Klippert, Heinz (2006): Methodentraining. Übungsbausteine für den Unterricht. Berlin.

Meyer, Hilbert (2007): Merkmale guten Unterrichts. In: Guter Unterricht. Friedrich Jahresheft XXV. Seelze, Velber, S. 64-65.

Müller, Walter (2008): Schnee von gestern. Was ist das Neue an der »Neuen Unterrichtskultur«? In: Vierteljahrsschrift für wissenschaftliche Pädagogik 3/2008. Paderborn, S. 323-335.

Odendahl, Johannes (2008): Karussell und Kugellager. In: Vierteljahrsschrift für wissenschaftliche Pädagogik 3/2008. Paderborn, S. 352-370.

Petzelt, Alfred (1964): Grundzüge systematischer Pädagogik. Freiburg.

Ruhloff, Jörg (1998): Lernen des Lernens? In: Rekus, Jürgen (Hrsg.): Grundfragen des Unterrichts. Weinheim, S. 83-94.

Schirlbauer, Alfred (2008): 37 Elefanten. Oder: Kann man ohne Lerntheorie unterrichten? In: Vierteljahrsschrift für wissenschaftliche Pädagogik 4/2008. Paderborn, S. 436-447.

Gegenwind für PISA

Ein systematisierender Überblick über kritische Schriften zur internationalen Vergleichsmessung

VOLKER BANK / BJÖRN HEIDECKE

I. Einleitung

Die OECD untersucht mit den seit 1997 laufenden PISA-Studien in drei Zyklen (2000, 2003 und 2006) unter jeweils anderer Schwerpunktsetzung die Kompetenzbereiche Lesekompetenz, mathematische Grundbildung und naturwissenschaftliche Grundbildung in den teilnehmenden Ländern. Mit der Lancierung der Studien und der Platzierung ihrer Ergebnisse in der breiten Öffentlichkeit gelang den am Konsortium beteiligten Wissenschaftlern ein nicht nur in der Pädagogik beispielloser Erfolg in Deutschland (vgl. Bayrhuber et al. 2004, Meding/Roe 2006, Ammermueller 2008 und Fertig 2004). Dieser betrifft gleichermaßen die Steigerung der Drittmittelumsätze sowie der Öffentlichkeitswirksamkeit. Das wurde bereits frühzeitig etwa von Karg (2005, S. 237) angesprochen, wenn sie vom ›Mythos PISA‹ spricht oder auch von Radtke (2005, S. 359), wenn er das von ihm so bezeichnete ›PISA-Event‹ mit dem Sputnik-Schock vergleicht. Sjøberg (2004) fand für einen Zeitraum von nur zwei Monaten rund 200 Artikel in deutschen Zeitungen, die auf PISA verweisen und zugleich die deutsche Bildungslandschaft kritisieren. Wiewohl die Studien nicht in allen beteiligten Staaten ähnlich weite Kreise wie in Deutschland gezogen haben, ließen sich auch in anderen Ländern ähnliche Resultate finden (vgl. Sjøberg 2004, S. 58; Dolin 2007, S. 93 sowie Bozkurt/Brinek/Retzl 2007, S. 326 beide erschienen in Hopmann/Brinek/Retzl 2007).

Es kann nicht ausbleiben, dass sich bei der großen Aufmerksamkeit, die das Schulvergleichsprogramm der OECD erregt hat, über kurz oder lang kritische Stimmen zu Wort melden. Diese wurden zumeist vereinzelt geäußert, mittlerweile gibt es jedoch auch zwei Sammelbände, in denen eine Reihe kritischer Überlegungen zusammengetragen worden ist und die Gegenstand dieser Besprechung sein sollen. Diese sind der Sammelband von Hopmann, Brinek und Retzl (2007) und jener von Jahnke und Meyerhöfer (2007). Letzterer hatte sich unter Hervorhebung messtheoretischer Aspekte schon zwei Jahre zuvor mit einer PISA-kritischen Veröffentlichung zu Wort gemeldet. Ferner liegt schon etwas länger eine Dissertation von Gaeth (2005) vor, die insbesondere auf methodologische Aspekte von PISA eingeht und die hier fallweise in die Besprechung mit einbezogen werden soll.

Anders als bei Rezensionen häufig üblich, sollen hier nicht die einzelnen Kapitel der Schriften durchgegangen werden, sondern es soll im Interesse einer größeren inhaltlichen Transparenz eine systematische Behandlung der Kritikpunkte versucht werden. Zur leichteren Rückverfolgbarkeit werden deshalb die Belege der Fundstellen mit den Bezeichnungen HBR (Hopmann/Brinek/Retzl 2007) und JM (Jahnke/Meyerhöfer 2007) ergänzt. Soweit es sich anbietet, werden auch vereinzelt Artikel angesprochen, die nicht in den gewählten Sammelbänden veröffentlicht sind, aber ebenfalls einzelne kritische Hinweise zu den PISA-Studien geben. Die Darstellung verfolgt das Ziel, die in den Sammelbänden vorgetragenen Kritikpunkte zunächst zu systematisieren, um dann erst im Anschluss die Bände zu kommentieren und Leseempfehlungen zu geben.

II. Aspekte der Kritik

Die Kritik an PISA behandelt eine ganze Bandbreite an Aspekten; diese sind vorrangig methodologischer Natur, beziehen sich aber auch auf die politischen, ökonomischen und kulturellen Vorbedingungen und Schlussfolgerungen der Studie, mithin auf die Testökologie und wissenschaftstheoretische Fundierung.

