Joachim Wuttke

Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung

> »One example for the need of mathematical literacy is the frequent demand for individuals to make judgements and to assess the accuracy of conclusions and claims in surveys and studies. Being able to judge the soundness of the claims from such arguments is, and increasingly will be, a critical aspect of being a responsible citizen.«
> *The PISA 2003 Assessment Framework* (OECD 2003 a, S. 27)

Übersicht

PISA ist auf eine bestimmte, extrem reduktionistische Auswertung hin angelegt. Das reichhaltige Datenmaterial, das sich aus der Bearbeitung von weit über hundert Testaufgaben durch jeweils viele zehntausend Schüler ergibt, wird im wesentlichen nur zur Bestimmung von zweierlei Kennzahlen genutzt. Der eine Satz Kennzahlen soll die Schwierigkeit der einzelnen Aufgaben, der andere bestimmte Kompetenzen der einzelnen Schüler beschreiben. Die Aufgabenschwierigkeiten werden in den offiziellen Berichten des PISA-Konsortiums hauptsächlich zur Konstruktion von »Kompetenzstufen« herangezogen; im übrigen konzentriert sich die Analyse auf die statistische Verteilung der Schülerkompetenzen. Die Ergebnisse werden durchgehend nach Staaten aufgeschlüsselt und sind in der Regel als Ranglisten formatiert (z. B. OECD 2001, 2004). Die öffentliche Wahrnehmung von PISA als Nationen-Wettkampf ist von daher durchaus sachgerecht.

Ohne die Suggestivkraft der Ranglisten wäre PISA nicht zum Großereignis geworden. Die Autoren der internationalen und nationalen Berichte konnten diese Wirkung vorhersehen und tragen deshalb eine Mitverantwortung für das, was die Medien aus ihren Ergebnissen gemacht haben. Sie haben zwei Maßnahmen ergriffen, um zu einfachen oder zu weitgehenden Interpretationen vorzubeugen: Erstens haben sie kein Gesamt-Ranking veröffentlicht, sondern sich stets auf einzelne Aufgabenfelder (Lesen, Mathematik, Naturwissenschaften; 2003 zusätzlich Problemlösen) bezogen. Zweitens haben sie zu allen Mittelwerten Standardfehler angegeben und darauf hingewiesen,

dass nicht jeder Leistungsunterschied statistisch signifikant ist.

Diese Maßnahmen haben nicht verhindern können, dass die Presse PISA als ein eindimensionales Ranking des Schülerkönnens dargestellt hat. Journalisten haben sich auf ein einzelnes Aufgabenfeld konzentriert oder durch Mittelwertbildung eine eigene Gesamtskala konstruiert. Solche Vereinfachungen haben ihre theoretische Rechtfertigung spätestens mit der Auswertung von Rindermann (2006) erfahren, der darauf hinweist, dass die Kompetenzwerte in den verschiedenen Aufgabenfeldern extrem stark miteinander korreliert sind – weitaus stärker als die Ergebnisse verschiedener Intelligenztests. PISA misst demnach in erster Linie nicht bestimmte fachliche Kompetenzen, sondern einen Generalfaktor kognitiver Fähigkeiten.

Fraglich ist jedoch, mit welcher Genauigkeit PISA diesen Faktor misst; mit anderen Worten, wie fair der Nationen-Wettkampf ist. Die offiziell mitgeteilten Standardfehler und Signifikanzbewertungen vermitteln den Eindruck erstaunlicher numerischer Präzision. Laut Ergebnisbericht (OECD 2004, S. 92) hat beispielsweise Island 2003 im Aufgabenfeld Mathematik den Leistungsmittelwert 515 erzielt, der mit einem Standardfehler von 1,4 behaftet ist. Eine Tabelle führt auf, dass Island mit 95%iger Sicherheit unter 29 OECD-Staaten einen Rang zwischen dem 10ten und dem 13ten einnimmt. Eine Kreuztabelle besagt, dass die isländischen Schülerleistungen signifikant schlechter als die australischen ($524 \pm 2{,}1$), aber signifikant besser als die schwedischen ($509 \pm 2{,}6$) sind. Sie besagt ferner, dass die Bonferroni-Korrektur, die 95%ige Sicherheit nicht nur für einzelne Vergleiche, sondern für das Ranking als ganzes gewährleisten soll, dem Unterschied zwischen Island und Schweden die Signifikanz nimmt; signifikant ist dann erst der Unterschied zwischen Island und Deutschland ($503 \pm 3{,}3$).

Die Standardfehler, auf denen solche Aussagen beruhen, berücksichtigen jedoch nur zwei bestimmte Quellen stochastischer Ungenauigkeit: die Stichprobenziehung und den Schluss von Aufgabenlösungen auf Kompetenzpunkte. Es liegt nahe, dass es in einer so komplexen Unternehmung wie PISA zahlreiche weitere Quellen von Ungenauigkeiten geben kann. Von der Auswahl und Übersetzung der Aufgaben über die Ziehung der Stichproben und die Durchführung in den Schulen bis hin zur Kodierung der Antworten und Aufbereitung der Daten hat jede Projektphase ihre eigenen Schwierigkeiten. Viele Probleme sind in Fachkreisen bekannt, bisher jedoch nur isoliert erörtert worden.

Beispielhaft sei eine Arbeit von Artelt und Baumert (2004) genannt, die untersuchen, ob die *fairness* des Tests durch die Herkunft der Aufgaben aus bestimmten Sprachräumen gefährdet wird. In PISA 2000 war mehr als die Hälfte der Leseaufgaben auf Englisch eingereicht worden. Die Autoren finden, dass allein dieser Einflussfaktor für zwei Staaten (Vereinigtes Königreich und Irland) eine Verzerrung von 6 bzw. 5 Punkten bewirkt. Das ist deutlich mehr als der stochastische Standardfehler (OECD 2001, S. 53) von 2,6 bzw. 3,2 Punkten. Für Staaten wie Korea oder Japan, die keine Aufgaben beigetragen haben, ist eine quantitative Abschätzung der möglichen Benachteiligung nicht möglich. Ich kann nicht nachvollziehen, wie Artelt und Baumert von diesen Befunden zu der Zusammenfassung kommen, einem potentiellen *cultural bias* sei in PISA »durch eine möglichst multi-kulturelle Zusammensetzung von Testaufgaben begegnet« worden.

Im folgenden möchte ich einen möglichst breiten Überblick über tatsächliche, wahrscheinliche und mögliche Verzerrungen in den PISA-Ergebnissen geben. Komplementär zu anderen Beiträgen in diesem Buch konzentriere ich mich auf Aspekte des Testgeschehens, die unmittelbar mit der numerischen Auswertung zusammenhängen. Ich zeige, wie die numerischen Ergebnisse zustande kommen, welche Informationen auf dem Weg zu diesen Ergebnissen verloren gehen, und mit welchen Ungenauigkeiten und Unsicherheiten sie behaftet sind. Dabei stütze ich mich auf die technischen Berichte (Adams/Wu 2002; OECD 2005 a), das Auswertungshandbuch (OECD 2005 b) und vor allem auf eigene, selbst programmierte Auswertungen des internationalen Datensatzes (ACER 2004, 2005). Auf inferenzstatistische Hypothesentests verzichte ich bewusst; die Originalautoren mögen selber prüfen, welche ihrer Schlussfolgerungen von den im folgenden dokumentierten Fehlerquellen unberührt bleiben.

Der Übersichtlichkeit halber beschränke ich mich auf PISA 2003. Mehrsprachige Staaten schlüssele ich, soweit möglich, nach Regionen auf.[1] Auf

[1] Der internationale Datensatz enthält über 500 Byte pro Schüler, darunter zum Beispiel die Auskunft, ob es im Haushalt eine Geschirrspülmaschine gibt, vermerkt aber nicht, in welcher Sprache der Test durchgeführt wurde. Nur für Belgien und die Schweiz ist die Sprachregion angegeben. Die Sprache der meisten Südtiroler Schulen lässt sich indirekt erschließen. Die Luxemburger Schüler haben sich ganz überwiegend auf Deutsch testen lassen. Für andere mehrsprachige Staaten, wie Kanada, Spanien, Finnland oder Lettland, bleibt die Testsprache Geheimnis der nationalen Projektleitungen.

nationale Ergänzungsstudien gehe ich nicht ein. Das Vereinigte Königreich (UK) wurde wegen verfehlter Teilnahmequoten in sehr inkonsequenter Weise von der offiziellen Auswertung ausgeschlossen: bei der Skalierung der Aufgabenschwierigkeiten und Schülerkompetenzen und bei der Berechnung von OECD-Mittelwerten wurden die britischen Daten noch einbezogen; nur in den Staaten-Ranglisten des Ergebnisberichts werden sie nicht aufgeführt (OECD 2004, S. 33). Um meine Daten möglichst vergleichbar mit denen der offiziellen Berichte zu halten, beziehe ich das UK durchgehend in meine Auswertung ein.

In Teil I fasse ich Einwände gegen die Repräsentativität der untersuchten Stichprobe zusammen. In Teil II weise ich nach, dass die Skalierung der Aufgabenschwierigkeiten und Schülerkompetenzen falsch durchgeführt und unzutreffend dokumentiert wurde. In Teil III untersuche ich die Lösungshäufigkeiten einzelner Aufgaben; ich zeige, dass das Antwortverhalten der Schüler teilweise erheblich von den Modellannahmen abweicht, die der offiziellen Auswertung zugrunde liegen, und schließe auf Dimensionen des Testgeschehens, die dort ausgeblendet oder nur beiläufig erwähnt werden.

Teil I: Wie repräsentativ ist PISA?

Schon bei der Definition der Grundgesamtheit und der Ziehung der Stichprobe zeigt sich, wie schwierig die einheitliche Durchführung eines internationalen Leistungsvergleichs und wie problematisch der Schluss auf einen ganzen Altersjahrgang ist. Bei einigen Ungenauigkeiten lässt sich größenordnungsmäßig abschätzen, dass sie möglicherweise mit fünf, zehn oder mehr Punkten auf nationale Kompetenzmittelwerte durchschlagen. Für solche quantitativen Angaben beziehe ich mich auf die bekannte Punkteskala mit Mittelwert 500 und Streuung 100, auf der, wie in der Einleitung erwähnt, Leistungsunterschiede ab ungefähr 10 Punkten als signifikant angesehen werden.

§ 1 Schulbesuchsquoten

Die Grundgesamtheit von PISA sind fünfzehnjährige Schüler ab dem 7. Schuljahr. Das Alter wurde gewählt, um zu untersuchen, wie gut junge Menschen gegen Ende der obligatorischen Schulzeit auf die Herausforderungen der heutigen Gesellschaft vorbereitet sind (OECD 2003 a, S. 8). Wie in

der vagen Formulierung »gegen Ende« anklingt, gehen manche Jugendliche schon vor oder während dem fünfzehnten Lebensjahr von der Schule ab. In der Türkei beträgt die Schulbesuchsquote nur 54%, in Mexiko 58%. Auch in vielen Nicht-OECD-Partnerländern sind die Schulbesuchsquoten so niedrig, dass die PISA-Ergebnisse in keiner Weise repräsentativ sind.

Das erklärte Ziel, durch internationalen Vergleich die leistungsmäßigen Ergebnisse der nationalen Schulsysteme (*outcomes in terms of student achievements*, OECD 2003 a, S. 6) zu beobachten, wird systematisch verfehlt: in PISA erscheint ein Erziehungssystem umso leistungsfähiger, je mehr schwache Schüler frühzeitig aus ihm herausfallen (vgl. Freudenthal 1975, S. 151). In Portugal gehen beispielsweise allein in den höchstens zwei Monaten, die zwischen der Stichprobenziehung und der Testung liegen (McKelvie 2006 a), über fünf Prozent des Jahrgangs von der Schule ab, wodurch die Schulbesuchsquote auf unter 86% sinkt (OECD 2005 a, S. 168 f.). Wenn man auch Schulabgänger zum Ergebnis eines Erziehungssystems zählt und konservativ abschätzt, dass diese im Mittel um nur eine Standardabweichung, also 100 Punkte, schwächer sind als die verbleibenden Schüler, dann überschätzt PISA die Leistung des portugiesischen Systems um mehr als 14 Punkte.

Dass PISA eine ungünstige Altersstufe testet, haben wohl auch einige Projektpartner eingesehen. Aus einem Sitzungsbericht der OECD (2005 c) geht trotz diplomatischer Sprache deutlich hervor, dass etliche Staaten in Zukunft lieber jüngere Schüler testen möchten. Um die begonnene Zeitreihe fortzusetzen, will eine Mehrheit jedoch an der Testung der Fünfzehnjährigen festhalten. So kam der Wunsch auf, PISA in Zukunft für zwei Altersklassen durchzuführen: es droht also eine Verdoppelung der Kosten.

§ 2 Stichprobenziehung

Für PISA 2003 wurde eine Mindeststichprobengröße von 4500 Schülern gefordert und überall erreicht.[2] In den meisten Staaten wurde die Schülerstichprobe in einem zweistufigen Verfahren gezogen: zunächst wurden Schulen ausgewählt; in Schulen, deren Jahrgangsstärke über einem Richtwert *n* (in

[2] Außer in Island und Luxemburg sowie außerhalb der OECD in Liechtenstein und Macao, wo das mögliche getan und ein ganzer Jahrgang getestet wurde (OECD 2005 a, S. 48, 173).

den meisten Staaten 35) lag, wurden anschließend *n* Schüler ausgewählt. Dieses Verfahren bringt mit sich, dass Probanden aus verschiedenen Schulen verschiedene statistische Gewichte zugeordnet werden müssen.[3] Die Gewichte können so weit auseinander liegen, dass sie durch einen willkürlichen *trimming factor* begrenzt werden müssen, damit nicht das Gesamtergebnis übermäßig von einigen wenigen Schülern abhängt (OECD 2005 a, S. 108ff.).

Für die Ziehung der Schulstichprobe und für die Festlegung der Gewichte ist im Prinzip eine Urliste erforderlich, die für alle Schulen des Teilnehmerstaats aufführt, wie viele Schüler dem Testjahrgang angehören. Eine solche, aktuelle Liste ist in den wenigsten Staaten verfügbar. Sie durfte deshalb durch eines von vier Schätzverfahren ersetzt werden. In Griechenland fehlten die Voraussetzungen selbst für das gröbste Schätzverfahren, so dass alle Schulen gleichgewichtet werden mussten (OECD 2005 a, S. 52.). Für Schweden wurde eine Schulbesuchsquote von 102,5% registriert, für die Toskana von 107,7%. Das lässt zumindest ahnen, wie inkonsistent mancherorts das verwendete Datenmaterial war (OECD 2005 a, S. 168, 183).

§ 3 Ausschlüsse

Die internationalen Regeln erlauben den Teilnehmerstaaten, bis zu 5% der Grundgesamtheit vom Test auszuschließen, und zwar bis zu 0,5% aus organisatorischen Gründen und bis zu 4,5% wegen geistiger oder körperlicher Behinderung oder wegen ungenügender Beherrschung der Testsprache (OECD 2005 a, S. 48). Diese Möglichkeiten sind sehr unterschiedlich ausgestaltet und ausgeschöpft worden. Der technische Bericht nennt keine Abgrenzungskriterien für »intellectual disabilities«; einigen Fußnoten (OECD 2005 a, S. 183 f.) ist jedoch zu entnehmen, dass einer von mehreren Ausschlussgründen in nationaler Verantwortung festgelegt wurde. Die Tschechische Republik hat Schüler ausgeschlossen, die längere Zeit nicht am Unterricht teilgenommen hatten; Luxemburg frisch Zugewanderte; Dänemark, Finnland, Griechenland, Irland und Polen Schüler mit Lese-Rechtschreib-Schwäche; Dänemark zusätzlich Schüler mit *acalculia*.

Innerhalb der OECD reicht die Ausschlussquote von 0,7% in der Türkei bis 7,3% in Spanien und den USA (OECD 2005 a, S. 169). Auch Kanada,

[3] Im folgenden nenne ich die ungewichteten Testteilnehmer »Probanden«, die gewichteten Testteilnehmer »Schüler«, weil sie repräsentativ für alle Schüler eines Landes sein sollen.

Dänemark und Neuseeland überschreiten die 5%-Grenze; das wird im Technischen Bericht vermerkt (OECD 2005 a, S. 241 ff.), hat aber keine weiteren Konsequenzen: die Daten aus diesen Staaten werden uneingeschränkt in die Auswertung einbezogen. Kleinere Regelverstöße bleiben völlig unbeanstandet.

§ 4 Sonderschüler

Unabhängig von allen Ausschlusskriterien sieht das Testdesign die Option vor, in Sonderschulen ein spezielles einstündiges Testheft einzusetzen. Diese Option wurde von nur sieben Staaten genutzt, und zwar in sehr unterschiedlichem Umfang: in Österreich wurden 0,9% aller Schüler mit dem Kurzheft getestet, in Ungarn dagegen 6,1%. Die nationalen Leistungsmittelwerte der Schüler, die das einstündige Heft bearbeitet haben, streuen enorm, im Lesen zum Beispiel von 215 in Österreich bis 397 in Ungarn. Das muss nicht überraschen, denn in Österreich wurden 1,6% der Grundgesamtheit, in Ungarn aber 3,9% *ganz* vom Test ausgeschlossen: Wer in Österreich den Kurztest bearbeitet hat, wäre in Ungarn wahrscheinlich überhaupt nicht getestet worden; wer in Ungarn zum Kurztest eingeteilt wurde, wäre in Österreich wahrscheinlich zum regulären Test herangezogen worden.

Wie der Vergleich zwischen Deutschland (1,9% Ausschluss, 3,6% Kurztest, 287 Leseleistung im Kurztest) und den Niederlanden (1,9%, 3,0%, 380) zeigt, erklären die offiziell registrierten Ausschlussquoten aber nicht jeden Leistungsunterschied: entweder sind die niederländischen Sonderschulen den deutschen haushoch überlegen, oder es gibt weitere, weniger offensichtliche Ungleichmäßigkeiten in Zielgruppendefinition, Stichprobenziehung oder/und Testdurchführung. Prais (2003), der die Uneinheitlichkeiten bei der Durchführung von PISA 2000 und die dafür gegebenen oder rekonstruierbaren Erklärungen »kafkaesk« (S. 149) nennt, weist darauf hin, dass *innerhalb* der Sonderschulen leistungsschwächere Schüler vom Test ausgeschlossen werden konnten (S. 158).

