit nicht anders testen. Aber vom Testen der ML ist man viel weiter entfernt, als man glaubt. Vielleicht sind solides Wissen und Können (natürlich inklusive des Lernstoffs "Modellbildung") viel bedeutsamer für die ML, als die "internationalen" PISA-Leute annehmen.

Hier ein Beispiel (das mit seiner Realitätswidrigkeit nicht besonders schlecht, sondern typisch ist) aus TIMSS II und TIMSS III, wo es u.a. für "Social Utility" steht (T2, 73, 675 Punkte, T3.1, 164, 554 Punkte) (ob diese Aufgabe 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?"

Vermutlich wird erwartet: Bei A muss man 12·800 = 9600 *zeds*, bei B 110·90 = 9900 *zeds* zahlen. Die P&P 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 wäre 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 475·12+35·90 = 8850 *zeds*!). — Solche Überlegungen wären Teil der ML. Genau diese werden hier nicht erwartet und können nicht erwartet werden; sie waren auch von den A&A offensichtlich nicht angestellt worden.

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 usw. 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-Aufgaben entsprechende Mängel herausgearbeitet (auch wenn Baumert u.a., 2000, einige Relativierungen anbringen konnten). Ähnliche Analysen zu PISA 2003 u.a. findet man bei (Braams 2005). Abgesehen davon, dass die A&A von PISA usw. anscheinend hin und wieder selbst Lücken bei ihrer "Scientific Literacy" haben, muss man ihnen zugute halten, dass sie zum Zwecke der Genießbarkeit durch die P&P 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 eines Interviews mit dem Sprecher des deutschen Konsortiums bei PISA 2003, Manfred Prenzel, wurde in der "Zeit" (09.12.04) folgende Aufgabe als beispielhaft vorgestellt (P2, 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 kommt hier sogar doch noch die Neigung der Erdachse ins Spiel: Ist sie nämlich nicht geneigt, dann hält sich die Sonne am Nord- und am Südpol immer am Horizont auf, und an diesen beiden Orten findet nie ein Wechsel zwischen Tageslicht und Dunkelheit statt, während an allen anderen Orten Tageslicht und Dunkelheit immer genau 12 Stunden lang sind. 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 Rede von "in unseren Breiten" ins Spiel, die ja von Vielen nicht verstanden würde.

Die IGLU-Stromkreis-Aufgabe (V&V sollen bei einigen Stromkreisen in der "üblichen" schematischen Darstellung feststellen, wo Strom fließt; I1, 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?"

Vergleichbare Schwierigkeiten wie bei der ML hat man sich bei PISA 2003 mit dem Abtesten eines eigenen Bereichs "Problemlösen" aufgeladen. Ich gehe nicht auf das läppische sog. *dynamische* Problemlösen ein (das sowieso nur Teil des deutschen Ergänzungstests war; P2, 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 Nut-

zung von Graphiken, beschriebenen Ausgangslage ... ergebenden Problemstellung" (P2, 148f) besteht, hilft nicht weiter, weil der Begriff "Problemlösen" da i.W. durch sich selbst erklärt wird. Der Bedingung, "dass der Lösungsweg nicht unmittelbar erkennbar ist" (P2, 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. "Kinobesuch" (P2, 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" (P2, 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 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 (P2, 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 (P2, 157f). "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" (P2, 170f). — 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 (P2, 75), also dort das Potenzial doch ausgeschöpft, sogar leicht "überkompensiert" haben. Die Mathematikdidaktik weiß schon lange, dass in den anderen drei Bereichen, zumal bei Geometrie und Stochastik, Intensivierungsbedarf besteht. — Mit der Überlegenheit des Lebens als Lehrmeister und der Schwäche des deutschen dreigliedrigen Schulsystems im Sinne Strucks hat das allerdings nichts zu tun.

Der *unmittelbare* Vergleich der PISA-Punktzahlen in Problemlösen und in Mathematik ist eigentlich unzulässig. Es könnte doch sein, dass, wenn man die Leistungen in den beiden Bereichen wirklich 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 nicht noch ganz andere Aufgaben dazu herangezogen werden? Und noch weiter gehend: so wenig wie ML lässt sich m.E. Problemlösen letztlich in Tests wie PISA usw. prüfen. Meine diesbezügliche Argumentation zu ML lässt sich wörtlich auf das Problemlösen übertragen.

