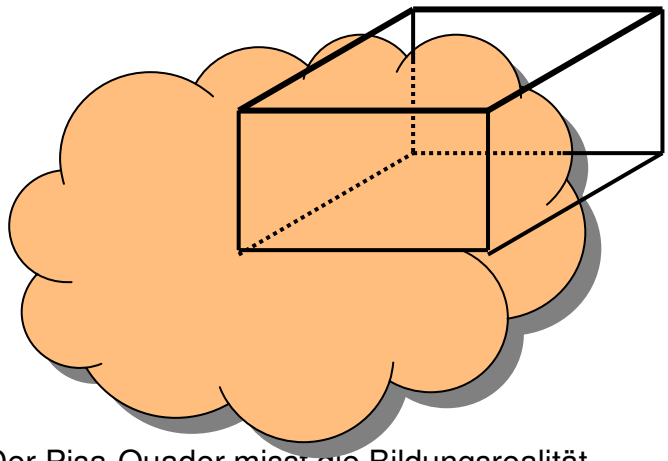
Peter Bender
Universität Paderborn EIM-Fak
bender@math.upb.de
03.09.2009
Berlin

Problematik der Mess-Instrumente am Beispiel jüngerer Schulstudien



Der Pisa-Quader misst die Bildungsrealität

"Just another opinion." Damit tun die Vertreterinnen & Vertreter der quantitativen Bildungsforschung gerne Ergebnisse ab, die nicht mit ihren Methoden erzielt werden. Stattdessen machen sie sich z.B. mit Pisa anheischig zu demonstrieren, wie man die Bildung der 15-Jährigen "misst". Mir kommt dieses Unternehmen allerdings so vor, wie wenn ein Erdölfeld erschlossen werden soll, die Ingenieurinnen & Ingenieure sich ein quaderförmiges Modell davon machen und ihre Aufgabe darin sehen, die Kanten dieses Quaders zu bestimmen. Am Schluss haben sie zwar nicht das Ölfeld, aber dafür einen schönen Quader nach ihren Vorstellungen genau vermessen.

Viele im Bildungsbereich mit quantitativen Methoden erzielte Erkenntnisse haben, im Gegensatz zum Aufwand für ihre Gewinnung, zur Überzeugung ihrer Protagonistinnen & Protagonisten und zu den wegen der Drittmittel-Förderung erlangten höheren Weihen im Wissenschaftsbetrieb, keine große Aussagekraft. Für diese Behauptung kann ich aus meinem Bereich der Mathematikdidaktik aus den letzten Jahrzehnten -zig Belege liefern.

Viele dieser Arbeiten sind mit methodischen Fehlern gespickt. Vor allem fehlt es immer wieder an der Repräsentativität der Stichproben; wichtige Einflussgrößen werden bei Durchführung und Interpretation außer Acht gelassen; die Unschärfe von Antworten auf "weiche" Fragen wird ignoriert; es steht immer wieder in Zweifel, ob die Forschungsfragen und die veröffentlichten Antworten einerseits sowie das Untersuchungsdesign andererseits sich wirklich entsprechen (Validität); u.v.a.m. – Ein anderer großer Mängelbereich tut sich bei der Interpretation der Ergebnisse i.w.S. auf, zu der auch schon die Auswahl der Literatur sowie das Verständnis von deren jeweiligen Aussagen gehört. – Nicht immer merken die Autorinnen & Autoren, wie sehr sie dabei Vorurteilen unterliegen, besonders wenn sie ihre Zahlenwerte, die ja oft das Ergebnis von weitgehenden Annahmen, stark vergröbern-

den Schätzungen und Wahrscheinlichkeitsbetrachtungen sind, auf fünf wesentliche Stellen angeben, als ob sie eine naturwissenschaftliche Messung durchgeführt hätten. Oft genug bedienen sie aber *bewusst* Ideologien, politische Ziele oder ganz utilitaristische Absichten. – Wer hier an die Objektivität von Wissenschaft glaubt, ist naiv.

Im Folgenden beziehe ich mich i.W. auf Pisa, das ja uns allen wohlbekannt und in der Bildungsdebatte hochrelevant ist. In Pisa werden zwar keine *primitiven* Fehler gemacht; bzw. die primitiven Fehler stammen von den manchmal naiven, oft eigennützigen Exegetinnen & Exegeten. Aber in subtilerer Form treten einige der genannten Fehlertypen sehr wohl auf, und ihre Analyse ist durchaus lehrreich.

Aus aktuellem Anlass gehe ich am Schluss noch auf eine Studie im Auftrag der Bertelsmann-Stiftung zu dem "teuren und unwirksamen" Sitzenbleiben ein, deren tendenziöse Aussagen kürzlich durch den deutschen publizistischen Blätterwald gejagt wurden.

Zunächst möchte ich aber noch einmal ausdrücklich das Paradigma in Frage stellen, das dem ganzen Unternehmen "Pisa" zugrunde liegt, nämlich die Meinung, man könne (und solle) Bildung *messen*; hier: die Bildung des Kollektivs der 15-Jährigen eines Lands. Pisa drückt sich da scheinbar bescheidener aus: man "untersucht, wie gut fünfzehnjährige Schülerinnen und Schüler auf die Anforderungen der Wissensgesellschaft vorbereitet sind" (Buchrücken von Pisa 2007 und 2008). Für mich ist das sehr wohl die Frage nach der Bildung der 15-Jährigen, und die Antwort in Form von gemessenen und abgeleiteten Zahlenwerten aus einem ganz schmalen Bereich halte ich für unangemessen und vermessen. – Trotz meiner grundsätzlichen Bedenken will ich mich aber im Folgenden auf den Mess-Ansatz von Pisa einlassen. Für viele Zitate verweise ich auf (Bender 2007).

1. Die Mathematikdidaktik in Pisa

Anders als noch bei Timss findet bei Pisa ein "Verzicht auf transnationale curriculare Validität" statt, stattdessen führen die Tests "ein didaktisches und bildungstheoretisches Konzept mit sich, das normativ ist", angelehnt an die NCTM-Standards aus den USA. Der Erfolg des deutschen Mathematikunterrichts wird also an einem US-amerikanischen Curriculumsentwurf gemessen.

Grundlegend ist dabei das Konstrukt der "Mathematical Literacy" (ML; "mathematische Grundbildung"): "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 Mathe-

matik zu befassen, die den Anforderungen des gegenwärtigen und künftigen Lebens einer Person als konstruktivem, engagiertem und reflektierendem Bürger entspricht."

Diese "Definition" passt durchaus zur Tradition der deutschen bildungstheoretischen Didaktik, wie sie z.B. vom *alten* Wolfgang Klafki (1958) verkörpert wird. Sie ist so gefasst, dass der reale Mathematikunterricht, wie er über weite Strecken in Deutschland und tendenziell wohl weltweit stattfindet, nämlich konzentriert auf das Ausführen von Verfahren und weniger auf Verstehen und Anwenden, ihr nur unzureichend gerecht wird.

Die Aufgaben, die in Pisa gestellt sind, entsprechen in ihrer Gesamtheit aber ebenfalls dieser Definition nicht, d.h. zu ihrer Lösung wird vielleicht die Kompetenz zum Entkleiden von eingekleideten Rechenaufgaben gebraucht, nicht aber ML. Wer viele Pisa-Punkte erzielt, kann gut Pisa-Aufgaben lösen, zeigt aber nicht notwendig ML (insbesondere den deutschen Jugendlichen fehlten da, zumindest in den ersten Durchgängen, auch gewisse Techniken und Strategien auf mehreren Ebenen). Diese ML-Ferne der Pisa-Aufgaben haben zahlreiche Kollegen im In- und Ausland (Bender, Braams, Gellert, Hagemeister, Kießwetter, Meyer, Meyerhöfer, Wuttke) in zahlreichen Analysen dargestellt. Bezogen auf die grundsätzliche Forschungsfrage von Pisa, nämlich nach dem Vorhandensein von ML, ist der Pisa-Aufgabensatz also *nicht valide*. – Hierzu ein typisches Beispiel, das Uwe Gellert aus einer OECD-Schrift von 2000 ausgegraben hat, von dem ich natürlich nicht weiß, ob es jemals in einem Pisa-Test eingesetzt wurde:

Beispiel A "Terrasse": Nick möchte die rechteckige Terrasse seines neuen Hauses pflastern. Die Terrasse ist 5,25 Meter lang und 3,00 Meter breit. Er benötigt 81 Pflastersteine pro Quadratmeter. – Berechne, wie viele Pflastersteine Nick für die ganze Terrasse braucht.

Gedacht ist an eine Lösung der Art 5,25x3x81 = 1275,75, und als korrekte Antworten sollen 1275, 1275,75 und 1276 akzeptiert werden. Klassifiziert wird diese Aufgabe so:

– "Kompetenzstufe 2: Beziehungen und Zusammenhänge zum Zwecke des Problemlösens" [wo kommt so etwas nicht vor?];
– "Fundamentale mathematische Ideen: Raum und Form" [eigentlich geht es um Arithmetik];
– "Erfahrungsbereich: Alltag" [na ja].

Angeblich könne einem eine solche Aufgabe in vielen Situationen des Alltags und der Arbeitswelt begegnen und passe sie gut zur Definition der ML, wofür ja die Anwendung von Mathematik in "authentischen" Situationen wesentlich sei.