Wenn Kurztestteilnehmer einheitlich von der Auswertung ausgenommen werden, steigt die Leseleistung in Deutschland um 7,6 Punkte; im OECD-Ranking verbessert sich Deutschland allein dadurch vom 18. auf den 12. Rang. Abbildung 1 zeigt, dass der Anteil besonders schwacher Schüler in

Deutschland mit Sonderschulen leicht über, ohne Sonderschulen leicht unter dem OECD-Mittel liegt.[4]

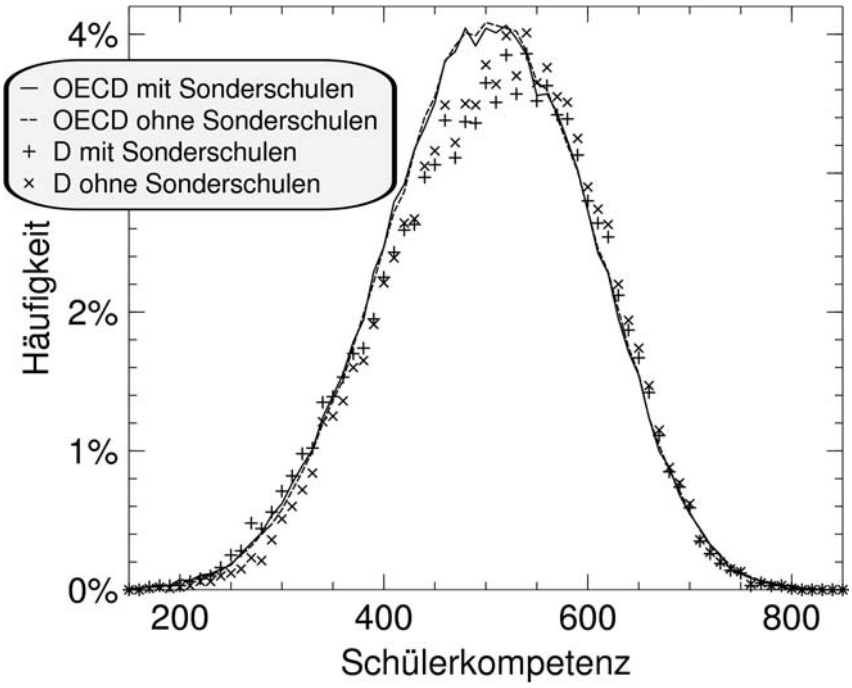

Abbildung 1: Die Mathematikkompetenz in Deutschland und im Mittel über alle 30 OECD-Staaten, jeweils mit oder ohne Einbeziehung der Kurzhefte. OECD-weit ist der Unterschied zwischen beiden Auswertungen sehr gering, weil nur wenige Staaten in nennenswertem Umfang Kurzhefte einsetzen. In Deutschland liegt der Anteil sehr schwacher Schüler (Kompetenz zwischen 250 und 350) nur dann über dem OECD-Mittel, wenn Kurzhefte einbezogen werden. – Alle Abbildungen in diesem Beitrag beruhen auf von mir selbst programmierten und durchgeführten Auswertungen des internationalen Datensatzes.

§ 5 Korrelation von Teilnahmequote und Testleistung

Auf beiden Stufen der Stichprobenziehung gibt es Ausfälle: Schulen lehnen die Teilnahme ab, und Schüler erscheinen nicht zum Test oder bleiben nicht bis zum Schluss. In PISA wird eine Teilnahmequote von 85% aller Schulen gefordert; Quoten zwischen 65% und 85% können durch Nachbenennen von Ersatzschulen geheilt werden, ohne dass dadurch 85% erreicht werden müssen. Schulen, in denen weniger als 50% der ausgewählten Schüler den kognitiven Test abgeschlossen haben, werden als nicht teilnehmend gezählt; sofern diese Quote über 25% lag, werden die Ergebnisse der teilnehmenden Schüler nichtsdestoweniger in den internationalen Datensatz aufgenommen

[4] Ausschlussquoten aus OECD 2005 a, S. 169; die übrigen Angaben aus eigener Filterung und Mittelung des internationalen Datensatzes.

und sogar hoch gewichtet. Landesweit wird eine Schülerteilnahmequote von 80% gefordert (OECD 2005 a, S. 48–50).

Das letztgenannte Quorum wurde 2003 von allen OECD-Staaten außer dem UK erreicht, von einigen aber nur knapp: in Australien, Österreich, Kanada, Irland, Polen und den USA lag die Schülerteilnahemequote unter 84%, in acht weiteren Staaten unter 90%. Die Schulteilnahmequote lag in der Mehrheit der Teilnehmerstaaten spätestens nach Heranziehung von Ersatzschulen zwischen 98% und 100%, in Belgien, Griechenland, Irland, Japan und Mexiko aber unter 96%, in Australien, Frankreich, den Niederlanden und Norwegen unter 91%. Das UK verfehlte auch dieses Kriterium (anfänglich 64,9%, mit Ersatzschulen 77,4%) und wurde letztlich aus allen OECD-Ranglisten ausgeschlossen (OECD 2005 a, S. 171–173).

Die niedrigen Schulteilnahmequoten der USA (64,9%, mit Ersatz 68,1%) und Kanadas (80,0%, 84,4%) wurden jedoch hingenommen, obwohl mehrere Verstöße gegen *technical standards* festgestellt wurden (OECD 2005 a, S. 238 ff.). Ähnlich war schon in PISA 2000 zugunsten der USA und des UK verfahren worden. Diese a posteriori und ad hoc beschlossene Missachtung selbstgegebener Regeln hat alsbald Kritik auf sich gezogen (von Collani 2001, S. 234 ff.; Prais 2003 S. 149 f.). Der PISA-Projektleiter warf Prais daraufhin »misunderstanding« und »a lack of research« vor. Er gestand zu, dass niedrige Antwortraten »a matter for concern« und »a threat of bias« seien, berief sich aber auf Zusatzuntersuchungen, die für das UK keine Korrelation zwischen Teilnahmequoten und Leistungsfähigkeit fanden. Für PISA 2003 kündigte er eine ergänzende Überprüfung anhand landesweiter Lernkontrollen an (Adams 2003, S. 383 ff.). Sie ergab, dass in der PISA-Stichprobe sowohl besonders schwache als auch besonders starke Schulen unterrepräsentiert sein könnten, und dass die Nichtteilnahme einzelner Schüler wahrscheinlich eine Verzerrung der Leistungsmittelwerte bewirkt, die nicht durch Höhergewichtung anderer Schüler ausgeglichen werden kann. Dieses Ergebnis, das Prais in einem wichtigen Punkt recht gibt, wurde an entlegener Stelle im Technischen Bericht veröffentlicht (OECD 2005 a, S. 246 f.). Für die USA wurde nichtsdestoweniger erneut mit der Nichtauffindbarkeit einiger Korrelationen argumentiert, aus der man entnehmen könne, dass der Datensatz »relativ wenig« durch Schulnichtteilnahme verzerrt sei (OECD 2005 a, S. 247 f.).

Ich vermute, dass – gemessen am Genauigkeitsanspruch der offiziellen Auswertung – nicht allein die Ausnahmen problematisch, sondern die Regeln selbst nicht streng genug sind, um Verzerrungen der Testergebnisse sicher auszuschließen. Für Frankreich erlaubt es eine technische Besonderheit des Datensatzes, diese Vermutung empirisch zu überprüfen.[5] In 115 Schulen mit einer Teilnahmequote zwischen 75% und 100% erzielen die Schüler eine mittlere Testleistung von 522 PISA-Punkten; in 22 Schulen mit 50–75% 494 Punkte; in 7 Schulen mit 25–50% 421 Punkte (Abb. 2).

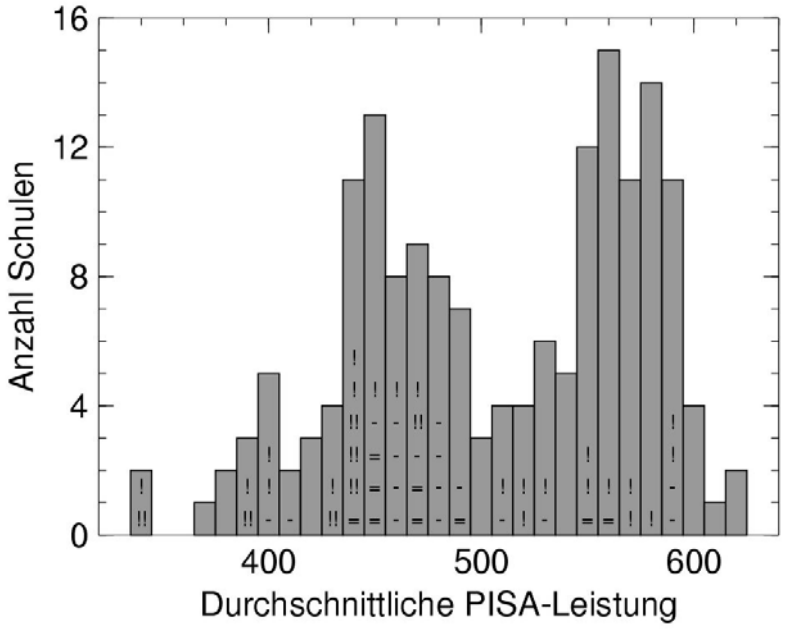

Abbildung 2: Das Säulenhistogramm zeigt die Testleistungen der 170 Schulen aus Frankreich, die an PISA 2003 teilgenommen haben. Die Testleistung wurde hier über alle vier Aufgabenfelder gemittelt (gewichtet im Verhältnis 2:7:2:2 gemäß der Anzahl der gestellten Aufgabenblöcke). Schulen aus den Strata »sehr kleine Schulen« und »ziemlich kleine Schulen« sind mit = bzw. - markiert. Unter allen übrigen Schulen sind die sieben mit der geringsten Schülerbeteiligung (zwischen 25% und 50%) mit !! markiert; Schulen mit einer Teilnahmequote zwischen 50% und 75% sind mit ! bezeichnet.

Es besteht also zumindest in Frankreich eine erhebliche Korrelation zwischen schulweiter Teilnahmequote und durchschnittlicher Testleistung. Die dadurch

[5] In Frankreich bilden »kleine« und »sehr kleine« Schulen eigene Strata. Laut technischem Bericht zählen 163 von 183 ausgewählten Schulen als teilnehmend (OECD 2005 a, S. 172). Der internationale Datensatz enthält jedoch 170 Schulen; die Diskrepanz erklärt sich durch genau sieben nicht kleine Schulen, in denen zwischen 8 und 15 Schüler teilgenommen haben. Das stützt die Annahme, dass alle nicht kleinen Schulen groß genug waren, um eine Stichprobe von 32 Schülern zu ziehen. Unter dieser Annahme ist es möglich, die ansonsten nicht veröffentlichten Teilnahmequoten der einzelnen Schulen zu erschließen.

entstehende Verzerrung würde nur dann durch die Höhergewichtung der teilnehmenden Schüler ausgeglichen, wenn es in den problematischen Schulen keine Korrelation zwischen Leistungsfähigkeit und Testkomplianz gäbe. Für diese Annahme besteht kein Anlass. Zusätzlich dürfte es eine Verzerrung durch Schulen geben, die wegen einer Schülerbeteiligung unter 25% als nicht teilnehmend gewertet werden. Aus diesen Gründen ist zu vermuten, dass das Leistungsmittel für Frankreich um etliche Punkte überschätzt wird. Von der Schul- und Schülerteilnahmequote her liegt Frankreich im Mittelfeld; in etlichen anderen Ländern sind die Ergebnisse vielleicht noch stärker verzerrt.

§ 6 Geschlechterverteilung

Die Repräsentativität der Stichprobe lässt sich auch anhand der Variablen Geschlecht und Geburtsdatum überprüfen. Dabei ist zu berücksichtigen, dass etwas mehr Jungen als Mädchen geboren werden. In der OECD liegt der Mädchenanteil in der Altersklasse der Zehn- bis Zwanzigjährigen zwischen 47,5% in Südkorea und 49,3% in Mexiko (U. S. Census Bureau 2006, vgl. NCHS 2006). Im PISA-Datensatz streut der Mädchenanteil hingegen zwischen 40,5% in Südkorea und 52,6% in Frankreich. Die konservative Abschätzung in Tabelle 1 zeigt, dass die Abweichung zwischen PISA-Stichprobe und Grundgesamtheit in fünf von dreißig OECD-Staaten mehr als 5σ beträgt. Dass diese Abweichungen allein auf stichprobenbedingte Variation zurückgehen, ist jenseits aller Plausibilität.

In Südkorea ist bereits das Geschlechterverhältnis im Altersjahrgang auffällig. Möglicherweise wird die Anomalie in den gemittelten Daten des U. S. Census Bureau sogar noch unterschätzt; andere Quellen deuten darauf hin, dass der Mädchenanteil im PISA-2003-Jahrgang zwischen 46% und 47% liegt.[6] Auch nach Berücksichtigung dieser Anomalie bleibt jedoch eine ganz erhebliche Diskrepanz von ca. 9σ zwischen dem Altersjahrgang und der PISA-Stichprobe.

[6] Weltweit liegt der Mädchenanteil bei der Geburt bei ungefähr 48,8%. Das Problem der »fehlenden Mädchen« in Asien ist Gegenstand intensiver Forschung (CEPED 2006). Zum Teil kann es möglicherweise als Auswirkung von Hepatis-B erklärt werden (Oster 2005); in Südkorea beruht es jedoch überwiegend auf selektiver Abtreibung namentlich ab der zweiten Schwangerschaft (Song 1998, Kim 2004). Diese Praxis hat Mitte der 1990er Jahre ihren Höhepunkt erreicht und zu einem Mädchenanteil von knapp unter 46,5% geführt (CEPED 2006).

Staat	Mädchenanteil			
	Bevölkerung	PISA 2003		Differenz
Südkorea	47,7%	40,5%	(0,7%)	–10,0 σ
Türkei	49,2%	45,0%	(0,7%)	–5,2 σ
Ungarn	48,9%	47,2%	(0,7%)	–1,6 σ
Schweden	48,6%	50,0%	(0,7%)	+1,1 σ
Finnland	48,9%	50,1%	(0,7%)	+1,2 σ
Luxemburg	48,7%	50,8%	(0,8%)	+2,0 σ
Dänemark	48,8%	50,9%	(0,8%)	+2,1 σ
Japan	48,8%	51,7%	(0,7%)	+3,3 σ
Spanien	48,6%	50,8%	(0,5%)	+3,4 σ
Griechenland	48,6%	51,7%	(0,7%)	+3,5 σ
Frankreich	48,9%	52,6%	(0,8%)	+4,2 σ
Kanada	48,9%	50,7%	(0,3%)	+4,3 σ
Portugal	48,0%	52,4%	(0,7%)	+5,3 σ
Italien	48,6%	51,9%	(0,5%)	+6,0 σ
Mexiko	49,4%	51,8%	(0,3%)	+6,9 σ

Tabelle 1: Mädchenanteil in der Bevölkerung und in der PISA-Stichprobe. Der Mädchenanteil in der Bevölkerung (U. S. Census Bureau o. J.) ist über die Altersklassen 10–14 und 15–19 Jahre gemittelt (die Abweichung beträgt nur in zwei Staaten mehr als 0,2% und überall weniger als 0,5%). Zum Mädchenanteil von PISA 2003 ist in Klammern der Standardfehler σ angegeben, der sich aus der Stichprobengröße ergibt. Die Differenz zwischen Bevölkerungsjahrgang und PISA-Stichprobe ist in Vielfachen von σ angegeben. Vor Normierung auf σ wurde der Differenzbetrag um 0,5% reduziert, um etwaige Ungenauigkeiten der Bevölkerungsdaten konservativ abzuschätzen.

Eine solche Abweichung kann drei verschiedene Ursachen haben: (1) geschlechtsabhängige Schulbesuchsquoten, (2) Fehler bei der Stichprobenziehung und (3) geschlechtsabhängige Teilnahmequoten. Laut Technischem Bericht (OECD 2005 a, S. 171 ff.) gehen in Südkorea 99,94% aller Fünfzehnjährigen zur Schule, die Schulteilnahmequote betrug 100%, und die Schülerteilnahmequote hatte den Spitzenwert von 98,81%. Wenn diese Angaben zutreffen, können die Ursachen (1) und (3) ausgeschlossen werden; dann muss es zu eklatanten Fehlern bei der Stichprobenziehung gekommen sein. Auffällig ist überdies die Altersverteilung: 29,7% der Schüler sind im ersten Drittel, 38,7% im letzten Drittel des getesteten Jahrgangs geboren. Bei funktionierender Plausibilitätskontrolle hätten die koreanischen Daten nicht in die Auswertung einbezogen werden dürfen.

Der Mädchenanteil von nur 45,0% in der Türkei dürfte hingegen einen realen Geschlechterunterschied in der Schulbesuchsquote wiederspiegeln und

bestätigt, dass die Wahl der PISA-Grundgesamtheit ungeeignet ist, das Gesamtergebnis des türkischen Erziehungssystems zu untersuchen. Schwerer zu beurteilen ist, wieweit der zu hohe Mädchenanteil der PISA-Stichprobe in etlichen entwickelteren Staaten auf unterschiedlichem Schulbesuch und wieweit auf unterschiedlicher Testteilnahme beruht. Bei beiden Ursachen ist eine Korrelation mit der Leistung zu vermuten.

§ 7 Umgang mit unvollständigen Testheften

PISA 2003 bestand aus einem zweistündigen »kognitiven« Test, gefolgt von einer knappen Stunde, in der mit Fragebögen verschiedenste Hintergrundvariablen erhoben wurden. Laut Regelwerk war als Teilnehmer zu werten, wer an der kognitiven Testung teilgenommen hatte. In den internationalen Datensatz sollten demnach auch Schüler aufgenommen werden, die den Test vor oder während der Fragebogen-Sitzung abgebrochen haben (OECD 2005 a, S. 50, 162).

Bei der Berechnung der Leistungskennzahlen werden fehlende »conditioning variables« (dazu unten § 14) durch Mittelwerte ersetzt (OECD 2005 a, S. 129, 402 ff.). Für weiterführende Untersuchungen aber, in denen kognitive Ergebnisse zum Beispiel auf den sozialen Hintergrund der Schüler bezogen werden sollen, sind Datensätze mit unausgefülltem Fragebogen wertlos. PISA-Auswertungen liegen deshalb, je nach Untersuchungsziel, unterschiedlich umfassende Stichproben zugrunde.

Der erste und umfangreichste Fragebogen, das *Student Questionnaire*, umfasst über 100 Teilfragen. Aus verschiedensten Gründen lassen manche Schüler einzelne Fragen unbeantwortet. In Polen sind jedoch sieben Fragen von keinem einzigen Schüler *nicht* beantwortet worden, und es findet sich kein einziger polnischer Schüler, der weniger als 25 Fragen gültig beantwortet hat. Solange keine andere Erklärung für diese Anomalie glaubhaft gemacht wird, muss, wenn die Verlässlichkeit der PISA-Ergebnisse beurteilt werden soll, als maximale Verzerrung unterstellt werden, dass Polen Schüler, die die Fragebögen nicht oder sehr unvollständig bearbeitet haben, nicht zum internationalen Datensatz beigetragen und dadurch das nationale Leistungsmittel geschönt hat.

In Kanada haben 9,7% aller Schüler das *Student Questionnaire* überhaupt nicht bearbeitet, in Deutschland 2,1% (nach Ausschluss der Sonderschüler, sonst 5,6%), überall sonst deutlich weniger. Auch das deutet auf eine uneinheitliche Umsetzung des Teilnahmekriteriums hin. Fast überall liegen die Testleistungen derjenigen Schüler, die das *Student Questionnaire* nicht bearbeitet haben, unter dem nationalen Durchschnitt, oft um 100 oder mehr Punkte. Wenn man Schüler, die das *Questionnaire* nicht bearbeitet haben und die nicht mit dem Kurzheft getestet wurden, ausschließt, steigt die mittlere Leistung in Kanada um 3,6 Punkte, in Deutschland um 1,5 Punkte, in den übrigen OECD-Staaten im Mittel um nur 0,2 Punkte.

§ 8 Testtermin

Bei der Ziehung der Schülerstichprobe wird das Lebensalter an einem bestimmten Stichtag zugrunde gelegt. Der Test darf bis zu einem Monat vor oder nach diesem Stichtag stattfinden (Adams/Wu 2002, S. 39; OECD 2005 a, S. 46). Das heißt, das mittlere Alter der Probanden kann um bis zu zwei Monate differieren. Der Leistungszuwachs in einem Schuljahr macht in der Deutung des deutschen Konsortiums 35 bis 40 Punkte aus (Prenzel *et. al.* 2004, S. 32). Wenn man dieser, durchaus angreifbaren, Umrechnung folgt und berücksichtigt, dass in einem Schuljahr höchstens zehn Monate lang gelernt wird, dann machen zwei Monate einen Leistungsunterschied von 7 oder 8 Punkten aus. Auf Grundlage der veröffentlichten Dokumentation lässt sich weder er-, noch ausschließen, dass Ergebnisse einzelner Länder systematisch verzerrt werden.