4. Das didaktischen Belangen fern stehende Testmodell von PISA usw.

Auf internationaler Ebene wurden die Punkteskalen für die Testleistungen der P&P (*Leistungsskalen*) in den verschiedenen Inhaltsbereichen in Stufen eingeteilt, und zwar im Prinzip willkürlich ("dividing ... into levels, though useful for communication ..., is essentially arbitrary", O1.2, 197). Bei PISA 2000 wurde dabei wegen der geringen Anzahl von Mathematikaufgaben auf eine inhaltliche Beschreibung der Stufen verzichtet und eine solche den nationalen Gruppierungen überlassen (P1.1, 159f). 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.

Aus den Testleistungen der P&P lassen sich leicht *Schwierigkeitsskalen für die Aufgaben* ermitteln. Für jede Aufgabe wird die relative Häufigkeit der P&P, 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 usw. 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 P&P-Leistungsskala und der Aufgaben-Schwierigkeitsskala lässt sich aber nicht ohne Weiteres auf diesen Schluss von der Stichprobe auf die Grundgesamtheit übertragen. Während die P&P-Grundgesamtheit bei PISA usw. jeweils feststeht, gilt das Entsprechende für die Aufgabengrundgesamtheit nicht. Mein Kapitel 3 behandelt genau diese Problematik. Über ML (bzw. "mathematische Grundbildung") gibt es wohl einen gewissen Grundkonsens, aber schon nicht mehr darüber, was vielleicht über die o.a. Definition von ML hinaus noch dazu gehört, und erst recht nicht darüber, *wie* bzw. *ob* überhaupt ML 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 P&P reden zu können. In der Filosofie von PISA usw. wird das anders gesehen, und es werden (unter dem Einfluss der testorientierten 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 4 (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 P&P-Leistungs- und die Aufgaben-Schwierigkeitsskala zu vereinigen. Dies wird bei PISA usw. in der Tat durchgeführt, und zwar nach folgendem Prinzip: Man zerlegt die Stichprobe S in nichtleere Schichten S_t, jeweils bestehend aus den P&P mit genau t Leistungspunkten. Für jede Aufgabe A wird nun für jede solche Schicht S_t der Anteil $p_A(t)$ der P&P 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 P&P 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 P&P 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.

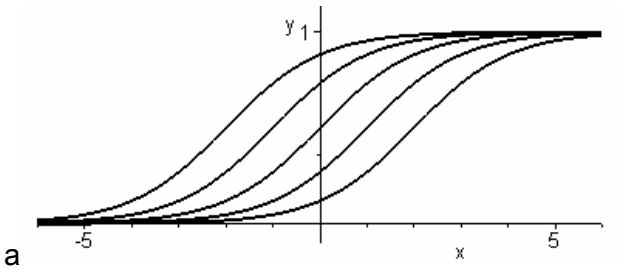
Aus mathematischer Sicht muss man bei dieser Definition noch Sorge tragen für mögliche Randfälle, die allerdings praktisch irrelevant sind. Der Schwellenwert von 62% geht auf einen Wert von 65% zurück, der "nach internationaler Absprache ... Lösen ... mit 'einiger Sicherheit'" bedeutet und deswegen z.B. in TIMSS verwendet wurde (T2, 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, für 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.

