Ein weiterer Aspekt des nationalen Ergänzungstests betraf die Erkennung leseschwacher Schüler durch Lehrkräfte. Hier konnte am Beispiel verschiedener systematischer Untersuchungen gezeigt werden, dass Lehrkräfte testleistungsstarke und weniger testleistungsstarke Schülerinnen und Schüler identifizieren können. Weiterhin wurde gezeigt, dass für die Bewertung eines Testverfahrens die Gleichsetzung von Testergebnis und Testeigenschaften nicht sinnvoll ist.

Es hat sich gezeigt, dass für den Nachweis der Koppelung von sozioökomomischem Hintergrund der Eltern und Testleistung des Schülers der Schulformvergleich ungeeignet ist. Die Entkoppelung von Schulform und Schulabschluss ist ein wesentliches Element des deutschen Schulsystems. Der an dessen Stelle in der vorliegenden Untersuchung eingesetzte Vergleich der geplanten Schulabschlüsse entspricht dem der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie (TIMSS).

5 Abschließende Bewertung

5.1 International

Die Generierung international vergleichbarer Stichproben angesichts strukturell unterschiedlicher Schulsysteme ist ein nach wie vor ungelöstes Problem. Insbesondere ist das Verhältnis von Schul- zu Lebensalter ein wesentlicher Faktor, der den direkten Vergleich unterschiedlicher Systeme zu einer Frage der Perspektive macht. Die Ergebnisse von Leistungstests variieren in dem Maße, in dem der einen oder anderen Variablen mehr Bedeutung beigemessen wird.

Weitere Probleme betreffen die praktische Umsetzung einer weltweit angelegten Untersuchung.[145] Teilnehmerstaaten in Asien mit hohen Testwerten folgten nicht den Vorgaben der Stichprobenziehung. Ein europäischer Teilnehmerstaat weitete das Zeitfenster der Erhebung aus. Die in dieser Zeitspanne getesteten Schulen wurden nachträglich wegen auffällig hoher Performanz disqualifiziert. Es ist nicht auszuschliessen, dass der besondere Leistungszuwachs dieses Teilnehmerstaats in der PISA 2003 Studie auf eine unterschiedliche Handhabung dieses Problems in der Folgestudie zurückzuführen ist. In einem weiteren Teilnehmerstaat wurden sämtliche Schulen eines Landesteils ausgeschlossen.

Die Auswahl der Schüler war bereits in der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie (TIMSS) Gegenstand der Diskussion. Dieser Sachverhalt gilt

145 Vgl. Adams (ACER), Wu: PISA 2000 Technical Report. S. 182 ff

ebenfalls für die in PISA gezogenen Stichproben der Teilnehmerstaaten. Die Testergebnisse sind gleicherweise vor diesem Hintergrund zu bewerten. Stichprobenspezifische Probleme hatte bereits Collani angesprochen.[146]

Es hat sich gezeigt, wie sehr die Bildung von Aggregaten letztlich Zusammenhänge aufzeigen, die lediglich statistische Effekte darstellen, etwa Stage Migration als Ergebnis einer über die Bundesländer unterschiedlichen Bildungsexpansion.

Die statistische ‚Genauigkeit', beispielshalber in Form sehr kleiner Konfidenzintervalle, ein Resultat des Stichprobenumfangs von etwa 180.000 Schülern, mindert inhaltliche Präzision. In dem Maße, in dem der zu erklärende Gegenstand präziser definiert wird, reduzieren sich die Vergleichsgruppen. Unscharfe Definition des Untersuchungsgegenstande ist gewissermaßen der Preis großer Stichproben.[147]