Teil II: Wo kommen die Punkte her?

Im folgenden rekonstruiere ich, wie die offizielle Auswertung aus den kognitiven Testergebnissen Kennzahlen bestimmt, die als Aufgabenschwierigkeiten und Schülerkompetenzen gedeutet werden. Dabei finde ich erhebliche Abweichungen zwischen der tatsächlich durchgeführten Auswertung und der im Technischen Bericht beschriebenen, die darauf hindeuten, dass der mathematische Kern des Item-Response-Modells nicht korrekt implementiert worden ist.

§ 9 Testdesign, Datenerfassung und -aufbereitung

Der kognitive Test ist in Aufgabenstämme gegliedert. Ein Aufgabenstamm umfasst ein bis sieben einzelne Aufgaben (*Items*). Jeder Schüler bearbeitet rund 50 Aufgaben, die sich auf vier Aufgabenfelder und innerhalb des schwerpunktmäßig untersuchten Feldes Mathematik auf vier Unterfelder verteilen. Um zu vermeiden, dass die Testergebnisse in extremer Weise vom »Funktionieren« einzelner Aufgaben abhängen, werden dreizehn verschiedene Testhefte eingesetzt. Um Ergebnisse aus verschiedenen Heften miteinander vergleichen zu können, wird jede der 165 Aufgaben in vier verschiedenen Heften gestellt, jeweils in einem anderen halbstündigen Block (Tab. 2).

Zeitablauf	Testheft												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Stunde	M1	M2	M3	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2
	M2	M3	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2	M1
Pause													
2. Stunde	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2	M1	M2	M3
	L1	L2	P1	P2	M1	M2	M3	M4	M5	M6	M7	N1	N2
Pause													
3. Stunde	Fragebögen												

Tabelle 2: Testdesign von PISA 2003: Jedem Schüler wird eines von dreizehn Testheften zugewiesen. Jedes Heft enthält vier Blöcke, mit Aufgaben aus je einem der vier Felder Lesen (L), Mathematik (M), Naturwissenschaften (N) und problemlösendes Denken (P).

Dieses Testdesign bringt mit sich, dass Ergebnisse aus verschiedenen Heften oder Blöcken nicht unmittelbar miteinander verglichen werden können. Die unterschiedliche Mischung leichter und schwerer Aufgaben kann zu ganz unterschiedlich verteilten Lösungshäufigkeiten führen (Abb. 3). Um dennoch zu einer eindimensionalen Bewertung der Schülerleistungen zu kommen, haben die Veranstalter von vornherein eine ganz bestimmte, modellabhängige Auswertung geplant. Immerhin stellt das Testdesign von PISA 2003 einen erheblichen Fortschritt gegenüber 2000 dar, wo die Verteilung der Aufgaben auf die Testhefte so asymmetrisch war, dass nahezu jede modellunabhängige Auswertung obstruiert wurde.

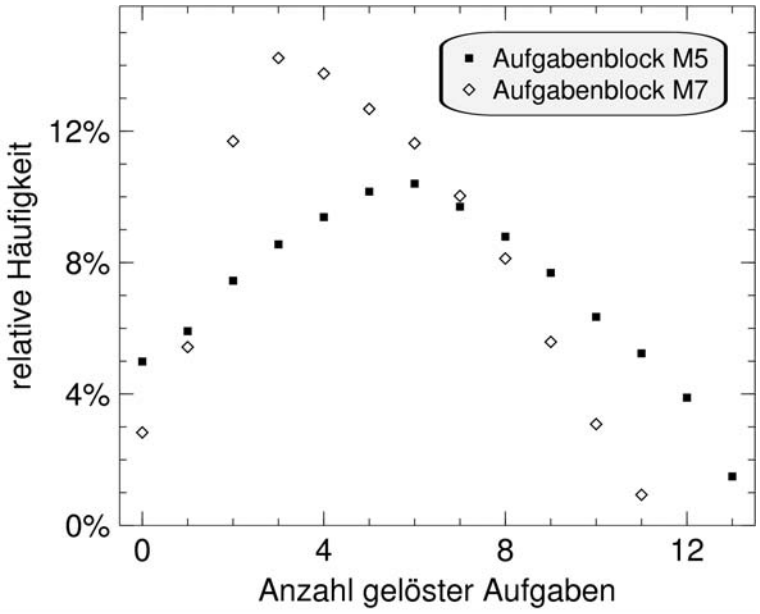

Abbildung 3: Diese Auftragung zeigt beispielhaft für zwei Aufgabenblöcke, wie viel Prozent der Schüler, im Mittel über 30 OECD-Staaten, wieviel Aufgaben richtig gelöst haben.

Als erster Schritt der Datenauswertung werden die Schülerhefte in nationaler Verantwortung von eigens geschulten Hilfskräften kodiert. Die kodierten Schülerantworten aus kognitivem Test und Fragebögen werden in eine Datei zusammengeführt und an das internationale Projektzentrum beim Konsortialführer ACER (Australian Council of Educational Research) übermittelt. Dort werden die nationalen Ergebnisse zum internationalen Datensatz zusammengesetzt.

Anschließend werden aus der Gesamtheit der Schülerantworten die Aufgabenschwierigkeiten und Schülerkompetenzen bestimmt. Dieser Schritt wird als die *Skalierung* des internationalen Datensatzes bezeichnet. Dabei wird ein Item-Response-Modell eingesetzt, dessen probabilistische Natur mit sich bringt, dass man für eine Kompetenz eines Schülers keinen eindeutigen Zahlenwert, sondern eine Wahrscheinlichkeitsverteilung erhält; diese Verteilung ist umso schärfer um ihren Mittelwert konzentriert, je modellkonformer der Schüler mit dem Test zurecht gekommen ist. Für die gesamte weitere Auswertung wird die Wahrscheinlichkeitsverteilung der Kompetenz eines Schülers pro Aufgabenfeld durch fünf unabhängige Zufallszahlen, sogenannte *plausible values*, repräsentiert. Für jeden Schüler und für jedes Aufgabenfeld werden diese je fünf Zahlenwerte dem internationalen Datensatz hinzugefügt. Der so skalierte Datensatz wird dann den nationalen Projektzentren zurückübermittelt; nach Veröffentlichung der offiziellen Berichte wird er auf der Website von ACER frei verfügbar gemacht.

§ 10 Skalierung: obskure Dokumentation

Zur Skalierung verwendet ACER proprietäre Software, die vermutlich nicht einmal den nationalen Projektpartnern offengelegt wird. Ich kann die Skalierung deshalb nicht im Detail nachvollziehen – aber ich kann die Konsistenz der Ergebnisse prüfen. Dabei finde ich verstörende Diskrepanzen zwischen dem internationalen Datensatz und den Angaben im Technischen Bericht (OECD 2005 a).

Dort wird die Skalierung in Kapitel 9 beschrieben. Der Text ist weitgehend unverändert aus dem Bericht zu PISA 2000 (Adams/Wu 2002) übernommen.[7] Damals waren noch die Autoren der einzelnen Kapitel angegeben; für Kapitel 9 zeichnete allein Ray Adams, der Projektleiter des internationalen Konsortiums, verantwortlich. Als Autoren der Software zitiert er Wu, Adams und Wilson. Wilson ist weder 2000 noch 2003 an der Auswertung von PISA beteiligt gewesen (OECD 2001, S. 320ff.; OECD 2004, S. 474ff.).

Kapitel 9 gibt keine nachprogrammierbare Beschreibung des verwendeten Algorithmus. Es gibt aber auch keinen allgemeinverständlichen Überblick über dessen Kerngedanken. Die letztlich simple Mathematik wird unter einer ungeschickten, inkonsequenten und unnötig allgemeinen Notation verschüttet.[8] Adams behandelt durchgehend den Fall, dass eine Aufgabe außer falschen und richtigen Lösungen auch verschiedene Abstufungen teilrichtiger Lösungen zulässt (*partial credits*). Das betrifft 23 der insgesamt 165 Aufgaben von PISA 2003. Zur Vereinfachung der Darstellung beschränke ich mich im folgenden auf die übrigen 142 nur als falsch oder richtig kodierbaren (zweiwertigen) Aufgaben.

Jede Aufgabe gehört genau einem Aufgabenfeld an. Das versteckt Adams hinter einer Menge linearer Algebra, die zunächst den Anschein erweckt,

---

[7] Lange Passagen lassen sich noch weiter zurückverfolgen; sie stammen wörtlich aus Adams *et al.* (1997 b) oder/und Adams *et al.* (1997 a). Auch diese Arbeiten sind nicht selbsterklärend, aber meine Geduld im Zurückverfolgen der Selbstzitate war dort erschöpft.

[8] Für Funktionen fällt Adams kaum ein anderer Buchstabe als f ein. Formel (9.10) verknüpft beispielsweise drei paarweise verschiedene Funktionen, die f_θ, f_x und f_x (sic!) heißen, wobei die Buchstaben θ und x in derselben Formel auch noch als Funktionsargumente auftreten und eine der Funktionen f_x zuvor ohne Subscript eingeführt worden ist.

dass jede Aufgabe zu mehreren latenten Variablen beiträgt (OECD 2005 a, S. 120, Formel 9.2). Aus anderen Angaben lässt sich zurückschließen, dass die Matrix, die diese Koppelung vortäuscht, pro Spalte außer *einer* Eins nur Nullen enthält (Adams *et al.* 1997 b). Wenn ich die multidimensionale Maskerade richtig durchschaue, dann besagt Kapitel 9, dass die offizielle PISA-Auswertung für die meisten Aufgaben auf dem einfachsten aller Item-Response-Modelle, dem einparametrigen Rasch-Modell beruht.

In diesem Modell wird jedem Schüler *i* pro Aufgabenfeld eine latente Variable θ_i zugeschrieben und als seine Kompetenz gedeutet; jede Aufgabe *j* wird durch einen einzigen Parameter ξ_j charakterisiert, der als ihre Schwierigkeit gedeutet wird.

Die Wahrscheinlichkeit, dass Schüler *i* Aufgabe *j* richtig löst, wird als

$$P_{ij} = \frac{1}{1+\exp(-(\theta_i - \xi_j)/100 - \ln(62/38))} \qquad (1)$$

angenommen. Der mathematische Gehalt dieser Gleichung besteht darin, dass sie glatt zwischen den Grenzfällen $P_{ij} \to 0$ für $\theta_i \ll \xi_j$ und $P_{ij} \to 1$ für $\theta_i \gg \xi_j$ interpoliert. Die willkürliche Konstante $\ln(62/38) \simeq 0{,}490$ hat das Konsortium so gewählt, dass eine Aufgabe mit der Wahrscheinlichkeit 62% gelöst wird, wenn $\xi_i = \theta_j$. Der Skalenfaktor 100 wurde aus unerklärten Gründen identisch mit der Standardabweichung des Bevölkerungsmodells gewählt.[9]

Das Skalierungsproblem besteht darin, aus der Gesamtheit aller Schülerantworten sowohl die Schwierigkeiten ξ_j aller Aufgaben *j* als auch die Kompetenzen θ_i aller Schüler *i* zu ermitteln. Im Rahmen des Modells (1) ist es mit Hilfe des Satzes von Bayes möglich, aus gegebenen ξ_j auf die Wahrscheinlichkeitsverteilung der θ_i oder aus gegebenen θ_i auf die Wahrscheinlichkeitsverteilung der ξ_j zu schließen.

Zu Beginn der Auswertung sind aber weder die Aufgabenschwierigkeiten noch die Schülerkompetenzen bekannt; es ist lediglich vorgegeben, dass die Schülerkompetenzen normalverteilt sein sollen. Deshalb ist das Skalierungsproblem keineswegs trivial.

[9] Der Technische Bericht springt unsystematisch zwischen einer internen Skala mit Mittelwert 0 und Standardabweichung 1 und der in den Ergebnisberichten verwendeten 500-100-Skala hin und her. Ich übersetze hier einheitlich in die 500-100-Darstellung.

Naheliegend fände ich, es iterativ zu lösen: man bestimmt aus den Lösungshäufigkeiten eine nullte Näherung der Aufgabenschwierigkeiten, daraus eine erste Näherung der Schülerkompetenzen, renormiert diese auf die gewünschte 500-100-Skala, berechnet damit eine verbesserte Näherung der Aufgabenschwierigkeiten, dann wieder eine verbesserte Näherung der Schülerkompetenzen, renormiert und wiederholt die Prozedur, bis die Lösung selbstkonsistent ist.[10]

Diesem Ansatz folgt ACER laut Technischem Bericht *nicht*. Zur Skalierung des PISA-Datensatzes wurden die Aufgabenschwierigkeiten nur in nullter Näherung und die Schülerkompetenzen in erster Näherung bestimmt; es wurde weder renormiert noch iteriert (OECD 2005 a, S. 122, 128 ff.). Von daher ist zu erwarten, dass die Schülerkompetenzen nicht genau der vorgegebenen Gaußverteilung folgen und dass die veröffentlichten Aufgabenschwierigkeiten nicht unbedingt mit Gleichung (1), geschweige denn mit einem detaillierteren Modell konsistent sind.

§ 11 Skalierung: fehlerhafte Durchführung

Aus unerklärten Gründen hat ACER zur Skalierung der Aufgabenschwierigkeiten nur einen Bruchteil der internationalen Schülerstichprobe, nämlich 500 Schüler pro Staat, herangezogen. In Widerspruch zur weiteren offiziellen Auswertung werden die Daten aus dem Vereinigten Königreich eingeschlossen, die Kurzhefte aber nicht (OECD 2005 a, S. 128 nebst Fußnote 1). Das Ergebnis der Skalierung ist in Anhängen des Technischen Berichts (a. a. O., S. 411 ff.) mitgeteilt.

Zu jeder Aufgabe *j* werden die prozentuale Lösungshäufigkeit ρ_j und die Schwierigkeit ξ_j genannt. Mir scheinen sowohl die Zahlenwerte der ρ_j als auch der Schluss von ρ_j auf ξ_j fehlerhaft.

Bei eigener Mittelung über den internationalen Datensatz kann ich die ρ_j aus dem Technischen Bericht nicht reproduzieren. Wenn ich genau den Vorgaben folge, also das UK einschließe, die Kurzhefte ausschließe und alle 30

[10] Die zahlreich verfügbaren Bücher über *Item Response Theory* behandeln überwiegend triviale Skalierungsprobleme, die aber in unübersichtlicher Ausführlichkeit. Soweit ich sehe, kommen Hambleton und Swaminathan (1985, S. 125 ff.) einer klaren Darstellung eines iterativen Lösungsverfahrens am nächsten.

Staaten gleich gewichte, finde ich für das Aufgabenfeld Mathematik Lösungshäufigkeiten, die durchschnittlich um knapp 1% unter den im Technischen Bericht tabellierten liegen. Bei anderer Gewichtung oder Einschluss der Kurzhefte werden die Abweichungen noch größer.

Die folgende Erklärung ist spekulativ, aber der Größenordnung nach plausibel. Der Technische Bericht deutet an (in der bereits zitierten Fußnote zu S. 128), dass bei der Aufgabenskalierung die stochastischen Gewichte der Probanden nicht berücksichtigt wurden. Wie aber kann eine Fehlgewichtung zu einer *systematischen* Überschätzung der Lösungshäufigkeiten führen? Die Antwort ergibt sich aus § 5: Die Probandengewichte sollen unter anderem unterschiedliche Teilnahmequoten ausgleichen. Wenn Probandengewichte ignoriert werden, dann werden Schüler aus Schulen mit hohen Teilnahmequoten überproportional berücksichtigt. Wenn die Schulteilnahmequote mit der Testleistung korreliert ist, dann erklärt das, warum die Vernachlässigung der Probandengewichte zur Überschätzung der Lösungshäufigkeiten führt.

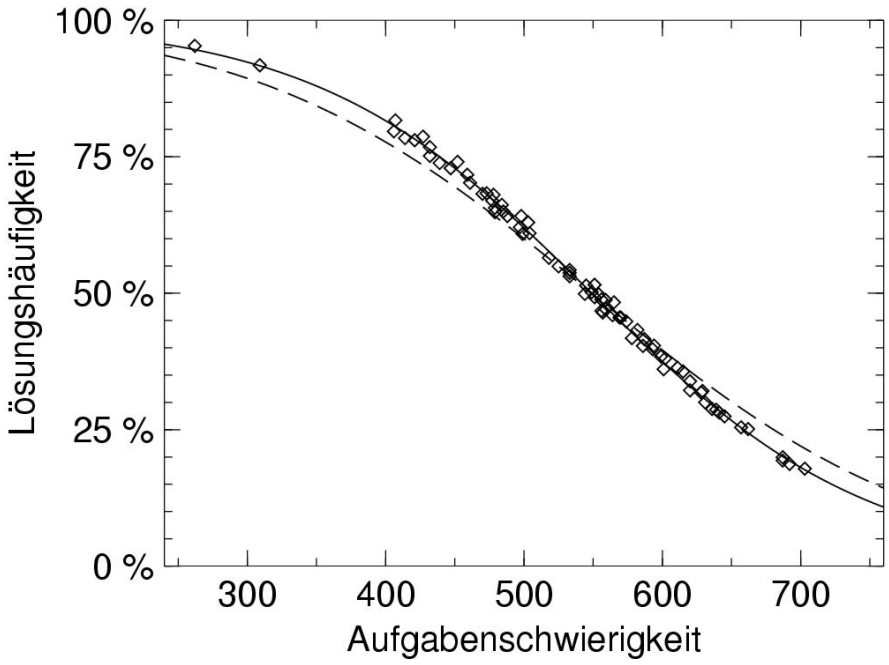

Abbildung 4: Lösungshäufigkeiten und Aufgabenschwierigkeiten der 76 zweiwertigen Mathematikaufgaben gemäß Technischem Bericht (OECD 2005 a, S. 412 f.). Die durchgezogene Kurve zeigt die vermutlich angewandte Gleichung (2), die gestrichelte Kurve den eigentlich zu erwartenden Verlauf (4). In vertikale Richtung ist die Abweichung nicht groß – horizontal aber entspricht sie je nach Bereich einer Streckung der Schwierigkeitsskala um bis 21%.

Um die Umrechnung zwischen Lösungshäufigkeiten und Aufgabenschwierigkeiten zu überprüfen, habe ich in Abbildung 4 die im Technischen Bericht

tabellierten ρ_j und ξ_j für die 76 zweiwertigen Mathematikaufgaben gegeneinander aufgetragen. Bei naher Betrachtung fällt auf, dass der Zusammenhang nicht monoton ist: Beispielsweise hat »Science Test Q1« die Schwierigkeit 556 und wird angeblich von 46,77% der Schüler gelöst; »Exports Q2« wird mit 565 Punkten als schwieriger bewertet, aber von 48,33% gelöst. Im Technischen Bericht finde ich keine Erklärung für dieses erwartungswidrige Verhalten.

