Bos, Wilfried; Voss, Andreas

Empirische Schulentwicklung auf Grundlage von Lernstandserhebung. Ein Plädoyer für einen reflektierten Umgang mit Ergebnissen aus Leistungstests

Die Deutsche Schule 100 (2008) 4, S. 449-458

Wilfried Bos/Andreas Voss

Empirische Schulentwicklung auf Grundlage von Lernstandserhebung

Ein Plädoyer für einen reflektierten Umgang mit Ergebnissen aus Leistungstests

Empirical School Development on the Basis of Performance Tests A Plea for a Reflected Handling of the Results from Large-Scale Assessments

Die Rückmeldungen von Ergebnissen aus empirischen Leistungsvergleichsstudien an die beteiligten Schulen werden mittlerweile routinemäßig durchgeführt. Die an die Kollegien zurückgespiegelten Untersuchungsergebnisse sollen zur Schul- und Unterrichtsentwicklung in den jeweiligen Schulen genutzt werden. Da jeder Schüler bzw. jede Schülerin mit einer überschaubaren Anzahl von Aufgaben eines Themengebietes getestet wird, handelt es sich bei den Testergebnissen aus Leistungsvergleichsstudien jedoch um fehlerbehaftete Schätzungen für die wahren Schülerkompetenzwerte. Das Ausmaß des Schätzfehlers ist jedoch quantifizierbar und sollte bei der Interpretation der Ergebnisse berücksichtigt werden, um Fehlschlüssen vorzubeugen. Es wird gezeigt, wie groß Messfehler aus Leistungsvergleichstudien bei Aussagen auf Individual-, Klassen- und Schulebene sind, von welchen Faktoren sie abhängen und wie sie mit alternativen Untersuchungsdesigns reduziert werden können.

Schlüsselwörter: Leistungsvergleichsstudien, Rückmeldung, Bildungsforschung, Unterrichts- und Schulentwicklung

Feedback of results from large-scale assessments to the schools participating in such studies is now part of the usual routine of empirical research. This feedback should be used by principals and teachers of the respective schools to enhance school development and instruction processes. As each student is assessed in such studies with a limited number of items for each competence domain tested, the test results are mere estimations of the students' true ability scores and contain a certain measurement error. However, the size of this error can be determined and should be considered when interpreting the results of such assessments in order to avoid drawing erroneous conclusions. This paper shows how large these measurement errors are when making statements on individual, class and school levels. In addition, this paper addresses the factors which influence the extent of measurement errors and presents a number of different research designs which can reduce such errors.

Keywords: standardized testing and reporting, educational research, instructional and school development

Zitat:

„Dieses Testverfahren [zentrale Tests in Schweden] ergab aus Sicht der Schulaufsicht und auch von Seiten der Konstrukteure nur einen von vielen Indika-

toren für die Leistung des einzelnen Schülers, der Schule und auch für den Leistungsstand des Faches bzw. der Schule insgesamt. Diese Einschränkung beruhte auf der Einsicht über die Begrenzungen standardisierter Testverfahren. So kann eine Reihe von Fähigkeiten und Aspekten nicht oder nur mit nicht vertretbarem Aufwand erfasst werden. Schulen und insbesondere Lehrerinnen und Lehrer sahen die Beschränkungen und Grenzen der zentralen Tests aber häufig nicht. Sie gaben nicht selten den Tests ein höheres Gewicht, als ihnen zukommen sollte, und konzentrierten sich im Unterricht stark auf Testvorbereitung (teaching to the test)" (Eikenbusch 1995, S. 22).

1. Einleitung

Die unter erheblichem Aufwand auf Bundeslandebene im Schuljahr 2004/2005 eingeführten schulübergreifenden Leistungsvergleichstests (Lernstandserhebungen) stellen alle daran Beteiligten vor neue Aufgaben. Von Seiten der Ministerien ist es angedacht, die aus diesen Untersuchungen gewonnenen Informationen zukünftig als zentrales Element einer systematischen Standardsicherung und ergebnisorientierten Unterrichtsentwicklung zu nutzen.