So weit meine Übersetzung aus dem Englischen. Es handelt sich um eine eingekleidete Aufgabe, bei der es nicht auf die Lösung eines Sachproblems ankommt, sondern auf das Erkennen und Ausführen der erforderlichen arithmetischen Operation (die zweifache Multiplikation). Dies wird besonders deutlich an der Zulässigkeit der Lösung 1275, die ja mit der Pflasterung der Terrasse nichts zu tun hat, sondern lediglich aus arithmetischer Sicht, aber auch da nur mit Mühe, akzeptiert werden kann.

Wenn man einmal unterstellt, dass die Pflastersteine quadratisch sind und die Seitenlänge 1/9 m haben, dann hat man, im Sinne der vorgegebenen Aufgabenlösungen, mit dem nicht-ganzen Teil der Terrassenlänge Probleme, weil man zu dessen Auslegen einige Steine noch vierteilen müsste, und zwar in Rechtecke mit Seitenlängen 1/36 m und 1/9 m.

In der Realität würde man jedoch beim Pflastern in der Länge einen kleinen Rand lassen oder aber die Fugen leicht verbreitern und dann wohl nur 47 Steine legen, wodurch dann 47x27 = 1269 Steine gebraucht würden. Diese Zahl erscheint mir, so gesehen, noch am "richtigsten".

Aber wofür ist sie überhaupt von Interesse? Nach meiner Erfahrung werden solche Pflastersteine nach Flächeneinheiten verkauft. Aber selbst wenn sie stückweise verkauft würden, dann bestimmt nicht einzeln, sondern vielleicht in 81-er- oder 100-er-Gebinden. Außerdem werden auf Terrassen üblicherweise viel größere Steine verwendet. Usw.

Unter sämtlichen Gesichtspunkten sind Situation und Fragestellung nicht authentisch. Darüber hinaus ist die Beschreibung der Kompetenzstufe nichtssagend, und was die angesprochenen mathematischen Ideen betrifft, so ist die Arithmetik von erheblich größerer Bedeutung als die Geometrie; – von "Raum" kann sowieso keine Rede sein.

Selbstverständlich haben solche Textaufgaben ihren Platz im Mathematik-Curriculum; aber ihre Funktion dort ist von ML im Sinne von Pisa himmelweit entfernt; und das Kritische ist: die Pisa-Expertinnen & -Experten haben offensichtlich dafür kein Gespür.

Das liegt aber nicht nur an deren mangelnden mathematikdidaktischen Expertise, sondern ist in der Sache selbst begründet: Natürlich kann ein Test mit weltweit 250.000 Probandinnen & Probanden (P&P) nur in Form von Häppchen-Aufgaben, wohl oder übel viele im

Multiple-Choice-Format, durchgeführt werden. Eigentlich kein einziger Aspekt der ML-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. Keine einzige dieser Aufgaben, sei sie noch so komplex aufgebaut, stellt ein authentisches Sachproblem dar, gar ein Problem der P&P selbst; denn Allen ist klar, dass es um einen Test geht. Natürlich ist keine Aufgabe wirklich offen; es ist lediglich immer wieder der Versuch erkennbar, ein direktes Anwenden von Faktenwissen und Fertigkeiten durch häufig textlastige Einkleidungen zu verhindern, wobei die Autorinnen & Autoren immer wieder über ihre eigenen Füße stolpern.

Beispiel B "Fläche eines Kontinents": Hier siehst du eine Karte der Antarktis. Schätze die Fläche der Antarktis, indem du den [mit abgedruckten] Maßstab der Karte benutzt.

Diese Aufgabe ist ja ganz nett. Aber sie ist symptomatisch für die ML-Ferne von Pisa. Wer den Flächeninhalt der Antarktis wissen will und nicht im Lexikon oder im Internet nachschaut, sondern anfängt, die Karte mehr oder weniger genau auszumessen, verfügt, mit Verlaub, über wenig ML! – Die Kompetenz zur Nutzung externer Informationsquellen kann mit einem Test à la Pisa eben nicht gemessen werden.

Die typische Pisa-Aufgabe entsteht offenbar am Schreibtisch eines männlichen gebildeten Bürgers im angelsächsischen oder niederländischen Raum mit wenig schulischen und anscheinend oft eigentümlichen alltagspraktischen Erfahrungen. Wer als P&P wenig Affinität zu diesem Autorentyp aufweist, hat es eben ein bisschen schwerer mit einem aus einer fremden Sprache übersetzten Aufgabentext, mit geringerer Vertrautheit mit der angelsächsischen Kultur und Mentalität sowie der gehobenen Schicht des Autors und nicht zuletzt mit dessen Bild von der Mathematik und der Realität.

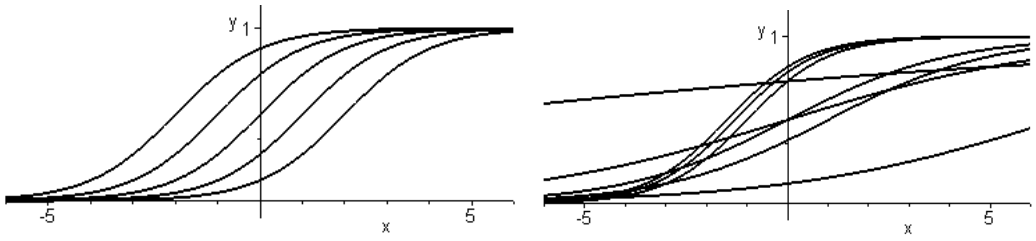
Alle diese aufgeführten mathematikdidaktischen Probleme sind bei einem Unternehmen wie Pisa vermutlich unvermeidlich, und das spricht gegen es.

2. Die Psychometrie und das unzulängliche Kompetenzstufenmodelle in Pisa

Allerdings kommt es in Pisa auf die mathematikdidaktische Qualität gar nicht so sehr an, sondern eher darauf, ob eine Aufgabe im psychometrischen Sinn "gut" misst. – Eine Aufgabe misst gut, wenn sie möglichst trennscharf ist, d.h. wenn es eine Zahl c ($0<c<100$)

gibt, so dass die c Prozent Pisa-schlechtesten Jugendlichen die Aufgabe alle nicht lösen und die anderen sie alle lösen.

Da immer nur Stichproben betrachtet werden, geht es um Lösungs*wahrscheinlichkeiten*, und man ist auch schon mit Aufgabencharakteristiken (Lösungswahrscheinlichkeit als Funktion der P&P-Testleistung) der Form 1/(1+exp(c-t)) zufrieden; jedenfalls dürfen sie nicht wie im rechten Bild aussehen; oder gar Bereiche mit negativer Steigung haben.



Letzteres gibt es gar nicht so selten. Ein schönes Beispiel stammt von Wartha (2009):

Beispiel C (nicht aus Pisa): Herr Brinkmeier hat bei einer Fernsehlotterie gewonnen. Er möchte den sechsten Teil seines Gewinns einem Kinderheim spenden. Sein Gewinn beträgt 2400 €. Wie viel Geld spendet er?

Im 5. Schuljahr betrug die Lösungshäufigkeit in einer Stichprobe in Bayern 76% (Gymnasium 89%, Realschule 81%, Hauptschule 59%), im 7. betrug sie nur noch 59% (G 76%, R 53%, H 45%). Im Text war allerdings für das 7. Schuljahr eine kleine Veränderung vorgenommen worden: "den sechsten Teil" war ersetzt worden durch "ein Sechstel". – Tatsächlich haben die Älteren häufig formalisierte, und damit oft fehlerbehaftete Bruchrechung eingesetzt, mit der sie ja in der Zwischenzeit intensiv konfrontiert worden waren, während die Jüngeren – adäquat – viel elementarer gerechnet haben. Wartha erklärt die unterschiedliche Vorgehensweise mit dem veränderten Text. Ich will das nicht komplett ausschließen, aber ich meine, dass – i.W. unabhängig von der Formulierung – die Jüngeren einfach unbefangener herangegangen sind.

Bei Pisa werden schlecht messende Aufgaben in Pilotstudien identifiziert und dann eliminiert. Ob das übrigbleibende Ensemble noch ein adäquates Bild von ML-Mathematik liefert, also valide für ML ist, ist offenbar zweitrangig. Die Mathematikdidaktikerinnen & -didaktiker im deutschen Pisa-Team haben sich jedenfalls beklagt, dass sie an dieser Stelle gegen das Diktat der Psychometrikerinnen & -metriker nicht ankommen.

Bei den verbleibenden Aufgaben wird jedenfalls unterstellt, dass ihre Charakteristik i.W. wie oben aussieht, d.h. man arbeitet nicht mit den realen Lösungswahrscheinlichkeiten, sondern mit einem mathematischen Modell, dem sog. Rasch-Modell. Joachim Wuttke (2007) hat festgestellt, dass bei vielen Aufgaben die realen Abweichungen von einer idealen Charakteristik jedoch erheblich sind, und es fragt sich, wie weit sie akzeptabel sind. Georg Rasch selbst hat übrigens erklärt, dass sein Modell lediglich für die Untersuchung ganz primitiver Items geeignet ist und nicht für komplexe Fragen (wie etwa Pisa-Mathematik-Aufgaben).