Die durchgezogene Kurve zeigt, dass der Zusammenhang im großen und ganzen der Funktion

$$\rho(\xi) = \frac{1}{1+\exp(-(500-\xi_j)/100 - \ln(62/38))} \tag{2}$$

folgt. Das ist Gleichung (1) mit θ = 500 – und an dieser Stelle völlig unerwartet, denn so müssten die Aufgabenschwierigkeiten modelliert werden, wenn alle Schüler einheitlich die Mathematikkompetenz 500 besäßen. Das Bevölkerungsmodell von PISA legt aber fest, dass Kompetenzen θ in der Schulbevölkerung gemäß

$$h(\theta) = \frac{1}{\sqrt{2\pi 100}} \exp(-[(\theta-500)/100]^2/2) \tag{3}$$

normalverteilt sind. Das müsste dazu führen, dass häufig gelöste Aufgaben als weniger schwierig, selten gelöste Aufgaben als weniger leicht eingestuft werden. Der modellgemäße Zusammenhang zwischen Aufgabenschwierigkeit und Lösungshäufigkeit ergibt sich durch Faltung [im Technischen Bericht Formel (9.10)] der Lösungswahrscheinlichkeit (1) mit der Bevölkerungsverteilung (3):

$$\tilde{\rho}(\xi) = \int \frac{1}{1+\exp(-(\theta-\xi)/100 - \ln(62/38))} h(\theta) d\theta. \tag{4}$$

Diese Funktion ist in Abbildung 4 gestrichelt gezeichnet. Sie ist im Bereich der größten Steigung gegenüber der durchgezogenen Kurve (2) um rund 21% gestreckt. Das bedeutet: Sämtliche Aufgabenschwierigkeiten sind in der offiziellen PISA-Auswertung auf einer fehlerhaften Skala bestimmt worden, die nicht zur vorgegebenen 500-100-Normalverteilung der Schülerkompetenzen passt.

Wie konnte es zu einem so gravierenden Fehler kommen, und wie konnte dieser so lange unbemerkt bleiben? Meine derzeitige Arbeitshypothese ist,

dass der Programmierer, auf dessen Werk sich die PISA-Auswerter verlassen haben, vor lauter linearer Algebra den mathematischen Kern der Item-Response-Theorie vielleicht nicht verstanden, jedenfalls aber falsch implementiert hat, und dass der Fehler unbemerkt geblieben ist, weil er durch eine weitere Inkonsitenz in der Datenauswertung verdeckt wird: der übernächste Paragraph wird zeigen, dass die Schülerdaten nachträglich umskaliert wurden.

§ 12 Umrechnung zwischen Prozenten und Punkten

Soweit man nicht die Gesamtleistung von Schülern, sondern die Lösungsstatistik einzelner Aufgaben untersucht, erhält man als quantitative Ergebnisse nicht PISA-Punkte, sondern prozentuale Lösungshäufigkeiten. Für eine kritische Bewertung der Testergebnisse ist es unerlässlich, eine ungefähre Umrechnung zwischen diesen beiden Skalen zu kennen. Diese Umrechnung erhält man als Nebenprodukt der vorstehenden Analyse.

Aus Gleichung (2) oder aus der Steigung der durchgezogenen Kurve in Abb. 4 kann man ablesen, dass an der steilsten Stelle (bei einer Lösungshäufigkeit von 50% und einer Schwierigkeit von 500+100 ln (62/38) $\simeq$ 549) ein Prozent in der Lösungshäufigkeit vier PISA-Schwierigkeitspunkten entspricht. Wenn meine Hypothese, dass die Schülerkompetenzen nachträglich reskaliert wurden, zutrifft, liegt der Kompetenzwertung letztlich nicht die durchgezogene, sondern die gestrichelte Kurve zugrunde. Dann macht ein Prozent in der Lösungshäufigkeit fast fünf Kompetenzpunkte aus.

Der offiziellen Auswertung zufolge können 9 Punkte bereits als »signifikanter« Leistungsunterschied zwischen zwei Staaten gelten. Staaten, die sich um nur 9 Punkte unterscheiden, können im OECD-Ranking um bis zu vier Plätze auseinander liegen. Auch der Unterschied von 10 Punkten zwischen den deutschen Ergebnissen in den Aufgabenfeldern Mathematik und Problemlösen ist ernst genommen und inhaltlich gedeutet worden (Prenzel *et. al.* 2004, S. 15).

Neun Punkte entsprechen, wie oben hergeleitet, einer Differenz der Lösungshäufigkeiten von 2%. Da im Mittel jedem Schüler knapp 26 Mathematikaufgaben gestellt wurden, entsprechen 9 Punkte ziemlich genau einer halben Aufgabe. Auf eine halbe Mathematikaufgabe entfallen 75 Sekunden Testzeit. Damit ist klar, dass die Gesamtergebnisse von PISA empfindlich

von der Validität jeder einzelnen Testaufgabe abhängen. Auch ein von Land zu Land unterschiedlich strenger Umgang mit der Testzeit kann zu erheblichen Verzerrungen führen.

Als Beispiel, wie sich mit Hilfe dieser Umrechnung eine Ungenauigkeit des Tests quantifizieren lässt, sei die handgreiflichste Manifestation des Übersetzungsproblems genannt: die unterschiedliche Textlänge. In PISA 2000 wurde für rund fünfzig Leseaufgaben und zwei Sprachen untersucht, wie sich die Länge der Einleitungstexte auswirkt. Die französische Version ist um durchschnittlich 12% länger als die englische. Die Auswirkung auf die Lösungshäufigkeit ist von der Größenordnung 2% (Adams/Wu 2002, S. 64 ff.), was gut 8 PISA-Punkten entspricht. Es ist kaum zu vermuten, dass der deutsche Sprachraum durch besonders kompakte Texte begünstigt wird.

§ 13 Schülerkompetenzverteilungen

Die Schwierigkeitsskalen in Lesen und Naturwissenschaften wurden 2003 an PISA 2000 angeschlossen; die Skalen in Mathematik und Problemlösen wurden neu normiert. Den Ergebnisberichten zufolge wurden diese Skalen so konstruiert, dass sie im Mittel über alle OECD-Staaten den Mittelwert 500 und die Standardabweichung 100 haben (OECD 2004, S. 45 nebst Fußnote 6; Prenzel *et al.* 2004, S. 5). Diese Aussage ist, wenn auf drei Dezimalstellen wörtlich genommen, nicht mit dem Technischen Bericht vereinbar. Das oben beschriebene, nicht iterative Skalierungsverfahren stellt nicht sicher und lässt auch nicht erwarten, dass die Schülerkompetenzen im Ergebnis exakt der für die Aufgabenkalibrierung vorgegebenen 500-100-Normalverteilung folgen. Aufgrund der fehlenden Selbstkonsistenz ist zu erwarten, dass Mittelwert und Standardabweichung der tatsächlichen Kompetenzverteilung um bis zu ein paar Prozent von der 500-100-Vorgabe abweichen. Aufgrund der bei der Skalierung der Aufgabenschwierigkeit nicht berücksichtigten Probandengewichte ist zu erwarten, dass die Standardabweichung sogar sehr deutlich von 100 abweicht.

Diese Vorhersagen habe ich am internationalen Datensatz empirisch geprüft. Um mit der offiziellen Aufgabenskalierung konsistent zu bleiben, habe ich erneut das Vereinigte Königreich eingeschlossen, Kurzhefte ausgeschlossen und alle 30 Staaten gleich gewichtet. Ich finde für die Schülerkompetenz im Aufgabenfeld Mathematik einen Mittelwert von 501,1 und eine Standard-

abweichung von 99,4. Wie erwartet, weicht die Kompetenzverteilung also von der in den Ergebnisberichten behaupteten 500-100-Normalverteilung ab; die Standardabweichung liegt aber erstaunlich nahe an der Vorgabe. Die Kreise in Abbildung 5 zeigen ein Histogramm der Schülerkompetenz. Es weicht geringfügig, aber systematisch von der vorgegebenen Normalverteilung ab: Kompetenzen zwischen 250 und 350 sowie zwischen 500 und 650 sind häufiger, Kompetenzen oberhalb von 700 seltener als nach der Normalverteilung zu erwarten.

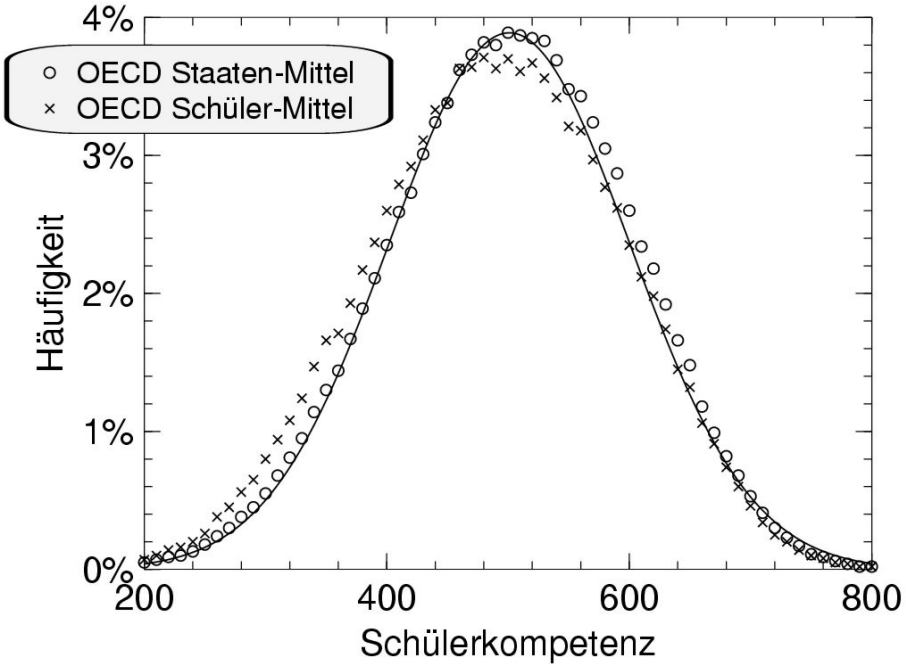

Abbildung 5: Die Kreise zeigen die Häufigkeitsverteilung der Mathematikkompetenz im Mittel über 30 OECD-Staaten (einschließlich UK, ohne Sonderschulen). Besonders zwischen 500 und 600 sind Abweichungen von der Gaußverteilung (durchgezogen) sichtbar. Wenn man die OECD-Staaten nicht alle gleich, sondern entsprechend der Anzahl fünfzehnjähriger Schüler gewichtet (Kreuze), dann verschiebt sich die Verteilung nach links und verbreitert sich.

Am Rande vermerkt sei, dass OECD-Mittelwerte in PISA zumeist Mittelwerte über 30 je gleichgewichtete Staaten bedeuten und somit nicht repräsentativ für die Gesamtheit aller Fünfzehnjährigen (»the stock of human capital« – OECD 2004, S. 33) innerhalb der OECD sind. Der Tendenz nach haben die kleineren Länder die besseren Testergebnisse; drei der vier bevölkerungsreichsten Staaten (Mexiko, Türkei, USA) befinden sich auf der schwächeren Seite. Deshalb liegt das OECD-Bevölkerungsmittel zum Beispiel in Mathematik mit 489,7 Punkten deutlich unter dem Staatenmittel; die Standardabweichung steigt bei dieser Betrachtung auf 103,2 Punkte. Dass das

PISA-Konsortium über die OECD-Staaten und nicht über die Zielpopulation mittelt, erklärt sich aus der Absicht, keine Aussagen über die OECD-Bevölkerung als ganze, sondern über nationale Schulsysteme machen zu wollen.

Soweit meine Analyse auf Grundlage des Technischen Berichts. Nun aber eine überraschende Entdeckung: Wenn ich die Schülerkompetenzen unter Einschluss des Vereinigten Königreichs, *unter Einschluss der Kurzhefte* und unter Gleichgewichtung aller 30 OECD-Staaten mittele, finde ich einen Mittelwert von 499,9998 und eine Standardabweichung von 100,0016. Diese völlig unerwartete, auf fünf bis sechs Stellen genaue Übereinstimmung mit der Vorgabe 500-100 kann keine zufällige Koinzidenz sein, zumal ich sie ganz ähnlich auch im Aufgabenfeld Problemlösen finde.

Ich muss daraus schließen, dass Kapitel 9 des Technischen Berichts die Skalierung nicht nur vernebelt, sondern manifest falsch beschreibt. Entgegen einer unmissverständlichen, abschließenden Aufzählung der Auswertungsschritte (OECD 2005 a, S. 122) hat es eine zusätzliche Transformation von Zwischenergebnissen ergeben: die Schülerkompetenzen sind nachträglich umskaliert worden, so dass ihre Verteilung in Mittelwert und Standardabweichung mit der 500-100-Vorgabe übereinstimmt. Wahrscheinlich wurde zur Umskalierung bloß eine lineare Transformation vorgenommen, denn auch in den umskalierten Daten (den offiziellen Schülerkompetenzwerten unter Einschluss des Vereinigten Königreiches, unter Einschluss der Kurzhefte und unter Gleichgewichtung aller 30 OECD-Staaten) finde ich noch Abweichungen von der Normalverteilung.

Dass die Kurzhefte letztlich doch – wenn auch nur in einer der beiden Skalen – berücksichtigt wurden, ist ein weiterer expliziter Verstoß gegen das, was im Technischen Bericht beschrieben wird (OECD 2005 a, Fußnote 1 zu S. 128). Die inkonsistente Durchführung der offiziellen Auswertung ist insoweit verständlich, als die Kurzhefte die Symmetrie von Tab. 2 durchbrechen und dadurch erhebliche organisatorische und grundsätzliche Probleme aufwerfen. Deshalb halte ich es für möglich, dass die Verantwortlichen mit der nachträglichen Umskalierung der Schülerkompetenzen keinen anderen Zweck verfolgt haben, als die Kurzhefte doch noch in das Nationen-Ranking einzubeziehen, nachdem es ihnen nicht gelungen war, sie schon bei der Bestimmung der Aufgabenschwierigkeiten zu berücksichtigen. Schon das allein hätte zwar dazu geführt, dass die veröffentlichten Schülerkompetenzen nicht mehr genau zu den veröffentlichten Aufgabenschwierigkeiten passen;

weil aber Kurzhefte nur in wenigen Staaten in nennenswertem Umfang eingesetzt waren, wäre das numerische Ausmaß der Verzerrung recht gering geblieben.

Wenn diese Erklärung zutrifft, dann haben die Verantwortlichen möglicherweise bis heute nicht bemerkt, dass sie mit der Umskalierung der Schülerkompetenzen zugleich die in § 11 beschriebenen Fehler unter den Teppich gekehrt haben: Dass die Probandengewichte in der Aufgabenskalierung nicht korrekt berücksichtigt wurden und dass das Faltungsintegral (4) nicht korrekt implementiert wurde, ist nach der Umskalierung nicht mehr an der Verteilung der Schülerkompetenzen zu erkennen, hat dafür aber zur Folge, dass die Skala, die diesen Kompetenzangaben zugrunde liegt, erheblich stärker von der Aufgabenschwierigkeitsskala abweicht, als allein aufgrund der inkonsistenten Berücksichtigung der Kurzhefte der Fall wäre.

§ 14 Lösungsprofile und Aufgabenschwierigkeiten

Nach Festlegung der Aufgabenschwierigkeiten wird die Skalierung des internationalen Datensatzes mit der Bestimmung der Schülerkompetenzen abgeschlossen. Dazu wird nicht das Bevölkerungsmodell (3) benutzt, sondern eine Verfeinerung, die mit Hilfe von »Kollateralinformation« über den einzelnen Schüler die Präzision der Parameterschätzung verbessern soll (Adams *et. al.* 1997 a, S. 50). Fünf »conditioning variables« (Testheft, Geschlecht, Berufe der Eltern, Mathematik-Mittelwert der Schule) werden herangezogen, um den Mittelwert der Normalverteilung von 500 auf einen wahrscheinlicheren Wert zu schieben (OECD 2005 a, S. 121, 129. f.).[11] Mit Hilfe der *conditioning variables* traut sich das Konsortium sogar zu, auch denjenigen Probanden *plausible values* in Lesen, Naturwissenschaften und Problemlösen zuzuschreiben, die in einem oder zwei dieser Aufgabenfelder gar nicht getestet wurden. Diese Rechnungen unabhängig nachzuvollziehen, ist unmöglich; auch hier kann ich nur die Plausibilität und Konsistenz der Ergebnisse prüfen.

[11] Sofern ich richtig verstehe, heißt das: wenn die Tochter der Bürgermeisterin und der Sohn eines Arbeitslosengeld-II-Empfängers eine identische Testleistung erbringen, dann werden der Bürgermeisterintochter tendenziell die besseren *plausible values* zugesprochen. Müssen Folgestudien, die soziale Bedingtheiten von Testleistungen untersuchen, das nicht herausrechnen? Mir ist unklar, wie das möglich sein soll, da die von Staat zu Staat verschiedene Gewichtung der *conditioning values* im Technischen Bericht nicht mitgeteilt wird.

Wäre die offizielle Auswertung in sich stimmig, dann müsste Gleichung (1), zumindest im Mittel über die 30 OECD-Staaten, für jede einzelne Aufgabe den stochastischen Zusammenhang zwischen Schülerkompetenz und Lösungshäufigkeit beschreiben. Um das zu prüfen, ordne ich sämtliche Schüler (mit UK, mit Kurzheften[12]) nach ihrer Mathematik-Kompetenz und bilde dann 25 Leistungsperzentile, die jeweils 4% der Schüler enthalten. Für jede dieser Gruppen berechne ich die mittlere Lösungshäufigkeit einzelner Testaufgaben. Die Auftragung der Lösungshäufigkeiten gegen die Kompetenzmittelwerte der Schülergruppen bezeichne ich im folgenden als *Lösungsprofil* einer Aufgabe.

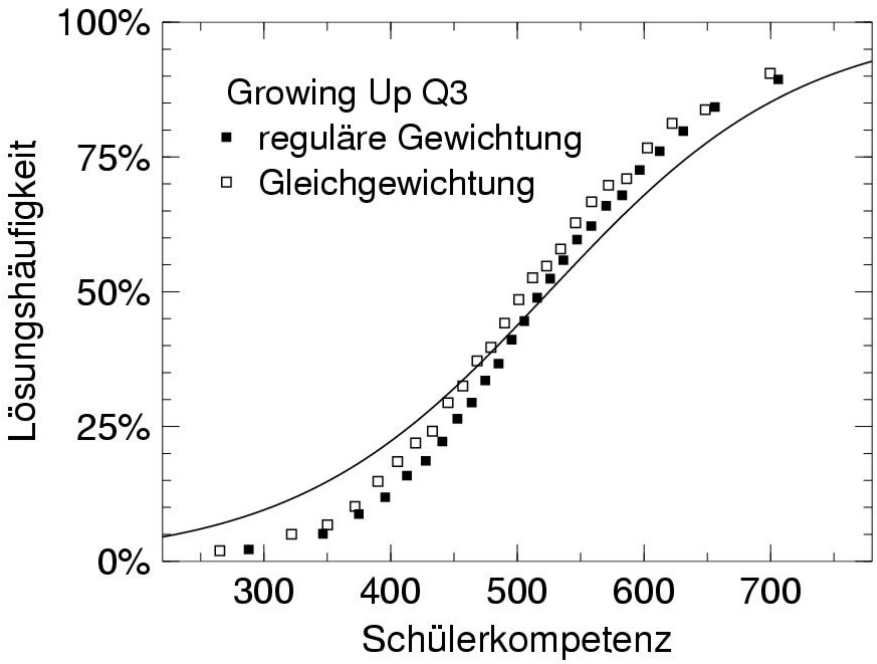

Abbildung 6: Das Lösungsprofil der Mathematikaufgabe »Growing Up Q3«, im Mittel über 30 OECD-Staaten, für zwei verschiedene Gewichtungen der Probanden innerhalb eines Staates. Die durchgezogene Kurve folgt dem Modell (1) mit dem im Technischen Bericht angegebenen Parameter ξ=574.