Es ist u.a. vorgesehen, dass

- die Fachlehrkräfte den einzelnen Schülerinnen und Schülern die von ihnen erzielten Ergebnisse in Verbindung mit dem Ergebnis für die jeweilige Klasse und die Schule zurückspiegeln. Zusätzlich sollen die Erziehungsberechtigten über ausgewählte Ergebnisse mit einem Formblatt informiert werden;
- in den Lehrer- bzw. den entsprechenden Fachkonferenzen über die erzielten Leistungen diskutiert wird und Konsequenzen für die schulische Arbeit hieraus abgeleitet werden,
- die Leitung der jeweiligen Schulen das Kollegium über landesweite Referenzwerte informiert und Lehrerinnen und Lehrer die Ergebnisse ihrer Klassen bzw. die Schulergebnisse in diesen landesweiten Ergebnissen verorten und hieraus Konsequenzen für die zukünftige schulische Arbeit ableiten,
- die Ergebnisse der Lernstandserhebungen im Rahmen von Leistungsbewertungen bei der Festlegung von Halbjahresnoten ergänzend berücksichtigt werden, um die Bedeutung der Lernstandserhebungen für Schülerinnen und Schüler sowie bei den Lehrkräften zu erhöhen, und
- die Testergebnisse zur Ermittlung des Förderbedarfes von Schülerinnen und Schülern herangezogen werden.

Es ist also beabsichtigt, die Ergebnisse der Vergleichsarbeiten für die Schulentwicklung auf Landes-, Schul-, Klassen- und Individualebene zu nutzen (vgl. MSJK 2005b, 2006).

Bei der konkreten Umsetzung dieser Pläne ergibt sich jedoch das Problem, dass die Daten aus Vergleichsarbeiten Schülerleistungen nur mit einer begrenzten Genauigkeit ermitteln und die Nutzung dieser Daten für die Rückmeldung auf Klassen- und Individualebene für die oben genannten Zwecke aus methodologischer Sicht problematisch ist. Bei der Quantifizierung von Schülerleistungen aus Vergleichsarbeiten sind zwei Arten von Messfehlern zu berücksichtigen.

Wird eine Schülerstichprobe genutzt, um beispielsweise eine Aussage über das Leistungsniveau der dieser Stichprobe zu Grunde liegenden Population vorzunehmen, so ist die Genauigkeit dieser Aussage vom Stichprobenumfang und der Heterogenität der Testleistungen in der Population bzw. der Stichprobe abhängig. Begreift man die Testerhebung in Vergleichsarbeiten als Vollerhebungen ganzer Klassen und Schulen, so ist der *Stichprobenfehler der Person* in der Regel gering und als solcher zu vernachlässigen. Nicht zu vernachlässigen ist hingegen bei der Vorgabe einer begrenzten Auswahl von Testaufgaben aus einer Gesamtheit von möglichen Aufgaben, die zu einem bestimmten Kompetenzbereich gestellt werden können, der *Standardschätzfehler der Personenparameter.* Mit dem Standardschätzfehler wird das Ausmaß an Unsicherheit quantifiziert, das mit der Aussage aus einer Aufgabenstichprobe verbunden ist. Der Standardschätzfehler reduziert sich mit zunehmendem Umfang der Aufgaben, die zur Leistungsmessung eingesetzt werden. In Lernstandserhebungen bearbeiten Schülerinnen und Schüler in der Regel zwischen 10 und 30 Aufgaben je Themengebiet. In individualdiagnostischen Untersuchungen bearbeiten Kinder ein Vielfaches dieser Aufgaben. Ergebnisse aus individualdiagnostischen Untersuchen sind daher genauer, als es Ergebnisse aus Lernstandserhebungen sein können.

Eine weitere Einschränkung ergibt sich aus der *Erfassungsvalidität* dieser Daten, da die Schülerlösungen durch die Lehrkräfte der betroffenen Schule bzw. Klasse eingegeben werden. In den großangelegten Leistungsvergleichsstudien wie PISA, TIMSS und IGLU werden die Schülerdaten durch externe Testleiter erhoben und die Dateneingabe bzw. -verarbeitung basiert auf automatisierten und standardisierten Routinen.

In diesem Beitrag soll am Beispiel von Daten zur Erfassung von Leseverständnis aufgezeigt werden, mit welchem Ausmaß an Unsicherheit Aussagen aus Schulleistungstests auf Individual-, Klassen- und Schulebene behaftet sein können.

2. Datengrundlage und Analyseverfahren

Für die folgende Darstellung zur Genauigkeit von Testergebnissen aus Vergleichsarbeiten wird auf Daten der im Jahr 2001 durchgeführten Internationalen-Grundschul-Lese-Untersuchung (IGLU) zurückgegriffen. Die Testzeit der Lesetests in IGLU lag bei insgesamt 80 Minuten. Die Kinder haben jeweils zwei Texte mit insgesamt rund 25 Aufgaben bearbeitet (Lankes et al. 2003, S. 21). In der im Jahre 2005 u.a. in Nordrhein-Westfalen durchgeführten Lernstandserhebung VERA lag die Testzeit bei 50 Minuten (vgl. MSJK 2005a). Es ist also davon auszugehen, dass mit den IGLU-Lesetests auf Individualebene mindestens soviel Informationen über das Leseverständnis eines Kindes zur Verfügung stehen wie aus den Daten der Vergleichsarbeit.