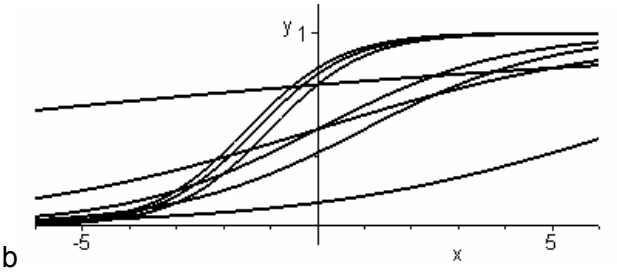
Um diese Ideen deutlich zu machen, habe ich mit Lösungshäufigkeiten usw. bei konkreten Tests argumentiert. Tatsächlich will man aber bei PISA usw. (nicht nur dort) Aufgaben unabhängig von realisierten Tests charakterisieren und verwendet dazu ein probabilistisches Testmodell, wo zu jeder Aufgabe eine bestimmte Funktion gehört, nämlich die, die jeder P&P-Punktzahl die Wahrscheinlichkeit zuordnet, dass P&P mit dieser Punktzahl die Aufgabe lösen (sog. Aufgabencharakteristik): Auch hier wird Monotonie unterstellt, d.h. zu höherer P&P-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 P&P häufiger gelöst wurden als von stärkeren, möglicherweise, weil sie unbefangener zu Werke gegangen sind und diese Unbefangenheit bei solchen Aufgaben hilfreich war (wieder ein Beispiel für Meyerhöfers, 2003ff, Konstrukt der inhaltsunabhängigen Testfähigkeit). (S. dazu auch die gut verständliche Darstellung in Kleine, 2004, 87ff.) Man wüsste gerne, ob solche Phänomene bei PISA usw. auch aufgetreten sind; Monotonie bei allen Aufgaben wäre angesichts der großen P&P-Zahlen aber durchaus plausibel.

In starker Idealisierung werden bei PISA usw. (auf Rasch zurückgehend) für alle Aufgaben logistische Charakteristiken mit einheitlichen Parametern, also der Form 1/(1+exp(c-t)), angesetzt (Lind 1994, 279ff), so dass sie (bis auf waagrechte Verschiebung) alle denselben Graf haben (s. Abb. a). Das Schaubild eines Bruchrechentests (Lind 1994, 317) zeigt, wie eine solche Kollektion realistischerweise auch aussehen kann (s. Abb. b) (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. Das Rasch-Modell jedenfalls war ursprünglich für Tests mit primitiv strukturierten Aufgaben und nicht für so etwas Komplexes wie die Untersuchung von ML gedacht. In Lind (1994, 283, 303ff, 316ff), Knoche & Lind (2000), Knoche u.a. (2002) wird die beschränkte Eignung des Rasch-Modells wiederholt angedeutet.



a



b

(Die beiden Abbildungen sind nicht Kopien der Originale, sondern diesen nur nachempfunden. Auch diese stellen ja starke Idealisierungen real aufgetretener Aufgabencharakteristiken dar.)

Konsterniert hat mich an der Kalibrierung der Aufgabenschwierigkeit die Tatsache, dass nicht i.W. sämtliche P&P-Punktzahlen herangezogen wurden, sondern aus jedem OECD-Land gleichgewichtig je 500 P&P (d.h. etwa ein Zehntel) (P1.1, 51, 520f, O1.2, 105). Hierbei haben also die Jugendlichen aus Island dasselbe Gewicht wie die aus den USA, obwohl diese für etwa 900-mal so viele stehen (O1.2, 135f), und z.B. die aus dem Nicht-OECD-Staat Brasilien haben das Gewicht 0. Ich nehme an, dass sich die resultierenden Über- und Unterschätzungen, jedenfalls innerhalb der OECD, ungefähr ausgleichen. Womöglich trifft diese Ausgleichsannahme auch auf die Verfälschung der "wahren" Aufgabencharakteristiken durch die Verwendung des Rasch-Modells zu.

Zweck dieser Identifizierung der P&P-Leistungsskala mit der Aufgaben-Schwierigkeitsskala war nicht zuletzt die simultane Verwendung der Stufung, das sog. Kompetenzstufenmodell, das die PDMG ja gerade für die Analyse der Aufgaben ausbeuten wollte. Die Stufen waren folgendermaßen festgesetzt worden. Zunächst wurde eine bestimmter Punktewert als Untergrenze von Stufe I ausgewählt, und von da aus nach oben vier Stufen folgendermaßen festgelegt: die Obergrenze einer Stufe ist dadurch bestimmt, dass die P&P, die auf der Untergrenze dieser Stufe angesiedelt sind, 50% aller Aufgabe 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 (P1.1, 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 "widerspenstige" Aufgaben auf, die mit ihrer tatsächlichen Schwierigkeitspunktzahl in einer offensichtlich falschen Stufe landen. Beispiel: Die IGLU-Naturwissenschaften-Aufgabe "Welches Tier säugt seine Jungen? Huhn, Frosch, Affe, Schlange?" gerät mit 474 Punkten in die Stufe III "Anwenden natur-

wissenschaftsnaher 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. I1, 156ff) (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 (P1.1, 168). Diese von mir wörtlich zitierten Schlagwörter stellen Kurzfassungen aus dem PISA-Bericht selbst dar. Die ausführlicheren Beschreibungen (P1.1, 160) wiederum sind nicht genügend operational. Verbindet man die Eigenschaften einer Stufe mit "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 Vergleichsuntersuchungenkommunität insgesamt so inflationär gebraucht, dass sie banal wird. Wie vom PISA-Ableger IGLU explizit eingeräumt (I2, 118f), wird sie nämlich mit jeglichem Bearbeiten von Aufgaben identifiziert.