Im Technischen Bericht wird beispielhaft ein einziges Lösungsprofil mitgeteilt (zur Aufgabe »Growing Up Q3« – OECD 2005 a, S. 127). Ich kann es nicht reproduzieren. Abbildung 6 zeigt das von mir aus dem internationalen Datensatz generierte Lösungsprofil, und zwar in zwei Varianten. Ich habe jeweils über 30 gleich gewichtete OECD-Staaten gemittelt, aber nur im tiefer liegenden Profil das offizielle statistische Gewicht der Probanden berück-

[12] Der Einschluss oder Ausschluss der Kurzhefte betrifft nur 27 der insgesamt 165 Aufgaben und hat im Mittel über alle OECD-Staaten keine nennenswerte Auswirkung auf die folgende Analyse.

sichtigt; im höher liegenden Profil sind alle Probanden gleich gewichtet, wie es ACER bei der Bestimmung der Aufgabenschwierigkeiten praktiziert hat. Die Abweichung beträgt im mittleren Bereich fast 5%, was demonstriert, dass die Vernachlässigung der Schülergewichte bei der Beurteilung einzelner Aufgaben zu ganz erheblichen Verzerrungen führen kann. Wenn die Erklärung aus § 11 zutrifft, dann äußert sich hier eine unerwartet starke Korrelation zwischen Schulteilnahmequote und Schulmittelwert.

Die durchgezogene Kurve zeigt das Modell (1). Für ξ habe ich die offizielle Aufgabenschwierigkeit 574 eingesetzt. Im Technischen Bericht stimmt diese Kurve über einen weiten Bereich recht gut mit dem gezeigten Lösungsprofil überein. In Abbildung 6 ist die Übereinstimmung zwischen den beiden Lösungsprofil-Varianten und dem Modell hingegen sehr mäßig. Das heißt, die von mir aus dem internationalen Datensatz bestimmten Lösungsprofile stimmen nicht mit dem im Technischen Bericht abgebildeten Profil überein.

Die Ursache für diese Abweichung dürfte in dem nicht iterativen Gang der offiziellen Auswertung (OECD 2005 a, S. 122) zu suchen sein: Die zitierte Abbildung (a. a. O., S. 127) gehört zur »national calibration«, einem Aufbereitungsschritt, der der Bestimmung der Aufgabenschwierigkeiten und der Schülerkompetenzen vorangeht. Deshalb vermute ich, dass das im Technischen Bericht gezeigte Lösungsprofil ohne Berücksichtigung von *conditioning variables* generiert wurde. Meinen Profilen liegen dagegen die endgültigen Schülerkompetenzwerte zugrunde, die, wie oben beschrieben, von ACER unter Verwendung der *conditioning variables* festgelegt wurden. Die *conditioning variables* führen demnach absichtsgemäß zu schärferen Lösungsprofilen.

Wie in § 12 beschrieben, beruhen die offiziell mitgeteilten Aufgabenschwierigkeiten auf einem fehlerhaft angewandten Modell, sind im einzelnen nicht nachvollziehbar und hängen nicht einmal monoton von den Lösungshäufigkeiten ab. Zur näheren Überprüfung bestimme ich deshalb Aufgabenschwierigkeiten direkt aus den Schülerdaten. Dazu passe ich für jede Aufgabe *j* das Rasch-Modell

$$P_j(\theta) = \frac{1}{1+\exp(-(\theta-\xi_j)/100 - \ln(62/38))} \qquad (5)$$

durch Optimierung des Parameters ξ_j nach der Methode der kleinsten Quadrate an das aus dem internationalen Datensatz berechnete Lösungsprofil

$\rho_j(\theta)$ an. Für die Aufgabe »Growing Up Q3« finde ich bei Gleichgewichtung aller Probanden in recht guter Übereinstimmung mit der offiziellen Auswertung $\xi_j = 577{,}6$; bei Verwendung der regulären Gewichte den deutlich niedrigeren Wert 563,7. Da es keinen theoretischen Grund für die Gleichgewichtung aller Probanden gibt, werde ich in den folgenden Analysen durchgehend regulär gewichtete Daten verwenden.

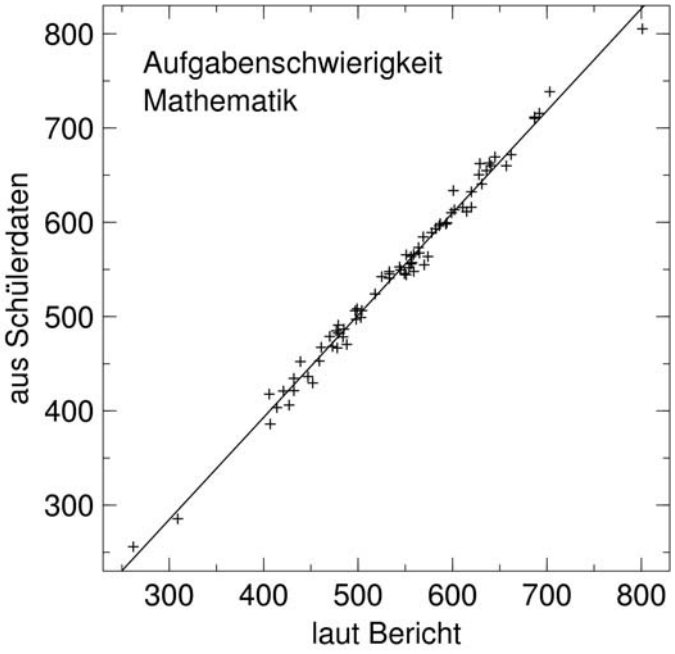

Abbildung 7: Schwierigkeit der 76 zweiwertigen Mathematikaufgaben. Horizontale Skala: tabellierte Punktwerte aus dem Technischen Bericht (OECD 2005 a, S. 412 f.). Vertikale Skala: Anpassung des Rasch-Modells an 4%-Leistungs-Perzentile, 30 gleichgewichtete OECD-Staaten, offizielle Probandengewichte, mit Kurzheften. Die Ausgleichsgerade hat die Steigung 1,08 und schneidet die Diagonale bei 482.

In Abbildung 7 trage ich die Schwierigkeiten ξ_j der 76 zweiwertigen Mathematikaufgaben so, wie sie im Technischen Bericht tabelliert sind, und so, wie ich sie durch eigene Anpassung des Rasch-Modells finde, gegeneinander auf. Die horizontale Skala zeigt die offizielle Bewertung der Aufgaben; die vertikale Skala ist aus der offiziellen Bewertung der Schülerkompetenzen zurückgerechnet. Die Abweichung der Ausgleichsgeraden von der Diagonalen bestätigt auf sehr direktem Wege, dass die Kompetenzskala gegenüber der Aufgabenskala verstimmt ist. Anders als Abbildung 4 vermuten ließ, beträgt die Verzerrung allerdings nur gut 8 %, nicht 16 % bis 21 %. Weitere numerische Inkonsistenzen scheinen die ursprüngliche Verzerrung teilweise zu kompensieren.[13]

[13] Zum kleineren Teil wird die ursprüngliche Verzerrung dadurch kompensiert, dass in der

Der verbindliche Bezugspunkt ist das Bevölkerungsmodell (3), das letztlich allen Ergebnisinterpretationen zugrunde gelegt wird und das, wie oben rekonstruiert, den Schülerkompetenzwerten sehr wahrscheinlich durch nachträgliche Normierung aufgezwungen wurde. Bezogen auf die Schülerkompetenzskala stuft die offizielle Auswertung schwierige Aufgaben zu niedrig und leichte Aufgaben zu hoch ein. Wenn anhand von Aufgaben, die Schüler typischerweise lösen oder nicht lösen, Unterschiede zwischen Subpopulationen beschrieben werden, dann führt die Verzerrung der Skalen zu einer leichten Überschätzung dieser Unterschiede.

Teil III: Was testen die einzelnen Aufgaben?

Die folgenden Bemerkungen stützen sich auf die Auswertung einzelner Aufgaben. Ich zeige, dass das Rasch-Modell keine adäquate Beschreibung der empirischen Daten ermöglicht und dass die vom Konsortium ermittelten Aufgabenschwierigkeiten auf zig Punkte ungenau sind. Bei der weiteren Analyse treten Dimensionen des Testgeschehens in den Vordergrund, die wenig mit Fachkompetenz zu tun haben.

§ 15 Trennschärfe

Das Rasch-Modell in der Gestalt (1) oder (5) enthält einen Skalenfaktor 100, der festlegt, dass eine Änderung der Schülerkompetenz um 100 Punkte erforderlich ist, um die Lösungswahrscheinlichkeit beispielsweise von 27% auf 50% oder von 50% auf 73% oder von 73% auf 88% steigen zu lassen. Er ist der Kehrwert der *Trennschärfe* der Testaufgaben: Je kleiner der Skalenfaktor, desto schärfer fallen die Lösungswahrscheinlichkeiten schwächerer und stärkerer Schüler auseinander.

Es gibt keinen Grund, warum die Trennschärfe ausgerechnet den Wert

---

weiteren Auswertung wieder die regulären Probandengewichte berücksichtigt werden. Eine andere numerische Ungenauigkeit der offiziellen Auswertung deutet sich bei dem im Technischen Bericht gezeigten Lösungsprofil an (OECD 2005 a, S. 127): Dort sind die Schülerergebnisse nicht als Perzentile, sondern in gleichmäßigen Leistungsintervallen histogrammiert. Wenn das bei einer *least squares*-Anpassung nicht berücksichtigt wird, kann die massive Übergewichtung der Extrempunkte zu einer systematischen Verzerrung führen.

1/100 haben sollte. In anderem Kontext ist es zwar möglich, über eine willkürliche Vorgabe der Trennschärfe die Schwierigkeits- und Kompetenzskalen festzulegen. In PISA sind die Skalen aber über das Bevölkerungsmodell (3) definiert, dessen vorgegebene Standardabweichung gerade 100 ist. Das zweimalige Auftreten des Zahlenwerts 100 ist gewiss kein Zufall – aber ich sehe keine theoretische Rechtfertigung dafür. Eine radikale Hypothese, die auch die fehlerhafte Implementierung im Auswerteprogramm erklären würde, wäre, dass die Verantwortlichen das Zusammenspiel von Bevölkerungsmodell und Rasch-Modell in letzter Konsequenz nicht verstanden haben.

Eine andere Hypothese wäre, dass die Testaufgaben so vorsortiert wurden, dass sie tatsächlich empirische Trennschärfen in der Nähe von 1/100 aufweisen. Die in PISA eingesetzten Aufgaben haben nämlich ein mehrstufiges Auswahlverfahren durchlaufen (OECD 2005 a, S. 23). Im Aufgabenfeld Mathematik wurden aus einer unbekannten Anzahl von Vorschlägen 512 soweit entwickelt, das sie den nationalen Projektpartnern zur Begutachtung vorgelegt werden konnten. 217 davon wurden soweit ausgearbeitet, dass sie in einem Feldtest erprobt werden konnten. 68 wurden für tauglich befunden und davon 65 im Haupttest eingesetzt, neben 20 Items, die aus PISA 2000 übernommen wurden.

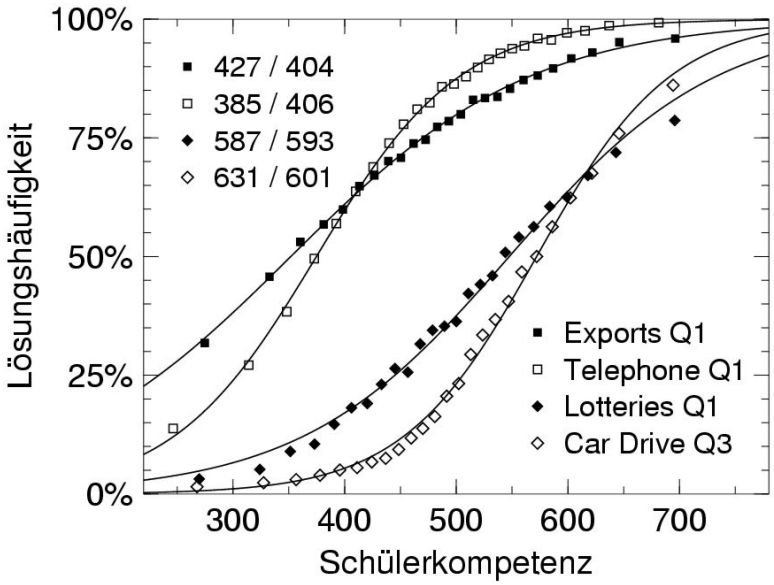

Abbildung 8: Lösungsprofile für zwei Aufgabenpaare mit jeweils ungefähr gleicher Schwierigkeit, aber sehr unterschiedlicher Trennschärfe. In der Legende sind unten rechts die englischen Namen der Aufgaben angegeben, oben links die Schwierigkeiten, und zwar zuerst laut Technischem Bericht und dann so, wie sie sich aus meiner Anpassung ergeben. Die durchgezogenen Kurven folgen dem zweiparametrigen Modell (6).

Spätestens nach dem Feldtest waren die psychometrischen Eigenschaften ein wichtiges Selektionskriterium (a. a. O., S. 27). Abbildung 8 widerlegt jedoch die Hypothese, dass nur Aufgaben mit ungefähr gleicher Trennschärfe eingesetzt wurden: man sieht Lösungsprofile mit recht steilem und recht flachem Anstieg. Für eine adäquate Auswertung solcher Aufgaben ist es erforderlich, das Rasch-Modell um einen zweiten Parameter zu erweitern,

$$P_j(\theta) = \frac{1}{1+\exp(-D_j(\theta-\xi_j)-\ln(62/38))} \tag{6}$$

und für jede einzelne Aufgabe *j* außer der Aufgabenschwierigkeit ξ_j auch die Trennschärfe D_j nach der Methode der kleinsten Quadrate durch Anpassung von P_j an die gemessenen Lösungshäufigkeiten $\rho_j(\theta)$ zu bestimmen.

Die in Abbildung 8 gezeigten Lösungsprofile werden auf diese Weise recht genau modelliert. Für die Trennschärfe finde ich hier Werte zwischen 1/108 (»Exports«) und 1/60 (»Car Drive«). Insgesamt variiert die Trennschärfe bei denjenigen Aufgaben, die halbwegs seriös mit dem Zweiparametermodell auswertbar sind, zwischen 1/196 (»South Rainea Q2«) und 1/44 (»Library System Q2«).

Die Schwierigkeit einer Aufgabe ist in PISA definiert als die Kompetenz eines Schülers, der die Aufgabe mit 62%iger Wahrscheinlichkeit löst. Die Lösungsprofile der beiden Aufgabenpaare in Abbildung 8 schneiden sich jeweils in der Nähe von 62%. Es ist deshalb zu erwarten, dass die Paare »Exports« und »Telephone« sowie »Lotteries« und »Car Drive« jeweils recht ähnliche Schwierigkeitswerte haben. In meiner Auswertung ist das der Fall (404 und 406 bzw. 593 und 601). Die offizielle Auswertung mit dem inadäquaten Rasch-Modell liefert hingegen Aufgabenschwierigkeiten, die sich um mehr als 40 Punkte voneinander unterscheiden.

Sobald man zugesteht, dass Aufgaben unterschiedliche Trennschärfen haben können, wird klar, dass die willkürliche Verankerung der Schwierigkeitsskala bei 62% unmittelbare Konsequenzen für die inhaltliche Interpretation der Testergebnisse hat: eine Verankerung bei einem anderen Prozentwert würde die Schwierigkeitswerte nicht einfach um einen konstanten Betrag verschieben, sondern sich je nach Trennschärfe unterschiedlich auswirken und letztlich zu einer anderen Schwierigkeits-Rangfolge der Aufgaben und in manchen Fällen zu einer anderen Zuordnung zu Kompetenzstufen führen.

§ 16 Teilschritte oder alternative Lösungswege?

Etliche Aufgaben können auch mit zwei Parametern nicht angemessen modelliert werden. Abbildung 9 zeigt einige Beispiele, denen gemeinsam ist, dass das Lösungsprofil zunächst recht steil ansteigt, dann aber flacher verläuft, als nach Gleichung (6) zu erwarten. Um das hervorzuheben, habe ich Gleichung (6) jeweils zweimal an ein Lösungsprofil angepasst: einmal im unteren und einmal im oberen Leistungsbereich. Die ξ-Werte aus den beiden Anpassungen liegen um 25 bis 91 Punkte auseinander, die Trennschärfen unterscheiden sich um Faktoren 1,7 bis 2,2. Die Schwierigkeitswerte der offiziellen Auswertung erweisen sich erneut als hochgradig unzuverlässig; sie liegen in keinem einzigen Fall innerhalb der Spanne, die meine beiden Anpassungen eröffnen.

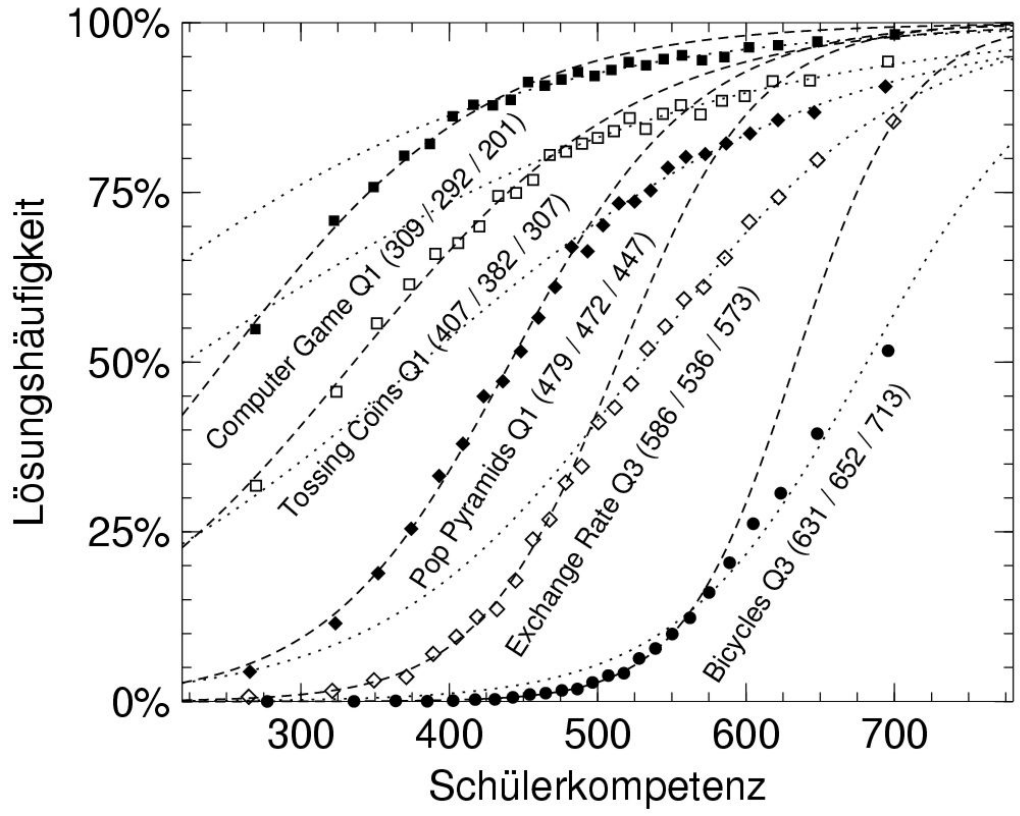

Abbildung 9: Zu fünf Lösungsprofilen wurden je zwei verschiedene Anpassungen des zweiparametrigen Modells (6) vorgenommen: die gestrichelte für die leistungsschwächere, die gepunktete für die stärkere Hälfte der Schüler. In Klammern nach dem Aufgabennamen jeweils die ξ-Werte der offiziellen Auswertung, der gestrichelten und der gepunkteten Kurve.