Die Datenanalyse basiert auf der Grundlage von probabilistischen Testverfahren. Diese Auswertungsverfahren haben sich im Bereich der empirischen Bildungsforschung als Standardverfahren etabliert und werden gewöhnlich auch für die Auswertung der Daten aus Vergleichsarbeiten eingesetzt.

In internationalen Schulleistungsstudien werden üblicherweise Informationen, die im Rahmen der eingesetzten Schüler- und Elternfragebögen erhoben werden, als zusätzliche Informationsquellen beim Schätzungsprozess der Schülerfähigkeiten in Form von so genannten Hintergrundmodellen eingebunden. Da in Vergleichsarbeiten hauptsächlich leistungsbezogene Daten von den Schülerinnen und Schülern erhoben und ergänzende Schüler- und Elternfragebögen in der Regel nicht eingesetzt werden, basieren die dargestellten Ergebnisse ausschließlich auf der Grundlage der erfassten Testdaten. Da Informationen zu Bildungshintergrund, Leseinteresse, Leseverhalten, familiärer Unterstützung etc. aus den Vergleichsarbeiten nicht zur Verfügung stehen, wurden die deutschen IGLU-Daten aus Gründen der Vergleichbarkeit mit einem einparametrigen logistischen Modell reskaliert.[1] Alle in diesem Beitrag berichteten Ergebnisse (Mittelwerte und deren Standardfehler – also ohne Hintergrundmodell) basieren auf diesen reskalierten IGLU-Daten.

3. Analysenergebnisse

In Tabelle 1 sind zentrale statistische Kennwerte für verschiedene Bezugsgruppen bzw. für drei Schüler mit unterschiedlichen Leseverständniswerten wiedergegeben. Für Deutschland ist der Mittelwert auf den Wert 500 zentriert. Der Standardfehler des Mittelwertes (S.E.) beträgt 2,1. Mit diesem statistischen Wert lässt sich das Ausmaß an Unsicherheit quantifizieren, das mit einer stichprobenbasierten Schätzung eines Populationsparameters verbunden ist: Für den deutschen Stichprobenmittelwert von 500 kann mit einer Wahrscheinlichkeit von 95 Prozent angenommen werden, dass der ‚wahre' Populationsmittelwert aller deutschen Grundschulkinder in einem Bereich zwischen 495,8 und 504,2 liegt. Dieser enge Wertebereich zeigt, welches hohe Maß an Präzision internationale Vergleichsstudien auf Ebene von Ländervergleichen erzielen. In den nächsten beiden Spalten ist ein Konfidenzintervall für wahrscheinliche Populationsparameter dargestellt, die den Stichprobenkennwert mit einer Wahrscheinlichkeit von 95 Prozent erzeugt haben können.

Die Konfidenzintervalle für die drei Schüler in den Zeilen fünf bis sieben mit den Leseverständniswerten 400, 500 und 600 verdeutlichen die Probleme, die sich mit dem Anspruch einer punktgenauen Verortung von einzelnen Schülern auf der IGLU-Kompetenzskala ergeben. Lediglich für die Leseverständnisleistungen der Schüler 1 und 3 lässt sich aus statistischer Sicht von einem bedeutsamen Leistungsunterschied sprechen (400 bzw. 600 Punkte). Die Konfidenzintervalle für die Schüler 1 und 2 bzw. 2 und 3 überschneiden sich, obwohl die Testleistungen dieser Schüler 100 Punkte auseinander liegen, was einer Leistungsdifferenz von gut zwei Schuljahren entspricht.

In den Spalten mit den Überschriften min (S.E.), max (S.E.) und Mittelwert (S.E.) sind beschreibende Funktionen für die Bezugsgruppen Bundesländer, Schulen und Klassen dargestellt. Wie zu erkennen ist, steigt der gemittelte Standardfehler in Abhängigkeit von der Bezugsgruppengröße von 6,8 im Falle der

1 Diese Modellklasse kommt auch in den PISA-Studien zur Anwendung.

Bundesländer bis 21,0 für die 390 gemittelten Klassenwerte. Zusätzlich sind die minimalen und maximalen Standardfehler für die jeweiligen Bezugsgruppen dargestellt. Für die sieben Bundesländer beispielweise, die an den IGLU-Bundeslandvergleichen teilgenommen haben (vgl. Bos u.a. 2004), beträgt der geringste Standardfehler 4,1 Punkte und der größte 11,7 Punkte.