Man braucht ein ganzes Ensemble gut messender Aufgaben, deren Trennpunkte c sich einigermaßen gleichmäßig über den Bereich von 0 bis 100 verteilen (wie oben im linken Bild), und hat eine Skala für die Aufgabenschwierigkeit: Je höher der Trennpunkt, desto weniger P&P lösen die Aufgabe, desto schwieriger ist sie also.

Die Dualität zu der entsprechenden Skala für die Testleistungen der P&P liegt auf der Hand.

Die Testpunktzahlen werden noch so normiert, dass der Mittelwert 500 und die Standardabweichung 100 beträgt. Diese Normierung wird allein auf der Basis der OECD-Länder vorgenommen, d.h. unter Ausschluss der Daten der Partnerländer wie Brasilien. Sie wird außerdem für bestimmte Berechnungen auf die Mittelwerte jeweils früherer Durchgänge bezogen und weicht dann für den jeweils aktuellen Durchgang vom Wert 500 ab.

Unbedingt ist zu beachten, dass die Pisa-Zahlen immer nur relativ zu verstehen sind. Wenn z.B. die deutschen Jugendlichen im Jahr 2003 im sog. Problemlösen 513 Punkte und in Mathematik 503 Punkte erzielten, heißt das nicht, dass sie in Problemlösen besser als in Mathematik sind (dieser Vergleich ist sowieso sinnlos), sondern nur, dass sie in Pisa-Problemlösen im Vergleich zu den anderen Ländern besser abgeschnitten haben als in Pisa-Mathematik im Vergleich zu den anderen Ländern.

Ein weiterer beliebter Fehlschluss besteht darin, die Länderpunktzahlen und damit die Rangplätze als exakt anzunehmen. Die Punktzahlen sind, als Ergebnis von Stichproben, aber immer mit dem sog. Standardfehler behaftet, der von Pisa auch stets angegeben wird. Daher muss man nahe beieinander liegende Länder zu Clustern zusammenfassen, weil bei Variation der Stichproben die Reihenfolge sich ohne Weiteres um mehrere Plätze verändern könnte, z.B. 2006 in Mathematik: Österreich 505, Deutschland 504, Schweden 502 und Irland 501.

Die Werte der P&P-Testleistungsskala werden nun folgendermaßen auf die Aufgabenschwierigkeitenskala übertragen: Für eine bestimmte Aufgabe wird für jede Testpunktzahl die Menge der P&P mit dieser Punktzahl betrachtet und ermittelt, wie hoch der Anteil derer, die die Aufgabe richtig gelöst haben, an *diesen* P&P ist. Es wird unterstellt – bei allen genannten Vorkehrungen wohl zu Recht –, dass mit zunehmender Testpunktzahl dieser Anteil wächst (je besser die P&P, desto eher lösen sie eine bestimmte Aufgabe). Dann wird diejenige Testpunktzahl, bei der der Anteil der Löserinnen & Löser erstmals 62% beträgt, als die Schwierigkeit dieser Aufgabe festgelegt.

Der Wert 62% ist willkürlich. Er gibt die Meinung eines anonymen, zufälligen, vorübergehenden Kollektivs von sog. Expertinnen & Experten (für: was weiß ich) darüber wieder, ab wann man mit der Lösungsquote eines Kollektivs wohl zufrieden sein kann.

Nun ist also eine gemeinsame Skala vorhanden. Zur weiteren Vereinheitlichung werden die P&P-Testleistungen und die Aufgabenschwierigkeiten unter den gemeinsamen Begriff "Kompetenzen" gefasst: die P&P *verfügen* über Kompetenzen, und die Aufgaben *erfordern* Kompetenzen, die mit der Skala simultan "gemessen" werden.

Sinnvollerweise hat das internationale Pisa-Konsortium diese Skala (für die Inhaltsbereiche Mathematik, Lesen usw. sowie für einzelne Teilbereiche jeweils separat) in Stufen eingeteilt und zwar erklärtermaßen willkürlich, lediglich zum Zweck des leichteren Redens darüber, und die inhaltliche Beschreibung den nationalen Gruppierungen überlassen. So gesehen, ist nichts dagegen einzuwenden, dass

(i) die Stufen alle gleich breit gemacht wurden,
(ii) die Lage der Stufen und die gemeinsame Breite von der (bei verschiedenen Inhaltsbereichen bzw. Teilbereichen sowie bei verschiedenen Durchgängen allerdings unterschiedlichen) Anfangs- und Endpunktzahl (ca. 300 und ca. 700) sowie von der Anzahl der Stufen abhängt und
(iii) P&P oder Aufgaben auf unterschiedlichen Stufen landen, je nach dem, wer alles am Test teilgenommen hat.

Die deutsche Pisa-Mathematik-Gruppe hat allerdings aus dieser ersichtlich zufälligen Stufeneinteilung Großes gemacht und ein ganzes Kompetenzstufenmodell darauf gegründet. Jeder Stufe wurden gewisse Kompetenzen zugewiesen, und idealerweise soll sich dann allein aus der Pisa-Punktzahl bzw. -Stufe quasi naturgesetzlich erschließen lassen, welche Kompetenzen ein Mensch hat bzw. eine Aufgabe erfordert.

Diese Gruppe geht anscheinend davon aus, dass

(i) sich die möglichen Kompetenzen überhaupt sinnvoll linear anordnen lassen und
(ii) der Aufgabensatz von Pisa geeignet ist, diese Ordnung treu auf die Punkteskala zu übertragen und sie damit zu metrisieren, bzw. dass ein solcher Aufgabensatz wenigstens denkbar ist.

Beide Annahmen sind höchst naiv und werden in den ausführlichen Analysen auch nicht wirklich substantiiert.

Ob die anderen Länder eigentlich zum selben Modell gekommen sind (was sie ja eigentlich wegen dessen Naturgesetzlichkeit müssten)? Es sollte im ureigenen Interesse von Pisa liegen, einmal die Kompetenzstufenmodelle der ca. 50 Pisa-Länder zu vergleichen. Derartige Vergleiche existieren m.W. nicht, vermutlich weil die anderen Länder sich nicht die Mühe gemacht haben, solche Modelle breit zu entwickeln, sondern bestenfalls ein paar (mehr oder weniger triviale) Stichworte aufgeschrieben haben.

Tatsächlich landen Menschen mit ähnlichen Pisa-Punktzahlen auf ein und derselben Stufe, und wenn sie noch so unterschiedliche Kompetenz*profile* besitzen.

Eine noch stärkere Mehrdeutigkeit besteht bei den Aufgaben: Schon verschiedene Aufgaben*teile* können unterschiedliche Kompetenzen erfordern und dadurch auf verschiedene Stufen gehören. Eine ähnliche Uneindeutigkeit wird vom spezifischen Wissen der P&P, von ihrer Vertrautheit mit der jeweiligen Aufgabe bzw. dem Kontext oder *vom jeweils eingeschlagenen Lösungsweg* erzeugt (Wolfram Meyerhöfer).

Beispiel D: Die Grundfläche einer **Pyramide** ist ein Quadrat. Jede Kante der skizzierten Pyramide misst 12 cm [Zeichnung]. Bestimme den Flächeninhalt einer der dreieckigen Seitenflächen.

Man muss das ebene Problem in der räumlichen Situation sehen, und das fällt einem umso leichter, je mehr man sich in der Schule mit räumlichen Sachverhalten befasst hat. Dann braucht man nur den Flächeninhalt eines gleichseitigen Dreiecks zu kennen, und es handelt sich um eine reine Wissensaufgabe. Aber auch wenn man diesen zuerst noch herleiten muss, liegt hier kein "komplexes Modellieren" vor, was aber für die Kompetenzstufe charakteristisch sein soll, auf der diese Aufgabe wegen ihres hohen Schwierigkeitsgrads von 810 Punkten landet.

Beispiel E: Wie kannst du einen Geldbetrag von genau 31 Pfennig hinlegen, wenn du nur 10-Pfennig-, 5-Pfennig- und 2-Pfennig-Münzen zur Verfügung hast?

Wegen ihrer hohen Punktzahl von 797 wird diese Aufgabe unter "begriffliches Modellieren und Problemlösen" eingeordnet. Ich kann diese Einordnung inhaltlich nicht nachvollziehen. Es handelt sich doch um eine begrifflich völlig anspruchslose Abzählaufgabe an vorgestellten oder aufgezeichneten konkreten Objekten, und die Schwierigkeit liegt nur in der Erfassung aller Fälle.

Am schönsten zeigt sich die Unzulänglichkeit dieses Kompetenzstufen-Ansatzes m.E. an folgender Aufgabe, die zwar dem Naturwissenschaftentest von IGLU 2001 (mit 10-jährigen P&P) entnommen ist, dessen Stufenmodell aber denselben Prinzipien gehorcht wie das von Pisa-Mathematik.

Es hat folgende Stufen: 0 "Vorschulisches Alltagswissen", I "Einfache Wissensreproduktion", II "Anwenden alltagsnaher Begriffe", III "Anwenden naturwissenschaftsnaher Begriffe", IV "Beginnendes naturwissenschaftliches Verständnis", V "Naturwissenschaftliches Verständnis und Lösungsstrategien".

Beispiel F: Welches Tier säugt seine Jungen? Huhn, Frosch, Affe, Schlange?"

