Ms. für Grundschule-aktuell 1/2005 (Februar) des Grundschulverbands

Hans Brügelmann

Wahrheit durch VERA?

Anmerkungen zum ersten Durchgang der landesweiten Leistungstests in sieben Bundesländern[1]

Die in am Anfang der vierten Klasse in Rheinland-Pfalz und sechs weiteren Bundesländern im September dieses Jahres durchgeführte Lernstandserhebung VERA (**VER**gleichs**A**rbeiten...[2]) hat unter LehrerInnen viel Aufregung verursacht. Meine Reaktion ist ambivalent: Als Forscher habe ich großen Respekt davor, was die VERA-KollegInnen unter den gegebenen Rahmenbedingungen geleistetet haben. Gleichzeitig frage ich mich, ob man sich auf die gegebenen politischen und organisatorischen Rahmenbedingungen hätte einlassen sollen. Funktion, Anlage und Durchführung der Studie haben darunter gelitten.

Die Erfahrungen aus diesem Durchgang sollten deshalb als Chance genutzt werden, um zu lernen. Das gilt für diejenigen, die die Tests entwickeln bzw. ihren Einsatz verordnen und durchführen, aber auch für die LehrerInnen, die kritisch prüfen müssen, was sie mit den Ergebnissen anfangen können – *und was nicht.*

Was *sollen* und was *können* zentrale Testprogramme leisten?[3]

Drei Funktionen von VERA sind in der Außendarstellung von verschiedenen Beteiligten unterschiedlich stark betont worden und faktisch auch unterschiedlich gut einlösbar:

- Bei der Bestandsaufnahme von grundlegenden Leistungen auf Landesebene („System-Monitoring") hat das deutsche Schulsystem tatsächlich einen Nachholbedarf. Diese Funktion können standardisierte Testprogramme und konkret auch VERA von ihrer Anlage her gut erfüllen (zu einigen spezifischen Vorbehalten s. u.).

- Auch die Rückmeldung der Ergebnisse an einzelne Schulen und LehrerInnen ist hilfreich. Zum einen können LehrerInnen genauer sehen, in welchen Bereichen ihre Klassen

---

[1] Ich danke verschiedenen KollegInnen (die z.T. ausdrücklich lieber ungenannt bleiben wollen....) für hilfreiche Anmerkungen zu einer Vorfassung dieses Papiers

[2] Informationen des Projekts unter: http://www.uni-landau.de/vera/

[3] Vgl. dazu ausführlicher meine Beiträge in Bartnitzky/ Speck-Hamdan (2004, i.D.) und in GSV-aktuell Nr. 79, 82, 83, sowie in Brügelmann (i. D., Kap. 47-51).

relativ zu anderen Lerngruppen Stärken und Schwächen haben – bedingt durch besondere Vorerfahrungen der Kinder, durch eigene Schwerpunkte im Unterricht, durch die Anlage der verwendeten Schulbücher oder andere Faktoren. Deren Bedeutung ist allerdings nur vor Ort von den Beteiligten selbst zu klären. Außerdem können LehrerInnen im Vergleich mit Klassen, die unter ähnlichen Bedingungen arbeiten („Referenzschulen"), ihre Anforderungen an die Kinder und ihre Maßstäbe bei der Leistungsbeurteilung überprüfen. Dies darf aber nicht einfach Anpassung an die externen Kriterien bedeuten, sondern verlangt eine Reflexion der externen *und* der eigenen Annahmen.

- Der Anspruch einer „Diagnose" des Lernstands einzelner Kinder allerdings überfordert die Instrumente. Eine punktuelle Messung muss inhaltlich auf wenige Ausschnitte begrenzt werden. Zudem ist sie immer fehlerbehaftet, d. h. der festgestellte Wert kann nur als Anhaltspunkt für eine Bandbreite, innerhalb derer der „wahre" Wert mit einiger Sicherheit liegt, genommen werden. Je mehr man das Fehlerrisiko minimieren will, umso breiter muss man die Bandbreite möglicher Schwankungen ansetzen – und umso weniger hilfreich ist dann das Ergebnis. In Kennwerten für Gruppen, also in Werten für ganze Klassen oder gar ein Bundesland, gleichen sich individuelle Schwankungen weitgehend aus, so dass das Fehlerrisiko der entsprechenden Durchschnittswerte entsprechend gering ist. Einmalige Tests bei einzelnen Personen bieten dagegen nur grobe Annäherungen an die tatsächliche Leistung. Sinnvoll nutzen lassen sich die Ergebnisse trotzdem, wenn man Abweichungen zur eigenen Einschätzung als „Warnlampe" nutzt, also als Anlass, um die Differenzen durch weitergehende Beobachtungen aufzuklären.

