Zentrale Lernstandsmessung in der Primarstufe – Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern

Jens Holger Lorenz, Heidelberg (Germany)

Abstract: The development of the central tests in mathematics which in the meantime are regularly conducted in seven Federal States (Bundeslaender) is outlined. The problems in the acceptance and the development of test items are described. The aims of VERA, specifically the system monitoring and the aid in the schools' selfevaluation with regard to the "standards" are given. The process of developing VERA as a tool in cooperation between school practice and science must be considered as successful.

Kurzreferat: Die Entstehung der inzwischen länderübergreifenden und flächendeckenden Vergleichsarbeiten (VERA) wird beschrieben und die dabei auftretenden Schwierigkeiten der Durchführung und Akzeptanz in der Schulwirklichkeit. Die Ziele des „Systemmonitoring" und der Hilfen für die Selbstevaluation im Sinne der Qualitätsentwicklung entlang der vereinbarten Standards werden entfaltet. Die Geschichte von VERA muss summarisch als durchaus erfolgreich eingeschätzt werden, wobei es sich um einen langwierigen Prozess handelt, in dem es zu einer engen Kooperation zwischen Praxis und Wissenschaft kommt.

ZDM-Classification: B12, C72, D62, D72

0. Vorbemerkung

Im Jahre 2003 wurden in Rheinland-Pfalz in den vierten Klassen flächendeckend die ersten Vergleichsarbeiten geschrieben. In den Fächern Mathematik und Deutsch waren zuvor an einem Aufgabenpool Normierungen vorgenommen worden, so dass direkt nach den Vergleichsarbeiten Rückmeldungen an die Lehrerinnen und Lehrer gegeben werden konnten. Rheinland-Pfalz betätigte sich insofern als Vorreiter, als die Erstellung der Aufgabensätze einer Expertengruppe und die Durchführung und Auswertung zentral einer Hochschule oblag. Dass die Absicht der Bildungsbehörde nicht ungeteilte Zustimmung erfuhr und immer noch erfährt sondern mit Misstrauen aufgenommen wurde, deutet auf die hohe emotionale Betroffenheit der Beteiligten hin. Inzwischen nehmen, durch entsprechende Weiterbildungsmaßnahmen flankiert, die Vergleichsarbeiten allerdings einen festen Platz zu Schuljahresbeginn ein.

1. Eine kurze Geschichte von VERA

Im Anschluss an die bundesweite Bildungsdebatte nach TIMSS und PISA sah die Koalitionsvereinbarung vom 25.04.02 im Land Rheinland-Pfalz zwischen der SPD und der FDP *Vergleichsarbeiten* in der 4. Klassenstufe der Grundschule in den Fächern Deutsch und Mathematik vor. Es handelte sich hierbei keineswegs umein neues, sondern ein durchaus auch in anderen, vor allem südlichen Bundesländern praktiziertes Verfahren. Die beiden Zielrichtungen waren klar und eigentlich konsensfähig:

– *Systemmonitoring*, d.h. Überprüfung der Gesamtentwicklung des Bildungsniveaus in den Grundschulen des Landes im Sinne des (damaligen) neuen Bildungsplanes und der Standards
– *Hilfen bei der Selbstevaluation der Schulen*, um daraus sich ableitenden Entwicklungsbedarf zu erkennen (Schul- und Unterrichtsentwicklung)

Es muss notwendig Misstrauen entstehen, wenn unvermittelt eine „Kontrolle" eigener Tätigkeit durchgeführt wird. Insbesondere handelte (und handelt) es sich um eine neue Maßnahme für Grundschullehrer, wobei es zu anfänglichen Reibungen in der Begründung kam. Aber es handelte sich um eine flankierende Maßnahme zur Realisierung der Bildungspläne, die mit bereitgestellten Multiplikatoren, Fachberatern und Ausweitung in der Lehrerfortbildung Früchte tragen sollte.

In der Schulpolitik und Bildungsforschung gab es bislang keinen feststehenden Begriff „Vergleichsarbeiten" mit genau umrissener Bedeutung. Und da es für Rheinland-Pfalz zu betretendes Neuland war, musste auch auf konzeptioneller Ebene Vorarbeit geleistet werden. Die Universität Landau, genauer: das Institut für Psychologie führt, unter der Verantwortung von Prof. Helmke und Prof. Hosenfeld, vertragsgemäß sämtliche theoretischen und empirischen Arbeiten durch und unterstützt die Schulen und die Bildungsverwaltung in der Entwicklung, Durchführung und Auswertung dieses Vorhabens. Hierzu werden regelmäßig von einer Expertengruppe, die sich inzwischen aus Vertretern aller beteiligten Bundesländer zusammensetzt, Aufgaben entwickelt, die in einer Stichprobe von Schulklassen normiert werden. Diejenigen Aufgaben, die die strengen Itemanforderungen erfüllen, werden in den Pool der Zentralaufgaben aufgenommen.

Vergleichsarbeiten (VAen) lassen sich wie folgt charakterisieren: Es handelt sich um schriftliche Arbeiten, die in einer größeren Anzahl von Schulen (ggf. landesweit) auf der Basis einer vorgegebenen Aufgabenstichprobe eingesetzt werden mit dem Ziel, die Leistungen der Schüler an einem über die einzelne Klasse und Schule hinausgehenden Vergleichsmaßstab, d.h. an einer klassen- und schulübergreifenden sozialen und/oder kriterialen Bezugsnorm zu messen. Die Bezugsnorm ist insofern „kriterial", als die Aufgaben an den Standards und den Bildungsplänen der jeweiligen Bundesländer als Kriterium (im Sinne der zu entwickelnden Fähigkeiten) orientiert sind und diese abzubilden versuchen. Sie sind „sozial", als die Rückmeldungen an die Mathematiklehrer den Vergleich ihrer Klasse mit (a) der gesamten Population, (b) den Schülern des eigenen Bundeslandes und (c) mit den Schülern bzw. Klassen mit ähnlichem sozialen Hintergrund umfasst.