Diese Aufgabe gehört ersichtlich auf Stufe I oder gar 0, je nach den Fernseh-Erfahrungen der Kinder, auf keinen Fall aber auf II oder gar III. – Aus welchen Gründen auch immer, sie fällt den Kindern recht schwer und landet mit 474 Punkten doch auf III.

Natürlich gibt es schon bei der Bewertung der Aufgabenlösungen Unklarheiten noch und noch. Z.B. würde ich bei der Terrassen-Aufgabe die Antwort 1275 als inkorrekt und dafür 1269 als korrekt bewerten.

Aber i.W. sind die P&P-Punktzahlen doch harte Daten, auch wenn sie nicht nur das Ergebnis von ML-Kompetenzen sind, sondern auch von Rate-Vermögen, Erfahrungen mit Tests (die in Deutschland – noch – geringer ausgeprägt sind), Abfolge der Aufgaben, Verteilung auf die Testhefte, Tageszeit, zu der der Test stattfindet, Konzentrationsfähigkeit usw.

3. Harte Soziometrie mit extrem weichen Daten in Pisa

Pisa hat aber höhere Ambitionen, und zwar sollen die Testergebnisse mit dem ökonomischen, sozialen und kulturellen Status der P&P verknüpft werden. Die Daten dazu hat man u.a. mit Fragebögen gewonnen, auf denen die Jugendlichen selbst über sich und ihre Familie Auskunft geben sollten.

Da hat man sich nach dem Vorhandensein gewisser Haushaltsgeräte erkundigt oder Fragen folgender Art gestellt: "Wie viele Bücher habt ihr zu Hause?" oder "Wie oft kommt es im Allgemeinen vor, dass deine Eltern mit dir über Bücher, Filme oder Fernsehsendungen diskutieren?" – Das ergibt offensichtlich extrem weiche Daten.

So haben z.B. – völlig neben der Realität – die Schweizerinnen & Schweizer ihre Lage schlechter eingeschätzt als die Deutschen.

Oder: Bei Pisa 2003 haben in Schleswig-Holstein 43,0%, in Bayern 24,8% und in Deutschland 23,9% der Jugendlichen angegeben, schon einmal eine Klasse wiederholt zu haben (Pisa 2005, 169ff, Klemm, 9, Abb. 2). Diese Werte spiegeln sich in den W&W-Quoten, die für die Schuljahre 1995/96 und 2000/01 in Klemm (19, Tab. 3) abgedruckt sind, völlig unterschiedlich wieder. Bei Schleswig-Holstein würde man besonders hohe Quoten erwarten; sie sind aber niedriger als die von Bayern. Wenn man die anderen Bundesländer einbezieht, wird das Bild noch viel uneinheitlicher. – Haben die Jugendlichen in Schleswig-Holstein ein anderes Verständnis vom Sitzenbleiben als die in Bayern? Bekennen sie sich eher zum Sitzenbleiben?

Oder: Nur 40% aller Jugendlichen in Deutschland haben als genauen Beruf ihres Vaters denselben angegeben wie dieser; und auch wenn sie ihn nur grob nennen sollten, lag die Übereinstimmung noch unter 70%. Jeweils etwas größer ist die Kohärenz bezüglich der Mutter, weil diese häufiger keiner Erwerbstätigkeit nachgeht und dieser Status von den Jugendlichen leichter erkannt wird.

3.1 Der "soziale Gradient" ist unbrauchbar

Der "höchste" Beruf in der Familie spielt aber eine ganz wichtige Rolle: der für Pisa relevante Status der Familie wird reduktionistisch i.W. in Form des sog. HISEI mit ihm identifiziert (ISEI: International Socio-Economic Index of Occupational Status; HISEI: Highest ISEI). Man hatte 2003 einen anderen Parameter verwendet, den ESCS (Economic, Social, and Cultural Status). Wegen eines total uneinheitlichen Bildes beim Vergleich von 2000

mit 2003 hat man im Bericht über 2006 wieder den HISEI für den Vergleich der *drei* Durchgänge herangezogen (Pisa 2007, 323).

Für jedes Land wird *seine* Abhängigkeit der Variablen "Pisa-Punktzahl" (und zwar beim Lesen) von der Variablen "sozialer Status der Familie" (dem HISEI) mittels einer linearen Regression dargestellt. Bei einer solchen Analyse mit zwei Variablen entsteht eine Punktwolke, und diese wird durch eine Gerade repräsentiert. Je größer deren Steigung (der sog. soziale Gradient) ist, desto ausgeprägter erscheint die Abhängigkeit.

Die beim Lesetest erzielten Punktzahlen sind zwar (mit gewissen Abstrichen) in voller Genauigkeit vorhanden. Die Variable "Sozialstatus" dagegen ist, wie gesagt, wachsweich. Deren Anordnung auf einer linearen Skala ist eine erneute fragwürdige Reduktion. Wie man dann noch darauf eine Metrik pfropfen kann, ist mir unbegreiflich. Da werden ja *Abstände* etwa zwischen Professor und Astrologe oder zwischen Botschafter und Tänzer (um einmal einige der Berufe zu nennen) als Zahlen definiert, und mit diesen wird auf zwei Stellen hinter dem Komma genau Regressionsrechnung getrieben, zwecks exakter Vermessung des Pisa-Quaders.

Wenn man sich aber einmal auf diese Vorgehensweise einlässt, dann ist klar, dass große Gruppen mit sehr niedrigen Punktzahlen in Verbindung mit niedrigem sozialem Status die Steigung des sozialen Gradienten erhöhen. Und daran haben unsere Jugendlichen mit Migrationshintergrund (MH; wenigstens ein Elternteil im Ausland geboren), und zwar vornehmlich die mit doppeltem MH aus bestimmten Ländern, zusätzlich zum schwachen Viertel unserer eingeborenen Jugendlichen einen erheblichen Anteil.

Während sich Deutschlands sozialer Gradient von 2000 bis 2006 von 45 über 38 bis 35 deutlich verringerte, hat es neben ähnlich starken positiven wie negativen Entwicklungen in anderen Ländern in diesen 6 Jahren auch ausgeprägte Berg- und Talfahrten gegeben, und das trotz der einheitlichen Verwendung des HISEI (ebenda, 323):

Island	19	12	18
Kanada	26	22	25
Korea	15	19	17
Österreich	35	40	35
Portugal	38	31	39
Schweiz	40	30	32
Tschechien	43	32	46

Hier hätte deutlich ausgesprochen gehört, dass dieser Gradient unbrauchbar, weil zu labil, ist. Das tun die Autoren nicht; sie geben (ebenda, 323) "Veränderungen in der Stichprobenausschöpfung und veränderte Anteile von fehlenden Werten" (ein Pisa-Euphemismus für "Mängel in der Erhebung") als mögliche Ursache an und stellen damit redlicherweise auch gleich noch ihr Stichprobenauswahlverfahren in Frage.

Dieser Gradient war ja in der deutschen Bildungsdiskussion nach 2000 von interessierten Kreisen als Grundlage für die Parole von der in Deutschland besonders großen Abhängigkeit der Schulleistungen vom sozialen Status benutzt worden, die bis heute als Begründung für die Einheitsschule herhalten muss. Es ist wohl zu viel verlangt, sich selbst und diesen Kreisen deutlich zu machen, dass man da i.W. einem Artefakt aufgesessen ist.

3.2 Die "relative Wahrscheinlichkeit des Gymnasialbesuchs" ist nichtssagend

Ein noch fragwürdigerer Parameter ist die sog. "relative Wahrscheinlichkeit des Gymnasialbesuchs" (rWG), sehr grob gesprochen: der Quotient aus dem Verhältnis der Anzahl der "reichen" 15-jährigen Gymnasiastinnen & Gymnasiasten (G&G) zu der Anzahl der "reichen" 15-jährigen Nicht-G&G und dem Verhältnis der Anzahl der "armen" 15-jährigen G&G zu der Anzahl der "armen" 15-jährigen Nicht-G&G.

Warum nicht Anteile (von G&G an *allen* "reichen" 15-Jährigen bzw. an *allen* armen 15-Jährigen), sondern diese Verhältnisse ("odds") zueinander in Bezug gesetzt werden ("odds ratios" gebildet werden), ist mir nicht klar. Vermutlich hat sich das irgendwo "bewährt" (eine Begründung, die in den Pisa-Berichten immer wieder einmal auftaucht). Jedenfalls wachsen die "odds" überproportional mit den Anteilen, und zwar zunehmend rasanter: wenn der Anteil gegen 100% geht, gehen die "odds" gegen ∞. Beispiel: Beträgt der Anteil bei den "Reichen" 60% und bei den "Armen" 20%, dann ist der rWG nicht 60/20 = 0,6/0,2 = 3, sondern die "odds" lauten 60/40 = 1,5 sowie 20/80 = 0,25 und die rWG = 1,5/0,25 = 6. Bei hoher Gymnasialbeteiligung der "Reichen" wird also durch Verwendung der "odds" der Zahlenwert der rWG deutlich erhöht.