Betont sei, dass es keinen theoretischen Grund dafür gibt, die Schülerschaft in zwei Hälften zu teilen und jede Hälfte mit einem eigenen Parametersatz zu beschreiben. Die beiden Anpassungen in Abbildung 9 sollen nur verdeutlichen, dass das zweiparametrige Modell (6) jedenfalls nicht geeignet ist, die gezeigten Lösungsprofile in Gänze zu beschreiben. Angemessene Modelle müssten zusätzliche Parameter einführen. Beispielsweise kann man Gleichung (6) multiplikativ zu einem Vier-Parameter-Modell oder additiv zu einem Fünf-Parameter-Modell zusammensetzen. Das multiplikative Modell ist angemessen, wenn die Schüler zur Aufgabenlösung nacheinander zwei Teilschritte unterschiedlicher Schwierigkeit bewältigen müssen; das additive Modell ist angemessen, wenn sich die Schüler stochastisch für unterschiedlich schwierige Lösungswege (Meyerhöfer 2004, Bender 2005) entscheiden. Dummerweise lassen sich *beide* Modelle gleichermaßen perfekt an Lösungsprofile wie die hier gezeigten anpassen, liefern aber völlig unterschiedliche Schwierigkeitswerte.[14] Es bestätigt sich, dass eine Beschreibung der hier gezeigten Aufgaben durch einen einzigen Schwierigkeitsparameter hoffnungslos inadäquat ist.

A priori ist nicht einmal sicher, dass die Lösungshäufigkeit eine monoton steigende Funktion der Schülerkompetenz sein muss. Die zweifelhafte inhaltliche Qualität einiger Aufgaben legt vielmehr die Vermutung nahe, dass ein Schüler, der sich fachlich auskennt und nicht mit der nächstliegenden, oberflächlichen Lösung begnügt, bei manchen Aufgaben benachteiligt sein könnte (Kießwetter 2002, Meyerhöfer 2005). Aufgaben, bei denen dies manifest in einzelnen Ländern der Fall ist, werden zwar als »psychometrisch nicht funktionierend« aus der Auswertung herausgenommen und nicht näher dokumentiert. In PISA 2003 haben die Kontrollen jedoch bei mindestens zwei Aufgaben versagt: Abbildung 10 zeigt Aufgaben, bei denen in mindestens drei Regionen die ansonsten leistungsstärksten fünf, zehn oder fünfzehn

[14] Beispielsweise lässt sich »Tossing Coins Q1« multiplikativ durch eine Abfolge von Teilschritten mit Schwierigkeit 98 (und Trennschärfe 1/270) und Schwierigkeit 331 (1/73) erklären – oder additiv, wenn 56% der Schüler einen Lösungsweg der Schwierigkeit 355 (1/59) und die übrigen 44% einen Weg der Schwierigkeit 458 (1/180) beschreiten. Man kann einwenden, dass schon die unterschiedlichen Trennschärfen durch Teilschritte oder konkurrierende Lösungswege verursacht werden, so dass die Gesamtheit aller Lösungswege ein ganzes Netz von teils in Reihe, teils parallel geschalteten Teilschritten bildet. Die in den Lösungsprofilen enthaltene Information ist zu unspezifisch, um eine nähere Klärung zugunsten eines bestimmten Modells herbeizuführen.

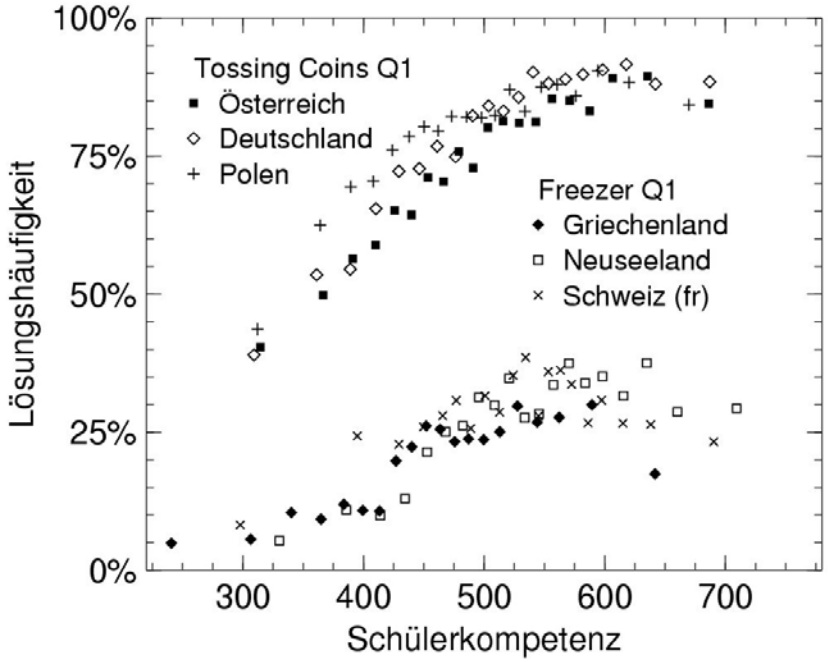

Abbildung 10: Nicht-monotone Lösungsprofile zweier Aufgaben in je drei Ländern/Regionen; ein Symbol entspricht hier jeweils 5% der Schüler.

Prozent der Schüler schlechter als die nächstschwächeren Perzentile abschneiden. »Freezer Q1« wird auch in den übrigen Ländern von nur einem Drittel Schüler im Sinne der Veranstalter gelöst – und das, im leistungsstärksten Drittel der Schülerschaft, nahezu unabhängig von der Schülerkompetenz. Dass diese offensichtlich untaugliche Aufgabe in die Auswertung einbezogen wurde, zeigt die Unzuverlässigkeit der verwendeten Prozeduren.

§ 17 Irgendetwas antworten

Die niederländischen Schüler ragen dadurch heraus, dass sie im Mittel weniger als 3,4% aller Aufgaben unbeantwortet lassen. Mit beträchtlichem Abstand folgen zwischen 6,3% und 8,0% die fünf englischsprachigen Staaten sowie Finnland und Südkorea. In Deutschland, Österreich und der Schweiz liegt der Anteil fehlender Antworten zwischen 10,9% und 11,3%, in Dänemark über 14%, in Italien über 19%.

Noch konturierter wird das Bild, wenn man nur diejenigen Aufgaben betrachtet, die von besonders vielen Schülern übersprungen werden. Diese Aufgaben sind nicht vom Multiple-Choice-Typ, sind fast alle ursprünglich in Englisch oder Niederländisch und überwiegend von den Konsortialinstituten CITO (Niederlande) und ACER (Australien) eingereicht worden, und sie sind fast alle unveröffentlicht.

Die acht auffälligsten Mathematik-Items wurden im Mittel von 10,8% der niederländischen Schüler unbeantwortet gelassen. Es folgen mit 20,9% die USA und bis 29,5% die anderen englischsprachigen Staaten nebst Finnland, Südkorea und Island. Die weitere Spanne reicht über 44,6% in Dänemark bis 55,6% in Italien. Man vergleiche damit das Gesamtergebnis in Mathematik: 514 Punkte für Dänemark, 483 für die USA. Amerikanische Schüler haben demnach trotz schlechter fachlicher Voraussetzungen ausgesprochen geringe Hemmungen, auf schwierige oder abstruse Testaufgaben irgendeine Antwort zu geben – und niederländische Schüler sind den speziellen Stil des CITO wohl schon gewohnt, zumal einige PISA-Aufgaben aus niederländischen Schulbüchern stammen (Meyerhöfer 2006 a).

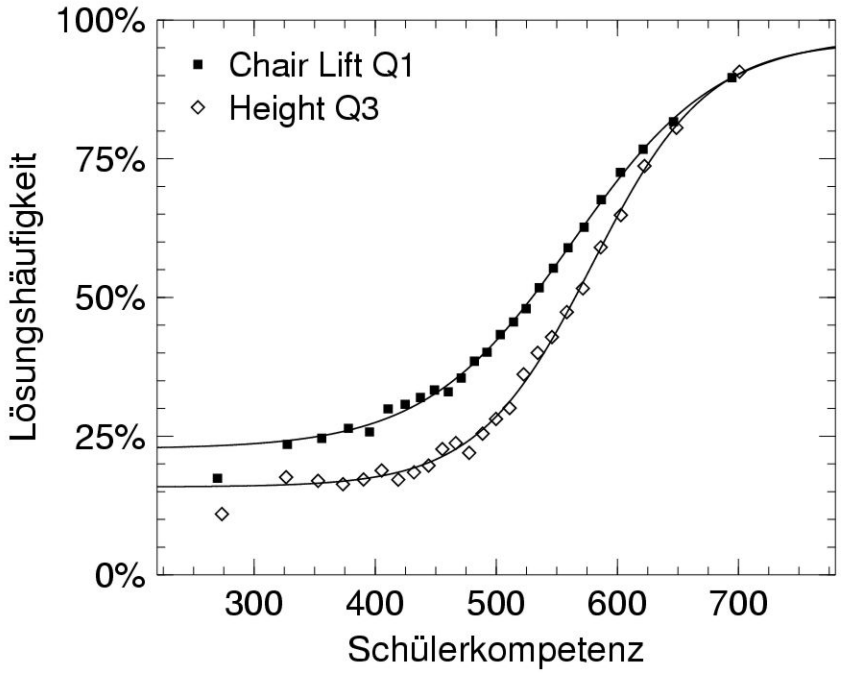

Abbildung 11: Lösungsprofile zweier Multiple-Choice-Aufgaben und Anpassung mit dem Vier-Parameter-Modell (7), das als alternativen Lösungsweg leistungsunabhängiges, qualifiziertes Raten annimmt.

Dass manche Aufgaben sehr einfache, abgekürzte Lösungswege zulassen, kann man auch aus den Lösungsprofilen ablesen. Abbildung 11 zeigt für zwei Multiple-Choice-Aufgaben jeweils ein Plateau der Lösungshäufigkeit im unteren Leistungsbereich. Bei »Chair Lift Q1« ist es besonders ausgeprägt: im Perzentilbereich zwischen 4% und 28% hängt die Lösungshäufigkeit nicht erkennbar von der Schülerkompetenz ab. Die Testleistung der schwächsten 4% aller Schüler liegt jedoch deutlich unter dem Plateau: vielleicht haben diese Schüler den Test nicht bis zu Ende bearbeitet oder aus anderen Gründen auf die Möglichkeit verzichtet, ihr Ergebnis durch Raten aufzubessern.

Ich sehe nicht, wie man aus solchen Lösungsprofilen auf *eine* latente Variable schließen und diese als Fachkompetenz deuten kann. Die ein- oder zweiparametrigen Modelle (5) und (6) sind jedenfalls inadäquat. Eine ungefähre Anpassung ist mit drei Parametern möglich, inhaltlich jedoch kaum interpretierbar. Um der vollen Komplexität des Testgeschehens wenistens nahezukommen, habe ich die beiden Lösungsprofile mit vier Parametern modelliert:

$$P_j(\theta) = c_j r_j + \frac{1 - c_j}{1 + \exp(-D_j(\theta - \xi_j) - \ln(62/38))}. \qquad (7)$$

Ein Bruchteil c_j aller Schüler bearbeitet die Aufgabe durch qualifiziertes Raten und hat dabei mit einer Wahrscheinlichkeit r_j Erfolg. Die übrigen Schüler wählen einen komplexeren Lösungsweg, der wie gehabt durch (6) beschrieben wird. Dass r_j nicht von θ abhängt, ist eine grobe Vereinfachung. Die schwächsten 4% der Schüler lasse ich bei der Anpassung unberücksichtigt. Die Ergebnisse sind in Abbildung 11 als durchgezogene Kurven gezeigt. Ich finde $c_j = 25\%$ bzw. 19% und $r_j = 89\%$ bzw. 80%. Der Parameter $\xi_j = 592$ bzw. 603 weicht im einen Fall um über 40 Punkte von der offiziellen Schwierigkeitsangabe (592) ab, im anderen Fall stimmt er fast mir ihr überein (603).

§ 18 Multiple Choice: Mehrfachantworten

42 der 165 Aufgaben von PISA 2003 sind im einfachen Multiple-Choice-Format mit jeweils vier oder fünf Antwortalternativen. Wenn ein Schüler mehr als eine Antwort markiert, wird seine Lösung als *multiple response* kodiert und in der weiteren Auswertung als falsch gewertet. Die Testteilnehmer aber werden nicht auf diese Spielregel hingewiesen,[15] und auch aus dem

[15] Ein Handbuch (OECD 2003 b) schreibt Wort für Wort vor, wie der Testleiter zu Beginn der Testsitzung mit den Schülern einige Beispielaufgaben durchzugehen hat. Darunter ist eine Aufgabe vom einfachen Multiple-Choice-Typ, bei der sich die Frage, ob mehr als eine Antwort richtig sein kann, nicht stellt: »Wo fanden 1972 die Olympischen Spiele statt? « Es folgt (natürlich in der jeweiligen Landessprache; ich zitiere ohne zu übersetzen, weil es hier auf den genauen Wortlaut ankommt) die Instruktion »*If you are not sure about the answer to a question, circle the answer that you think is best and continue with the next question on the test.*«. Dieser Hinweis impliziert zwar, dass nur eine Lösung richtig ist – aber nur für den, dem die Möglichkeit sich nicht logisch ausschließender Antwortalternativen bereits bewusst ist.

Text der einzelnen Multiple-Choice-Items geht nicht hervor, dass nur eine Antwort richtig sein kann. Die Unkenntnis der Spielregel verzerrt den Test über die Multiple-Choice-Aufgaben hinaus, denn sie bewirkt einen erheblichen Zeitverlust: es ist viel aufwendiger, vier oder fünf Antworten jeweils auf zutreffend oder nicht zutreffend zu prüfen, als eine einzige Alternative auszuwählen (vgl. Meyerhöfer 2004).

Bei der unveröffentlichten Leseaufgabe »Optician Q1« haben 10,5% aller antwortenden österreichischen Schüler mehr als eine Alternative angekreuzt (Frankreich 8,6%, Deutschland 8,1%), während dieser Anteil in Australien, Island, Japan, Kanada, Südkorea, Mexiko, Neuseeland, den Niederlanden und den USA zwischen 0,0% und 0,2% liegt. Bei »Daylight Q1«[16] wurden die meisten Mehrfachantworten in Luxemburg (9,3%), Österreich (8,5%) und Deutschland (7,5%) gegeben. Insgesamt haben in Deutschland, Luxemburg, Österreich bei 11, 12 bzw. 13 Aufgaben mehr als 4% der Schüler eine Mehrfachantwort gegeben.

Singulär ist die unveröffentlichte Problemlöse-Aufgabe »Cinema Outing Q2«, bei der selbst in den Niederlanden, Neuseeland, Australien über 10% und in Katalonien volle 30% mehr als eine Alternative für zutreffend gehalten haben. Dass auch diese offenkundig missratene Aufgabe in die offizielle Auswertung einbezogen wurde, zeigt erneut, wie unempfindlich die Kontrollprozeduren des Konsortiums sind.

§ 19 Weltwissen statt Leseverständnis?

»Optician Q1« ist eine Multiple-Choice-Aufgabe aus dem Aufgabenfeld Leseverständnis. Tabelle 3 zeigt, mit welcher relativen Häufigkeit die Schüler zweier Staaten die vier Antwortalternativen gewählt haben. In beiden Staaten hat knapp die Hälfte der Schüler die als korrekt gewertete Antwort gegeben. Die übrigen Schüler haben im wesentlichen zwei andere Alternativen angekreuzt, und zwar mit beinahe spiegelbildlichen Häufigkeiten: in der Slowakei im Verhältnis 18:33, in Schweden im Verhältnis 37:14. Das heißt, bei formal nahezu identischer Testleistung unterscheiden sich die Präferenzen für die am häufigsten gewählten Distraktoren um rund 20 Prozentpunkte.

[16] Dies ist eine der wenigen veröffentlichten Naturwissenschaftsaufgaben. Die vier Antwortalternativen sind *alle* falsch (Bender 2006, in diesem Band).

	1	2	3	4
Slowakei	3,1%	46,1%	17,5%	33,3%
Schweden	3,1%	46,2%	37,0%	13,7%

Tabelle 3: Relative Häufigkeit der vier Antwortalternativen der Multiple-Choice-Aufgabe »Optician Q1«. Die als korrekt gewertete Alternative »2« wurde in den beiden aufgelisteten Staaten fast gleich häufig gewählt. Die Präferenzen für die Alternativen »3« und »4« unterscheiden sich hingegen um fast 20 Prozentpunkte.

Ähnliche Verwerfungen finden sich, wenn auch nicht ganz so prägnant, bei einer ganzen Reihe anderer Aufgaben. Solange die Aufgaben nicht veröffentlicht sind, ist eine Ursachenforschung kaum möglich. Ich zitiere deshalb aus dem Gedächtnis eine Analyse der Aufgabe »Flu« aus PISA 2000, die ich nicht weiterverfolgt und nicht näher dokumentiert habe. Als Textgrundlage sollte ein Firmenrundschreiben gelesen werden, das für eine Grippeschutzimpfung wirbt. In einer Multiple-Choice-Aufgabe sollten die Schüler dann angeben, wie das Rundschreiben die Schutzimpfung in Beziehung zu körperlicher Ertüchtigung und gesunder Ernährung setzt. Die korrekte Antwort entsprach einer Mittelposition; die Distraktoren gaben die Möglichkeit, der Schutzimpfung zuviel oder zuwenig zuzutrauen. Aus der Häufigkeitsverteilung der falschen Antworten ließ sich klar ablesen, dass französische Schüler die Impfung, deutsche Schüler gesunde Lebensführung für das wirksamere Mittel halten.

Oberflächliche Kenntnis der respektiven nationalen Mentalitäten (nicht der tatsächlichen Lebensführung!) hätte genügt, um dieses Ergebnis vorherzusagen. Aufgrund dieses Beispiels vermute ich, dass ein erheblicher Teil der Schüler manche Multiple-Choice-Aufgaben allein auf Grundlage allgemeinen Weltwissens beantwortet, ohne sich im geringsten auf den vorgelegten Text zu beziehen – was in Anbetracht des enormen Zeitdrucks, unter dem die Testung stattfindet, sogar eine vernünftige Strategie sein dürfte. Dann aber ist zu vermuten, dass auch die Häufigkeiten *richtiger* Antworten nicht allein das Leseverständnis, sondern ebensosehr Teststrategie und Weltwissen wiederspiegeln.

§ 20 Sprachgruppen

Nach dem bis hierhin Gesagten ist klar, dass national unterschiedliche Voraussetzungen, die nichts mit fachlicher Kompetenz zu tun haben, einzelne Aufgaben erschweren oder erleichtern. Um zu prüfen, wie sehr das auf die

Aufgabenbatterie als ganze durchschlägt, habe ich die Testergebnisse der einzelnen Staaten oder Regionen durch 660-komponentige Vektoren dargestellt. Jeder solche Vektor enthält die Lösungshäufigkeiten für alle 165 Aufgaben, getrennt nach den vier unterschiedlichen Positionen im Testverlauf. Anschließend habe ich die Korrelationskoeffizienten dieser Vektoren berechnet.

Innerhalb der OECD finde ich Korrelationen zwischen 0,979 (Australien – Neuseeland) und 0,743 (Japan – Mexiko).[17] Hohe Korrelationen werden vor allem durch eine gemeinsame Testsprache begünstigt; die staatliche Zusammengehörigkeit ist demgegenüber einflusslos. Zum Beispiel zeigt die deutschsprachige Schweiz die stärksten Korrelationen mit Deutschland (0,959), Luxemburg (0,956) und Österreich (0,954). Erst hinter dem niederländischsprachigen Belgien (0,937), Dänemark (0,930), Südtirol (0,926) und neun weiteren Staaten oder Regionen folgt die französischsprachige Schweiz mit einem Korrelationskoeffizienten von nur 0,910.