II.1 Kritik der Testökologie und wissenschaftstheoretischen Fundierung

Wuttke (HBR 2007, S. 261), wie schon zuvor Radtke (2005, S. 357), kritisiert die Vorgehensweise, anhand von nur wenigen Indikatoren – i.W. der Leseleistung, der Rechenleistung und der Naturkundekenntnisse von 15-Jährigen – auf die Leistung des gesamten Schulsystems schließen zu wollen. Da man kaum von einem parametrischen Messmodell in diesem Zusammenhang sprechen kann, das auf die Messung latenter Variablen verwiesen wäre, wird hier offenbar ein problematisches induktives Schlussverfahren vom Speziellen auf das Allgemeine angewandt. In diesem Kontext konstatiert Radtke (2005), dass sich die Betrachtungsperspektive von der pädagogischen Betrachtung des einzelnen Schülers fort und zur Organisation des Systems als Ganzen hin verschiebt. Auch Hörmann (HBR 2007, S. 71) bemängelt den Umgang mit benachteiligten Schülern. Durch den Ausschluss von Schülern aus solchen Tests würden sie möglicherweise weiter stigmatisiert und in einer benachteiligten Position fixiert.

Eine ähnliche Stoßrichtung wählt auch Sjøberg (2004, S. 53ff), wenn er bemängelt, dass nur wenige Aspekte bei wenigen Subjekten als Maßstab für die Gesamtheit gesehen werden. Andere Subjekte werden dadurch ignoriert. Demgegenüber hält Wuttke (HBR 2007, S. 261) auf einer ganz anderen Argumentationsebene die großen Stichprobenumfänge zur Reduzierung von statistischen Fehlern angesichts der sonst dem Test innewohnenden Schwächen für unökonomisch. Hier lassen sich durchaus auch unterschiedliche Positionen bei den Kritikern entdecken.

Ökonomische Fragen nach der Finanzierung der PISA-Erhebungen werden bei Langfeldt (HBR 2007, S. 238) gestellt. Nach seinem Dafürhalten kann durch die Beteiligung von großen Unternehmen wie etwa der Citogroup oder Weststat, die eigene Ziele verfolgen, eine Beeinflussung der Ergebnisse nicht ausgeschlossen werden. Man wird auch davon ausgehen müssen, dass die Schlussfolgerungen für die Umgestaltung der Erziehungssysteme in einzelnen Ländern ebenfalls interessengeleitet sein können. Es wird etwa von Keitel (JM 2007, S. 47) argumentiert, dass bei der Übernahme von Methoden und Techniken aus Erziehungssystemen, die nach PISA erfolgreich erscheinen, die in der nationalen Diskussion angebrachten Kritikpunkte vernachlässigt werden sowie die kulturellen und sozialen Bedingungen im vermeintlich ›schlechten‹ System außer Acht bleiben. Der Aspekt der kulturellen Einbettung schulischer Systeme ist in grundsätzlicher Hinsicht nach Sjøberg (2004, S. 53) bedenklich. Er weist darauf hin, dass in erster Linie nur reiche Länder an der Studie teilgenommen haben und es sich anmaßten, gleich Maßstäbe für die ganze Welt zu setzen. Er sieht PISA als: »[e]ducational bulldozer in the service of globalization and common norms« (Sjøberg 2004, S. 53). Die Vielfalt der Werte und Kulturen wird somit nicht nur vernachlässigt, sondern letztlich auch ausgedünnt. Diese Kritik führt bereits zu methodologischen Überlegungen, denn beispielsweise ist das Konzept der ›Real Life Challenges‹ auch insofern problematisch, als lebensweltliche Herausforderungen eben mit der Kultur variieren und sich keineswegs als gleichförmiges Messkonzept für alle eignen.

Nach diesen grundsätzlichen Kritikpunkten sollen im Folgenden das Vorgehen bei der Messung, die Konzeptionierung der Aufgaben und schließlich die motivationalen Aspekte der Probanden behandelt werden.

II.2 Messdesign

Da das zugrundegelegte Messmodell und das Prozedere von zentraler Bedeutung für die Qualität einer Messung sind, kann und muss eine Kritik zunächst hierauf Bezug nehmen. Die besondere Rolle der Aufgaben bzw. Testitems für die Qualität der Abbildungsrelation wird in einem eigenen Abschnitt gewürdigt. Die in diesem Abschnitt zu besprechenden Kritikpunkte am Messdesign betreffen also die Eignung der gewählten Instrumente und Daten für die Messung der untersuchten Größen.

Wie bei jedem Messvorgang muss es auch bei PISA das Ziel sein, ein empirisches Relativ unter Einhaltung des Eindeutigkeitssatzes und des Repräsentationssatzes in ein numerisches Relativ zu überführen. Da das empirische Relativ – wie etwa die Problemlösefähigkeit der Schüler – aber nur latent beobachtbar vorliegt und sich erst definitorisch ergibt, handelt es sich bei dem Messmodell um ein Parametermodell (Fricke 1972; Jongebloed 2005, S. 340f). Eine solche Messung ist in besonderer Weise auf eine theoretische Rückbindung verpflichtet, welche hinter dem empirischen Relativ stehen muss (Jongebloed 2005, S. 349; Jablonka 2007, S. 250).