Sie ähneln insofern *Parallelarbeiten*, als es vermutlich immer eine bestimmte Anzahl von Klassen bzw. Schulen gibt, die den identischen Aufgabensatz simultan einsetzen. Sie gehen in ihrem Vergleichsanspruch jedoch in dem Maße über *Parallelarbeiten* hinaus, in dem klassen- und schulübergreifende Normen ins Spiel kommen.

Sie ähneln auch *standardisierten Leistungstests*, als die Aufgaben von Experten entwickelt bzw. ausgewählt und

im Hinblick auf Testgütekriterien geprüft werden. Als „Experten“ entsenden die beteiligten Bundesländer Vertreter aus der Lehrerfortbildung, den Hochschulen und an der Entwicklung der Bildungspläne Beteiligte.

Allerdings sind die Anforderungen an die Testgütekriterien bei Vergleichsarbeiten geringer als bei standardisierten Tests. Dies wird etwa in Ausnahmefällen dadurch bedingt, dass die Objektivität leidet, wenn die auswertenden Lehrkräfte eine Kontrolle befürchten und deshalb die Ergebnisse „schönen“).Anders als Klassenarbeiten, die sich in der Regel auf einen bestimmten, zuvor durchgenommenen Unterrichtsstoff beziehen, umfassen Vergleichsarbeiten den Stoff des gesamten Schuljahres oder auch entsprechende Vorkenntnisse. Von anderen Formen von *Lernstandserhebungen* unterscheiden sich Vergleichsarbeiten dadurch, dass keine Aussage über die Leistungen einer ganzen Region (z.B. eines Bundeslandes) beabsichtigt ist. (Dass VAen später regional und vereinzelt von unteren Bildungsbehörden so missbraucht werden wollten, ist bedauerlich und verweist auf die Gefahr, dass jede Messung dieser Art gegen ihre Absicht gewendet werden kann. Die Kontrolle der „Kontrolleure“ kann aber nur auf der politischen Schiene erfolgen, nicht innerhalb von VERA.)

Im Jahre 2004 hat sich die KMK u. a. auf die Bildungsstandards für den Primarbereich (Jahrgangstufe) geeinigt. Auch wenn dies ein zähes Ringen um Formulierungen war, so war doch bereits vorher deutlich, dass die meisten Bundesländer diese in der ein oder anderen Form bereits in ihre neuen Bildungspläne aufgenommen hatten. Das Projekt VERA als Systemmonitoring in Rheinland-Pfalz rief entsprechende Begehrlichkeiten bei anderen Bundesländern hervor, die sich VERA anschlossen (parteipolitisch motiviert die damals unter SPD-Beteiligung regierten Länder Berlin, Brandenburg, Bremen, Mecklenburg-Vorpommern, Nordrhein-Westfalen und Schleswig-Holstein). So wurde VERA 2004 in sieben Bundesländern durchgeführt. Hieran soll sich nach derzeitigem Stand auch nach den Regierungswechseln in Schleswig-Holstein und Nordrhein-Westfalen nichts ändern.

2. Ziele von VERA

Mit Vergleichsarbeiten lassen sich verschiedene Ziele erreichen:

– *Verbesserung von Chancengleichheit durch Objektivierung (Hilfe zur Selbstevaluation):* Noten, aber auch andere Leistungsdiagnosen und -prognosen werden vielfach auf der Basis einer klassen- oder schulinternen „sozialen Bezugsnorm“ vergeben, d.h. der Leistungsbewertung wird die Leistungsverteilung innerhalb einer Klasse (oder Schule) zugrunde gelegt. Die Verwendung eines klassen- oder auch schulinternen Bezugssystems kann jedoch zu Fehlbeurteilungen, d.h. zu systematischen Unter- oder Überschätzungen führen. Dies hat bereits Ingenkamp (1971) in seinem inzwischen klassischen Buch über die „Fragwürdigkeit der Zensurengebung“ empirisch nachgewiesen. Die *Lernausgangslagenuntersuchung* (LAU) in Hamburg, eine Totalerhebung, die am Ende der Grundschulzeit erstmalig durchgeführt und mittlerweile bis in die Jahrgangstufe 11 fortgesetzt wurde, hat dramatisch deutlich gemacht, dass die „Leistungsstandards, die für eine Empfehlung für die Realschule oder das Gymnasium erfüllt sein müssen, erheblich zwischen den Grundschulen und auch einzelnen Grundschulklassen“ variieren (Lehmann 2001, S. 139). Zum anderen zeigte sich, dass sowohl das Entscheidungsverhalten der Eltern als auch die Empfehlungspraxis der Grundschulen im Sinne der Benachteiligung von Unterschichtkindern sozial beeinflusst ist: Die Übergangschancen für Kinder gleicher Fähigkeitsstufe hängen von der sozialen Schichtzugehörigkeit (bei LAU: Bildungsabschluss der Eltern) ab.