In der äußeren rechten Spalte sind die Verfahren, mit denen diese Standardfehler aus den reskalierten IGLU-Daten geschätzt wurden, wiedergegeben. Die Schätzungen für das Bundesgebiet bzw. die sieben deutschen Bundesländer basieren auf Schätzungen mit so genannten Replikationsverfahren (Programm WesVar), womit dem Erhebungsdesign dieser Schulvergleichstudie Rechnung getragen wird. Die Berechnungen für die Schul- bzw. Klassenschätzungen wurden mit einem Standardprogramm durchgeführt; die Schätzung auf Individualebene erfolgte mit einer Skalierungssoftware für probabilistische Verfahren (Conquest).

Tabelle 1: Standardfehler für verschiedene Referenzeinheiten im Vergleich

	Mittelwert	(S.E.)	Konfidenzintervall Mittelwert +/-2(S.E)	min (S.E.)	max (S.E.)	Mittelwert(S.E.)	berechnet mit
Deutschland	500	2,1	495,8 - 504,2	—	—	—	WesVar
Bundesländer (n=7)*	—	—	— —	4,1	11,7	6,8	WesVar
Schulen (n=211)	—	—	— —	9,0	50,3	16,0	SPSS
Klasse (n=390)	—	—	— —	5,8	50,3	21,0	SPSS
Schüler 1	400	42,2	315,6 - 484,4	—	—	—	Conquest
Schüler 2	500	39,2	421,7 - 578,3	—	—	—	Conquest
Schüler 3	600	47,3	505,4 - 694,6	—	—	—	Conquest

* Bundesländer, die am IGLU Bundeslandvergleich teilgenommen haben.

Den Angaben in Tabelle 1 kann die Systematik entnommen werden, nach der sich das Ausmaß der Standardfehler ergibt. Die nominelle Größe dieser Fehler hängt zum einem von der Stichprobengröße der Bezugsgruppen ab und zum anderen von der Passung aus Personenfähigkeit und Schwierigkeit der gestellten Aufgaben (vgl. Rost 1996, S.352). Der geringe Standardfehler Deutschlands ergibt sich durch den Umstand, dass für die Berechnung dieses Mittelwerts die Information von 7.633 Kindern verdichtet wurde, die jeweils rd. 25 Aufgaben bearbeitet haben. Auf Ebene der Bundesländer wächst die Unsicherheit der Aussagen im Durchschnitt bereits um den Faktor drei, für Aussagen auf Schulebene im Mittel um den Faktor acht und auf Klassenebene um den Faktor zehn. Es ist davon auszugehen, dass die Standardfehler und damit das Ausmaß an Unsicherheit insbesondere für große Schulen mit mehr als zweizügigen Jahrgängen günstiger ausfällt als für die hier dargestellten IGLU-Schulen, in denen jeweils zwei vierte Klassen eines Jahrgangs getestet wurden.

Die Vergleiche der drei Schüler verdeutlichen, dass vor allem die leistungsbezogene Verortung von leistungsstarken und leistungsschwachen Schülerinnen und Schülern mit einem erheblichen Unsicherheitsfaktor behaftet ist, da hier mit den standardisierten Leistungstests durch die ungünstige Passung von Personenfähigkeit und Testschwierigkeit auf Individualebene zu wenig Information generiert wird.