Dass es jedoch nicht unproblematisch ist, das empirische Relativ zu definieren, zeigt sich schon an den oben ausgeführten Überlegungen zur Testökologie. Rindermann (2006, S. 75 und S. 84) untersucht ebendiese Bestimmung des empirischen Relativs mit dem Ergebnis, dass die verschiedenen Kompetenzkonstrukte nicht aus-

reichend trennscharf voneinander abgegrenzt sind. ›Mathematische Grundbildung‹, ›naturwissenschaftliche Grundbildung‹, ›Problemlösefähigkeit‹ und ›Lesekompetenz‹ seien nur »global und diffus« (Rindermann 2006, S. 71) bestimmt. Oftmals wird einzig logisches Schlussfolgern sowie »allgemeines Weltwissen und/oder Schulwissen« (Rindermann 2006, S. 71) zum Beantworten der Fragen ausreichen. Angesichts der mangelnden Abgrenzung der gewählten Konstrukte im Rahmen der PISA-Untersuchungen kommt er zu der Einschätzung, dass PISA im Wesentlichen nicht ohne Erfolg die Intelligenz abprüfe, dies aber nicht expressis verbis angibt. Keitel (JM 2007, S. 55) ergänzt dazu, dass die Intelligenztests im Vergleich zu PISA auf eine längere Periode technischer Fortschritte verweisen können.

Im Interesse einer kritischen Würdigung des Messdesigns ist auch an die obigen wissenschaftstheoretischen Einwendungen anzuknüpfen, denn auf der Grundlage einer sich kritisch-rational verstehenden Wissenschaft ist die bei Sjøberg (HBR 2007, S. 212) und auch bei Langfeldt (HBR 2007, S. 239) angesprochene Geheimhaltung der Aufgaben (i.W. ›No data is published‹; Langfeldt [HBR 2007], S. 239) überaus problematisch zu bewerten. Rindermann (2006, S. 71f) führt hierzu aus, dass die üblichen statistischen und testbezogenen Informationen oft gar nicht oder nur bruchstückhaft und sehr schwer zugänglich sind. Es fehlten einige Korrelationsangaben, Aussagen über die Reliabilitätsermittlung und Faktorenanalysen, so dass die Tests insgesamt nur schwer bewertbar sind. Gaeth (2005) schließlich fordert einen freien Zugang zu den Aufgaben als Grundlage einer Validitätsüberprüfung.

Die Validität — also die Frage, ob das gemessen wird, was auch gemessen werden soll — wird von verschiedenen Autoren ausdrücklich im Wort kritisiert (Gaeth 2005; Dolin [HBR 2007], S. 106ff). Man kann sogar so weit gehen, dass auch manch andere Messprobleme implizit auf eine fehlende Validität zurückgehen. Auch Meyerhöfer (HBR 2007, S. 88; JM 2007, S. 65) zieht die Validität der Messungen erheblich in Zweifel, indem er nach der Testfähigkeit für PISA-Tests fragt, also nach den »Kenntnissen, Fähigkeiten und Fertigkeiten, die in einem Test miterfasst bzw. mitgemessen werden« (Meyerhöfer [HBR 2007], S. 88). Um diese Vermutung zu erhärten, untersucht er den Bereich der mathematischen Leistungsfähigkeit. Da die Tests Testfähigkeit mit erfassen (wie eine routinierte Bearbeitung von Multiple-Choice Aufgaben, die Ratefähigkeit oder Fähigkeiten, sich in die Realitätsvorstellung des Aufgabenstellers hineinzuversetzen), wird diese zum Teil der gesetzten Standards. MC-Tests spiegeln überdies nicht immer eine tiefgehende und mannigfaltige inhaltliche Auseinandersetzung mit den Problemen wider.

Wuttke konstatiert: »Several sources of systematic bias and uncertainty are quantitatively more important than the standard errors communicated in the official reports« (Wuttke [HBR 2007], S. 241) Ein Problem ist die Signifikanz der Standardfehler. 9 Punkte Differenz reichen demnach aus, um zwei Länder als unterschiedlich einzustufen. Bei 26 Items pro Schüler und einem Mittelwert von 500 entspricht dies aber lediglich einer halben falschen Antwort pro Schüler, was auch durch ganz andere Gründe verursacht sein kann. Weiterhin kritisiert der Münchner Physiker, dass zu einem falschen Zeitpunkt gemessen wird, da die Entwicklung mit 15 Jahren

noch nicht abgeschlossen sei und somit eine Unterschätzung auftreten kann. Ebenso argumentiert Gaeth (2005, S. 29ff), der zur Überprüfung eine lineare Regression von der mittleren Klassenstufe auf die mittlere Lesefähigkeit durchführt. Die sich ergebende Funktion verläuft mit positiver Steigung. Eine Zunahme der Klassenstufe lässt also eine Zunahme der mittleren Lesefähigkeit erwarten. In einer zusätzlichen Regressionsanalyse mit einem normalverteilten Störterm bestimmt er den Einfluss von Geschlecht, Klassenstufe und sozioökonomischem Index auf die Lesefähigkeit. Er stellt jeweils einen signifikanten Einfluss fest, wobei er für das Geschlecht hinsichtlich der Mädchen sowie für die anderen beiden Variablen positive Koeffizienten ermittelt. Eine Zunahme der Klassenstufe führt also nach beiden Verfahren zu einer Zunahme der Lesefähigkeit. Das impliziert, dass jene Länder, die mehr Schüler in höheren Klassenstufen haben, auch eine bessere Lesefähigkeit vorweisen, die Ergebnisse also verzerrt sind.