Da dieser Parameter ersichtlich nicht für Vergleiche mit dem Ausland gedacht ist, stellt die Verwendung der "odds ratios" ein einfaches Mittel dar, höhere Zahlenwerte für die "sozialen Disparitäten" im deutschen Bildungssystem zu erhalten. Das Hervorbringen "schlechter" Nachrichten gehört ja zum Erfolgsrezept von Pisa, und die Veröffentlichung der Werte des rWG von 2003 führten zu Schlagzeilen wie "Chancenungleichheit in Deutschland wächst" bzw. "Chancenungleichheit in Bayern am größten". Der damals für Bayern beson-

ders hohe Wert geht übrigens nicht auf das eben beschriebene Phänomen mit den "odds ratios" zurück, da in Bayern ja die Gymnasialbeteiligung in fast allen sozialen Schichten niedriger als in Deutschland insgesamt ist, insbesondere auch bei den "Reichen". – Aber eben auch bei den "Armen", und das führte, jedenfalls 2003, zu dem hohen Wert.

In den Pisa-Durchgängen 2000 und 2003 war Bayern das beste Pisa-Bundesland, eines der besten Länder der Welt und bei Kontrolle der Migrationsquote (MQ) sogar mit das beste Land der Welt überhaupt. Außerdem hatte Bayern damals unter den alten Bundesländern den zweitniedrigsten sozialen Gradienten, und das alles mit einem konservativen, betont dreigliedrigen Schulsystem. Aber bei der rWG war Bayern 2003 in der Variante "mit Kontrolle der Lese- und Mathematikkompetenz" das schlechteste Bundesland. Obwohl es in der Variante "ohne Kontrolle von Kovariaten" in der Nähe des Durchschnitts lag und Sachsen-Anhalt sowie Bremen Spitze waren und im Pisa-Bericht konzediert wurde, dass die Wahrheit wohl irgendwo zwischen den beiden Varianten liegt (Pisa 2005, 262), wurden an die breite Öffentlichkeit nur die Werte der erstgenannten Variante gebracht, und zwar mit den o.a. reißerischen Aufmachern.

Nun war in gewissen Kreisen der Jubel groß, als 2006 Sachsen mit seinem zweigliedrigen Schulsystem Bayern als Pisa-Spitzenbundesland ablöste, hatte man doch endlich einen Beleg für die Überlegenheit der Einheitsschule (jedenfalls wollte man dieses Ergebnis so interpretiert wissen). Allerdings ließ man außer Acht, dass die neuen Bundesländer allesamt eine viel geringere MQ als die alten Bundesländer haben (ca. 5% gegenüber 22% bis 41%) und dass außerdem dort die Migrations*struktur* erheblich günstiger ist. Bei Kontrolle der MQ liegt natürlich nach wie vor Bayern deutlich vorne, und mit der Zweigliedrigkeit seines Schulsystems hat der Erfolg Sachsens herzlich wenig zu tun.

Nachdem nun also die Ergebnisse des Pisa-Durchgangs 2006 vorlagen, wurde für die Bundesländer die rWG für die Durchgänge 2000 und 2006 explizit verglichen (Pisa 2008, 338). Zusammen mit den Werten für 2003 ergibt sich folgendes Bild:

	2000		2003		2006	
	o.Kontr.	(mit K.)	o.Kontr.	(mit K.)	o.Kontr.	(mit K.)
Baden-Württemberg	5,8	(3,2)	8,41	(4,40)	5,6	(4,0)
Bayern	10,5	(6,5)	7,77	(6,65)	4,3	(2,7)
Berlin			4,45	(2,67)		
Brandenburg	3,2	(1,9)	3,71	(2,38)	4,8	(4,3)
Bremen	6,1	(3,0)	9,06	(2,83)	4,8	(3,2)

Hamburg			7,53	(3,55)		
Hessen	6,5	(2,7)	5,70	(2,71)	5,6	(3,4)
Mecklenburg-Vorpommern	6,0	(4,0)	7,96	(3,47)	3,2	(2,3)
Niedersachsen	7,8	(5,0)	6,45	(2,63)	4,8	(4,8)
Nordrhein-Westfalen	6,5	(3,1)	8,07	(4,35)	6,7	(4,5)
Rheinland-Pfalz	9,1	(5,1)	8,28	(4,60)	4,0	(2,6)
Saarland	6,0	(3,5)	6,71	(3,48)	5,5	(4,1)
Sachsen	3,1	(2,2)	4,49	(2,79)	3,9	(2,8)
Sachsen-Anhalt	4,4	(3,1)	10,44	(6,16)	3,3	(3,0)
Schleswig-Holstein	8,1	(5,8)	6,24	(2,88)	5,0	(2,9)
Thüringen	4,0	(3,2)	5,13	(3,23)	3,0	(2,2)
Deutschland	6,0	(3,2)	6,87	(4,01)	4,6	(3,2)

(O. Kontr. = ohne Kontrolle von Kovariaten; mit K. = mit Kontrolle der Lesekompetenz und 2003 zusätzlich mit Kontrolle der Mathematikkompetenz. Da Hamburg und Berlin wegen fehlender Repräsentativität im Durchgang 2000 nicht extra ausgewiesen worden waren, wurden auch ihre Ergebnisse aus 2006 nicht publiziert. Warum eigentlich nicht?)

In den Durchgängen 2000 und 2006 einerseits sowie 2003 andererseits wurden die Klassen, aus denen dann jeweils zwei für den Vergleich zwischen "reich" und "arm" ausgewählt wurden, unterschiedlich definiert. 2000 und 2006 ging es nach sog. EGP-Klassen (ebenda, 322f), und aus den 7 Klassen wurden zunächst die Klassen V und VI (Facharbeiter und Arbeiter mit Leitungsfunktion) zusammengefasst und dann alle anderen Klassen mit dieser verglichen. In der o.a. Tabelle sind die Zahlen für den Vergleich der Oberen Dienstklasse (I) mit dieser zusammengesetzten Klasse angegeben. Das sind genau die Zahlen, die von Pisa und von der Öffentlichkeit als die entscheidenden angesehen werden. 2003 ging es nach Quartilen der sog. ESCS-Klassifikation, und alle Quartile wurden mit dem dritten Quartil verglichen. Auch hier wird nur der Vergleich des ersten mit dem dritten Quartil als wesentlich angesehen, und diese Zahlen sind oben aufgeführt.

Für eine Einschätzung der Nützlichkeit der rWG ist es prinzipiell unerheblich, nach welchen Gesichtspunkten die Klassen festgelegt sind. Wichtig ist, dass es da stark unterschiedliche Möglichkeiten gibt, auf deren Basis jedes Mal ein scheinbar sinnvoller Parameter definiert werden kann, der "rWG" genannt werden kann (und bei Pisa auch so genannt wird). Dass dann unterschiedliche Zahlenkollektionen herauskommen, ist zu erwarten und spricht noch nicht gegen die unterschiedlichen Definitionen. Aber wenn man sich den fast

zufälligen Zahlensalat in der obigen Tabelle bei waagrechten und lotrechten Vergleichen anschaut, stellen sich doch erhebliche Zweifel an irgendeiner Aussagekraft dieses Parameters ein, wohlgemerkt, auch wenn man die Unterschiedlichkeit der Definitionen 2000 und 2006 einerseits sowie 2003 andererseits ins Kalkül zieht.

In (Bender 2007, 291) habe ich, z.T. die Argumente aus dem Pisa-Bericht aufnehmend, diesen Parameter bereits inhaltlich kritisiert. Z.B. besitzen mehrere Bundesländer (darunter häufig solche mit niedrigem Wert) in nennenswertem Umfang Gesamtschulen, die zum Abitur führen, die aber bei diesen Rechnungen nicht berücksichtigt sind (Pisa 2005, 262, Fußnote 5), und dass es zusätzlich auf den Expansionsgrad der Gymnasien ankommt (ebenda, 263), der ebenfalls nicht einbezogen wurde. Schon dort habe ich die Willkür in der Wahl der ESCS-Skala und der Einteilung in Quartile bemängelt. Der jetzt vorliegende Vergleich der drei Pisa-Durchgänge bestätigt meine Skepsis voll und ganz.

Die Übersicht über die Gymnasialbeteiligung in jeder einzelnen Klasse (Pisa 2008, 336) wäre wohl ergiebiger. Die durchweg niedrigeren Werte von Bayern i.V.m. seinen hohen Leistungspunkten sind Ausfluss eines bis vor kurzem noch intakten dreigliedrigen Schulsystems mit wenigen Schulabgängerinnen & -abgängern ohne Abschluss sowie geringerer Jugendarbeitslosigkeit und höchstem Leistungsniveau bei allen Schulformen. Mit seiner geringeren Migrationsquote hat es Bayern leichter als die anderen alten Bundesländer. Aber es hat sich auch bis vor kurzem erfolgreich dagegen gewehrt, dass die Hauptschule schlecht gemacht bzw. geredet wird.

Leider verbreitet sich in der deutschen Bildungsdiskussion zunehmend die Auffassung, dass der Mensch erst mit dem Abitur anfängt (so zugespitzt, würden sich natürlich Alle distanzieren), und zwar mit dem Abitur des allgemeinbildenden Gymnasiums (obwohl etwa die Hälfte aller Hochschulzugangsberechtigungen auf anderem Weg erworben werden). Da ist auch der Duktus des Pisa-Berichts verräterisch, wenn er (ebenda, ab S. 335) praktisch "Bildungsbeteiligung" mit "Gymnasialbeteiligung" gleichsetzt. Gewiss, diese beiden Parameter sind halt entsprechend definiert; – und genau das ist zu kritisieren. Auch nichtgymnasiale Bildung ist Bildung!