Abbildung 1: Genauigkeit von Testergebnissen auf Individualebene

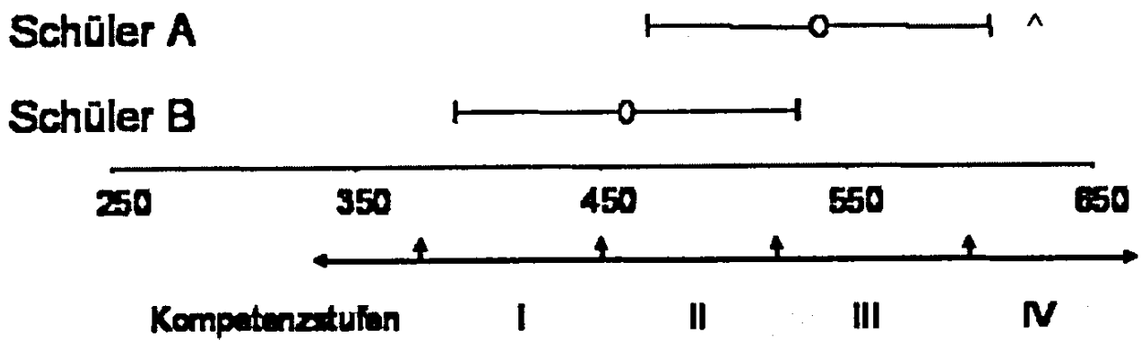


Abbildung 1 verdeutlicht darüber hinaus, dass die Verortung von einzelnen Schülerinnen und Schülern auch auf Kompetenzstufen nur ungenau möglich ist. Die Testleistung von Schüler A liegt bei rund 540 Punkten, die Leistung von Schüler B bei 460 Punkten. Diese Differenz entspricht einem Leistungsunterschied von knapp zwei Schuljahren. Die für die jeweiligen Schüler dargestellten Konfidenzintervalle verdeutlichen, dass diese Differenz von 80 Punkten auf Individualebene nicht annähernd ausreicht, um aus statistischer Sicht von einem abgesicherten Leistungsunterschied sprechen zu können.[2] Eine Verortung der jeweiligen Schülerleistung in den Kompetenzstufen macht zudem deutlich, dass Schüler A seiner Testleistung entsprechend in den Kompetenzstufen II bis IV und Schüler B in den Kompetenzstufen I bis III zu verorten ist.

Diese beiden Beispiele zeigen, dass eine punktegenaue Verortung von Schülerleistungen auf Individualebene mit einem so erheblichen Maß an Unsicherheit behaftet ist, dass sehr spezifische, auf den individuellen Lerner zielende Rückmeldungen an Schüler und Eltern methodisch und aus erziehungswissenschaftlicher Sicht *kaum* zu verantworten sind.

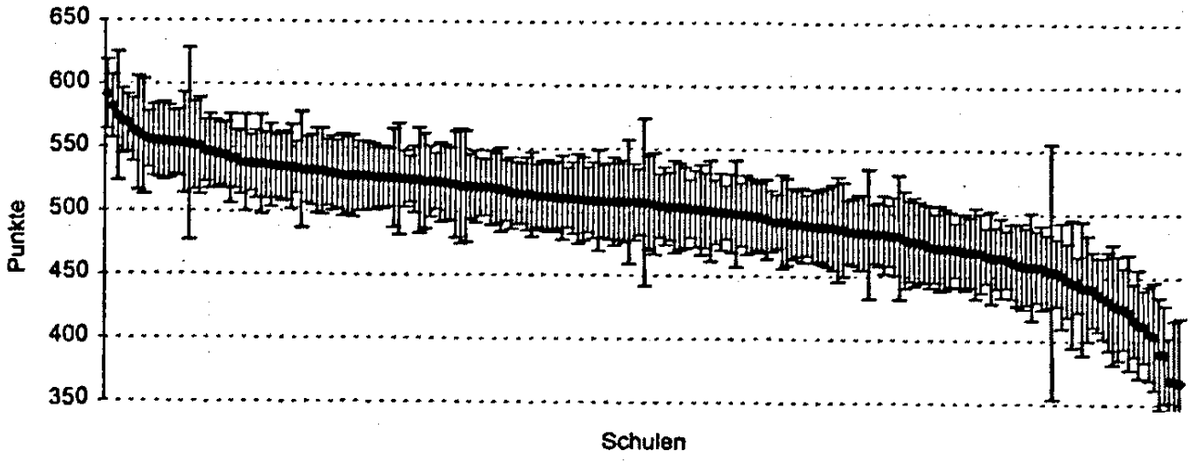
Rückmeldungen aus Leistungsstudien auf Klassen- und Schulebene sind – wie die Ergebnisse in Abbildung 2 und 3 verdeutlichen – ebenfalls mit Vorsicht zu interpretieren. Die Standardfehler für die 211 in IGLU getesteten Schulen liegen in einem Bereich zwischen neun und 50 Punkten. In Abbildung 2 sind die Mittelwerte und deren 95-prozentige Konfidenzintervalle für alle 211 Schulen dargestellt. Die vereinzelten Extremwerte in Abbildung 2 resultieren für Schulen, in denen nur sehr wenige und zudem Schüler mit heterogener Leistung am Test teilgenommen haben. Die Schulleistungen variieren im Bereich von 367 bis 591. Wie die Konfidenzintervalle zeigen, lassen sich die Schulen im mittleren Leistungsbereich aus statistischer Sicht nicht gegeneinander abgrenzen.

Aus statistischer Sicht lassen sich lediglich Schulen aus dem unteren und oberen Leistungsfünftel – also im Leistungsbereich größer als rd. 540 Punkte und kleiner als rd. 460 Punkte – zufallskritisch voneinander abgrenzen.