– *Zusätzliche Orientierungshilfe für die Schullaufbahnberatung:* Die Ergänzung der Grundschulempfehlung und der Noten durch Vergleichsarbeiten in den Fächern Deutsch und Mathematik ist eine Verbesserung der herkömmlichen Diagnosepraxis. Eltern können auf diese Weise fundierter beraten werden, was die Schullaufbahn ihrer Kinder anbelangt. Einerseits können Leistungsreserven erkannt werden; andererseits können Vergleichsarbeit im Falle überehrgeiziger Eltern, die ihre Kinder mit einer zu hoch gewählten Schullaufbahn möglicherweise überfordern, ein kritisches Korrektiv darstellen. Die Vergleichsarbeit erhält auf diese Weise die Funktion einer zusätzlichen *Orientierungshilfe*.
Allerdings: Als alleinige Entscheidungsgrundlage für die Grundschulempfehlung wäre das Ergebnis einer Vergleichsarbeit aus methodischen und inhaltlichen Gründen völlig ungeeignet; es kann lediglich ergänzenden Charakter haben! VAen sind punktuelle Aufnahmen und als solche diversen Fehlerquellen unterworfen (Tagesform, durchgenommener Schulstoff, Vertrautheit mit den Anforderungen etc.). Sie ergänzen das Lehrerurteil, dass sich auf der Erfahrung mehrerer Schuljahre gründet, es kann dieses aber auch relativieren und korrigieren helfen. Es bleibt ein Instrument in der Hand des Lehrers.

– *Stärkung der diagnostischen Kompetenzen der Lehrkräfte:* Bei der diagnostischen Kompetenz handelt es sich um ein Bündel von Fähigkeiten, um den Kenntnisstand, die Lernfortschritte und die Leistungsprobleme der einzelnen Schüler im Unterricht fortlaufend beurteilen zu können, so dass das pädagogische Handeln auf diagnostischen Einsichten aufgebaut werden kann (Weinert 1998) Diagnostische und fachdidaktische Kompetenzen sind wichtig für eine effektive Unterrichtsgestaltung. Die Vergleichsarbeiten bieten eine Gelegenheit zur Einschätzung und zur Stärkung dieser Kompetenzen und können somit – richtig genutzt – einen Hebel für die Verbesserung der Unterrichtsqualität darstellen – dies um so mehr, wenn die Ergebnisse der Vergleichsarbeiten und die dort aufgetretenen Typen und Muster von Fehlern zum Gegenstand schulinterner Diskussion z.B. in der Fachkonferenz werden.

- *Standardsicherung:* Mit Vergleichsarbeiten lässt sich feststellen, in welchen Bereichen und bei welchen Kompetenzen Abweichungen vom Standard vorkommen bzw. auf welcher Kompetenzstufe in jedem der drei erfassten Bereiche sich das einzelne Kind befindet (kriteriale Bezugsnorm). Das Konzept sieht vor, dass jeweils eine Hälfte der Items von der Schule gewählt,

die andere Hälfte von der Expertengruppe, die die Aufgaben entwickelt hat, vorgegeben wird.

3. Kritische Punkte von Vergleichsarbeiten

Prinzipiell wohnen zentralen Tests immer eine Reihe von Gefahren inne, die nicht ausgeschlossen werden können:

- das unerwünschte *„coaching on the test" und „backwash"-Effekte*. Solche Effekte sind nie ganz vermeidbar oder auch nur kontrollierbar. Andererseits können (und sollten) die *innovativen* Aufgaben in der Auswahl-Datenbank durchaus auch positive Rückkopplungseffekte auf den Unterricht haben im Sinne der Umsetzung des Bildungsplanes.
- eine *Überbetonung schriftlicher Leistungsprodukte* als Basis für Schullaufbahnempfehlungen. Dies ist im Rahmen von Vergleichsarbeiten allerdings unvermeidbar. Umfassende Evaluationen, die auch kommunikative Kompetenzen umfassen, sind extrem aufwendig und teuer und müssen regulären und breitbandigen standardisierten Testbatterien und einer Individualdiagnose vorbehalten bleiben (wie z.B. im Projekt DESI der KMK).
- eine *Vernachlässigung motivationaler Aspekte* wie Anstrengungsbereitschaft und Ausdauer, Lernkompetenz und Lernstrategien, Selbstkonzept und Selbstvertrauen. Dies wird dadurch gemildert, dass die Ergebnisse der Vergleichsarbeiten die Grundschulempfehlung lediglich ergänzen und nicht ersetzen. Andernfalls hätte man die Vergleichsarbeiten – wie bei den Untersuchungen PIRLS/IGLU (vgl. Bos et al. 2003, 2004) oder SCHOLASTIK (vgl. Weinert & Helmke 1997) – mit eine entsprechenden Schülerbefragung koppeln müssen, was den Rahmen einer Vergleichsarbeit sprengt.
- die *Messfehler*. Ein Nachteil der Vergleichsarbeiten gegenüber standardisierten Tests, ist die größere Messfehlerbelastetheit der Erhebungen und die eingeschränkte Verlässlichkeit der Angaben für eine individuelle Diagnose. Um die Qualitätsentwicklung der einzelnen Schule zu stärken, sieht das Konzept der Vergleichsarbeiten vor, dass Schulen sich selbst ein „Menu" von Items aus der Auswahl-Itembank zusammenstellen können. Erst in der Zusammenschau mit der Grundschulempfehlung und den Zeugnisnoten gewinnen Vergleichsarbeiten ihren Wert als *zusätzliche* Orientierungshilfe.
- die *Gefahr unerwünschter Rankings und Vergleiche*. Während schulinterne Vergleiche durchaus erwünscht sind, weil sie der didaktischen Diskussion dienen, sind Vergleiche von Schulen häufig kontraproduktiv. Auch auf Elternseite könnten Begehrlichkeiten („Wie gut ist die Grundschule, die mein Kind besucht?") geweckt werden, an entsprechende Informationen heranzukommen. Bei dem in VERA vorgelegten Konzept der Vergleichsarbeiten sind solche Vergleiche jedoch aus drei Gründen nicht möglich:
  a) Da jede Schule die Hälfte der Aufgaben selbst bestimmt, ist ein unmittelbarer Vergleich nur auf der Grundlage der von der Expertengruppe vorgegeben Hälfte der Aufgaben möglich. Ergebnisse verschiedener Schulen sind also schon deshalb nur eingeschränkt miteinander vergleichbar.
  b) Zu jedem Durchführungstermin werden etwa 60 Schulen zufällig gezogen, deren Ergebnisse eingeschickt und zentral ausgewertet werden. Schwerpunkt dieser Auswertung ist der Vergleich zwischen den vorgegebenen und den von den Schulen selbst gewählten Aufgaben. Ein unerwünschtes Ranking kann nicht stattfinden, da nur von einem kleineren Teil der Schulen Daten zentral erfasst werden und darüber hinaus diese nicht individualisiert berichtet werden.
  c) Die Vergleichsarbeit resultiert für den einzelnen Schüler weder in einer Note, einem Punkt- oder sonstigem Gesamtwert. Vielmehr sollten Art und Ausmaß der Abweichung der individuellen Schülerleistung von den Ergebnissen der Normierung, von den Lernzielen, für die die betreffenden Aufgaben stehen, in ein qualitatives, die Noten und die Grundschulempfehlung ergänzendes Urteil münden. Erwünscht ist also ein ganzheitliches Urteil, das die Leistung in der Vergleichsarbeit (wie viele und welche Aufgaben werden gelöst?) sowohl auf das erwünschte Kompetenzniveau als auch auf das Gesamtleistungsniveau der Klasse (wie es sich durch den Vergleich mit den geeichten Aufgaben ergibt) bezieht.