Es konnte erwartungsgemäß ein starker Effekt schulischer, nomineller Lernzeit auf die Testleistungen von Schülern nachgewiesen werden. Eine vergleichende Untersuchung, die Aussagen über Schulsysteme treffen will, muss daher zumindest die Verweildauer in dessen Institutionen berücksichtigen. Die Frage liegt deshalb nahe: Kann man unter dieser Voraussetzung überhaupt Schülerleistungen in verschiedenen Ländern vergleichen und sind dann gefundene Unterschiede nicht trivial.[148]
Die Vergleiche über einzelne Klassenstufen sowie die kovarianzanalytischen Berechnungen zeigen einen geringen Grenzertrag besonders früher Einschulung in England oder Neuseeland. Die englischen Schülerinnen und Schüler der IGLU-Studie erreichen trotz höheren Schulalters mit $\mu=553$ ein Niveau, das nur knapp über der deutschen Vergleichsgruppe liegt ($\mu=539$). Dass in der internationalen Grundschulstudie (IGLU) die Mehrzahl der Schülerinnen und Schüler der Klassenstufe 4, in England, Neuseeland und Schottland jedoch die Klassenstufe 5 getestet wurde, ist zu bemerken. Die englischen, australischen und neuseeländischen Schülerinnen und Schüler erreichen in PISA unter Berücksichtigung des hohen Schulalters lediglich ein niedriges Niveau.

---

146 Collani, E.v.: OECD PISA – An Example of Stochastic Illiteracy?, Economic Quality Control Vol 16 (2001), No.2, S. 240

147 Wößmann berichtet vom Poolen zweier TIMSS-Datenstätz zu einem Datensatz von N = 447.089. Wößmann, L.: How Central Exams Affect Educational Achievement: International Evidence from TIMSS and TIMSS-Repeat S. 19 Paper prepared for the conference Taking Account of Accountability: Assessing Politics and Policy John F. Kennedy School of Government Harvard University June 10 – 11, 2002

148 Baumert, Bos, Watermann: Mathematische und naturwissenschaftliche Grundbildung im internationalen Vergleich. In: Baumert, Bos, Lehmann (Hrsg.): TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie 2000, Band 1. S. 192

Nachdem in vorliegender Arbeit ein Einfluss des zwischen den Teilnehmerstaaten variierenden Schulalters auf die Höhe der Testergebnisse nachgewiesen und damit die Einschätzungen von Mullis, Martin, Beaton, Gonzalez, Kelly und Smith[149] bestätigt werden konnten, stellt sich die Frage nach den Ursachen der hohen finnischen Testwerte angesichts später Einschulung und hoher Rückhaltequote (Retentivität).

Das finnische Schulsystem wird in der fachwissenschaftlichen Diskussion sowohl als hochgradig integriert als auch im Gegensatz dazu als hochgradig selektiv dargestellt, wobei die Darstellung der hohen Testwerte als Folge eines integrierten Systems mit später Einschulung und hoher Rückhaltequote nicht kompatibel zu den vorliegenden Daten wäre.[150] [151] Eine Stratifikation des Schulsystems hätte Konsequenzen für die Stichprobenziehung und infolgedessen für die Testergebnisse. Diese würde für das im Test eingesetzte Stichprobenverfahren erhebliche, kaum zu lösende Probleme aufwerfen, insofern die Schichten nicht explizit ausgewiesen sind. Eine genauere Prüfung dieser Frage bleibt weiteren Untersuchungen vorbehalten. Es liegt auch in Hinblick auf vorliegende Untersuchungsergebnisse nahe, den je nach Teilnehmerstaat erheblich unterschiedlichen Freiheitsgraden der Stichprobenziehung insbesondere der Anzahl an Stratifikationsvariablen einen Einfluss auf die Testergebnisse einzuräumen.