Für alle drei Ebenen gilt gleichermaßen: Die Ergebnisse sind als *ein* Element in einem umfassenderen Rechenschaftssystem zu sehen[4], als ein wichtiges und bisher unterrepräsentiertes Evaluationsinstrument, aber auch als ein in seiner Aussagekraft und Geltung begrenztes. Das größte Problem in der seit TIMSS öffentlich geführten Schuldebatte ist, dass sie auf den Vergleich von Punktwerten und die Überhöhung der Testautorität schrumpft. PolitikerInnen und Medien schauen nur noch auf den Output. Berichte der Schulaufsicht, Forschungsergebnisse zu Lehr-/ Lern-Prozessen und ihren Bedingungen verlieren an Bedeutung. Auch viele LehrerInnen nehmen die Vergleichsdaten nicht als nützliche Zusatzinformation über Stärken und Schwächen ihrer Klasse, sondern trauen oft ihrem eigenen Urteil nicht mehr, obwohl es aus einer längerfristigen Erfahrung erwächst. Testergebnisse einzelner Kinder werden für Eltern (und oft auch für LehrerInnen) zum „wahren" Wert für ihre Leistung, statt zu einem Indikator, der interpretationsbedürftig ist.

[4] Vgl. Bartnitzky u. a. 1999

Konkrete Anmerkungen zu den Aufgaben und Verfahren von VERA

Analysiert man die Test- und die Begleitunterlagen, unterhält man sich mit KollegInnen aus der Grundschulpädagogik bzw. der Fachdidaktik und befragt man LehrerInnen aus den beteiligten Bundesländern nach ihren ersten Erfahrungen mit VERA, so stößt man auf sehr unterschiedliche Reaktionen. Manche KollegInnen, die VERA im oben genannten Sinn *als ein Element im Rahmen verschiedener Informationsquellen* verstehen, finden die Tests nützlich. Auch viele Kinder haben die Aufgabe, „zu zeigen, was ihr könnt", als Herausforderung positiv angenommen. Andere, vor allem leistungsschwächere SchülerInnen und Kinder mit Migrationshintergrund, erlebten viele Aufgaben und den Umfang insgesamt als völlige Überforderung. Darüber hinaus gibt es eine Reihe von Problemen, die bei der Weiterentwicklung des Instrumentariums und seinem zukünftigen Einsatz bedacht werden müssen:

Zum inhaltlichen Ansatz

Einige Aufgaben bieten interessante Anregungen für Leistungskontrollen, die LehrerInnen nutzen sollten, um ihr eigenes Repertoire zu erweitern. Die Formate bringen aber – wegen der notwendigen Standardisierung von Durchführung und Auswertung – unvermeidlich auch Einschränkungen mit sich[5]:

- Aufgaben ohne Kontextbezug, wie er z. B. in den Lehrplänen für Deutsch gefordert wird, sind für viele SchülerInnen überraschend. Sie verändern auch das Lösungsverhalten, z. B. wenn in 20 min. ein Text zu einem Thema geschrieben werden soll, obwohl den SchülerInnen beigebracht worden ist, dass das Schreiben guter Texte ein Brainstorming, eine Textplanung, mehrere Zyklen der Überarbeitung (z. B. in Schreibkonferenzen) und insgesamt eine rege Kommunikation mit anderen voraussetzt. Dagegen untersagen die Instruktionen bei VERA ausdrücklich Fragen an die Lehrerin und Gespräche der Kinder untereinander.
- Die Richtigkeits-Orientierung, wie sie für eine standardisierte Auswertung erforderlich ist, gerät in Konflikt mit der Mehrdeutigkeit von Verhalten, insbesondere von sprachlichen Vorlagen einerseits und Lösungen andererseits, und sie fördert bei SchülerInnen eine Haltung, die nach gewünschter Lösung sucht, statt dem eigenem Denken zu trauen.
- Werden Hilfsmittel, deren Gebrauch die SchülerInnen nicht nur gewohnt sind, sondern deren Beherrschung auch explizit als Lernziel von den Lehrplänen eingefordert wird (z. B.