4. Aufgabenentwicklung und Eichung

Die Entwicklung prototypischer Aufgaben steht im Mittelpunkt des Projektes. Grundlage der Aufgabenentwicklung ist ein fachdidaktischer und pädagogischer Konsens darüber, welches die zentralen Bildungsziele und Lernziele der Grundschule sind, und welches Kompetenzprofil von Abgängern der Grundschule erwartet werden kann („minimum competency"). Für die Aufgabenentwicklung wurde eine Gruppe von Fachleitern, Schulräten und Fachdidaktikern (aus der Hochschule) aus den sieben beteiligten Bundesländern eingesetzt. Für jede Aufgabe wird angegeben,

- welche Vorkenntnisse für ihre Lösung erforderlich sind,
- welcher Bezug zu den Bildungsstandards besteht,
- welche übergeordneten und fachlichen Anforderungen (Qualifikationen) sie stellt,
- welche Parameter für ihre Schwierigkeit maßgeblich sind und
- welche typischen Missverständnisse und Fehler bei dieser Aufgabe gemacht werden und welche fachdidaktische Bedeutung ihnen zukommt.

Diese Informationen sind unabdingbar, um (leichtere und schwierigere) Varianten von Aufgabenprototypen zu entwickeln. Das Konzept geht damit deutlich über das früher in NRW mit den „Aufgabenbeispielen" realisierte Konzept von Vergleichsarbeiten hinaus. Insbesondere die Analyse von *Schwierigkeitsparametern* (in Mathematik z.B. ob es sich um bloße Prozeduren, um ein- oder mehrschrittige Modellierungen oder um Reflexion handelt; Niveau der sprachlichen Komplexität; innermathematischer, realitätsbezogener oder authentischer Kontext; Art und Zahl zu berücksichtigender Größen; Vorhandensein

multipler Lösungswege etc.) ist für die mathematikdidaktische Diskussion und die Lehrerprofessionalisierung besonders fruchtbar. Das gleiche gilt für die Analyse von *Fehlermustern*, die möglicherweise Aufschluss über zugrunde liegende „misconceptions“ oder prozedurale Defizite seitens der Schüler und über spezifische Stärken oder Schwächen der eigenen Didaktik oder des verwendeten Lehr-Lern-Materials gibt.

Die Aufgabestellung und Kodierung typischer Fehler wird im Folgenden anhand der Zentralaufgabe „Ufos im All!“ exemplarisch konkretisiert[1]:

Aufgabe 6

Ufos im All!

Welche Rechengeschichte passt zu der Aufgabe **8·24**?

☐ Von 24 Ufos sind schon 8 abgestürzt. Wie viele sind übrig?

☐ 8 Ufos fliegen über Deutschland, 24 über Frankreich. Wie viele sind das zusammen?

☐ In 8 Tagen muss ein Ufo wieder zurück zu seinem Planeten. Wie viele Stunden kann es noch über Deutschland fliegen?

☐ 24 Außerirdische fliegen in 8 Ufos. Wie viele Außerirdische sitzen in jedem Ufo?

☐ Keine der Rechengeschichten passt!