Dolin (HBR 2007, S. 100f) problematisiert die Wahl des Messmodells. Als Messmodell wird bei PISA das Rasch-Modell verwendet, welches der Item Response Theory zuzuordnen ist. Das Modell misst nur eindimensional, sodass Unterschiede zwischen Ländern außerhalb der Skala übersehen werden. Es werden Items, die in mehr als acht Ländern schlechte psychometrische Charakteristiken haben, aus der weiteren Betrachtung ausgeschlossen, so dass die kulturspezifischen Unterschiede nicht analysiert werden. Als einen weiteren kritischen Aspekt hinsichtlich der Eindimensionalität (welche als Annahme für das Rasch-Modells notwendig ist), nennt Jablonka (JM 2007, S. 261ff), dass die Aufgaben immer auf eine Hauptfähigkeit bezogen werden müssen, die abgeprüft wird. Bei Analyse der Aufgaben stellt sie allerdings fest, dass man oft mehrere Fähigkeiten zum Lösen braucht. Das Modell ist darum nicht problemlos anwendbar. Wuttke (HBR 2007, S. 252f) bemängelt, dass die einzige zulässige Abweichung bei einem solchen Modell die Verschiebung des Funktionsverlaufes entlang der Ordinate ist, nicht aber eine gänzlich andere Verlaufsform, wie sie sich indes bei verschiedenen Aufgaben bei PISA ergibt.

Hinsichtlich der Aufgabenverteilung stellt er ein mit den Testheften zusammenhängendes Problem fest. Damit der Aufwand pro Schüler gering gehalten werden kann, wurden die gesamten Aufgaben in Testhefte aufgeteilt und verschiedenen Schülern zugeteilt (sog. multi-matrix-sampling). Die Schüler erhalten dann jeweils ein aus vier Aufgabenblöcken bestehendes Testheft. Jedes Item findet sich in vier verschiedenen Testheften und hier in verschiedenen Blöcken. Aus den länderspezifischen Prüfungskulturen ergeben sich so weitere Validitätsprobleme: Unter Zeitdruck raten Dänen am Testende schnell und hastig und versuchen alles zu beantworten, deutsche Schüler hingegen gehen gründlicher mit den Aufgaben um, riskieren so allerdings, nicht alles zu schaffen.

Allerup (HBR 2007, S. 183ff) beschäftigt sich mit der Homogenität der Items, die er auf die Frage der relativen Schwierigkeit der Aufgaben bezieht. Dabei zeigt sich, dass sowohl hinsichtlich der Geschlechter und der Herkunftsländer der Schüler als auch der Untersuchungsjahre Inhomogenitäten vorliegen. Durch die Inhomogenität der Items hinsichtlich der Parameter kommt es insofern zu Verschiebungen, weil

unterschiedliche Skalen angelegt werden, aber nur ein Durchschnitt ermittelt wird. Die Ergebnisse sind dadurch verfälscht. Bei manchen Aufgaben liegt die Abweichung bei mehr als 50 Punkten. In Zusammenhang mit der Feststellung, wonach nicht in allen Ländern nach dem gleichen Verfahren Schüler ausgewählt wurden und es zu unter- bzw. überrepräsentierten Gruppen kommt, scheint dieses Problem ernst zu sein. Dieses Validitätsproblem behandeln Hörmann (HBR 2007, S. 158) sowie Wuttke (HBR 2007, S. 246f). Es zeigt sich, dass die strengen Vorgaben der Messbedingungen nicht immer eingehalten werden konnten, wie die Einbeziehung beeinträchtigter Schüler zeigt. Sie werden unterschiedlich berücksichtigt: Die Türkei hat demnach weniger als ein Prozent der benachteiligten Schüler ausgeschlossen, Spanien mehr als sieben Prozent. Es ist liegt auf der Hand, dass diese unterschiedliche Implementation eines strenge Einheitlichkeit suggerierenden Messvorganges zu einer Verzerrung der Ergebnisse führt.

Schließlich konnte Gaeth (2005, S. 62ff) hinsichtlich des Zusammenhangs zwischen Klassengrößen sowie Ergänzungs- und Nachhilfeunterricht und den Leistungen jeweils signifikante Korrelationen nachweisen, so dass auch diese unterschiedlichen Voraussetzungen in den Ländern zu unterschiedlichen Ergebnissen führen müssen. Dadurch wird eine valide Interpretation der Vergleichbarkeit weiter erschwert.

Aber auch Reliabilitätsprobleme zeigen sich trotz der Zugangsschwierigkeit zu umfassenden Daten. Immerhin lassen sie offenbar die Aussage zu, dass die Reliabilitätskoeffizienten innerhalb einzelner Länder zwar recht hoch sind, zwischen den Ländern aber stark schwanken (vgl. Dolin [HBR 2007], S. 103). Zudem ergeben sich für die soft data (Hintergrundvariablen) wie etwa die Schulausstattung oftmals geringere Reliabilitäten.

II.3 Konzeption der Aufgaben

Da sich die Validität als ein besonderes Problem darstellt, scheint es ausdrücklich geboten, die Aufgabenkonzeption in Bezug auf die damit in Verbindung gebrachten Konstrukte zu analysieren – also zu fragen, ob die Aufgabenkonzeptionen geeignet sind, die gewünschten empirischen Relative (so z.B. Lesefähigkeit, Problemlösefähigkeit oder naturwissenschaftliche Grundbildung) in ein numerisches Relativ (also den Skalenwert) abzubilden. Tatsächlich konzentrieren sich weite Bereiche der an den Leistungsvergleichsstudien geübten Kritik auf Probleme der Itemformulierung.