149 Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., Smith, T.A.: Mathematics Achievement in the Primary School Years (TIMSS), Boston College 1997, S. 30

150 Linnakylä, Pirjo, Arbeitsgruppe „Internationale Vergleichsstudie“: Berlin Juni 2003 Vertiefender Vergleich der Schulsysteme ausgewählter PISA-Teilnehmerstaaten. S. 40

151 von Freymann: PISA in Finnland. In: MUT 09/2003. S. 74 ff

5.2 National (Bundesrepublik)

Der Vergleich innerhalb Deutschlands war nicht Schwerpunkt der hier durchgeführten Untersuchung. Die Testleistungen der Bundesländer ergeben sich aus der unterschiedlichen Zusammensetzung der Stichproben sowie unterschiedlicher Bildungsexpansion. Hier kommen statistische Effekte zum Tragen.[152]

Der Vergleich zwischen Gesamtschule und gegliedertem Schulsystem zeigt etwas höhere Korrelationen von sozialer Herkunft und Testleistung im gegliederten System. Insgesamt haben sich die in PISA berichteten hohen *Disparitäten* jedoch lediglich als statistischer Effekt erwiesen.

Es ist die Eigenschaft einer Vielzahl von Bildungssystemen, u.a. auch des deutschen, in Anerkennung unterschiedlicher Entwicklungsniveaus, Schul- und Lebensalter in engen Grenzen individuell zu differenzieren. Diese Flexibilität schafft zwar eine gößere Streuung der Testergebnisse bei einer am Lebensalter orientierten Perspektive, jedoch eine verhältnismäßig geringe Streuung bei einer am Schulalter orientierten Perspektive, wie die Autoren der IGLU-Studie, Bos, Valtin, Lankes, Schwippert, Voss, Badel und Plaßmeier bestätigen.

> *Für Deutschland insgesamt gilt: Die Streuung der Leistungswerte ist am Ende der vierten Jahrgangsstufe klein. Nur wenige andere Staaten erreichen eine geringere Streuung und übergeben somit eine in ihren Leseleistungen insgesamt homogenere Schülerschaft an nachfolgende Klassen.*[153]

---

152 Fertig und Wrigth beschreiben statistische Effekte, die im Rahmen von Auswertungen von PISA-Daten auftreten und lediglich verschiedenen Aggregatniveaus zuzuschreiben sind (Aggregation Bias).

> *As the level of aggregation at which class size is measured increases, the effect changes from being small, positive and statistically significant (elasticity=0.04) to being small, negative and statistically significant (elasticity =-0.07).*

Fertig, Wright: School Quality, Educational Attainment and Aggregation Bias, IZA DP No. 994, Januar 2004

153 Bos, Valtin, Lankes, Schwippert, Voss, Badel, Plaßmeier: Lesekompetenzen am Ende der vierten Jahrgangsstufe in einigen Ländern der Bundesrepublik Deutschland im nationalen Vergleich. In: Bos, Lankes, Prenzel, Schwippert, Valtin, Walther (Hrsg.): IGLU Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich (II) 2004. S. 85

5.3 Weiterführende Perspektiven

5.3.1 Testkonstruktion

Es wäre sicherlich von Nutzen, die in PISA eingesetzten Aufgaben zu betrachten, sobald diese der Öffentlichkeit zugänglich werden. Die individuellen Testwerte der Schüler in Mathematik und Naturwissenschaften, kovariieren mit dem Lesetestwert mit $r_p = 0{,}839$ und $r_p = 0{,}872$, wodurch sich die Frage der Abgrenzbarkeit stellt.[154]

Die bislang veröffentlichten Testaufgaben werfen meßtheoretische Fragen auf: Neben dem hohen Sprachanteil des Bereichs *Scientific Literacy*[155] wären insbesondere die Präzision der Fragestellungen, Kongruenz zu wissenschaftlichen Erklärungen,[156] Eindeutigkeit des Explanandums und eine Reihe weiterer Probleme anzusprechen. Messener hatte in diesem Zusammenhang nach dem Gegenstand der Untersuchung gefragt und auf die Praxis der Testentwicklung hingewiesen.[157] Hagemeister hatte diese Problematik bereits in Bezug auf die TIMSS angesprochen.[158] Eine breite wissenschaftliche Öffentlichkeit mit Zugang zu den Testaufgaben scheint die dringend notwendige Voraussetzung valider Testkonstruktion zu sein.[159]