Abb. 1: Zentralaufgabe „Ufos im All!“

Für eine Rückmeldung der Häufigkeiten von typischen Fehlern in einzelnen Klassen bzw. Schulen und im schulübergreifenden Vergleich wird bei VERA nicht nur „richtig/falsch“ kodiert, sondern es werden einschlägige Fehlertypen bei der Kodierung mit erfasst:

Aufgabe 6

Korrekte Lösung

Die *richtige* Lösung ist:

- *In 8 Tagen muss ein Ufo wieder zurück zu seinem Planeten. Wie viele Stunden kann es noch über Deutschland fliegen?*

Fehlerhafte Lösungen

Die Aufgabe gilt als nicht richtig gelöst, wenn ein Kreuz bei einer der anderen vorgegebenen Lösungsmöglichkeiten gesetzt wurde. Für die Übertragung der fehlerhaften Lösungen verwenden Sie bitte folgende Zuordnung:

- *Von 24 Ufos sind schon 8 abgestürzt. Wie viele sind übrig?* → falsche Antwort 1 (F1)
- *8 Ufos fliegen über Deutschland, 24 über Frankreich. Wie viele sind das zusammen?* → falsche Antwort 2 (F2)
- *24 Außerirdische fliegen 8 Ufos. Wie viele Außerirdische sitzen in jedem Ufo?* → falsche Antwort 3 (F3)
- *Keine der Rechengeschichten passt.* → falsche Antwort 4 (F4)
- *Mehrfachantworten* → and. Fehler

Abb. 2: Korrekturanweisung zu „Ufos im All!“

[1] Weitere Zentralaufgaben sowie die zugehörigen Korrekturanweisungen, didaktische Erläuterungen und Hinweise zu Lehrplanbezügen stehen auf der VERA-Projekthomepage (http://www.uni-landau.de/vera/).

Vergleichsarbeiten erfordern normative Daten. Diese werden in Pilotschulen erhoben. Diese Pilotschulen sollen die gesamte Bandbreite des Schulkontextes abdecken, d.h. sowohl Schulen in Ballungsgebieten mit hohem Anteil an Kindern mit Deutsch als Fremdsprache und geringer Bildungsnähe der Eltern als auch kleinstädtische und ländliche Areale umfassen. Auf der Basis präziser Angaben zu den Parametern der Stichprobe bestimmte die Bildungsverwaltung spezifische Schulen als Pilotschulen. Aus pragmatischen Gründen ist es vorzuziehen, dass es sich um Pilotschulen für *beide Fächer* (Mathematik und Deutsch) handelt und dass die Zusammensetzung der Pilotschulen während des gesamten Zeitraums so konstant wie möglich bleibt. Aus psychometrischer Sicht ist es unabdingbar, dass zu jedem Item der Aufgabenbank empirische Angaben von mindestens 10 Klassen vorliegen, weil sonst der Messfehler zu groß ist und eine Normierung, auf deren Basis die Rückmeldungen an die Schulen geschieht, zu ungenau würde. Diese Angaben müssen aus Schulen mit unterschiedlichem Kontext entstammen. Es muss rotierte Ankeritems geben, die in einer größeren Zahl von Schulen gleichermaßen eingesetzt werden. Es wird also die Abfolge der Aufgaben variiert, damit Reihungseffekte ausgeschlossen werden können. Wegen der zu erwartenden Klumpungseffekte in Schulen, insbesondere wegen schulspezifischer Schwerpunkte der Stoffbehandlung erhalten verschiedene Klassen unterschiedliche Aufgaben. Es wird davon ausgegangen, dass Parallelklassen innerhalb einer Schule ähnlichen Unterricht abhalten und die Inhalte abgesprochen sind.

Anders als bei den folgenden „echten“ Vergleichsarbeiten, wo Fachkonferenzen sich ein schulspezifisches Menu aus vorgegebenen Aufgabenklassen zusammenstellen können, müssen im Rahmen der Normierungsstudien den jeweiligen Parallelklassen der jeweiligen Schule deshalb unterschiedliche Aufgaben zugewiesen werden. Hieraus ergeben sich aus den falschen Antworten intelligente „Distraktoren“ für das dann in der Hauptstudie eingesetzte multiple-choice Format.

Das Design sieht vor, dass 200 Klassen aus Schulen mit unterschiedlichem Kontext (~ etwa 100 Schulen) während der gesamten Projektdauer an den Normierungsstudien teilnehmen. In diesen Klassen werden die vorgeschlagenen Aufgaben empirisch überprüft; darüber hinaus beteiligen sich die Lehrkräfte an der schöpferischen *Weiterentwicklung* der Aufgaben (quantitative und qualitative Verbesserung des ursprünglichen Aufgabenpools und seiner zunehmenden Erweiterung).

Durch die Ergebnisse in den Fächern Deutsch und Mathematik kann die brisante Frage, welche Rolle sprachliche Kompetenzen für mathematische Fähigkeiten (ein zentraler Aspekt von PISA 2000, vgl. Deutsches PISA-Konsortium 2001) im Grundschulalter spielen, untersucht werden. Dort zeigte sich bekanntlich, dass das Leseverständnis auch für die mathematische Kompetenz von erheblicher Bedeutung ist.

5. Durchführung

Grundsätzlich wird die eine Hälfte der Aufgaben von der Steuergruppe VERA festgelegt, während die andere Hälfte von den Grundschulen (bzw. den Fachkonferenzen) aus

dem vorgegebenen Aufgabenpool ausgewählt wird. Dies geschieht zu einem festgelegten Termin, in der Regel zwei Wochen vor Durchführung der VAen. Dabei folgt die Auswahl jeweils einem Schlüssel, der gewährleistet, dass die unterschiedlichen fünf Kompetenzklassen, die sich an die Standards anlehnen, die Lehrplanbereiche und Aufgabentypen über alle Schulen hinweg in gleicher Weise repräsentiert sind, d.h. aus Aufgaben, deren Schwierigkeiten (Lösungshäufigkeiten) aus den vorausgegangenen Normierungsuntersuchungen bekannt sind. Dadurch wird vermieden, dass auf die eigenen vorherrschenden Unterrichtsschwerpunkte zugeschnittene „leichte“ Vergleichsarbeiten konstruiert werden können. Der Zugang und sämtliche Angaben im Netz, die sich auf die Vergleichsarbeiten beziehen, sind passwortgeschützt, wofür jede Schule ein eigenes Passwort erhält.