---

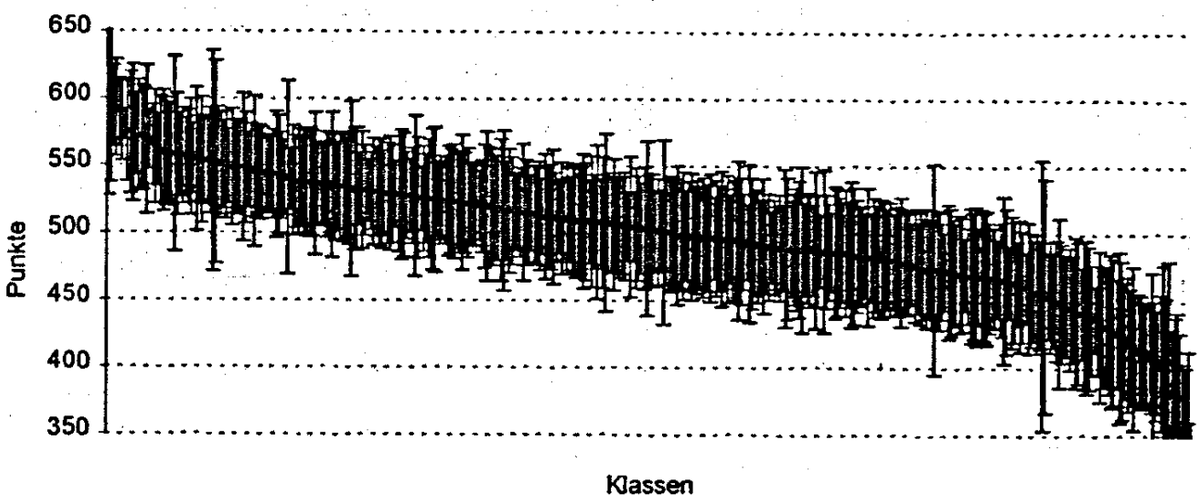
2 Der Standardschätzfehler für diese beiden Lesefähigkeitswerte liegt bei jeweils 39 Punkten. Die Breite der Konfidenzintervalle beträgt entsprechend 156 Punkte.

Abbildung 2: Standardfehler für die 211 in IGLU getesteten Schulen



In Abbildung 3 sind die entsprechenden Daten für die 393 Klassen, die an IGLU beteiligt waren, dargestellt. Die Breite der Konfidenzintervalle und damit das Ausmaß an Unsicherheit nehmen gegenüber den Schuldaten weiter zu. Wie bei den in Abbildung 2 dargestellten Schuldaten lässt sich maximal das obere Leistungszehntel vom unteren Leistungszehntel der Klassen – also Klassen im Leistungsbereich größer als rd. 560 Punkte und kleiner als rd. 440 Punkte – aus statistischer Sicht einwandfrei unterscheiden.

Abbildung 3: Standardfehler für die 393 in IGLU getesteten Klassen



Sicherlich gibt es zu bedenken, dass nicht unbedingt ein 95-prozentiges Konfidenzintervall zu Grunde gelegt werden muss – in individualdiagnostischen Verfahren werden häufig 90-prozentige Sicherheitsintervalle akzeptiert (vgl. Westhoff 1992; Tewes 2002). Darüber hinaus mag man berücksichtigen, dass der Standardfehler selbst in gewisser Weise normalverteilt ist – also die Wahrscheinlichkeit einer Abweichung nahe beim angegebenen Mittelwert höher liegt als eine größere Abweichung.

Fazit und Ausblick

Wie die dargestellten Ergebnisse auf Individual-, Klassen- und Schulebene verdeutlichen, lässt sich mit den Daten aus Leistungsvergleichsstudien eine punkt-

genaue Verortung von Testleistungen auf den zugrundeliegenden Skalen nur vornehmen, wenn man ein erhebliches Maß an Unsicherheit akzeptiert.

Auf Individualebene ist die Unsicherheit so hoch, dass die Ergebnisse der Vergleichsarbeiten in dieser Form für eine individuelle Diagnostik und daraus abzuleitende Fördermaßnahmen keinesfalls ausreichen.

Auf Klassenebene lässt sich zwar keine punktgenaue Verortung vornehmen, aber auf Basis der Kompetenzstufen können erste Hinweise auf die Lesekompetenz der Klasse abgeleitet werden. Diese Informationen lassen sich ergänzend zur Lehrerbeurteilung für die Unterrichtsplanung heranziehen.