[5] Vgl. zur inhaltlichen Kritik die konkreten Anmerkungen von Bartnitzky und Selter „Grudnshcule aktuell" H. 89/2005 sowie von Metzger u. a. (2004, i.D.).

Wörterbücher zur Kontrolle der Richtigschreibung von Wörtern) vorenthalten, entsteht eine Konkurrenz zu den Prinzipien des Unterrichts und den Erfahrungen der SchülerInnen.

- Auch das dicht konzentrierte Abarbeiten von unzusammenhängenden Aufgaben unter Zeitdruck (z. B. in Form von Diktaten) steht im Widerspruch zu Leistungssituationen, wie sie neuere Lehrpläne fordern und wie die Kinder in vielen Klassen sie gewohnt sind. Das ganze „Setting" wird von den Grundschulen, die nach den Richtlinien und Lehrplänen der letzten Jahre arbeiten, als Fremdkörper empfunden. Und diejenigen, die gerade die ersten Schritte machen, werden eher entmutigt – oder sogar in ihren alten Vorstellungen bestätigt.
- Die unterstellte Eindimensionalität der Fähigkeitsniveaus innerhalb z. B. von Arithmetik, Geometrie und Sachrechnen/ Größen in Mathematik, wird dem nicht gerecht, was zunehmend über „eigene Wege" des mathematischen und schriftsprachlichen Lernens bekannt ist[6].
- Über die Angemessenheit einzelner Aufgaben, z. B. der Deutung von wenig gängigen Sprichwörtern im Deutschtest, wird nach der Auswertung zu diskutieren sein.

Diese kritischen Anmerkungen stellen nicht in Frage, dass die Ergebnisse als wichtige Indikatoren für die angepeilten Leistungen dienen können. Aber die gewonnenen Daten müssen entsprechend *interpretiert* werden[7], Punktwerte dürfen nicht *at face value* zu Urteilen über Kinder oder LehrerInnen werden.

Damit sind wir bei der Durchführung von VERA und ihren Folgen:

Zur Wirkung der Erhebungssituation

Die Schulen sind sehr unterschiedlich mit den Vorgaben umgegangen, so dass sowohl das Durchschnittsniveau als auch die Vergleichbarkeit der Ergebnisse einzelner Klassen ernsthaft in Frage zu stellen sind. Zudem deuten sich schon jetzt ambivalente bis problematische Auswirkungen auf den Unterricht an.[8]

---

[6] Vgl. anschaulich, auch für Eltern, die Beiträge in Brügelmann (1998).

[7] Die Landauer Forschungsgruppe hat dazu unter dem Titel „Pädagogische Nutzung der Vergleichsarbeiten" und „Handreichung zur Analyse der Falschlösungen" wichtige Hinweise gegeben, deren Berücksichtigung helfen könnte, das vielerorts beklagenswerte Niveau der Feststellung, Interpretation und Bewertung von Schülerleistungen anzuheben. Zu befürchten ist auf der anderen Seite, dass viele LehrerInnen und Eltern sich mit Summenwerten begnügen und ihre Urteile auf oberflächliche Vergleiche stützen werden.

[8] Das ging bis zu Diskussionen, ob die Kinder während der vorgesehenen 10-minütigen Pause während der Deutscharbeit die Klasse verlassen und miteinander sprechen dürften – oder ob die Pause schweigend im Klassenraum zu verbringen sei, unterbrochen allenfalls vom Gang zur Toilette! Wenn die Klausurlogik sich entfaltet ...