Den sozialen Gradienten könnte man einer ähnlichen innerdeutschen Analyse unterziehen und käme vermutlich zu ähnlichen Inkonsistenzen wie beim internationalen Vergleich und bei der rWG. Bei beiden Parametern ist der einstige Musterknabe "Brandenburg" trotz seiner Zweigliedrigkeit 2006 weit nach "vorne" gerückt.

Es ist amüsant zu lesen, wie sich der glühende Einheitsschulverfechter Christian Füller in der "taz" am 30.06.2009 an Erklärungen für diese Entwicklung abmüht, statt einfach festzustellen, dass die Parameter wertlos sind, auch wenn man den einen früher gern gegen Bayern in Stellung gebracht hat. Immerhin hat Füller dabei die Ausbreitung der Privatschulen in den Blick genommen. Vielleicht gelingt ihm da ja noch der Transfer zum Szenario der Einheitsschule.

4. Beispiele für Mängel bei der Datenerhebung von Pisa

Das Image der Objektivität und Neutralität, das Pisa so sehr pflegt, ist ein, allerdings nicht leicht zu durchschauender, Mythos. So sind bei den Erhebungen Unzulänglichkeiten noch und noch vorgekommen, manchmal sogar von Pisa selbst vorgesehen, manchmal gegen die Pisa-Absicht von Schulbehörden, -lehrerinnen & -lehrern, Schülerinnen & Schülern (S&S) fabriziert, manchmal ungeplant;
z.T. repariert, z.T. nicht;
z.T. veröffentlicht, z.T. nicht;
z.T. selbst, z.T. von anderen, z.T. von niemandem bemerkt.

Z.B. hat man sich in *Südtirol* nach 2003 für eine der besten Pisa-Regionen der Welt gehalten, bis kritische Geister recherchiert haben, dass zahlreiche Schulen, vor allem Berufsschulen, mit einer schwächeren Klientel vom Test ausgeschlossen worden waren. – Ich bin überzeugt, dass weltweit immer wieder solche Manipulationen auf Schul- und auf S&S-Ebene vorkommen.

In vielen Ländern, vor allem in *Entwicklungsländern*, geht ein großer Teil der 15-Jährigen (die und nur die von Pisa untersucht werden) gar nicht (mehr) zur Schule und wird nicht erfasst. Aber auch in den *OECD-Ländern* Türkei, Mexiko und Portugal beträgt die Quote der erfassten S&S nur 54%, 58% und 86%, dagegen etwa in USA (trotz der vielen illegalen Migrantinnen & Migranten dort) genau 100,0%, in Schweden gar 102% und in der Toskana satte 108%.

Nach dem starken Absinken *Österreichs* von 2000 bis 2003, in den drei Inhaltsbereichen durchschnittlich um 11 Pisa-Punkte, ließ die Regierung die Ergebnisse von 2000 nachrechnen und erklärte schließlich, dass diese auf unzulässigen Stichproben beruhten und nach unten korrigiert werden müssten.

Als in den *Niederlanden* 2000 und in *Großbritannien* 2003 die vorgeschriebenen Teilnahmequoten von Schulen sowie von S&S nicht erreicht wurden, wurden diese Länder aus

dem Ranking, ihre Daten jedoch nicht aus den Auswertungen ausgeschlossen. Die *USA* wiederum wurden gar nicht ausgeschlossen, obwohl ihre Teilnahmequoten noch viel niedriger lagen.

2000 haben in *Berlin* und in *Hamburg* im Gesamtschulbereich so viele Schulen ihre Mitarbeit verweigert, dass diese beiden Länder aus dem innerdeutschen Vergleich herausgenommen werden mussten.

Das schlechte Ergebnis von *Luxemburg* (446) im Jahr 2000 wurde damit erklärt, dass "Unterschiede [ein Pisa-Euphemismus für "Fehler"] in der … Zuordnung der Testhefte nach Sprachgruppen" vorgekommen waren. Bei den nächsten Durchgängen wurde dieser Fehler vermieden. In der Folge stieg Luxemburg auf 493 und sank dann wieder leicht auf 490.

Starke Punktzahlveränderungen bei vielen Ländern innerhalb von drei Jahren oder gar die Berg- und Talfahrt von *Tschechien* von 498 über 516 bis 510 (alles noch viel gehäufter und extremer bei einem angepassten Einbezug der Timss-Ergebnisse von 1995 bis 2007) werden nicht als Anlass für Skepsis gegenüber der Statistik, sondern als Ausfluss und damit als Indikator einer entsprechenden Leistungsentwicklung genommen. Da wird sogar dem OECD-Pisa-Verantwortlichen Andreas Schleicher widersprochen, wenn er den Anstieg der Deutschen von 2003 bis 2006 in Naturwisssenschaften von 502 auf 516 Punkte mit der Unvergleichbarkeit der beiden Durchgänge relativiert. Hier pflichte ich Schleicher ausnahmsweise bei.

In *Südkorea* waren nur 40% der Teilnehmenden Mädchen, obwohl ihr Anteil an allen Jugendlichen 48% lautet. Diese Differenz beträgt 10 Standardabweichungen und ist damit ganz gewiss nicht zufällig.

Die Definition sowie der Ein- oder Ausschluss von fremdsprachlichen, behinderten oder legasthenischen Jugendlichen wird immer wieder unterschiedlich gehandhabt, was zu ausgeprägten Verzerrungen bei den Länderpunktzahlen führt. Solcherart begründete Ausschlüsse waren bis zu 5% statthaft. Obwohl 2003 die Länder *Dänemark, Kanada, Neuseeland, Spanien, USA* diese Grenze z.T. deutlich (bis zu 7,3%) überschritten, wurden deren Daten ohne Weiteres in die Auswertung mit aufgenommen.

Für die Kalibrierung der Aufgabenschwierigkeiten wurden aus jedem OECD-Land vorab 500 Jugendliche ausgewählt. Die Jugendlichen aus den *USA* haben dabei also etwa 1/1000 des Gewichts der Jugendlichen aus *Island*, und die Jugendlichen aus *Brasilien* haben gar kein Gewicht. – Hätte man die einbezogenen Jugendlichen aller OECD-Länder al-

le bevölkerungsproportional gewichtet, dann hätten alle Punktzahlen aufgrund dieser scheinbar geringfügigen Änderung jedes Mal ca. 10 Punkte besser ausgesehen (die deutschen in Mathematik also 500, 513, 514).

Im Vergleich zu diesen bekannt gewordenen Unzulänglichkeiten dürfte die Dunkelziffer weltweit viel, viel größer sein. – Solche Mängel sind bei einem Unternehmen wie Pisa unvermeidlich; und das spricht, wie gesagt, gegen es.

5. Was hat die Studie der Bertelsmann-Stiftung zum Sitzenbleiben wirklich gezeigt?

Nachdem das Instrument der Kassandra-Rufe zum deutschen Schulsystem in den letzten Jahren von der OECD gepachtet schien, hat sich im Sommer 2009 die Bertelsmann-Stiftung zu Wort gemeldet, deren Feld traditionell eher der Hochschulbereich ist, und die in ihrem Auftrag von Klaus Klemm erarbeitete Studie "Klassenwiederholungen – teuer und unwirksam" als wichtige Nachricht in der deutschen Öffentlichkeit publik gemacht. Der Titel suggeriert, dass die Kosten und die Wirksamkeit von Klassenwiederholungen untersucht worden seien. Tatsächlich stellt die Arbeit den Versuch einer naturgemäß groben und reduktionistischen Kostenberechnung dar, während für die behauptete Unwirksamkeit auf ein paar ältere untaugliche Arbeiten zurückgegriffen wird.

5.1 Wie ermittelt man die Kosten von Klassenwiederholungen?

Dass das Bildungssystem viel billiger wäre, wenn es keine "schlechten" Lernenden gäbe, ist eine Binsenweisheit. In viel größeren Klassen könnte in viel kürzerer Zeit das Nötige gelernt werden. Eigentlich bräuchte man gar keine Klassen (Schulgebäude!) mehr und kaum noch Lehrerinnen & Lehrer, weil die Kinder und Jugendlichen ja bequem zuhause an ihren Rechnern mit perfekter Software (z.B. Logo!) lernen könnten. Die Realität ist aber nun einmal nicht so; es gibt "schlechte" Lernende, und sie verursachen erhebliche Kosten, indem sie den Betrieb aufhalten, zusätzliche Maßnahmen verursachen, usw. bis hin zum frühen Ausscheiden aus dem Dienst einer manchen Lehrperson.

Selbstverständlich können "schlechte" Lernende den Unterricht auch bereichern. Das wären dann Leistungen, die von den Kosten wieder zu subtrahieren wären. Die Kostenstruktur ist jedenfalls viel komplexer, als sie wirkt, wenn man sie auf reale Zahlungen (hier: durch die öffentlichen Haushalte) reduziert. Das hat man in der Volkswirtschaftslehre schon lange erkannt, sieht sich aber mit der Bewertung nicht bezifferter Kosten und Leistungen und deren Zuordnung zu Verursacherinnen & Verursachern immer wieder vor erhebliche Probleme gestellt.