Innerhalb jeder Schule werden die Vergleichsarbeiten in allen Parallelklassen in identischer Form geschrieben. Die Schulen übernehmen die Organisation und Auswertung der Vergleichsarbeiten, insbesondere übernehmen die Lehrkräfte die Auswertung nach richtig/falsch und nehmen bei falschen Antworten eine Zuordnung zu den vorgegebenen (häufigen) Fehlertypen vor. Diese Auswertung erfolgt auf einen digitalisierten Erhebungsbogen auf dem Internetserver des Projekts VERA.

Die Leistung der Klasse kann für jede einzelne Aufgabe mit den bekannten Aufgabenschwierigkeiten der Gesamtgruppe verglichen werden. Auf diese Weise erhält die Lehrkraft Anhaltspunkte, wie die Leistung der Klasse im Vergleich zu einer größeren Gesamtheit von Klassen ausgefallen ist, und kann so ein genaueres Bild der Leistungsfähigkeit ihrer Klasse und einzelner Schüler gewinnen. Neben dieser Vergleichsinformation soll die Vergleichsarbeit *kriteriale Information* liefern: Die Aufgabentypen repräsentieren unterschiedliche Aspekte der zu Beginn des 4. Schuljahres erwartbaren Kompetenzen.

Die Ergebnisse stellen Informationen über klassen- oder schulspezifische Fehlermuster (z.B. Häufung bestimmter Fehlertypen) dar, sie liefern damit einen Anlass für didaktische Diskussionen innerhalb des Kollegiums und ermöglichen damit die Entwicklung eines veränderten Unterrichts. Der *Lösungsprozentsatz* und die *Fehlerverteilung* wird den Lehrkräften rückgemeldet. Für jedes Item wird angegeben, wie viel Prozent der Schüler es lösten oder nicht. Der Zweck liegt in der Orientierung des Leitungsstandes der eigenen Klassen, verglichen mit einem Referenzwert.

6. Interne Auswertung

Der Vorteil der Vergleichsarbeit liegt in der Ermöglichung eines „fremden Blicks“ auf das Leistungsprofil der eigenen Klasse: durch Vergleichsmöglichkeiten, die im Schulalltag normalerweise nicht zur Verfügung stehen. Dies kann Grundlage für gezielte unterrichtliche Schwerpunktsetzungen sein. Hierbei ist zu beachten, dass die Rückmeldungen lediglich an die Schulen, nicht an die Schulbehörde gehen, so dass die Auswertung bei den Kollegien verbleibt. Vier Vergleichsebenen sind hervorzuheben:

– *Innerschulische Profile*: Wo hat meine Klasse Stärken und Schwächen, verglichen mit den Parallelklassen?
– *Vergleich mit dem Durchschnitt aller Schülerinnen und Schüler der gesamten Jahrgangsstufe*: Wo weicht meine Klasse - nach oben oder unten - vom Durchschnitt ab? Vergleichsbasis war bis 2003 die rheinland-pfälzische „Normierungsstichprobe“; ab 2004 liegt eine bundeslandübergreifende Normierung zugrunde.
– *Fairer Vergleich*: Wie groß sind die Abweichungen, wenn ich das Ergebnis meiner Klasse mit einer Gruppe von vergleichbaren Klassen (ähnliches Einzugsgebiet, ähnliche Schülerzusammensetzung) vergleiche?
– Inzwischen gibt es eine *Orientierung an den Standards*: In welchen Kompetenzbereichen zeigen sich Anfang der 4. Klasse Schwächen, die im Hinblick auf die Erreichung der Standards (Ende der 4.Klasse) Anlass für gezielte Unterrichts- und Fördermaßnahmen sein müssen?

Jenseits der globalen Abschätzung des Leistungsstandes der Klasse kann der Blick auf einzelne Aufgaben didaktische Impulse liefern:

– *Aufgabenschwierigkeiten*: Bei welchen Aufgaben gibt es zwischen meiner Klasse und (a) den Parallelklassen, (b) dem Gesamtdurchschnitt besonders große Abweichungen in der Lösungshäufigkeit?
– *Gewählte vs. vorgegebene Aufgaben*: Zeigen sich Unterschiede in der Lösung zwischen den 10 zentralen Aufgaben (die erst am Vortag der Vergleichsarbeit verfügbar waren) und den 10 Aufgaben, die im Zeitraum von 2 Wochen vor der Vergleichsarbeit ausgewählt wurden? Wie ist es zu interpretieren, wenn z.B. die Klasse bei den zentralen Aufgaben erheblich schlechter abschneidet als bei den selbst gewählten Aufgaben - oder umgekehrt?
– *Auffällige Fehlermuster*: Bei welchen Aufgaben zeigen sich in meiner Klasse markante Abweichungen vom Durchschnitt, was die Häufigkeit bestimmter Falschlösungen anbelangt?