- In einer Reihe von Schulen wurden Aufgaben(typen) geübt[9], so dass die Ergebnisse über Klassen hinweg nicht vergleichbar sind.
- Manche LehrerInnen haben ihren Klassen oder einzelnen Kindern entgegen den ausdrücklichen Anweisungen geholfen, so dass die empirisch gewonnenen Daten nur schwer als „Referenz"-Daten zu etablieren sind.
- Zudem wurde bei der Bewertung von Lösungen unterschiedlich „kulant" mit Lösungen umgegangen – zum Schutz einzelner SchülerInnen, aber auch im Interesse des eigenen Unterrichtserfolgs...
- Testformate beginnen darüber hinaus, die Aufgabenformate im Unterricht zu bestimmen, obwohl Lernsituationen anderen Prinzipien gehorchen als Leistungskontrollen.
- Es zeichnet sich eine Einengung des Unterrichts auf die in den Tests geforderten Inhalte und Kompetenzen ab, z. B. auf technische Informations"entnahme" aus Textstücken statt persönlicher Auseinandersetzung mit der inhaltlichen „Botschaft".
- Es ist absehbar, dass Eltern und die lokale Presse Informationen aus der verpflichtenden schulinternen Diskussion für ein informelles äußeres Ranking nutzen werden, ohne dass die notwendigen Einschränkungen mit bedacht werden (können).
- Der Zeitpunkt der Erhebung und die Dreistufigkeit des Kompetenzmodells legen eine Zuordnung zu den Schularten der Sekundarstufe nahe und werden vor allem von Eltern oft so missverstanden. Die ministeriell immer wieder beschworene individuelle Förderung ist im letzten Halbjahr der Grundschulzeit kaum mehr möglich

Im Vergleich zu den in NRW in den Vorjahren eingeführten Parallelarbeiten mindert der weniger enge Lehrplanbezug den Wert der Ergebnisse, ebenso die nicht mehr notwendige Zusammenarbeit von KollegInnen den Ertrag für die innerschulischen Entwicklungsprozesse.

Zu den organisatorischen Rahmenbedingungen:

Auch wenn die logistische Leistung der Landauer Forschergruppe und der Projektleitungen in den Bundesländern bewundernswert ist: Viele Schulen klagen über den Zeitdruck, die unzureichende Vorbereitung und den hohen personellen und finanziellen Aufwand bei der Durchführung. Als abträglich für die alltägliche Arbeit unter sowieso schon schwierigen Bedingungen werden insbesondere genannt:

- hoher Materialaufwand für das Kopieren bei begrenztem Etat (in defr Sekundarstufe wurden die Materialien zentral vervielfältigt...);

[9] ... in mindestens einer mit bekannten Klasse sogar zum Üben mit nach Hause gegeben!

- hoher Zeitaufwand, einsetzbare Testhefte zusammen zu stellen, und hoher Korrekturaufwand[10] – beides zieht Zeit von anderen Aktivitäten ab;
- sehr kurzfristige Anforderung der Ergebnisse aus den Referenzschulen und damit immenser Zeitdruck bei der Auswertung (Herbstferien als Korrekturzeit...);
- die Fehleranfälligkeit bzw. unzulängliche Passung der komplizierten Auswertungsvorgaben und des Computersystems;
- die nicht immer abgestimmten Informationen von verschiedenen Quellen und an verschiedene Zielgruppen, so dass Unsicherheit und zum Teil auch Argwohn entstanden ist;
- die als nicht repräsentativ wahrgenommene Auswahl der Beispielaufgaben, die den Schulen im Frühjahr zur Vorinformation zugeschickt worden waren;
- der geringe Ertrag für die Weiterentwicklung des Unterrichts und die Förderung einzelner Kinder.