Bei der Analyse jeweils einer exemplarischen Aufgabe aus den Bereichen ›Lesen‹, ›Mathematik‹, ›Naturwissenschaft‹ und ›Problemlösen‹ aus dem Jahr 2000 bzw. zu letzerem aus 2003 stellt Rindermann (2006, S. 72ff) fest, dass die Aufgaben oft sehr lang sind und lösungsrelevante Informationen erst gesucht werden müssen. Selten wird spezifisches schulisches Wissen abgefragt, was allerdings in der erklärten Absicht der PISA-Tester liegt. Bodin (HBR 2007, S. 31) kommt an dieser Stelle zu einem ähnlichen Schluss, wenn er sowohl Lehrer als auch Mathematiker zitiert, die angeben, dass das Abgefragte sich nicht primär auf mathematische Inhalte bezieht. Das Konstrukt der mathematischen Grundbildung ist nicht ausreichend abgegrenzt, wodurch der »Bedeutungsgehalt [...] nicht eruiert werden kann« (Jablonka 2007, S. 261).

Gellert (JM 2007, S. 376ff) fragt nach der Verträglichkeit des mathematischen Konstrukts mit den Ausarbeitungen von Freudenthal, welcher von Gellert als einer der bedeutendsten Mathematikdidaktiker des 20. Jahrhunderts gesehen wird. Er vertritt die Auffassung, dass PISA Freudenthal nicht gerecht werden könne, obwohl es auf seinen Theoriekonzeptionen zur didaktischen Phänomenologie basiert. So verweist der Fachdidaktiker nicht auf das zukünftige Leben, sondern schlicht auf die Erkenntnis der mathematischen Genese und Konstruktion. Gellert zufolge wird keine Grundkompetenz, sondern einzig die Frage untersucht, inwieweit die einzelnen Länder mit ihrem Unterricht dazu beitragen, Aufgaben zu lösen, die sich an einem mehr oder weniger sinnvollen internationalen Curriculum orientieren.

Bodin (HBR 2007, S. 23) untersucht aus französischem Blickwinkel, wie die externe Validität in Bezug auf den mathematischen Bereich gegeben ist. Er vergleicht hierfür das Curriculum des französischen College (»from grade 6 to grade 9«; ebd., S. 24) mit den mathematischen Anforderungen aus PISA. Er kommt zu dem Schluss, dass die PISA-Fragen, welche ihm vorlagen, etwa 15% des französischen Stoffes abdecken. Diese 15% sind aber nur für 75% der PISA-Fragen relevant, was bedeutet, dass 25% der Fragen nicht Bestandteil des französischen Curriculums sind. Bei Gegenüberstellung einer französischen Vergleichsarbeit der Mittelstufe mit den PISA-Fragestellungen kommt Bodin zu dem Resultat, dass unterschiedliche Dimensionen abgefragt werden. So legt die Vergleichsarbeit einen Schwerpunkt auf den Bereich ›Knowing and recognising‹ (Bodin 2007, S. 26; etwa 65% der dortigen Fragen stellen auf diesen Punkt ab), bei den PISA-Fragestellungen sind es hingegen nur etwa 12%. Im Bereich des ›Understanding‹ und ›Creating‹ findet sich ein umgekehrtes Bild. Dolin (HBR 2007, S. 110ff) kommt für Dänemark zu einer ähnlichen Einschätzung, wenngleich hier keine vergleichbar strenge Quantifizierung erfolgt.

Die Aufgabenstellung und die Aufgabenformulierung werden gleich von mehreren Kritikern aus kultureller Perspektive angegriffen. Langfeldt (HBR 2007) gibt unter Rückgriff auf Nari (2002) bzw. Jablonka (2006) zu bedenken, dass zum Beispiel in PISA 2003 für den Teil der Mathematik 13 der 54 Aufgaben aus den Niederlanden, 15 aus Australien, 7 aus Kanada und die restlichen 19 aus den verbleibenden Ländern kommen. Es liegt nicht fern anzunehmen, dass Schüler aus diesen Ländern mit den Inhalten und den Aufgabenkontexten besser vertraut sind.

Es ist zudem auffällig, dass vier der sechs besten Länder anglophon sind (vgl. Langfeldt [HBR 2007], S. 232). Dolin zufolge kommt es zu einem ›cultural bias‹, weil der kulturelle Hintergrund der Teilnehmer sowie das Geschlecht offenbar eine Rolle spielen (HBR 2007, S. 110f). Als Beispiel gibt er eine Aufgabe an, die auf den Motorsport abstellt, in welcher Jungen deutlich bessere Ergebnisse erzielt haben. Ähnlich richtet Puchhammer (HBR 2007, S. 134) den Gang seiner Untersuchung aus, wenn er die Rangfolgen der Wortverwendungen in Deutschland und England vergleicht und feststellt, dass dieselben Wörter in beiden Ländern oft eine sehr unterschiedliche Relevanz haben (beispielweise weist ›average‹ eine Wortfrequenz auf,

die es auf Platz 388 hebt, das deutsche Wort ›Durchschnitt‹ bringt es im täglichen Gebrauch hingegen nur auf Rang 3259). Die englische Übersetzung ist insgesamt leichter zu verstehen, da der durchschnittliche Rang der englischen Worte bei 2770 liegt, wohingegen jener der deutschen 5133 beträgt. Es kommt somit also zu einer systematischen Benachteiligung der deutschen Schüler.