Abbildung 12 visualisiert die Korrelationen durch Klammern, die jedes Stratum mit demjenigen verbinden, mit dem es am stärksten korreliert ist. Der Betrag der Korrelationskoeffizienten lässt sich an der horizontalen Skala ablesen. Diese Auftragung führt zu Clustern von zwei bis sechs Regionen. Für die meisten Cluster sind die übereinstimmenden oder eng miteinander verwandten Testsprachen konstitutiv.[18]

Im wesentlichen übereinstimmende Cluster hat Rocher (2003) aus den Leseergebnissen von PISA 2000 abgeleitet. In PISA 2003 spielt das Aufgabenfeld Lesen jedoch nur eine untergeordnete Rolle. Wenn man die Leseaufgaben aus der Auswertung herausnimmt, ändert sich Abb. 12 nur minimal; kein einziger Staat wechselt seine Clusterzugehörigkeit. Demnach wirkt sich die Sprache in wesentlich fundamentalerer Weise auf den Schwierigkeitsgrad von

[17] Zum Vergleich: innerhalb der einzelnen Staaten liegt der Korrelationskoeffizient zwischen Jungen und Mädchen zwischen 0,963 (Kanada) und 0,920 (Türkei).

[18] Der Klarheit halber sind die zweisprachigen Regionen Luxemburg, Baskenland und Katalonien ausgespart. Die zwei Fälle, in denen gleichsprachige Regionen keine bevorzugte Korrelation aufweisen (Italien/Schweiz, Mexiko/Spanien), erklären sich vermutlich mit dem großen Leistungsunterschied, der feinere Muster in den Lösungshäufigkeiten überdeckt; überdies wurden die Übersetzungen für Mexiko und Spanien unabhängig voneinander erstellt.

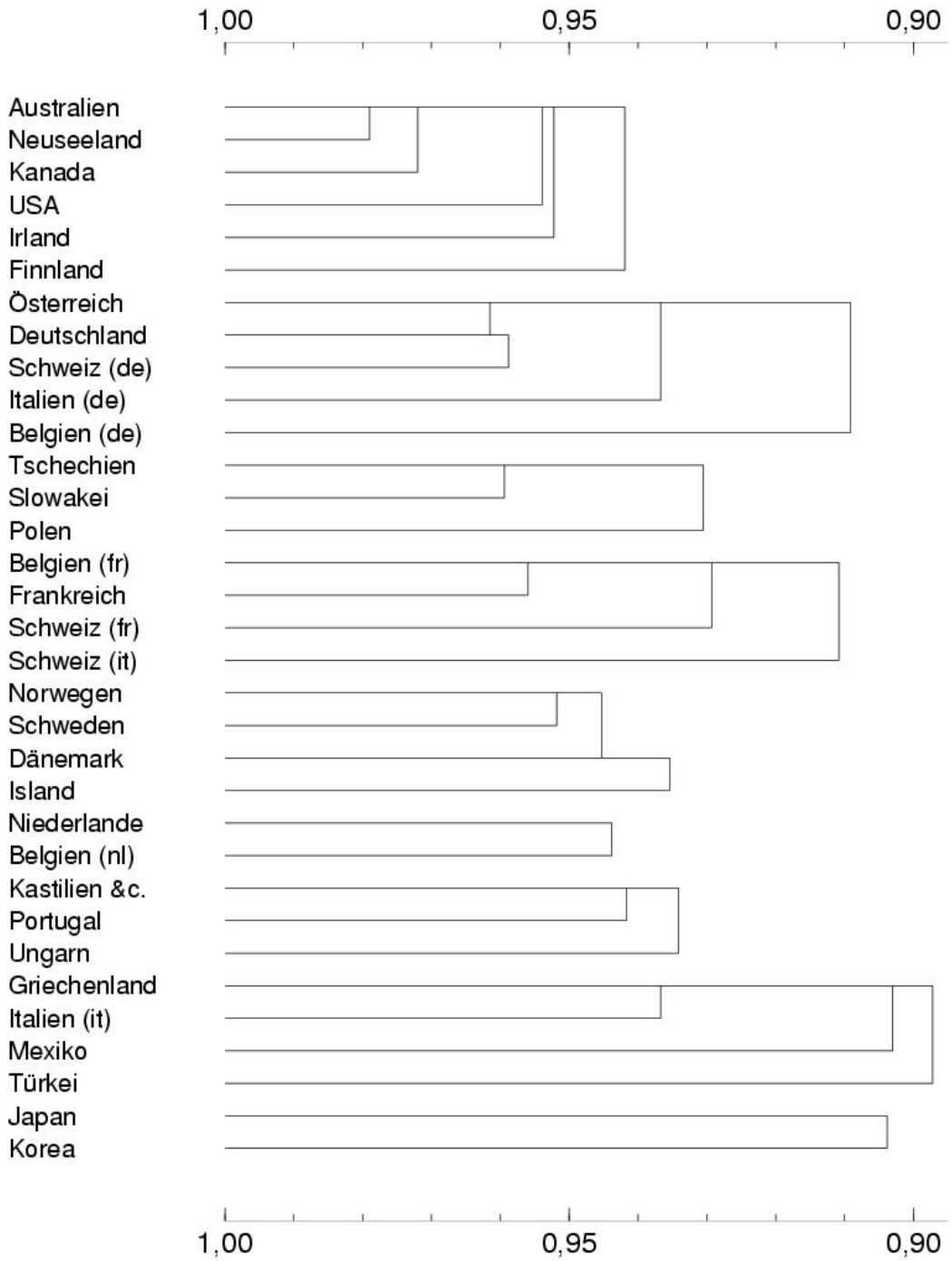

Abbildung 12: Dieses Diagramm veranschaulicht, in welchem Maße die Schüler verschiedener Staaten oder Sprachregionen dieselben Aufgaben mehr oder weniger erfolgreich lösen. Jede Region ist mit derjenigen verbunden, mit der ihr Lösungshäufigkeitsvektor die höchste Korrelation aufweist. Lesebeispiel: Den polnischen Ergebnissen stehen die tschechischen am nächsten (Korrelationskoeffizient ρ =0,931), den tschechischen aber die slowakischen (ρ =0,959).

Testaufgaben aus als allein durch die Beschaffenheit von Lesetexten – und schon deren Länge hat, wie in § 12 zitiert, einen quantitativ bedeutsamen Einfluss auf nationale Mittelwerte.

§ 21 Leistungsabnahme im Testverlauf

Eine elementare Dimension des Testgeschehens, die in der offiziellen Auswertung erwähnt, aber nicht ernsthaft berücksichtigt wird, ist das Nachlassen der Schülerleistungen im Verlauf der zwei Stunden. In der ersten halben Stunde werden OECD-weit 52,7% aller Aufgaben richtig gelöst. In der zweiten halben Stunde beträgt die Lösungshäufigkeit noch 51,6%, in der dritten 48,5%, in der vierten 43,6%. Dieser Leistungsabfall um 9,1% entspricht gut 40 Punkten.

Nach Staaten aufgeschlüsselt streut er von 4,7% in Österreich bis 17,7% in Griechenland. Neben Österreich zeichnen sich Südkorea, die Niederlande, Irland, Deutschland und die USA durch vergleichsweise geringes Nachlassen aus (zwischen 5,9% und 6,2%). Es besteht eine gewisse Korrelation ($\rho = -0{,}746$) zwischen dem relativen Leistungsabfall und der anfänglichen Testleistung, aus der jedoch einige Staaten deutlich herausfallen. So haben Österreich und die Slowakei anfänglich die gleiche Lösungshäufigkeit 51,1%; in der vierten halben Stunde liegt Österreich dagegen bei 46,4%, die Slowakei acht Rangplätze weiter unten bei 40,1%. Auch Irland und Ungarn liegen in der ersten halben Stunde gleichauf, in der vierten um acht Plätze auseinander. Frankreich fällt im Testverlauf gegenüber Österreich sogar um 13 Plätze zurück – im Klassement von nur 29 Staaten.

Das bedeutet, dass das PISA-Länder-Ranking bei einer anderen Testdauer völlig anders ausfallen könnte. Verschiedenste Ursachen für die nationalen Unterschiede sind denkbar: Gewohnheit kürzerer oder längerer Schulstunden; Gewohnheit kürzerer oder längerer Leistungskontrollen; Frustrationstoleranz und Kritikfähigkeit gegenüber dem Test; Übung im Zeitmanagement; Länge und Ausgestaltung der Pause zwischen den beiden Teststunden.

Insoweit die Leistungsabnahme im Testverlauf auf Ermüdung zurückgeht, ist zu vermuten, dass ähnlich stark auch die zeitliche Nähe des Testtags zu Klassenarbeiten, Zeugnisterminen und Ferien die Ergebnisse beeinflussen dürfte. Vielleicht genügt ein unterschiedliches Osterdatum, um »signifikante« Unterschiede zwischen verschiedenen Testjahrgängen vorzutäuschen.

Zusammenfassung und Bewertung
Zur Stichprobenqualität

Zur guten Praxis in den messenden Wissenschaften gehört, dass ein mit öffentlichen Mitteln gefördertes Forschungskonsortium nach Publikation der eigenen Auswertung seine Rohdaten zur Verifikation und als Basis für weiterführende Arbeiten verfügbar macht. Das hat ACER in anerkennenswerter Weise getan. Das Auswertehandbuch (OECD 2005 b) und der Technische Bericht (OECD 2005 a) sind zwar in manchen Details unklar oder unzureichend, aber über einen »helpdesk« konnte ich Projektmitarbeiter erreichen, die e-Mail-Anfragen in der Regel zügig beantworteten[19] – jedenfalls solange meine Fragen nicht auf ein gründliches Verständnis des Auswerteverfahrens zielten.

Der internationale Datensatz ist ein wertvolles Korpus, das anhand der über hundert Fragen des *student questionnaire* die Lebensumstände Fünfzehnjähriger in weiten Teilen der Welt erschließt. Die Beschränkung auf Jugendliche, die noch zur Schule gehen, bringt allerdings mit sich, dass die Daten nicht voll entwickelter Länder wenig repräsentativ für die Gesamtbevölkerung sind (§ 1). Unabhängig vom Entwicklungsstand wird die Repräsentativität durch den uneinheitlichen Ausschluss behinderter Schüler (§ 3), durch die uneinheitliche Einbeziehung von Sonderschulen (§ 4) sowie durch Schwänzen und Testverweigerung (§ 5) beeinträchtigt. Die Ergebnisse aus den USA wurden trotz eines verfehlten Teilnahmequorums in den Datensatz aufgenommen. Die Stichprobenziehung beruht auf problematischen Ausgangsdaten (§ 2) und hat in Südkorea zu einer unglaubwürdigen Ungleichverteilung der Geschlechter und Geburtsmonate geführt (§ 6). Bei der Datenerfassung ist es in Kanada und Polen (§ 7) zu Unregelmäßigkeiten gekommen. Wer PISA als Sozialerhebung auswertet, sollte sich dieser Vorbehalte bewusst sein.

[19] Meine Fragen haben ACER zu einer Korrektur des internationalen Datensatzes (Berezmer 2005) und zur Herausgabe eines Erratums (McKelvie 2006 b) veranlasst, was darauf hindeutet, dass eine unabhängige Analyse der Originaldaten bisher nicht von vielen unternommen wurde.

Zu den Testaufgaben

Für die fachdidaktische Forschung ist es möglicherweise von Interesse, wie Schüler aus verschiedenen Staaten mit gegebenen Testaufgaben umgehen. PISA ist allerdings nicht auf eine solche Auswertung hin angelegt. Von den 217 Mathematikaufgaben, die im Feldtest erprobt wurden, wurden nur 65 in den Haupttest übernommen. Die Ergebnisse aus dem Feldtest sind für ein Staaten-Ranking nicht zu gebrauchen; das Konsortium hat sie deshalb nicht näher dokumentiert. Vielleicht ließe sich für einen qualitativen Unterrichtsvergleich aus dem Nicht-Funktionieren bestimmter Aufgaben in bestimmten Ländern etwas lernen; solch ein Forschungsansatz wird von PISA jedoch nicht unterstützt.

Von den 85 im Haupttest eingesetzten Items sind über 60% nicht interpretierbar, weil noch geheim gehalten. Unter den veröffentlichten Aufgaben sind diejenigen mit offenem Antwortformat kaum interpretierbar, solange nicht auch typische Schülerantworten veröffentlicht werden.[20] Weiterhin sind bei der Interpretation der Lösungshäufigkeiten eine ganze Reihe von Faktoren zu berücksichtigen, die sich nicht unter eine eindimensional bewertete Fachkompetenz subsumieren lassen: Die Möglichkeit verschiedener Lösungswege (§ 15, § 16). Die Vertrautheit mit den Aufgabenformaten, die sich besonders deutlich am weitverbreiteten Missverstehen des Multiple-Choice-Typs zeigt (§ 18). Testroutine, die sich darin äußert, dass Schüler in manchen Staaten fast keine Aufgabe unbearbeitet lassen, also im Zweifel auf gut Glück eine Minimalantwort versuchen (§ 17). Das von Land zu Land sehr unterschiedliche Nachlassen der Leistung im Testverlauf (§ 21). Und schließlich der ganze Komplex der kulturellen Randbedingungen, dessen Tragweite die Korrelationsanalyse von § 20 summarisch andeutet: dazu gehört die unfair verteilte Herkunft der Testaufgaben (Einleitung), die je nach Sprache unterschiedliche Länge der Texte (§ 12), Hintergrundwissen, das entgegen der Intention einer Aufgabe in das Antwortverhalten einfließt (§ 19) und schließlich die hier nicht angesprochene, aber wichtige Frage der Übersetzungsqualität.

[20] Die Luxemburger Projektleitung hat einige eingescannte Schülerantworten veröffentlicht (Blanke *et al.* 2004). Die Tendenz der wenigen Beispiele ist deutlich: aus Antworten, die als falsch gewertet wurden, sollte man nicht auf mathematisches Nicht-Können schließen; manchmal haben Schüler lediglich die Intention der Tester nicht verstanden.

Gleich mehrere dieser Faktoren begünstigen die englischsprachigen Staaten: die Herkunft der Hälfte aller Aufgaben, die Kürze der Texte, die Vertrautheit mit Multiple-Choice-Tests. Dazu kommen in den USA und im UK die niedrigen Teilnahmequoten. Als PISA 2000 das Vereinigte Königreich im oberen Viertel des OECD-Vergleichs einstufte, wies der Abgeordnete Nick Gibb (2002) das Unterhaus darauf hin, dass dieses mit der tatsächlichen Misere des britischen Schulwesens nicht vereinbare Ergebnis »misleading and truly incredible« sei. Ein wenig von diesem Realitätssinn könnte auch der Bildungsforschung nicht schaden.

Zur Punkteberechnung

Ganz PISA ist darauf ausgelegt, dass die Antworten der Schüler unter Berücksichtigung der unterschiedlichen empirischen Aufgabenschwierigkeiten in Kompetenzpunkte umgerechnet werden. Fast die gesamte offizielle Auswertung und fast alle Sekundärarbeiten stützen sich auf diese Kompetenzpunkte, nicht auf die Lösungshäufigkeiten einzelner Aufgaben.

Bei der Skalierung der Schwierigkeiten und Kompetenzen ist es zu einer erstaunlichen Häufung gravierender Fehler gekommen, deren Ineinandergreifen zu schwer durchschaubaren Inkonsistenzen führt. Die Aufklärung ist mir nur gelungen, indem ich eine völlig eigenständig programmierte Auswertung in unzähligen Varianten durchprobiert und immer wieder mit den offiziellen Ergebnissen verglichen habe. Im einzelnen werfe ich den Veranstaltern vor (§ 10 ff.):

(1) Die Daten aus dem UK, die wegen verfehlter Teilnahmequoten letztlich aus der Ergebnisdarstellung entfernt wurden, wurden gleichberechtigt in die Skalierung der Schwierigkeiten und Kompetenzen einbezogen. (2) Die Kurzhefte, die in den Ergebnisdarstellungen einbezogen sind, wurden bei der Skalierung der Aufgabenschwierigkeiten nicht berücksichtigt. (3) Bei der Skalierung der Aufgabenschwierigkeiten wurde das statistische Gewicht der Probanden nicht berücksichtigt. (4) Der Algorithmus, der von den Schülerantworten auf die Aufgabenschwierigkeiten führt, ist nicht dokumentiert und führt zu einem nicht-monotonen Zusammenhang. (5) Vermutlich wurde bei der Festlegung der Aufgabenschwierigkeiten einheitlich eine Schülerkompetenz von 500 zugrunde gelegt. Entgegen der Beschreibung im Technischen Bericht wurde das Bevölkerungsmodell, das eine 500-100-Normalverteilung

vorgibt, in diesem zentralen Auswerteschritt nicht berücksichtigt. (6) Die gesamte Auswertung beruht auf dem Rasch-Modell mit der Trennschärfe 1/100. Es gibt keinen theoretischen Grund, warum deren Kehrwert mit der Standardabweichung des Bevölkerungsmodells übereinstimmen soll. (7) Der Algorithmus, mit dem nach Festlegung der Aufgabenschwierigkeiten von den Schülerantworten auf *plausible values* für die Schülerkompetenzen geschlossen wurde, ist nicht dokumentiert. (8) Die Kompetenzskala wurde nachträglich unter Einbeziehung der Kurzhefte renormiert. Dieser Schritt ist nicht dokumentiert; er ist nicht in den drei Auswerteschritten enthalten, die der Technische Bericht unmissverständlich und abschließend aufzählt. (9) Das Rasch-Modell mit fixer Trennschärfe passt nur für einen Bruchteil aller Testaufgaben zum empirischen Zusammenhang zwischen Schülerkompetenz und Lösungshäufigkeit. Bei vielen Aufgaben genügt es, die Trennschärfe variabel zu wählen; andere Lösungsprofile können aber nur mit drei oder mehr Parametern beschrieben werden, was dann keinen eindeutigen Rückschluss auf eine Aufgabenschwierigkeit mehr zulässt. (10) Die Aufgabenschwierigkeiten der offiziellen Auswertung verstoßen gegen die vorgegebene Verankerung bei 62% Lösungswahrscheinlichkeit.

Vermutlich trifft die Mehrzahl dieser Vorwürfe auch auf PISA 2000 zu. In summa ergibt sich der Eindruck, dass die Verantwortlichen von ACER an der Datenaufbereitung spektakulär gescheitert sind.

Um plausibel zu machen, wie leicht es in einem komplexen Projekt zu einem solchen Versagen kommen kann, möchte ich eine Hypothese vorschlagen, die zwar nichts entschuldigt, aber manches erklärt: Wie oben vermerkt, ist Ray Adams alleiniger Autor des einschlägigen Kapitels 9 im Technischen Bericht. Zugleich aber ist Adams *project director* des internationalen Konsortiums. Er hat über zehn eigene Mitarbeiter, koordiniert die Zuarbeit der anderen Konsortialinstitute, ist Mitglied der PISA Technical Advisory Group und Herausgeber der vielhundertseitigen Technischen Berichte. Es ist denkbar, dass Herr Adams gar keine Zeit für numerische Detailfragen hat. Möglicherweise ist das Auswerteprogramm (das Adams als Wu/Adams/Wilson 1997 zitiert) allein von Wilson implementiert worden. Dass Wilson nur als dritter Autor genannt wird, widerspricht dem nicht: die Geringschätzung der intellektuellen Leistung eines wissenschaftlichen Programmierers könnte vielmehr Teil des Problems sein. An der PISA-Auswertung war Wilson jedenfalls nicht mehr beteiligt. Wenn meine Hypothese zutrifft, hatte

Adams beim Verfassen der Technischen Dokumentation weder Zugriff auf den Programmierer, noch Zeit, den Code im Detail zu analysieren, und musste sich daher am grünen Tisch zusammenreimen, was das Auswerteprogramm eigentlich rechnet. Das würde sowohl die Unklarheiten als auch die manifesten Fehler im Technischen Bericht erklären.