Z.B. müssten in die Berechnungen die (zumindest eine Zeit lang ja vorhandenen) Vorteile der abgebenden Klasse, die der aufnehmenden Klasse (durch die Bereicherung) sowie die der Wiederholerinnen & Wiederholer (W&W) (durch die Chance des Neubeginns usw.) als Leistungen eingehen. — Natürlich weiß man zu wenig über diese Effekte, um sie bewerten zu können. Und genau das spricht überhaupt gegen eine solche Kostenrechnung, wie sie die Bertelsmann-Stiftung in Auftrag gegeben hat.

Wenn man sich aber einmal auf sie einlässt, dann stellt es durchaus eine wichtige Erkenntnis dar, dass die W&W nicht kostenneutral im System mitlaufen, sondern dass Klassenwiederholungen Kosten verursachen. Diese kann man *überschlägig* leicht ermitteln: Einmal Sitzenbleiben verlängert die durchschnittliche Schulzeit von ca. 11 auf ca. 12 Jahre, also um etwa 9%. Im letzten Jahrzehnt blieben vielleicht 30% aller S&S irgendwann einmal sitzen (Mehrfachfälle mehrfach gerechnet). Also wird die Gesamtschulzeit aller S&S durch das Instrument des Sitzenbleibens um 2,7% erhöht. Entsprechend geringer wären also die Personal- (und verwandte) Kosten des Schulsystems, wenn es dieses Instrument nicht gäbe.

Mit Recht weist Klemm (13) darauf hin, dass diese Rechnung nicht nur in denjenigen Bundesländern so anzustellen ist, in denen die Zuweisung von Lehrerstellen (und entsprechender Mittel) pro S&S-Kopf erfolgt, sondern auch in denjenigen, in denen sie klassenweise erfolgt. Wohl ändert sich in einer Schule oft an der Klassenfrequenz eines Jahrgangs nichts, wenn W&W hinzukommen, aber manchmal, nämlich wenn die Höchstgrenze für die Klassenfrequenz überschritten wird, eben doch, und dann muss gleich eine ganze Klasse zusätzlich eingerichtet werden. Man wird diesen Fall in Zeugniskonferenzen tunlichst verhindern; aber das wird nicht immer gelingen. Jedoch auch wenn keine zusätzlichen Klassen eingerichtet werden müssen, entstehen durch W&W zusätzliche Ausgaben.

Klemm setzt zu Recht in den Bundesländern mit klassenbezogener Lehrerstellenzuweisung pro W&W pauschal nur 50% der Kosten an, die in den Bundesländern mit kopfbezogener Zuweisung pro W&W anfallen, die er bei seinem Vorgehen wiederum sehr genau berechnen kann. – Der Prozentsatz von 50% ist extrem willkürlich. Außerdem liegt hier eine Argumentationslücke vor, indem nämlich nicht berücksichtigt wird (zumindest finde ich in der Arbeit keine Silbe dazu), dass in den abgebenden Klassen die Kosten entsprechend geringer werden. – Letztlich ist das aber egal; denn die ganze Rechnung ist voll von Annahmen, Schätzungen, Setzungen. Auch wenn statt des Werts von 50% für diese Kosten

nur 0% angesetzt würde (was bestimmt zu niedrig wäre), käme man noch auf einen Gesamtbetrag von über 700 Mio Euro.

In *mathematisch idealisierter* Form kann man sich die zahlenmäßige Wirkung des Sitzenbleibens so vorstellen, dass jede aufnehmende Klasse ihrerseits wieder W&W nach unten abgibt, ihr Umfang also (bei konstanter W&W-Quote!) über die gesamte Schulzeit hinweg konstant ist, und sich dann fragen, wo der W&W-Überschuss bleibt. Dieser entsteht im ersten Schuljahr, weil dort die Klassen keine W&W mehr nach unten abgeben können. Die ersten Klassen sind folglich alle um 2,6% größer, als sie wären, wenn das Instrument der Klassenwiederholung nicht existieren würde, bzw. es gibt entsprechend 2,6% mehr erste Klassen. Dieser Überhang zieht sich durch die ganze Schulzeit; jeder Jahrgang hat schließlich 2,6% S&S mehr, und in den Abgangsklassen sind dann jedes Jahr 2,6% der S&S ein Jahr länger in der Schule gewesen als vorgesehen.

5.2 Worauf stützt sich die Behauptung von der fehlenden Wirksamkeit von Klassenwiederholungen?

Ihre Eignung als Sensationsmeldung für die Presse zieht die Studie einerseits aus dem scheinbar hohen Kostenbetrag von knapp 1 Mrd Euro (der viel weniger eindrucksvoll ist, wenn man ihn als unter 2% der Gesamtausgaben für die allgemein bildenden Schulen angibt) und andererseits in Verbindung damit aus der Behauptung, dass Klassenwiederholungen unwirksam seien.

Diese Behauptung widerspricht den positiven Erfahrungen ganzer Heerscharen von Lehrerinnen & Lehrern (vgl. a. Tietze & Roßbach 1998, 467) mit den auf einmal wieder möglichen Fortschritten in der abgebenden Klasse, mit der Bereicherung der aufnehmenden Klasse (wenigstens für ein Jahr, nicht nur, aber vor allem, in den Fächern, in denen die W&W nicht so schwach waren) und der Chance eines Neubeginns für einen Teil der W&W. Allerdings handelt es sich bei diesen Erfahrungen wieder nur um die Anhäufung von "opinions", die zwar von *den* Expertinnen & Experten für das Lehren & Lernen stammen, aber eben keine statistisch "belastbare" Forschungsergebnisse darstellen.

Wie belegt nun Klaus Klemm die Behauptung der Unwirksamkeit? Da es zu den Wirkungen auf Klassen verständlicherweise keine brauchbaren Untersuchungen gibt, beschränkt er sich auf die Zielgruppe der W&W selbst und gründet seine Behauptung auf eine Handvoll Arbeiten aus der Literatur (S. 7, 2. und 3. Absatz).

Eine davon ist die Habilitationsschrift von Karlheinz *Ingenkamp* "Zur Problematik der Jahrgangsklasse" von 1967 in der 2., unveränderten Auflage von 1972. Diese bezieht sich auf eine Untersuchung von 1962 an Grundschulkindern im 6. Schuljahr in Berlin-Tempelhof, die wiederum mit einer entsprechenden Untersuchung von 1949 verglichen wird. Ingenkamp bedauert zwar, dass in seinem Buch nicht genug Raum für eine "ideologiekritische Analyse dieses Organisationssystems" (der Jahrgangsklassen) vorhanden ist (ebenda, 29, s.a. 290 u.a.), aber er nimmt über weite Strecken diese Kritik doch vor und plädiert vehement für eine "Integrierte Gesamtschule".

Dieses Plädoyer ist mehrfach widersprüchlich. Der Gegenstand seiner empirischen Forschung, die 6. Klasse, ist ja Teil einer Gesamtschule, nämlich der Berliner Grundschule von 1962, die die ersten sechs Schuljahre umfasste. Seine (negativen) Befunde legen doch gar nicht nahe, die Gesamtschule auf die nächsten Jahrgänge auszudehnen. Mit der Vereinheitlichung der Schulen wiederum fordert er zugleich eine ausgeprägte Differenzierung des Unterrichts, weil ja niemand mehr der Einheitsschule entrinnen kann. Man muss ihm zugute halten, dass er, anders als die heutigen Protagonistinnen & Protagonisten, wenigstens noch die organisatorischen Probleme eines solchen Systems erkennt (ebenda, 301).

Es ist klar, dass Klassenwiederholungen als Kulmination des von ihm kritisierten Systems erst recht in seine Schusslinie geraten. Zur damaligen Zeit wurden ja noch Intelligenztests durchgeführt, und es wundert nicht, dass die W&W bei diesem und bei fachspezifischeren Tests schlechter abschnitten als die "glatt Versetzten", erst recht, wenn die W&W mehrfach wiederholten (ebenda, 106, als Haupt-Argument von Klemm zitiert, und vor allem 275).

Wieso allerdings dieser Umstand gegen die Wirksamkeit von Klassenwiederholungen sprechen soll, erschließt sich mir nicht. Eine *Verringerung* des Leistungsrückstands ist doch auch eine (positive) Wirkung. Dass eine solche vorliegt, ist dermaßen plausibel, dass nicht die Frage, ob, sondern in welchem Ausmaß sie vorhanden ist, interessant gewesen wäre. Dieser Frage ist Ingenkamp allerdings nicht nachgegangen. Die Klassenwiederholung wäre wohl berechtigterweise dann als nicht wirksam zu bezeichnen, wenn die Verringerung des Leistungsrückstands zu schwach ausfiele. Dass der Leistungsrückstand aber dauerhaft auf Null zurückgehen muss, ehe man, konkludent gemäß Ingenkamp und Klemm, von Wirksamkeit sprechen können soll, ist jedoch nicht einzusehen.