Die PDMG hat die gesamte 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.

So hat die PDMG 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, 88ff). 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.

Ein Beispiel für die Problematik ist 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 (P1.1, 151ff, 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 P&P 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 (P2, 51). Deswegen wurde dieses Fach noch einmal zerlegt, allerdings nicht etwa in die Typen mathe-

matischen Arbeitens (die tauchen im internationalen Bericht gar nicht auf), sondern in die vier "Big Ideas" (≈ "inhaltliche Teilbereiche") "Quantität", "Veränderung und Beziehungen" (V&B), "Raum und Form" und "Unsicherheit" mit vier eigenen Länderrangfolgen, wenn auch (ohne inhaltliche Begründung) mit gemeinsamer Stufung (komplett in O2 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 mir diese Unterteilung in "Big Ideas" erscheint, 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 Anfertigen oder Interpretieren eines Funktionsgrafs geht. Vergleichbares gilt sowohl für "Quantität", als auch für "V&B". — 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 P2, 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. "V&B". 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 "V&B" 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" (P2, 54f).

Unter den kognitionsbezogenen Ansätzen erscheint mir der von Cohors-Fresenborg, Sjuts & Sommer (Neubrand 2004, 109ff) am fundiertesten. Wie weit er zum "Kompetenzstufenmodell" passt, steht dahin; einen direkten Niederschlag konnte ich kaum ausmachen (P2, 61).

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 PDMG 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 hoch-interessant wären). Meyerhöfer (2004a, 2005) hat für viele Aufgaben aus PISA 2000 dargestellt, wie stark Aufgabenanforderungen infolge unterschiedlicher Lösungswege differieren (die Analyse zu den geheimen Aufgaben darf er nicht publizieren). Dies 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.

Die PDMG (Lind u.a. 2005) stellt sich die Berücksichtigung unterschiedlicher Lösungswege bei der Testauswertung offenbar so vor, dass die Gesamtpopulation (nachträglich) gemäß den Lösungswegen in Teilpopulationen zerlegt wird. Die Lösungshäufigkeiten sagen dann nur etwas über die Leistungsfähigkeit der Teilpopulationen aus (S. 83). Dies hätte allerdings paradoxe Konsequenzen, etwa dass eine Teilpopulation mit einem Lösungsweg mit geringen Anforderungen sich als viel leistungsfähiger als eine mit einem fachlich anspruchsvolleren Lösungsweg erweist. Solche fachdidaktische Probleme hat man nicht, wenn man bei der Testauswertung auf die Differenzierung nach Lösungswegen verzichtet. Dieser Verzicht wird explizit schon bei der vorgängigen Zuordnung jeder Aufgabe zu *einem* Typ mathematischen Arbeitens (sogar zu *einer*, ja noch feineren, Kompetenzklasse) "über Experten-Einordnungen" (S. 82) vorgenommen. Die PDMG bewertet Meyerhöfers Ausführungen anscheinend ausschließlich innerhalb ihres eigenen Paradigmas und spricht deswegen davon, dass er einen bestimmten Grundsatz von PISA "unterschlägt", eine falsche "Behauptung" aufstellt (S. 83) und "in seiner Argumentation zwei entscheidende Fehler auftreten" (S. 85).

In der Tat versucht Meyerhöfer, 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 P&P 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 usw. auf den Punkt bringen.