7. Pädagogischer Nutzen

Die Optimierung des Lehrens und Lernens ist untrennbar verbunden mit einer verbesserten Fehler- und Aufgabenkultur und einer verbesserten diagnostischen Kompetenz von Lehrkräften (vgl. das BLK-Gutachten zur Vorbereitung des Programms „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“ – SINUS, BLK 1997). Nicht ohne Grund kommt beiden Aspekten sowohl bei den KMK-Projekten DESI und PISA als auch in Projekten des DFG-Schwerpunktprogramms „Bildungsqualität von Schule“ eine Schlüsselrolle zu. Die Vergleichsarbeiten werden mit diesen beiden Aspekten verknüpft. Insbesondere kommt der diagnostischen Kompetenz innerhalb VERAs eine Schlüsselrolle zu. Sie umfasst neben der

(1) *Diagnosegenauigkeit* (Fähigkeit, Schüler und Aufgaben zutreffend einzuschätzen) verschiedene Formen von Wissen und darauf basierendem Können auch

(2) *methodisches* Wissen (diagnostische Methoden; Urteilsfehler und –tendenzen),
(3) *lehrstoffbezogenes* Wissen (Anforderungen von Lehrstoffen und Aufgaben; mögliche Lösungswege, alterstypische Missverständnisse und Fehler) und
(4) *spezifische Kenntnisse* (über einzelne Schüler, Schülergruppen und Schulklassen, ihre Stärken und Schwächen). Bei PISA waren Lehrkräfte beispielsweise gefragt worden, wie viele Schüler ihrer Klasse der niedrigsten Stufe der Lesekompetenz zuzuordnen sind. Geprüft wurde, wie gut die Lehrerurteile mit der tatsächlichen Lesekompetenz übereinstimmten.

Bei VERA machen die Lehrkräfte auf einem Fragebogen folgende Angaben:

– *Unterschiede zwischen Aufgaben*: Es soll vorhergesagt werden, welche Aufgaben für die Klasse insgesamt am schwierigsten, welche leicht sind. Dies erfordert eine Facette der Diagnosekompetenz, die eine stark fachdidaktische Komponente hat.
– *Interindividuelle Unterschiede:* Lehrer können vor der VA eine Vorhersage darüber machen, wer innerhalb der Klasse welches Ergebnis erzielt, z.B. welche Schüler(gruppe) alle/fast alle Aufgaben lösen, welche weniger als die Hälfte etc. Diese Prognose kann leicht mit den empirischen Ergebnissen verglichen werden. Die Prognose kann sich auch auf eine Teilgruppe besonders „schwieriger" Schüler beziehen, oder auf Schüler, die neu in der Klasse sind. Interessant ist dann die Erklärung von Diskrepanzen zwischen Vorhersage und tatsächlich erzielten Leistungen. Hierbei können Hilfestellungen und Anregungen seitens der Pädagogischen Diagnostik und der Lehr-Lern-Forschung gegeben werden, so dass die diagnostische Sensibilität wirkungsvoll trainiert werden kann.

Das SINUS-Gutachten (BLK 1997) und die didaktische Diskussion im Anschluss an TIMSS (vgl. Blum & Neubrand 1998; Henn 1999) weisen der *Aufgabenkultur* eine wichtige Rolle für die Verbesserung des mathematisch-naturwissenschaftlichen Unterrichts zu. Dies gilt auch für die Grundschule: Beachtung von Aufgaben, die mehrere Vorgehensweisen und Lösungsmöglichkeiten zulassen, Einsatz von Übungsaufgaben, die nicht nur Vorgehensweisen einschleifen und automatisieren, sondern auch horizontalen Transfer erfordern. Dies wird mit den in VERA verwendeten Aufgaben versucht.

8. Zukünftige didaktische Probleme

Es scheint, dass VERA der Ausgangspunkt für eine didaktische Diskussion darstellt, da hier prinzipielle, von der Fachdidaktik noch zu lösende Probleme deutlich werden.

– *Standards: Di*e aktuelle Entwicklung innerhalb von VERA orientiert sich an den neuen, von der KMK formulierten Bildungsstandards und den Leitideen. Dies macht eine Schwierigkeit insofern aus, als es erstens (marginale) Unterschiede zwischen den anfangs verwendeten Standards und den neuen bestehen, so dass eine Zuordnung der vorhandenen Aufgaben aus dem Pool notwendig ist, und zweitens naturgemäß die Leitideen (Zahl und Operationen, Raum und Form, Muster und Strukturen, Größen und Messen, Daten, Häufigkeit und Wahrscheinlichkeit) keineswegs disjunkt sind und daher eine Zuordnung jeder Aufgabe zu nur einer dieser Leitideen problematisch erscheint. Für eine Auswertung wäre dies aber wünschenswert, um einerseits interne Kompetenzstufen operationalisieren zu können und andererseits beschreiben zu können, auf welcher Kompetenzstufe sich ein Schüler befindet.
– *Itemzahl:* Will man bezüglich der fünf Leitideen und den jeweiligen fünf Kompetenzstufen eine valide Zuordnung treffen, also fundiert aussagen, auf welchem Niveau sich ein Schüler befindet, dann sind eine Vielzahl von Items notwendig, die sich wiederum in der verfügbaren Zeit nicht bearbeiten lassen. Weder die Alternative, VAen zeitlich zu verlängern, etwa an zwei Tagen durchzuführen, noch auf eine Zuordnung zu verzichten, erscheint wünschenswert.
– *Allgemeine Lernziele*: Prinzipiell scheint es ausgeschlossen, die Fähigkeit „zu kommunizieren" qua Multiple-Choice-Test zu erfassen. Diese gewünschte Kompetenz ist durch das Instrument VA nicht erhebbar und muss daher innerhalb der Schulklassen anders erfasst werden.
– *Schwierigkeitsgrad*: Es ist von didaktischer Seite keine valide Angabe über den Schwierigkeitsgrad einer Aufgabe (bzgl. einer Schulstufe) zu machen, selbst schwierigkeitsbestimmende Parameter spielten in der didaktischen Diskussion bislang kaum eine Rolle. Aus diesem Grund ist in VERA und vergleichbaren Projekten die Beschreibung von Aufgabensets erschwert.
– *„partial credits":* Was in Klassenarbeiten möglich ist, nämlich mehrstufige Aufgaben zu stellen und das richtige Lösen von Teilaufgaben dem Schüler positiv anzulasten, ist in Multiple-Choice-Tests erschwert. Dahinter steckt weniger ein inhaltliches als ein psychometrisches Problem, das die Verrechnung teilrichtiger Antworten zu kaum handhabbaren Schwierigkeiten führt (inwieweit allerdings die Abbildung der Schülerlösungsqualität auf eine arbiträre Punktskala bei Klassenarbeiten dem Gerechtigkeitsideal entspricht, soll hier nicht diskutiert werden).