Ganz in diesem Sinn spricht die Arbeit von Klemms nächsten Kronzeugen, *Belser & Küsel* (1976), (entgegen deren Tenor) sogar eher *für* den Erfolg von Klassenwiederholungen. Untersucht wurde die "Schullaufbahn von Volksschulabgängern an 26 [von 313] zufällig ausgewählten Schulen Hamburgs" (ebenda, 103) von 1963 bis 1966, also von Jugendlichen, die etwa 1948 bis 1952 geboren sind. Wohl ist "ganz allgemein zwar im Wiederholerjahr eine Leistungsverbesserung zu beobachten, aber schon im nächsten Schuljahr, in dem neue und höhere Anforderungen gestellt werden, sinken die Leistungen wieder ab" (ebenda, 105, zitiert von Klemm, 7, als Beleg für die Unwirksamkeit von Klassenwiederholungen). – Wie weit die Leistungen "absinken", ist ein paar Seiten später ausgeführt: "Insgesamt erweisen sich dabei 75% aller zum Zeitpunkt des Sitzenbleibens ungenügenden Zensuren nach 3 Jahren als dauerhaft, mindestens auf 'ausreichend' verbessert" (ebenda, 111). Die Leistungen sind also vielleicht nicht mehr gut oder befriedigend wie im ersten Jahr, aber eben dauerhaft ausreichend. Dieser Erfolg schlägt sich auch in der Quote der W&W nieder, die den Abschluss erreichen. Während die Hälfte der W&W die Schule mit dem Ende der Schulpflicht, also vor dem Ende der 8. Klasse, verlässt, geht die andere Hälfte ein Jahr länger zur Schule, und davon erreichen 86% (der nur einmal Sitzengebliebenen) den Abschluss. Dieses Faktum wird übrigens nur wenige Zeilen vor dem o.a. scheinbar kritischen Zitat auf Seite 105 mitgeteilt.

Unter den von Klemm zitierten Arbeiten ist (Belser & Küsel 1976) die einzige, die den *langfristigen Ertrag* des Klassenwiederholens kontrolliert. Auch wenn die Autorin und der Autor es nicht wahr haben wollen, hat sich in ihrer Untersuchung das Sitzenbleiben als eine erfolgreiche Maßnahme erwiesen, auch "schwächere" S&S zu einem Abschluss zu führen. – Man darf allerdings nicht außer Acht lassen, dass die Umstände von vor fast einem halben Jahrhundert sich nicht ohne Weiteres auf die heutigen Gegebenheiten übertragen lassen.

Auch beim Lexikon-Artikel von *Tietze & Roßbach* (1998) wird der Charakter des Zitats, das Klemm anführt, erheblich verändert, wenn man es fortsetzt. Trivialerweise schneiden die W&W bei Schulleistungstests nach einem Jahr schlechter ab als diejenigen (gleichschwachen) S&S, die nicht wiederholen und also eine Klasse höher sind. Aber direkt danach folgt: "Werden Sitzenbleiber jedoch mit (leistungsschwachen) Schülern in der gleichen Klassenstufe verglichen (die nicht-versetzten Schüler sind dann mindestens ein Jahr älter), so zeigen sich (geringe) Leistungsdifferenzen zugunsten der Sitzenbleiber" (ebenda 467). – Na also.

Warum die W&W schlechter abschneiden als Diejenigen, die nicht wiederholen, bringen *Tillmann & Meier* (2001) auf den Punkt: "Zum einen sind Wiederholer im Durchschnitt mit weniger guten kognitiven Voraussetzungen ausgestattet …, zum zweiten wird ihnen aber auch die Befassung mit den anspruchsvolleren fachlichen Inhalten der nächsten Klassenstufe verwehrt" (475). Während bei Timss die 9. Schuljahre betrachten werden (mit gewissen Nachteilen), untersucht Pisa die 15-Jährigen und handelt sich dadurch andere Nachteile ein: Z.B. wird, vor allem in Entwicklungsländern, nur eine nicht-repräsentative Stichprobe erfasst, nämlich die der *beschulten* 15-Jährigen. Pisa kann zwar feststellen, dass in gewissen Populationen die Quote der W&W größer ist als in anderen oder dass die W&W weniger Punkte erreichen. Um aber Aussagen zur Wirksamkeit der Klassenwiederholung machen zu können, müsste Pisa auch die 16-Jährigen untersuchen, und es müssten Kriterien festgelegt werden (vielleicht Punktedifferenzen gegenüber wohl bedachten Bezugspopulationen). Wie bei allen von Pisa festgelegten Grenzen (z.B. bei den Kompetenzstufen u.v.a.) wäre das allerdings wieder eine subjektive Angelegenheit, die nur auf "opinions" beruhen würde.

Zusammenfassend lässt sich feststellen, dass die von Klemm herangezogene Literatur, *mit einer Ausnahme*, keine empirisch fundierten Aussagen über die Wirksamkeit von Klassenwiederholungen macht, auch wenn sie, inklusive dieser Ausnahme, sich zu Klassenwiederholungen (mehr oder weniger deutlich) kritisch äußert. Die Ausnahme (Belser & Küsel 1976) hat herausgefunden, dass unter den W&W, die 1965 in Hamburg ein Jahr länger zur Volksschule gingen, 86% den Abschluss schafften. Ob die Autorin & der Autor sich mit ihrer (sachten) Distanzierung vom "Sitzenbleiben" (im Zuge einer recht ausgewogenen Diskussion z.B. auf S. 113) dem Zeitgeist der universitären Pädagogik (nicht nur) der 1970-er Jahre anpassten?

Man muss Klemm zugute halten, dass er Auswahl und Interpretation seiner Literatur i.W. komplett dem Diskussionsbeitrag von *Krohne & Tillmann* (Mitarbeiterin und Leiter der Bielefelder Laborschule) (2006) entnommen hat.

6. Schlussbemerkung

Mit Pisa kann man messen, welche *Pisa-Aufgaben* von wie vielen Jugendlichen gelöst werden. Mit den Zahlenkolonnen kann man allerlei Statistik treiben und dadurch auf manche Tendenz aufmerksam machen. So hat Pisa durchaus seinen Nutzen. (Statt "Pisa" kann man hier viele andere Projektnamen aus der empirischen Bildungsforschung einsetzen.) Vom Anspruch, mit dem Pisa-Quader eine hinter diesen Zahlen stehende umfassen-

de kognitive, soziale und kulturelle Realität abzubilden, ist man jedenfalls weit entfernt, und zwar nicht, weil man noch nicht gut genug ist, sondern weil dieser Anspruch prinzipiell nicht einzulösen ist. Es ist mir darüber hinaus unbegreiflich, wie man mit Hilfe von Pisa-Zahlen die Überlegenheit von Schulsystemtypen beweisen will. Jedenfalls ist die Behauptung, dass die Einheitsschule dem gegliederten Schulsystem überlegen sei, "just another opinion".

Literatur

Belser, Helmut & Gabriele Küsel (1976): Zum Sitzenbleiber-Problem an Volksschulen. In: Rudolf Biermann (Hrsg.): Schulische Selektion in der Diskussion. Bad Heilbrunn: Klinkhardt, 101-115

Bender, Peter (2007): Was sagen uns PISA & Co, wenn wir uns auf sie einlassen? In: Jahnke & Meyerhöfer, 281-337

Ingenkamp, Karlheinz (1972): Zur Problematik der Jahrgangsklasse. 2. Aufl. Weinheim: Beltz

Jahnke, Thomas & Wolfram Meyerhöfer (Hrsg.) (2007): Pisa & Co. Kritik eines Programms. 2. Auflage. Hildesheim & Berlin: Franzbecker

Klemm, Klaus (2009): Klassenwiederholungen – teuer und unwirksam. Gütersloh: Bertelsmann-Stiftung

Krohne, Julia & Klaus-Jürgen Tillmann (2006): "Sitzenbleiben" – eine tradierte Praxis auf dem Prüfstand. In: SchulVerwaltung Spezial 4/2006, 6-9

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

Prenzel, Manfred, Cordula Artelt, Jürgen Baumert, Werner Blum, Marcus Hammann, Eckhard Klieme & Reinhard Pekrun (PISA-Konsortium Deutschland) (Hrsg.) (2007): PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie. Münster u.a.: Waxmann (zit. als Pisa 2007)

Prenzel, Manfred, Cordula Artelt, Jürgen Baumert, Werner Blum, Marcus Hammann, Eckhard Klieme & Reinhard Pekrun (PISA-Konsortium Deutschland) (Hrsg.) (2008): PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich. Münster u.a.: Waxmann (zit. als Pisa 2008)

Tietze, Wolfgang & Hans-Günther Roßbach (1998): Sitzenbleiben. In: Rost, Detlef H. (Hrsg.): Handwörterbuch Pädagogische Psychologie. Weinheim: Beltz, 465-469

Tillmann, Klaus-Jürgen & Ulrich Meier (2001): Schule, Familie und Freunde – Erfahrungen von Schülerinnen und Schülern in Deutschland. In: Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Petra Stanat, Klaus-Jürgen Tillmann & Manfred Weiß (Deutsches PISA-Konsortium) (Hrsg.): PISA 2000. Basiskompetenzen von Schülerinnen & Schülern im internationalen Vergleich. Opladen: Leske + Budrich, 468-509

Wartha, Sebastian (2009): Zur Entwicklung des Bruchzahlbegriffs ─ Didaktische Analyse und empirische Befunde. In: Journal für Mathematik-Didaktik 30, 55-79

Wuttke, Joachim (2007): Die Insignifikanz signifikanter Unterschiede. Der Genauigkeitsanspruch von PISA ist illusorisch. In: Jahnke & Meyerhöfer, 99-246