---

154 vgl. dazu auch die Diskussion um die in der TIMSS eingesetzten Items von Hagemeister. Hagemeister, V.: Was wurde bei TIMSS erhoben? Die Deutsche Schule, 91.Jg.1999, S.160-177

155 Baumert, Klieme, Lehrke, Savelsbergh hatten in Bezug auf das Argument der Sprachlastigkeit der TIMSS (Hagemeister) hingegen die hohe Mathematiklastigkeit des Physikunterrrichts angesprochen. (Jürgen Baumert, Eckhard Klieme, Manfred Lehrke & Elwin Savelsbergh: Konzeption und Aussagekraft der TIMSS-Leistungstests. Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik S. 19ff)

156 Als Beispiel aus der PISA-Studie: S129Q01 (DAYLIGHT/TAGESLICHT) Frage: *Welche Aussage erklärt, warum es auf der Erde Tageslicht und Dunkelheit gibt?* Antwort A (–als *vollständig gelöst* bewertete Antwort- *Die Erde rotiert um ihre Achse*) ist notwendige (jedoch nicht hinreichende) Voraussetzung für den Wechsel von Tag und Nacht. Antwort D (–als *nicht gelöst* bewertete Antwort- *Die Erde dreht sich um die Sonne*) ist ebenfalls ein Teil der Voraussetzung, da nicht die Erdrotation allein ursächlich für den Wechsel von Tag und Nacht ist, sondern das Verhältnis von Erdrotation (um ihre Achse) und der Rotation der Erde um die Sonne.
Nikolas Knake (---.dip.t-dialin.net) hat folgende Interpretation für die Erklärung, *warum es auf der Erde Tageslicht und Dunkelheit* gibt, die weder Erdrotation noch Rotation der Erde um die Sonne notwendig macht (und der Fragestellung möglicherweise näher kommt): *Tageslicht gibt es nämlich weil die Sonne elektromagnetische Strahlen eines bestimmten Frequenzbereiches emittiert, die wir Menschen als Tageslicht bezeichnen* (http://www.skh.de/phorum/read.php?f=8&i=70&t=70#reply_70).

157 Messener, R.: PISA und Allgemeinbildung In: Zeitschrift für Pädagogik 3/2003, S. 402

158 Hagemeister: Was wurde bei TIMSS erhoben? Die Deutsche Schule, 1999

159 Baumert, Klieme, Lehrke, Savelsbergh: Konzeption und Aussagekraft der TIMSS-Leistungstests. MPIB 1999

5.3.2 Curriculare Validität

Curriculare Validität wäre ein weiterer Gegenstand von wissenschaftlichem Interesse. In der Bundesrepublik Deutschland beziehen sich schulische Curricula vorwiegend auf Klassen-, jedoch kaum auf Altersstufen. Curriculare Validität im Rahmen einer Altersstichprobe von 15-Jährigen würde sich in der Bundesrepublik Deutschland unter Bezug auf die PISA-Daten gleichermaßen auf 1,2% Schüler der Klasse 7, 13,2% Schüler der Klasse 8, 63,3% Schüler der Klasse 9, 22,3% Schüler der Klasse 10 und 0,1% Schüler der Klasse 11 beziehen. Dieses Problem stellt sich in ähnlicher Weise für mehr als zwei Drittel der teilnehmenden Staaten und verweist auf den explizit normativen Charakter der Studien.[160]

5.4 Resümee

PISA wirft grundlegende Fragen auf, die über den wissenschaftlichen Rahmen hinaus gehen. Ergebnisse international vergleichender Bildungsstudien sollten nicht ungeprüft übernommen werden. Eine sachgerechte Behandlung der Thematik wäre zu wünschen.

---

[160] Baumert, Stanat, Demmrich: PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In: Deutsches PISA-Konsortium (Hrsg.): PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. S.19