Unbeschadet dieser Einschränkungen, die für die Fachdidaktik insgesamt und nicht nur für die beteiligten Didaktiker eine Herausforderung darstellen, scheinen die bisherigen Ergebnisse von VERA ermutigend. Nicht, dass der Beifall inzwischen auch aus den Lehrerzimmern oder den Standesverbänden dröhnte (vgl. die eher polemische Kritik des Grundschulverbandes im diesem Jahr). Es lässt sich aber, möglicherweise auch in Folge von VERA, eine Zunahme der Begehrlichkeit nach Lehrerfortbildung konstatieren mit Themen, die mit der Qualitätsentwicklung, „guten Aufgaben", Problemlösen etc. zusammenhängen und für die offenbar, in Einklang mit TIMSS und PISA, ein Nachholbedarf besteht.

Dass dies von der Bildungsverwaltung auch entsprechend angeboten wird, ist zu hoffen, liegt aber jenseits der Verantwortung der an VERA Beteiligten.

Literatur

Arnold, K.-H. (1999). Diagnostische Kompetenz erwerben. Wie das Beurteilen zu lernen und zu lehren ist. *Pädagogik*, *51 (7/8)*, S. 73-77.

Arnold, K.-H. (2001). Beurteilungskompetenz. *unterrichten/erziehen, 20 (1)*, S. 12-15.

Arnold, K.-H. (2001). Qualitätskriterien für die standardisierte Messung von Schulleistungen. Kann eine (vergleichende) Messung von Schulleistungen objektiv, repräsentativ und fair sein? In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 117-130). Weinheim: Beltz.

BLK (1997). *Gutachten zur Vorbereitung des Programms „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“*. Bonn: Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung.

Blum, W. & Neubrand, M. (Hrsg.) (1998). *TIMSS und der Mathematikunterricht: Informationen, Analysen, Konsequenzen.* Hannover: Schroedel.

Bos, W., Lankes, E. M., Prenzel, M., Schwippert, K., Walther, G. & Valtin, R. (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* Münster: Waxmann.

Bos, W., Lankes, E. M., Prenzel, M., Schwippert, K., Valtin, R. & Walther, G. (2004). *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich.* Münster: Waxmann.

Deutsches PISA-Konsortium (Hrsg.) (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske + Budrich.

Heller, K. A. & Hany, E. A. (2001). Standardisierte Schulleistungsmessungen. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 87-101). Weinheim: Beltz.

Helmke, A. (2004). *Unterrichtsqualität: Erfassen, Bewerten, Verbessern* (2. Aufl.). Seelze: Kallmeyersche Verlagsbuchhandlung (Kapitel 3.7 „Diagnostische Expertise“, S. 84-104).

Helmke, A. & Hosenfeld, I. (2003). Vergleichsarbeiten (VERA): Eine Standortbestimmung zur Sicherung schulischer Kompetenzen - Teil 2: Nutzung für Qualitätssicherung und Verbesserung der Unterrichtsqualität. *SchulVerwaltung, Ausgabe NRW (5)*, S. 143-145.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung*. Hohengehren: Schneider-Verlag.

Henn, H.-W. (Hrsg.) (1999). *Mathematikunterricht im Aufbruch.* Hannover: Schroedel.

Ingenkamp, K. (Hrsg.). (1971). *Die Fragwürdigkeit der Zensurengebung.* Weinheim: Beltz.

Lehmann, R. H. (2001). Messung von Schulleistungen im Primar- und Sekundarbereich. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 134-141). Weinheim: Beltz.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59-71). Weinheim: Beltz.

Schrader, F.-W. (2001). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 68-71). Weinheim: Psychologie Verlags Union.

Schrader, F.-W. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 45-58). Weinheim: Beltz.

Weinert, F. E. (1998). Guter Unterricht ist ein Unterricht, in dem mehr gelernt wird asl gelehrt wird. In J. Freund, H. Gruber & W. Weidinger (Hrsg.*), Guter Unterricht - Was ist das? Aspekte von Unterrichtsqua*lität (S. 7-18). Wien: ÖBV Pädagogischer Verlag.

Weinert, F. E. (Hrsg.). (2001). *Leistungsmessungen in Schulen.* Weinheim: Beltz.

Weinert, F. E. & Helmke, A. (Hrsg.) (1997). *Entwicklung im Grundschulalter.* Weinheim: Beltz PVU.

Autor

Lorenz, Jens Holger, Prof. Dr., Fakultät III, Pädagogische Hochschule Heidelberg, Keplerstraße 87 , D-69120 Heidelberg
E-Mail: jens.lorenz@ph-heidelberg.de