Die von der Politik erzeugte Hektik hat sowohl der Testzentrale in Landau als auch anderen Beteiligten die Arbeit erschwert und manchen Goodwill in den Schulen verschenkt. Insbesondere das Versprechen, es werde kein Ranking geben, wird durch die Verpflichtung, allen Eltern nicht nur die Ergebnisse ihres Kindes, sondern ebenso die Durchschnittswerte seiner Klasse und der Schule mitzuteilen, faktisch wertlos. Wer in Ländern wie England und den USA beobachtet hat, welche Konsequenzen die Publikation von globalen Testdaten einzelner Schulen z. B. auf die Immobilienpreise von Stadtteilen hat[11], wird sich keine Illusionen machen, was den verständigten Umgang mit solchen Daten in der Öffentlichkeit betrifft. Zu hoffen ist, dass aus diesen Schwierigkeiten für weitere Erhebungen gelernt wird

Fazit:

- Für ein regelmäßiges *System-Monitoring*, bei dem die Entwicklung des Schulsystems insgesamt erfasst werden soll, würde es reichen, alle vier bis sechs Jahre Erhebungen durchzuführen. Außerdem könnte man sich (wie bei PISA und IGLU) auf repräsentative Stichproben beschränken und anderen Schulen eine freiwillige Teilnahme (wie bei LUST[12]) ermöglichen, was die Belastungen solcher Testprogramme mindern und ihre Akzeptanz beträchtlich erhöhen dürfte. Zugleich würden damit Mittel frei für andere Evaluationsaktivitäten (s. u.).

[10] Mindestwert ca. ein Halbtag pro Fach und Klasse bis hin zu einem Tag schriftliche Auswertung in den Ferien zu Hause und einem weiteren weiterer Tag Eingabe am PC für die zentrale Auswertung in Landau.

[11] S. aktuell http://www.mercurynews.com/mld/mercurynews/living/education/10045224.htm?1c [Abruf: 4.11.2004]

[12] Vgl. meinen Beitrag in GSV-aktuell Nr. 84

- Um *LehrerInnen* hilfreiche Informationen zur Kalibrierung ihrer Maßstäbe und für den Vergleich der eigenen mit anderen Klassen zu geben, wäre bei den nächsten Terminen ein Wechsel auf andere Kompetenzbereiche von Deutsch bzw. Mathematik und auch auf andere Lernbereiche wie Sachunterricht und die musisch-ästhetischen Fächer sinnvoll. Dabei sollten generell auch weniger standardisierte Formate erprobt werden.

- Für die Individualbeobachtung müssten Instrumente zur *Lernbegleitung* entwickelt werden, um differenziertere Einschätzungen anzuregen und zu unterstützen, als sie durch eine punktuelle Messung möglich sind.

Der letzte Punkt hat aus meiner Sicht in der nahen Zukunft Priorität, soll das Evaluationssystem nicht Schlagseite bekommen. Der Grundschulverband hat eine Arbeitsgruppe eingesetzt, die entsprechende Hilfen für Sprache, Mathematik und Sachunterricht entwickeln und in Form eines Fortbildungspakets publizieren soll. Es ist zu hoffen, dass auch staatliche Institutionen zumindest einen Teil ihrer Ressourcen in solche Aktivitäten investieren.

Literatur

Bartnitzky, H./ Speck-Hamdan, A. (Hrsg.) (2004, i.D.): Pädagogische Leistungskultur: Leistungen der Kinder wahrnehmen – würdigen – fördern. Beiträge zur Reform der Grundschule, Bd. 118. Grundschulverband: Frankfurt.

Bartnitzky, H., u. a. (1999): Zur Qualität der Leistung – 5 Thesen zu Evaluation und Rechenschaft der Grundschularbeit. Grundschulverband – Arbeitskreis Grundschule e. V.: Frankfurt. Auch in: Schmitt, R. (Hrsg.) (1999): An der Schwelle zum dritten Jahrtausend. BundesGrundschulKongress 1999. Grundschulverband – Arbeitskreis Grundschule: Frankfurt, 165-196.

Brügelmann, H. (Hrsg.) (1998): Kinder lernen anders: vor der Schule – in der Schule. Libelle: CH-Lengwil.

Brügelmann, H. (i. D.): Schule verstehen – Forschungsbefunde zu Kontroversen über Erziehung und Unterricht. Libelle: CH-Lengwil (erscheint im Sommer 2005).

Metzger, K., u. a. (2004): Sprachbezogene Leistungen würdigen. Die aktuelle Leistungsdiskussion und erfolgreicher Deutschunterricht. Ms. für: Bartnitzky/ Speck-Hamdan (2004, i.D.).