Auf Schulebene ergibt sich zwar eine größere Genauigkeit durch die zusätzlichen Informationen aufgrund der Mehrzügigkeit. Vergleiche im Sinne eines Schulrankings auf Einzelschulebene sind dennoch gewagt. Wie die dargestellten Analysen gezeigt haben, lassen sich die schulbezogenen Leistungen in den Kompetenzstufen, die bei IGLU eine Breite von 75 Punkten haben, mit hinreichender Genauigkeit verorten. Auf dieser Grundlage ließen sich Schulen Leistungsgruppen zuordnen, die miteinander verglichen werden könnten. Bei der leistungsbezogenen Verortung von großen Schulen mit mehr als zweizügigen Jahrgängen ist zudem davon auszugehen, dass eine Verortung dieser Schulen ein höheres Maß an Genauigkeit aufweisen wird als die hier dargestellte Verortung für die getesteten IGLU-Schulen mit jeweils zwei Klassen pro Jahrgang.

Das Ziel, die Daten dieser Studien in ihrer jetzigen Form für einen empirisch basierten Schulentwicklungsprozess zu nutzen, muss hinterfragt werden, da vor allem für die Schulen im mittleren Leistungsbereich aus den empirischen Daten keine präzise Information über den tatsächlichen Leistungsstand zu entnehmen ist. Es ist vielmehr wahrscheinlich, dass Schulen bei einer wiederholten Testung, allein durch zufällig wirkende Einflüsse bedingt, eine unterschiedliche Verortung erfahren können.

Um Ergebnisse aus den Vergleichsstudien dennoch für die empirische Schulentwicklung zu nutzen, müssen andere Verfahren gefunden werden. Allein eine Erhöhung der zufallskritischen Absicherung der Ergebnisse von fünf auf zehn Prozent erschiene als ‚rechnerischer Trick', der sich zwar aus statistischer Sicht argumentieren ließe, das Problem aber eher verdeckt als es bearbeitet. Aus unserer Sicht bedarf es daher weiterer, die Validität stützender Verfahren. Hier sehen wir bspw. die folgenden Möglichkeiten:

Zur Präzision von Vergleichen auf Klassen- und Schulebene könnten – wie in den internationalen Schulvergleichsstudien – Schülerfragebogen erhoben und die erfassten Informationen in entsprechenden Hintergrundmodellen ergänzend zur Skalierung der leistungsbezogenen Daten und damit zur Genauigkeit der Messung optimierend genutzt werden.

Alternativ könnte durch ein zyklisches Datenerhebungsdesign, in dem in den einzelnen Schuljahren jeweils ein Schwerpunkt für die Leistungsmessung gesetzt wird (z. B. Deutsch, Mathematik, Englisch), die Informationsgrundlage in Form einer erweiterten Testlänge erhöht werden. So werden z.B. für den Cito-

Abschlusstest in Holland 260 Aufgaben eingesetzt, darunter 100 Aufgaben für den Teilbereich Sprache (BMBF 2007, S. 225-226). Mit diesem Ausmaß an Informationen lassen sich auch Aussagen und Konsequenzen auf Individualebene ableiten. Ein Testintervall von drei Jahren je Fach entspricht darüber hinaus auch eher den Erfahrungen der Schulentwicklungsforschung, wonach die Erwartung, durch einen datengestützten Input in Ein- bzw. Zweijahresintervallen einen nachhaltigen Entwicklungsprozess in den Schulen initiieren zu können, unrealistisch erscheint. Zusätzlich würden sowohl die Schulen als auch die Testentwickler durch ein Testintervall von drei Jahren je Fach erheblich entlastet. Zur Präzisierung von Individual- und Klassenrückmeldungen müssten also umfangreichere Tests verwendet werden.

Eine weitere Möglichkeit besteht darin, die erhobenen Informationen durch ein sogenanntes Computer-Adaptive-Testing-Verfahren (tailored testing) zu verbessern (Cito 2002; Embretson/Reise 2000). Durch den Einsatz von computerbasierten Testverfahren kann die Passung von Personenfähigkeit und Aufgabenschwierigkeit in einer Testsituation individuell gestaltet werden.

Ein Gewinn an Genauigkeit lässt sich auch durch eine geeignete Auswahl des Schätzverfahrens bei der Parametrisierung des Rasch-Modells erreichen. Klassisch stehen dabei drei Ansätze zur Auswahl. Beim "unconditioned maximum likelihood"-Ansatz (UML) werden aus dem Datensatz Item- und Personenparameter geschätzt. Durch die große Anzahl zu schätzender Parameter sind die Schätzungen sehr ungenau. Dies zeigt sich durch breite Konfidenzintervalle. Beim "marginal maximum likelihood"-Ansatz (MML) werden neben den Itemparametern nur die Parameter der Populationsverteilung geschätzt. Im eindimensionalen Fall ist diese bei Normalverteilungsannahme durch die zwei statistischen Konzepte Erwartungswert und Varianz vollständig beschrieben und kann sehr effizient aus den Daten geschätzt werden. Der MML- ist daher dem UML-Ansatz für Rückmeldungszwecke vorzuziehen. In der praktischen Umsetzung ist dieser Ansatz jedoch sehr arbeitsintensiv, da bei dieser Schätzvariante für jede Schule beziehungsweise Klasse ein eigenständiger Skalierungslauf durchzuführen ist (vgl. Voss/Strietholt, in Vorb.; Strietholt/Voss, in Vorb.).