Meyerhöfer (HBR 2007, S. 90), aber auch Wuttke (HBR 2007, S. 257) kritisieren, dass eine unterschiedliche Vertrautheit mit den Aufgabenformaten gegeben ist, sowie durch die Übersetzungen der Aufgaben in die jeweilige Sprache Verzerrungen auftreten, die eine Aufgabe möglicherweise schneller erfassbar machen oder aber gegenteilig wirken.

II.4 Probandenmotivation

Ein weiterer vieldiskutierter Aspekt ist in einer möglicherweise kulturalistisch gebundenen unterschiedlichen Motivation der Schüler, an der Untersuchung teilzunehmen, zu erkennen. Auch hieraus könnten sich Verzerrungen ergeben und damit Validitätsprobleme einstellen.

Sjøberg (HBR 2007, S. 221) führt in diesem Kontext zwei Kritikpunkte an. Zum einen werden die Schüler in manchen Ländern wie beispielsweise Taiwan oder Singapur anders auf Tests wie PISA oder TIMMS eingestimmt als etwa in Norwegen. Ein Beispiel ist der Appell des Schulleiters, das Beste zu geben. Zudem wird die Nationalhymne beim Betreten des Prüfungsraumes gespielt. Die grundsätzliche Haltung solchen Tests gegenüber ist dort ebenfalls eine andere; so zählt exemplarisch in Singapur ›Be best - teach to the test!‹ (Sjøberg [HBR 2007], S. 221).

Neben diesen eher kontextuellen Gegebenheiten fragt Sjøberg (HBR 2007, S. 222f) nach dem ›task value‹, also danach, warum man eine Aufgabe nach eigenem Dafürhalten ausführen sollte. In Anlehnung an Rhee et al. (2005) nennt er drei verschiedene Ausprägungen des task values. Der ›attainment value‹ stellt auf die Wichtigkeit ab, die die Schüler für die Aufgabe empfinden. Der ›intrinsic value‹ fragt nach der grundsätzlichen Freude an einer Aufgabe. Schließlich betont der ›utility value‹ den Nutzen im Sinne etwa von Karrierezielen. Sjøberg führt hier an, dass in Norwegen und in Dänemark kaum einer dieser Werte zutrifft, was zu einer Verzerrung führt, weil in anderen Ländern andere Motive relevant sind, die dann auch die Leistungsbereitschaft bei der Teilnahme beeinflussen.

III. Kritik der Kritik

Die kritisch-distanzierte Auseinandersetzung mit den PISA-Studien ist bislang eher in verstreuten Randbemerkungen zu finden denn in geschlossen formulierten Gegenpositionen. Doch auch da, wo dieses geschieht, kann von einem ernsthaften wissenschaftlichen Diskurs zwischen den Mitgliedern des PISA-Konsortiums und ihren Kritikern keine Rede sein: Wenn es zur Ausformulierung kritischer Positionen gegen PISA kommt, wird dieses in der Regel souverän ignoriert. Sicherlich finden sich in den kritischen Veröffentlichungen kaum Bemerkungen über die positiven Aspekte der PISA-Untersuchungen, andererseits wird seitens des PISA-Konsortiums nur bedingt auf die Kritik eingegangen. Nur vereinzelt wird mit eher knappen, oberflächlichen und häufig ungenauen Gegenkritiken geantwortet, was nicht immer ohne schulmeisterliche Untertöne oder ›persönliche Diffamierungen‹ abgeht, wie Bender (2007, S. 333) feststellt. So qualifiziert Prenzel (2005) in einem Zeitungsinterview die Dissertation von Gaeth (2005) mit den Worten ab, dass die Kritik, die »auf viel Papier ausgeführt wird« (Prenzel 2005, o.S.), an der Studie vorbeigehe. Gaeth steht am Ende des Interviews als anmaßender Ahnungsloser dar, der als Nachwuchswissenschaftler die Frechheit besessen hat, sich mit einem hochkarätigen internationalen Expertengremium anzulegen. Neben einzelnen stichhaltigen Argumenten kommt es zu performativen Selbstwidersprüchen des Interviewten. Es wird eingestanden, dass es bei PISA Aufgaben mit negativer Trennschärfe gibt. Auch Wuttkes Kritiken werden von Prenzel damit schnell abgetan, dass dieser »vieles von PISA nicht verstanden« hätte (Prenzel zitiert in Meyerhöfer 2006, o.S.). Das skurrilste Ereignis in der Geschichte der Abwehr von PISA-Kritik ist bei Köller (2006, o.S.) nachzulesen: Hier wurde ganz offiziell von der KMK die Kritik Wuttkes von Köller untersucht und selbstverständlich abgeschmettert. Wie sollte es auch anders sein, wenn man – wie es redensartlich heißt – den Bock zum Gärtner macht: Köller ist als langjähriger Mitarbeiter bei beiden bisherigen Konsortialführern Max-Planck-Institut für Bildungsforschung (PISA 2000) und Institut für Pädagogik der Naturwissenschaften (PISA 2003 und 2006) in die Studien involviert. Seine Kritik an der Kritik weist kein einziges stichhaltiges Argument auf und konzentriert sich auf die persönliche und fachliche Abqualifikation Wuttkes.

Wuttke (JM 2007, S. 99f) selbst gibt eingangs der Neuauflage seines Artikels einen Überblick über die Resonanz zu seinem Artikel in der ersten Auflage. Die Bezugnahmen von den PISA-Vertretern auf seine Kritikpunkte sind knapp ausgefallen und insgesamt wohl als eher dürftig zu bezeichnen. Karg (2005, S. 36) liefert an dieser Stelle einen Versuch und zeichnet unter Rückgriff auf Interviews in Tageszeitungen eine Kontroverse zwischen Meyerhöfer (2004, o.S.) auf der einen sowie Blum und Neubrand (2004, o.S.) auf der anderen Seite nach.