Die PDMG stellt fest, "dass aus der Lösungshäufigkeit einer Aufgabe nicht auf die möglichen Lösungswege der Schüler zurück geschlossen werden kann" (S. 80, 86), macht aber Halt vor dem Gedanken, dass dann ja auch wesentliche Anforderungsmerkmale nicht identifiziert werden können, mit denen man aber die "inhaltliche Beschreibung der [Kompetenz-] Stufen" bewerkstelligen möchte. Diese "Füllung solcher Stufen mit Leben" (S. 84) ist ja ein zentralen Anliegen der PDMG (s. z.B. Neubrand, 2004, 87ff, der allerdings von *Aufgaben*merkmalen spricht, was, so hoffe ich, dasselbe bedeutet, Knoche u.a., 2002, u.v.a. schriftliche und mündliche Verlautbarungen) innerhalb ihrer von mir oben skizzierten "ehrgeizigeren Plänen". Wäre die Fragestellung insgesamt wirklich "schlicht 'Können die Schüler die Aufgabe lösen?'" (S. 86) und würde die Aufgaben-Schwierigkeitsskala mit den Stufen und den eingeordneten Aufgaben wirklich nur als Trendmesser und als Redeerleichterung behandelt, gäbe es an dieser Stelle keinen Dissens.

Insgesamt sind die Gesichtspunkte, die von der PDMG 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 PDMG 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. Eine der zentralen Parolen des PISA-2000-Berichts, nämlich dass nur 44% der 15-Jährigen in Deutschland über den mathematischen Grundbildungsstandard verfügen (P1.1, 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.

Dieses Ergebnis ist weit von den Ansprüchen der mathematikdidaktischen Kommunität entfernt. Es stellt allerdings keine Überraschung dar. 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 P&P nicht viel, aber in der Mitte, und zwar im unteren Gymnasialbereich, scheint da noch Verbesserungspotential zu existieren (so Klaus Klemm vom deutschen PISA-Beirat, FR, 08.12.04, allerdings bezogen auf den Vergleich von PISA 2000 mit 2003, wie er in P2, 86ff, dargestellt ist).

Literatur

Berichte

(T1; TIMSS I) 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. TIMSS International Study Center. Boston College. Chestnut Hill, MA

(T2, TIMSS II) Jürgen Baumert, Rainer Lehmann u.a. (1997): TIMSS — Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen: Leske + Budrich

(T3.1, TIMSS III) 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

(T3.2, TIMSS III) 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

(P1.1, 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

(P1.2, 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

(P1.3, 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

(P2, 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

(O1.1, PISA 2000) OECD (Hrsg.) (2002): Manual fort he PISA 2000 Database. Paris: OECD

(O1.2, PISA 2000) Ray Adams & Margaret Wu (Hrsg.) (2002): PISA 2000 Technical Report. Paris: OECD

(O2, PISA 2003) OECD (2005): Lernen für die Welt von morgen — erste Ergebnisse von PISA 2003. Erscheint in Heidelberg u.a.: Spektrum

(I1, IGLU) Wilfried Bos, Eva-Maria Lankes, Manfred Prenzel, Knut Schwippert, Gerd Walter & Renate Valtin (Hrsg.) (2003): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster u.a.: Waxmann

(I2, 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

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

Bender, Peter (2005): Die etwas andere Sicht auf PISA sowie TIMSS und IGLU. Erscheint in: Beiträge zum Mathematikunterricht 2004. Hildesheim: Franzbecker

Bender, Peter (2005): Neue Anmerkungen zu alten und neuen PISA-Ergebnissen und -Interpretationen. Erscheint in: Beiträge zum Mathematikunterricht 2005. Hildesheim: Franzbecker (auf CD)

Engström, Arne (2005): Medelsta 1977–1986–2002. Untersuchung mathematischer Fähigkeiten in Kl. 1–9. Erscheint in: Beiträge zum Mathematikunterricht 2005. Hildesheim: Franzbecker (auf CD)

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

Jahnke, Thomas (2005): Ideologiekritisches und Versöhnliches zu PISA & Co. Erscheint in: Beiträge zum Mathematikunterricht 2005. Hildesheim: Franzbecker

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

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: 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

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: 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

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

Internet-Adressen

Bender, Peter (17.04.05 gültig): math-www.uni-paderborn.de/~bender

Braams, Bas (17.04.05 gültig): math.nyu.edu/mfdd/braams

Hagemeister, Volker (17.04.05 gültig): www.lisum.de

Meyerhöfer, Wolfram (17.04.05 gültig): www.math.uni-potsdam.de/prof/o_didaktik/a_mita/ac/Veroe