Wie die langjährigen Erfahrungen aus Ländern mit einem outputorientierten Schulsystem, wie z.B. Holland und Schweden, zeigen, können regelmäßig durchgeführte Schulleistungsmessungen mittelfristig zu einer messbaren Verbesserung der Schülerleistungen führen. Damit jedoch das Ziel erreicht wird, eine gerechtere Verteilung von Bildungschancen und eine optimale Kompetenzentwicklung bei den Schülern auf Grundlage dieser empirischen Daten zu erzielen, muss sichergestellt sein, dass Lehrerinnen und Lehrer, die mit diesen Daten arbeiten sollen, die Reichweite und Belastbarkeit dieser Daten rational einschätzen können.

Literatur

BMBF (Hrsg.) (2007): Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintergrund. Schriftenreihe Bildungsreform, Band 11. Bonn/Berlin: BMBF.

Bos, W./Lankes, E.-M./Prenzel, M./Schwippert, K./Valtin, R./Walther, G. (Hrsg.) (2004): IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. Münster: Waxmann.
Cito (2002): Balans van het taalonderwijs aan het einde van de basisschool 3. Arnhem: Cito.
Eikenbusch, G. (1995): Tendenzen und Entwicklungen der Schulentwicklung in Schweden. In: Landesinstitut für Schule und Weiterbildung (Hrsg.): Schulentwicklung und Qualitätssicherung in Schweden. Entwicklungen – Erfahrungen – Materialien. Reihe Lehrerfortbildung in Nordrhein-Westfalen. Bönen/Westfalen: Druckverlag Kettler, S. 7-40.
Embretson, S.E./Reise, S.P. (2000): Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.
Lankes, E.-M./Bos, W./Mohr, I./Plaßmeier, N./Schwippert, K./Sibberns, H./Voss, A. (2003): Anlage und Durchführung der Internationalen Grundschul-Lese-Untersuchung (IGLU) und ihre Erweiterung um Mathematik und Naturwissenschaften. In: Bos, W./Lankes, E.-M./Prenzel, M./Schwippert, K./Walther, G./Valtin, R. (Hrsg.): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster: Waxmann.
MSJK (2005a): VERA-Schulmail 8 vom 07.06.2005. URL: http://www.learn-line.nrw.de/angebote/vergleichsarbeiten4/download/VERA_Schulmail8.pdf, Zugriffsdatum: 17.04.2007.
MSJK (2005b): Zentrale Lernstandserhebungen: Materialien zur Unterstützung der Schulaufsicht beim Umgang mit Lernstandserhebungen vom 15. Juni 2005. URL: http://www.learn-line.nrw.de/angebote/lernstand9/download/schulaufsicht_ lernstand.pdf, Zugriffsdatum: 17.04.2007.
MSW (2006): Zentrale Lernstandserhebungen (Vergleichsarbeiten). Runderlass des Ministeriums für Schule und Weiterbildung. Düsseldorf.
Rost, J. (1996): Lehrbuch Testtheorie. – Testkonstruktion. Bern: Huber.
Strietholt, R./Voss, A. (in Vorb.): Schulische Leistungsmessung und individuelle Förderung. In: Praxis Deutsch.
Tewes, U. (Hrsg.) (^{3}2002): HAWIK-III: Hamburg-Wechsler Intelligenztest für Kinder. Bern: Huber.
Voss, A./Strietholt, R. (in Vorb.): Was können Lehrerinnen und Lehrer aus standardisierten Ergebnisrückmeldungen lernen?
Westhoff, K. (Hrsg.) (1992): Entscheidungsorientierte Diagnostik. Bonn: Dt. Psychologen Verlag.

Wilfried Bos, geb. 1953, Dr. phil., Professor für Bildungsforschung und Qualitätssicherung an der Technischen Universität Dortmund. Direktor des Instituts für Schulentwicklungsforschung
Anschrift: Vogelpothsweg 78, 44227 Dortmund
E-Mail: officebos@ifs.uni-dortmund.de

Andreas Voss, geb. 1969, Dr. phil., Wissenschaftlicher Mitarbeiter am Institut für Schulentwicklungsforschung (IFS) an der Technischen Universität Dortmund
Anschrift: Vogelpothsweg 78, 44227 Dortmund
E-Mail: voss@ifs.uni-dortmund.de