Es ist positiv zu resümieren, dass ansonsten die eher politische als wissenschaftliche Abwehr der Kritik die Kritiker ihrerseits nicht dazu bringt, die politisch-persönliche Ebene zu suchen. Die Auseinandersetzung ist zumeist akzentuiert um die Sache bemüht, was namentlich für den vorzugsweise (wenngleich nicht ausschließlich) methodologisch angelegten Sammelband von Hopmann, Brinek und Retzl (2007) zu konstatieren ist. In dem Sammelband von Jahnke und Meyerhöfer (2007) finden sich darüber hinaus Artikel, welche neben den Auseinandersetzungen mit den Konstrukten an sich auch nach den Wirkungen von PISA sowie den wissenschaftstheoretischen und historischen Bezügen solcher Tests fragen. Diese Artikel wirken unstrukturiert und fußen auf keiner ausreichend klaren Fragestellung. Die Argumente werden nicht immer in der wünschenswerten Unvoreingenommenheit eingebracht.

Die Sammelbände liefern ein breites Spektrum an kritischen Auseinandersetzungen, wobei die meisten Kritikpunkte entweder auf den Messvorgang an sich oder auf die

Aufgabenkonzeption zielen. Die Dissertation von Gaeth (2005) hingegen verharrt bei der Frage, »ob mit dem erhobenen Datenmaterial die Kernthesen der beiden PISA-Studien zu belegen sind.« (Gaeth 2005, S. 16). Dies ist zwar bewusst so angelegt, behandelt dennoch nur einen kleinen Ausschnitt möglicher PISA-Kritik.

Hinsichtlich der Anlage der Sammelbände fällt auf, dass die Texte jeweils zum Teil redundant sind. Dies scheint zwar kaum vermeidbar, weil verschiedene Autoren mit der Abfassung betraut waren, führt aber zu einer gewissen Unübersichtlichkeit der Argumente. Eine stärkere Systematisierung wäre wünschenswert gewesen, um leichter einen Überblick über die Kritikpunkte gewinnen zu können. Aufgrund der Anlage als Sammelband kann es nicht ausbleiben, dass manche Quellen in den Texten leider gänzlich fehlen – wie etwa ausgerechnet die Arbeit von Gaeth (2005), welche die einzige von uns aufgefundene Monographie war, die sich kritisch mit der Methodologie der PISA-Studien beschäftigt. Sie wird in keinem der beiden Sammelbände aufgegriffen und diskutiert.

Durchaus berechtigte Kritikpunkte wie etwa der des cultural bias oder solche hinsichtlich der Messmethodologie scheinen nicht nur für PISA spezifische Gültigkeit beanspruchen zu können, sondern sind wenigstens zum Teil der kaum zu lösenden Schwierigkeit des internationalen Vergleichs und der Schwierigkeit des Messens an sich geschuldet. Daher kann eine Vermischung von Kritikpunkten an PISA als Studie und solchen, die allgemein gegen Vergleichsstudien gerichtet sind, nicht ausgeschlossen werden. Gleichwohl hat man sich auch bei der Einrichtung der PISA-Studie von der unbezweifelbaren methodologischen Schwierigkeit des internationalen Vergleichs nicht anfechten lassen.

Die Beurteilung der Sinnhaftigkeit internationaler Vergleichsstudien dürfte wesentlich auch durch die wissenschaftstheoretische Position des Urteilenden bestimmt sein. Hier wäre eine größere Transparenz in den Kritiken hinsichtlich der Spezifizität der Kritikpunkte auf PISA wünschenswert gewesen. Der Induktionsvorwurf ist mitunter auch den Kritiken selbst vorzuwerfen, schließen sie doch bei der Beurteilung der Itemqualität von den wenigen publizierten Aufgaben auf die PISA-Studie insgesamt. Allerdings ist ja auch von den Machern der Studie hier eine konsequente Geheimhaltungspolitik betrieben worden, die von den Kritikern der Studien wiederholt angesprochen werden. Man kann nur mutmaßen, woran hier sich das Misstrauen gegen dritte Wissenschaftler nährt, jedenfalls scheint die Sorge des Verrats vertraulich bereitgestellter Unterlagen unverhältnismäßig; wiewohl ein unkontrolliertes Bekanntwerden der Items natürlich die kostspielige Konsequenz ihrer Unbrauchbarkeit hätte. Vor allem was die Aufgabenkonzeption betrifft, ist eine Würdigung nur dann möglich, wenn man als Hilfsannahme die Repräsentativität der veröffentlichten Aufgaben unterstellt.

Mit Uljens (HBR 2007, S. 302) ist abschließend zu fragen, warum wir scheinbar zwei Parallelwelten im Diskurs der Erziehung vorfinden – den wissenschaftlichen Diskurs, von dem wesentliche Teile Gegenstand dieser Besprechung sind, und den öffentlichen Diskurs. Letzterer greift nach dem Verlust der Bildungsideale kaum einmal kritische Aspekte auf.

Ein auch kritisch akzentuierter Diskurs kann durch die beiden untersuchten Sammelbände fundiert und forciert werden. Insofern kommt beiden Publikationen ihr Rang zu. Beide Bände enthalten eine Reihe sehr lesenswerter und relevanter Beiträge; insgesamt jedoch leistet der Sammelband von Hopmann, Brinek und Retzl (2007) den substantielleren Beitrag. Eine weitergehende Systematisierung und Intensivierung der kritischen Auseinandersetzung mit den Leistungsvergleichsstudien bleibt allerdings auch nach der Herausgabe der beiden wichtigen Bände noch Desiderat.