Was bleibt von den Kompetenzstufen?

Die kurze Antwort: Nichts. Die lange Antwort:

Um die Skala der Aufgabenschwierigkeiten über ihre Hilfsfunktion im Staaten-Ranking hinaus »mit Leben zu füllen« (Lind *et. al*. 2005, S. 84), wurde sie willkürlich (OECD 2005 a, S. 268) in sechs »Kompetenzstufen« (*proficiency levels*) und eine darunter liegende Inkompetenzstufe geteilt. Anhand der auf einer bestimmten Stufe typischerweise lösbaren Aufgaben wurde eine verbale Beschreibung der erforderlichen Schülerkompetenz erarbeitet (OECD 2004, S. 45 ff.). Ähnliches war unter Hinzuziehung nationaler Ergänzungsaufgaben bereits in einer deutschen Analyse von PISA 2000 unternommen worden (Klieme/Neubrand/Lüdtke 2001, S. 167–185).

Diese Auswertungen sind Makulatur, da sie auf den Schwierigkeitsangaben der offiziellen Auswertung beruhen. Diese Schwierigkeitsangaben sind, wie oben gezeigt, anhand einer falsch gewichteten Stichprobe in nicht nachvollziehbarer Weise mit modellwidrigem Ergebnis berechnet worden; sie hängen nicht monoton von den Lösungshäufigkeiten ab und sind nicht mit den im internationalen Datensatz veröffentlichten Schülerkompetenzen kompatibel.

Aber auch bei einer kompletten Neuskalierung der Datensätze von PISA 2000 und 2003 wären die Kompetenzstufen nicht zu retten, da das einparametrige Rasch-Modell nur für eine Minderheit aller Testaufgaben in vernünftiger Näherung den Zusammenhang zwischen Schülerkompetenz und Lösungshäufigkeit beschreibt. Die Lösungsprofile anderer Aufgaben erfordern zwei oder mehr Parameter; dann aber lassen sich die Aufgaben nicht mehr ohne Willkür eindimensional anordnen, womit die Grundlage für die Einteilung in Stufen entfällt.

Die Mehrdimensionalität des Testgeschehens, die sich in der Verschiedenheit der Lösungsprofile äußert, ist nicht überraschend. Mehrere Autoren haben darauf hingewiesen, dass eine eindimensionale Interpretation über

wichtige Variablen wie kulturelle Voraussetzungen und Testsprache (Bonnet 2002, S. 394 f.; Goldstein 2004) und unterschiedlich schwierige Lösungswege (Meyerhöfer 2004; Bender 2005) mittelt. Meine Analyse bestätigt, dass diese Variable eine erhebliche Rolle spielen und zeigt weitere Einflüsse, wie namentlich die Position einer Aufgabe im Testheft und die Vertrautheit mit dem Aufgabenformat auf.

PISA findet unter erheblichem Zeitdruck statt. PISA misst nicht Wissen und Können, sondern allenfalls Leistung im Sinne von Arbeit durch Zeit. PISA misst nicht, ob ein Schüler in der Lage ist, eine bestimmte Aufgabe zu lösen, sondern wie er in einer Prüfungssituation mit einer ganzen Aufgabenbatterie umgeht. Wenn schwache Schüler nicht darauf vorbereitet sind, zu schwierige Aufgaben zu überspringen (oder besser noch: durch Raten zu bearbeiten), haben sie nicht mehr genug Zeit (und Selbstvertrauen?), die einfacheren Aufgaben zu lösen. Die unterschiedliche Vertrautheit mit dem Testformat verzerrt die internationale Skalierung deshalb vor allem im unteren Bereich. Wer PISA-Ergebnisse wörtlich nimmt, muss erheblichen Bevölkerungsteilen funktionalen Analphabetismus unterstellen. Deshalb ist eine Kompetenzstufen-Interpretation nicht nur unscharf, sondern führt der Tendenz nach dazu, dass die tatsächlichen Fähigkeiten der Schüler unterschätzt werden.[21] Das Nachlassen im Testverlauf (§ 21) gibt zumindest einen Anhaltspunkt, wieviel mehr die Schüler unter entspannten Rahmenbedingungen leisten könnten.

Was bleibt vom Staaten-Ranking?

Die kurze Antwort: Nicht genug, um einen so aufwändigen Test zu rechtfertigen. Die lange Antwort:

Ohne eine vollständige, in allen Details durchdachte und dokumentierte Neuauswertung kann nicht beurteilt werden, wie sehr die verschiedenen, ineinander greifenden Fehler die Ranglisten verzerrt haben. Eine Neuauswertung sollte allerdings nicht nur neue Ranglisten liefern, sondern zugleich

[21] Dies dürfte die Tendenz vieler Studien sein. Unplausibel schlechte Ergebnisse der Lesestudie IALS haben zur Aufdeckung etlicher methodischer Mängel geführt (Bottani/ Vrignaud 2005, S. 39 ff.). Kießwetter (2002) zieht aus seiner Erfahrung in der Begabtenförderung den Schluss, einen Testerfolg als hinreichendes, aber nicht notwendiges Kriterium für eine besondere Begabung anzusehen.

darlegen, wie verschieden diese ausfallen können, je nachdem, wie man verschiedene Verzerrungen auszugleichen versucht und wie man mit Aufgaben umgeht, die nicht dem Rasch-Modell folgen. Zwar würde man ebensowenig über Schulsysteme lernen wie aus der bisherigen Auswertung, aber immerhin einiges über die Grenzen von Schulleistungsstudien.

Es ist *nicht* zu erwarten, dass eine Neuauswertung die bisherigen Ranglisten auf den Kopf stellt. Im mittleren Bereich würden Korrekturen um ±20 oder auch nur ±10 Punkte zwar erhebliche Veränderungen mit sich bringen, aber dass Finnland vor Deutschland und Deutschland vor Mexiko liegt, wird sich nicht ändern. Nur: um herauszufinden, dass Finnland vor Deutschland liegt, würde ein wesentlich unaufwändigerer Test genügen. In PISA sind das Testdesign mit den dreizehn verschiedenen Heften und die Stichprobengröße von 5000 Schülern pro Staat von Anfang an darauf ausgelegt, sieben verschiedene Kompetenzen (vier Aufgabenfelder, davon eines mit vier Subskalen) mit ungefähr *der* Genauigkeit zu messen, die den Signifikanzkriterien der Ergebnisberichte zugrunde liegt. Wenn dieser Genauigkeitsanspruch in Anbetracht der Nichtbeherrschbarkeit gewisser systematischer Fehler auf ein realistisches Maß zurückgefahren würde, könnte man ohne nennenswerte Einbußen an Aussagekraft den Aufwand erheblich verringern, indem man zum Beispiel auch etwas größere stochastische Fehler in Kauf nimmt und den Stichprobenumfang entprechend reduziert.

Eine noch viel weitergehende Vereinfachung erscheint angebracht, wenn man berücksichtigt, dass sowohl die hohen Korrelationen zwischen den verschiedenen Aufgabenfeldern als auch Analysen der veröffentlichten Aufgabenbeispiele darauf hindeuten, dass es PISA nicht gelingt, sieben wohldefinierte Fach- und Subkompetenzen zu differenzieren. Um aber einen »Generalfaktor kognitiver Fähigkeiten« (Rindermann 2006), kurz gesagt also ein unspezifisches Gemisch aus Intelligenz, Wissen und *Testfähigkeit* (Meyerhöfer 2005) zu messen, könnte man sich zwölf der dreizehn Hefte und Tausende Probanden sparen.

Während sich die Ranglisten bei einer Neuauswertung als in ihren groben Zügen stabil erweisen könnten, halte ich Aussagen über die schwächsten 10% oder 25% der Schülerschaft (»Risikogruppen«, Prenzel *et. al.* 2004, S. 21) für völlig unsicher. Eine ganze Reihe von Fehlerquellen dürfte in diesem Bereich verstärkt durchschlagen: Frühzeitige Schulabgänge, uneinheitliche Berücksichtigung von Sonderschulen, Testverweigerung und -abbruch, Frus-

tration nach ein paar unzugänglichen Aufgaben, Leseprobleme auch in Nicht-Lese-Aufgaben und vieles andere mehr.

Nichts Neues unter der Sonne

Der spezielle Stil der Mathematik-Aufgaben in PISA beruht maßgeblich auf dem Konzept der »realistic mathematics education« aus dem Freudenthal-Institut in Utrecht (OECD 1998). Die zumindest indirekte Berufung auf Hans Freudenthal ist nicht ohne Ironie. Freudenthal war es in den 1970er Jahren gelungen, »new math« aus den Niederlanden herauszuhalten – die er zuallererst mit fachlichen Argumenten als schlechte Mathematik entlarvte. Aus dem gleichen Impetus heraus unterzog er zwei der ersten großen internationalen Schulleistungsstudien der IEA[22] einer vernichtenden Kritik (Freudenthal 1975) – in deren statistischem Bombast er ebenfalls zuallererst »schlechte Mathematik« sah.

»If time permits, errors can be traced in every big statistical survey. The time needed can be a measure of trustworthiness. In the IEA studies it is too easy; interior and exterior inconsistencies are blatant.« (a. a. O., S. 176). Seitdem hat sich erschreckend wenig geändert. Viele methodische Mängel von PISA liegen zum Greifen nah. Ursprünglich hatte ich mich nur für die kulturelle Prägung von Antwortmustern interessiert. Bei der dazu nötigen Aufbereitung des internationalen Datensatzes stolperte ich dann von einer Überraschung zur nächsten.

Auf die naheliegende Frage, wie es möglich ist, dass solche Mängel jahrelang übersehen werden, gab Freudenthal zwei Antworten:

> »blind faith in mathematics« (S. 131).
>
> Und:
>
> »What happens in educational research looks as though in natural science it would have become a habit that – because of the importance of mathematics as a tool – all research is done by mathematicians, who for experiments, if need be, would hire some analysts, laboratory assistants, and stablemen. Fortunately science is not run this way. Otherwise instead of science we would have orgies of bad mathematics« (S. 178).

[22] Damals noch »Council of the International Project for the Evaluation of Educational Achievement«, später umbenannt in »International Association for the Evaluation of Educational Achievement«. Bevor sich mit PISA die OECD eingeschaltet hat, hat die IEA alle großen Studien organisiert, zum Beispiel auch TIMSS und PIRLS.

Genau diese ins Absurde gesteigerte Arbeitsteilung, die Jahnke (2006) als organisierte Verantwortungslosigkeit beschreibt, könnte auch Ursache dafür sein, dass ein Auswerteprogramm anderes rechnet, als der Chef der Auswertung zu Papier bringt.

Seine Schlussfolgerungen hat Freudenthal als Frage-Antwort-Spiel formuliert (S. 175, 178):

> »Question: What effects will this failure produce?
>
> Answer: Powerful groups such as the IEA usually gain by failure, which is turned into an argument for further existence. From the beginning the IEA studies were described as first attempts. They will justify the need for second attempts«
>
> »Question: Could the objectives have been attained if the studies had been designed differently?
>
> Answer: Certainly not on so broad a front as was attempted [damals waren 12 Staaten beteiligt]. A much more modest design would have been more promising.«
>
> »Question: Would a more modest undertaking have been better organised?
>
> Answer: Financial support would not have been available for a more modest undertaking. People trust global dimensions and huge figures. They are part of the official IEA advertising and have certainly influenced financial supporters. Modesty engenders distrust, but a more modest undertaking would certainly have worked better.«

Danksagung

Ich danke Alla Berezmer, Eveline Gebhardt, Pippa McKelvie, Sheila Krawchuk, Martin Murphy (Australian Council for Educational Research), Detlef Lind (Univ. Wuppertal) und einer Mitarbeiterin eines nationalen Projektzentrums, die nicht genannnt werden möchte, für Hilfe bei der Exegese der Technischen Berichte. Für Anmerkungen zu Vorformen dieses Manuskripts danke ich Katharina Inhetveen (Univ. Siegen) und Wolfram Meyerhöfer (Univ. Potsdam).

Literatur

ACER [Australian Council of Educational Research] (2004): PISA 2000 International Database. Online-Resource http://pisaweb.acer.edu.au/oecd/oecd_pisa_data_s1.html, Datenstand vom 18. Feb. 2004.

ACER (2005): PISA 2003 International Database. Online-Resource http://pisaweb.acer.edu.au/oecd_2003/oecd_pisa_data_s1.html, Datenstand vom 9. Nov. 2005.

Adams, R. J.; Wilson, M.; Wu, M. (1997 a): Multilevel Item Response Models: An Approach to Errors in Variables Regression. J. Educational and Behavioral Statistics 22 (1) S. 47–76.

Adams, R. J.; Wilson, M.; Wang, W.-C. (1997 b): The Multinomial Random Coefficients Multinomial Logit Model. Applied Psychological Measurement 21 (1) S. 1–23.

Adams, R.; Wu, M. (Hrsg.) (2002): PISA 2000 Technical Report. Paris: OECD.

Adams, R. J. (2003): Response to »Cautions on OECD's Recent Educational Survey (PISA)«. Oxford Review of Education 29 (3) S. 377–389.

Artelt, C.; Baumert, J. (2004): Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs. Z. Päd. Psych. 18 (3/4) 171–185.

Bender, P. (2005): PISA, Kompetenzstufen und Mathematik-Didaktik. J. Math.-did. 26 (3/4) S. 274–281.

Bender, P. (2006): Was sagen uns Pisa & Co., wenn wir uns auf sie einlassen? [dieses Buch].

Berezmer, A. (2005): persönliche Mitteilung, e-Mail vom 24. Jan.

Blanke, I.; Böhm, B.; Lanners, M. (2004): Beispielaufgaben und Schülerantworten. Le Gouvernement du Grand-Duché de Luxembourg. Ministère de l'Éducation nationale et de la Formation professionelle.

Bonnet, G. (2004): Evaluation of education in the European Union: policy and methodology. Assessment in Education 11 (2) S. 179–191.

Bottani, N. ; Vrignaud, P. (2005): La France et les évaluations internationales. Rapport établi à la demande du Haut Conseil de l'évaluation de l'école. Online-Resource http://lesrapports.la documentation francaise.fr/BRP/054000359/0000.pdf [26. Jan. 2006].

CEPED [Centre français sur la population et le développement] (2006): Le déficit des femmes en Asie: Tendances et perspectives. Chronique no. 51.

von Collani, E. (2001): OECD PISA - An Example of Stochastic Illiteracy? Economic Quality Control 16 (2) S. 227–253.

Freudenthal, H. (1975): Pupils achievements internationally compared – the IEA. Educational Studies in Mathematics 6, S. 127–186.

Gibb, N. (2002): Rede im Unterhaus. Commons Hansard Debates, Friday 15 Nov 2002.

http://www.publications.parliament.uk/pa/cm200203/cmhansrd/vo021115/debtext/21115-17.htm.

Goldstein, H. (2004) International comparisons of student attainment: somme issues arising from the PISA study. Assessment in Education 11 (3) S. 319–330.

Hambleton, R. K. / Swaminathan, H. (1985): Item response theory. Boston: Kluwer, Nijhoff.

Jahnke, T. (2006): [Dieses Buch]

McKelvie, P. (2006 a,b): persönliche Mitteilungen, e-Mails vom 23. Jan. und 9. Feb.

Kießwetter, K. (2002): Unzulänglich vermessen und vermessen unzulänglich: PISA u. Co. Mitteilungen der Deutschen Mathematiker-Vereinigung, H. 4, S. 49–58.

Kim, D.-S. (2004): Le déficit de filles en Corée du Sud: évolution, niveaux et variations régionales. Population [Paris] 59, S. 982–997.

Klieme, M.; Neubrand, J.; Lüdtke, O. (2001): Mathematische Grundbildung: Testkonzeption und Ergebnisse. In: Deutsches PISA-Konsortium (Hrsg.): PISA 2000. Opladen: Leske & Budrich.

Lind, D.; Knoche, N.; Blum, W.; Neubrand, M. (2005): Kompetenzstufen in PISA. — eine Erwiderung auf den Beitrag von W. Meyerhöfer [...] J. Math.-did. 25 (1) S. 80–87.

Meyerhöfer, W. (2004): Zum Kompetenzstufenmodell von PISA. J. Math.-did. 25 (3/4) S. 294–305.

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

Meyerhöfer, W. (2006 a): persönliche Mitteilung vom 27. Feb.

Meyerhöfer, W. (2006 b): PISA & Co. als kulturindustrielle Phänomene [dieses Buch]

NCHS [U. S. Department of Health and Human Services, Centers for Desease Control and Prevention, National Center for Health Statistics] (2005): More Boys Born Than Girls. New Report Documents Total Gender Ratios At Birth From 1940 to 2002. Online-Resource http://www.cdc.gov/nchs/pressroom/05facts/moreboys.htm, [2. Jan. 2006].

OECD [Organisation for Economic Co-operation and Development] (1998): Fourth Meeting of the Board of Participating Countries (Paris, 6–7 July).

OECD (2001): Knowledge and Skills for Life. First Results from the OECD Programme for International Student Assessment (PISA) 2000. Paris: OECD.

OECD (2003a): The PISA 2003 Assessment Framework. Mathematics, Reading, Science and Problem Solving Knowledge and Skills. Paris: OECD.

OECD (2003b): Test Administrator's Manual – PISA 2003. Paris: OECD.

OECD (2004): Learning for Tomorrow's World. First Results from PISA 2003. Paris: OECD.

OECD (2005 a): PISA 2003 Technical Report. Paris: OECD.

OECD (2005 b): PISA 2003 Data Analysis Manual. SPSS Users. Paris: OECD.

OECD (2005 c): Longer Term Strategy of the Development of PISA. 20th meeting of the PISA Governing Board. 3–5 October, Reykjavik, Iceland. Paris: OECD.

Oster, E. (2005): Hepatitis B and the Case of the Missing Women. Journal of Political Economy 113 (6) 1163–1216.

Prais, S. J. (2003): Cautions on OECD's Recent Educational Survey (PISA). Oxford Review of Education 29 (2) 139–163.

Prenzel, M.; Baumert, J.; Lehmann, R.; Leutner, D.; Neubrand, M.; Pekrun R.; Rolff, H.-G.; Rost, J.; Schiefele, U. [PISA–Konsortium Deutschland] (Hrsg.) (2004): PISA 2003. Ergebnisse des zweiten internationalen Vergleichs. Zusammenfassung. Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften.

Rindermann, H. (2006): Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? Psych. Rundschau 57 (2) 69-86.

Rocher, T. (2003): La méthodologie des évaluations internationales de compétences. Psychologie et Psychométrie 24 [Numéro spécial : Mesure et Éducation], 117–146.

Song, S.-Y. (1998): The Problem of Sex Ratio in Asia. S. 188–190 in Fujiki, N., Macer, D. R. J. (Hrsg.): Bioethics in Asia. Eubios Ethics Institute.

U. S. Census Bureau (2006): International Data Base. Population Pyramids. Online-Resource http://www.census.gov/ipc/www/idbpyr.html [2. Jan. 2006].