Literatur

Ammermueller, A. (2008): PISA: What Makes the Difference? Explaining the Gap in PISA Test Scores Between Finland and Germany. In: Dustmann, C./Fitzenberger, B./Machin, S. (Hrsg.): The Economics of Education and Training. Physica-Verlag: Heidelberg, S. 241-266.

Bayrhuber, H./Ralle, B./Reiss, K. et al. (Hrsg.) (2004): Konsequenzen aus PISA. Perspektiven der Fachdidaktiken. StudienVerlag: Innsbruck.

Blum, W./Neubrand, M. (2004): Der schiefe Blick auf Pisa. In: Süddeutsche Zeitung vom 11.12.2004.

Deutsches PISA Konsortium (Hrsg.) (2004): PISA 2003: der Bildungsstand der Jugendlichen in Deutschland. Ergebnisse des zweiten internationalen Vergleichs. Waxmann: Münster.

Deutsches PISA Konsortium (Hrsg.) (2007): PISA 2006: die Ergebnisse der dritten internationalen Vergleichsstudie. Waxmann: Münster.

Fertig, M. (2004): What can we Learn from International Student Performance Studies? Some Methodological Remarks. In: RWI Discussion Papers 23/2004. Essen.

Fricke, R. (1972): Über Meßmodelle in der Schulleistungsdiagnostik. Schwann: Düsseldorf.

Gaeth, F. (2005): PISA (Programme for International Student Assessment). Eine statistisch-methodische Evaluation. Diss. Freie Universität Berlin.

Hopmann, S. T./Brinek, G./Retzl, M. (Hrsg.) (2007): PISA zufolge PISA – PISA According to PISA. Hält PISA, was es verspricht? – Does PISA Keep What It Promises? LitVerlag: Wien, Münster. Zugleich abrufbar unter http://www.univie.ac.at/ pisaaccordingtopisa/pisazufolgepisa.pdf, Stand: 02.02.2009.

Hörner, W. (2004): ›Europa‹ als Herausforderung für die Vergleichende Erziehungswissenschaft – Reflexionen über die politische Funktion einer pädagogischen Disziplin. In: Tertium Comparationis. Journal für International und Interkulturell Vergleichende Erziehungswissenschaft 2/2004, S. 230-244.

Jahnke, T./ Meyerhöfer, W. (Hrsg.) (²2007): PISA & Co. Kritik eines Programms. Fraunzbecker: Hildesheim.

Jongebloed, H.-C. (2005): Die Messung schulischer und betrieblicher Leistungen in bildungsökonomisch-modellhafter Sicht. In: Bank, V. (Hrsg.): Vom Wert der Bildung. Bildungsökonomie in wirtschaftspädagogischer Perspektive neu gedacht. Haupt: Bern, Stuttgart, Wien: Haupt, S. 331-354.

Karg, I. (2005): Mythos PISA. Vermeintliche Vergleichbarkeit und die Wirklichkeit eines Vergleichs. V&R unipress: Göttingen.

Köller, O. (2006): Kritik an PISA unberechtigt. IQB-Direktor Olaf Köller zu den Vorwürfen gegenüber der Studie. In: http://bildungsklick.de/a/50155/kritik-an-pisa-unberechtigt/, Stand: 30.01.2009.

Mejding, J./Roe, A. (Hrsg.) (2006): Northern Lights on PISA 2003. A Reflection from the Nordic Countries. Nordic Council of Ministers: Kopenhagen.

Meyerhöfer W. (2004): Und wieder sehen wir betroffen die Studie an und alle Fragen offen. In: Süddeutsche Zeitung vom 07.12.2004.

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

Meyerhöfer, W. (2006): Statistische Ungereimtheiten. Daten ohne Aussagekraft. In: Freitag 46 vom 17.11.2006. Zugleich abrufbar unter http://www.freitag.de/2006/46/06460401.php, Stand: 19.01.2009.

OECD (Hrsg.) (2001): Lernen für das Leben. Erste Ergebnisse der internationalen Schulleistungsstudie PISA 2000. OECD: Paris.

Prenzel, M. (2005), Viel gerechnet, aber wenig nachgedacht. Interview von Schlicht, U. In: Der Tagesspiegel vom 01.09.2005. Abrufbar unter http://www.tagesspiegel.de/magazin/wissen/gesundheit/;art300,1884136, Stand: 19.01.2009.

Radtke, F.-O. (2005): Die Schwungkraft internationaler Vergleiche. In: Bank, V. (Hrsg.): Vom Wert der Bildung. Bildungsökonomie in wirtschaftspädagogischer Perspektive neu gedacht. Haupt: Bern, Stuttgart, Wien, S. 355-386.

Rhee, C. B./Kempler, T./Zusho, A. et al. (2005): Student learning in science classrooms: what role does motivation play? In: Alsop, S. (Hrsg.): Beyond Cartesian Dualism. Encountering Affect in the Teaching and Learning of Science. Springer, Science and Technology Education Library: Dordrecht.

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

Sjøberg, S. (2004): Internationale Vergleichsstudien – ihre guten und schlechten Seiten. In: Bayrhuber, H./Ralle, B./Reiss, K. et al. (Hrsg.) (2004): Konsequenzen aus PISA. Perspektiven der Fachdidaktiken. StudienVerlag: Innsbruck, S. 51-61.