Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenzen[1]

Prof. Dr. Andreas Helmke und Dr. Ingmar Hosenfeld, Landau

erschienen in "SchulVerwaltung Hessen, Rheinland-Pfalz, Saarland" 1/2003 (S. 10-13) und 2/2003 (S. 41-43)

Seit PISA 2000 gibt es verstärkte Bemühungen, bereits in der Grundschule - also zu einem Zeitpunkt, bei dem es noch nicht zu spät ist - eine *Standortbestimmung* zur Sicherung grundlegender schulischer Kompetenzen vorzunehmen. Die KMK hat auf ihrer *298. Plenarsitzung* am 23./24.05. in Eisenach „gemeinsame Standards für die Schulbildung" beschlossen; zur Überprüfung der Einhaltung dieser Standards sollen künftig in Verantwortung der Länder Orientierungs- und Vergleichsarbeiten geschrieben werden, die bereits in der Primarstufe beginnen sollen. Diese Planung wurde in einer *Sitzung der KMK vom 25. Juni 2002* bestätigt und präzisiert: Insbesondere wird auf die Wichtigkeit einer Verzahnung der Bildungsstandards mit Maßnahmen zur Verbesserung der Unterrichtsqualität hingewiesen: „Die Qualität von Unterricht und Schule muss gesichert und konsequent weiterentwickelt werden. Grundlagen dafür sind verbindliche Standards sowie eine ergebnisorientierte Evaluation. Wir wollen die Professionalität der Lehrertätigkeit verbessern, insbesondere im Hinblick auf diagnostische und methodische Kompetenz als Bestandteil systematischer Schulentwicklung." (D. Schipanski, Präsidentin der KMK, 25. Juni 2002). Den vorläufigen Abschluss bilden die Beschlüsse der *299. Plenarsitzung* der KMK (17./18.10. in Würzburg), in der die „länderübergreifende Orientierungs- oder Vergleichsarbeiten", verbunden mit dem Aufbau von „Aufgabenpools" nochmal ausdrücklich bekräftigt werden (s. SchVw HRS 12/2002).

Neben das Ziel der Standortbestimmung tritt das Bedürfnis insbesondere der *Eltern* nach ergänzenden Informationen, die bei einer realitätsangemessenen Planung der Schullaufbahn ihrer Kinder berücksichtigt werden können. Bisherige Studien, insbesondere PISA 2000 (Baumert, Klieme, Neubrand, Prenzel, Schiefele, Schneider, Stanat, Tillmann & Weiß, 2001) und LAU (Lehmann, Peek & Gänsfuß, 1997) haben gezeigt, dass sowohl das elterliche Entscheidungsverhalten als auch die Empfehlungspraxis der Grundschulen zu einer Benachteiligung von Kindern aus bildungsfernen Schichten führt. Zusätzliche Hinweise auf die Leistungsfähigkeit der Kinder, wie sie im Rahmen der Vergleichsarbeiten gewonnen werden, können eine wertvolle ergänzende Orientierungshilfe für die Schullaufbahnberatung sein.

Nach unserer Einschätzung hängt der Erfolg eines solchen Vorhabens nicht zuletzt davon ab, ob es gelingt, die Messung der Schülerleistungen mit Maßnahmen zur *Verbesserung didaktischer und diagnostischer Kompetenzen von Lehrkräften* und damit der *Verbesserung der Unterrichtsqualität* zu verknüpfen (Helmke, 2003; Peek, 2001; Ditton, Arnoldt & Bornemann, 2002; Schrader & Helmke, in Druck). Deshalb kommt den „pädagogischen Optionen" unseres Konzeptes ein herausragender Stellenwert zu.

---

[1] Die hier wiedergegebene Darstellung ist eine überarbeitete und aktualisierte Version des gleichnamigen Kapitels aus dem Buch „Unterrichtsqualität: Erfassung, Bewertung, Verbesserung" (Helmke, 2003).

1 Begriffsbestimmung

Die Orientierung über den aktuellen Stand von Orientierungs- und Vergleichsarbeiten in der Grundschule wird dadurch erschwert, dass es bisher keine einheitliche Begrifflichkeit gibt. Die Aktivitäten und Initiativen in verschiedenen Bundesländern laufen vielmehr unter sehr unterschiedlichen Begriffen (Vergleichsarbeiten, Orientierungsarbeiten, Diagnosearbeiten, Parallelarbeiten, zentrale Klassenarbeiten etc.), was die Vergleichbarkeit nicht erleichtert. Es ist in diesem Rahmen nicht möglich, eine detaillierte Synopse darzustellen (für Details vgl. Orth, 2002), stattdessen beschränken wir uns darauf, das in Rheinland-Pfalz realisierte Konzept der Vergleichsarbeiten zu skizzieren.

Vergleichsarbeiten lassen sich wie folgt charakterisieren und von anderen Typen der Leistungsmessung abgrenzen: Es handelt sich um *schriftliche Arbeiten*, die in einer größeren Anzahl von Schulen (ggfs. landesweit) auf der Basis einer vorgegebenen Aufgabenstichprobe eingesetzt werden mit dem Ziel, die Leistungen der Schüler an einer klassen- und schulübergreifenden sozialen und/oder kriterialen Bezugsnorm zu messen. Im hier dargestellten Konzept wird dabei grundsätzlich die eine Hälfte der Aufgaben der Vergleichsarbeiten zentral (vom Ministerium) festgelegt, während die andere Hälfte von den Grundschulen aus dem vorgegebenen Aufgabenpool ausgewählt wird. Dabei folgt die Auswahl einem Schlüssel, der gewährleistet, dass unterschiedliche Lehrplanbereiche und Kompetenzklassen über alle Schulen hinweg in vergleichbarer Weise repräsentiert sind.

- Vergleichsarbeiten dieses Typs ähneln insofern *Parallelarbeiten*, als (in mehrzügigen Schulen) die Parallelklassen einer Schule jeweils identische Aufgabensätze bearbeiten und so ein Vergleich der Leistungsstände über die Klassen hinweg möglich ist. Das Konzept der Parallelarbeiten wurde bisher u.a. in Nordrhein-Westfalen in den Klassenstufen 3, 7 und 10 realisiert (Ministerium für Schule und Weiterbildung, Wissenschaft und Forschung des Landes Nordrhein-Westfalen 2000), vgl. Thürmann (2002), der auch auf die Schwachstellen der Parallelarbeiten hinweist. *Vergleichsarbeiten* gehen in ihrem Vergleichsanspruch über Parallelarbeiten hinaus, indem sie sowohl einen Vergleich mit landesweiten oder auch länderübergreifenden Normwerten erlauben als auch zeitliche Trends identifizieren können. Sie bieten somit Informationen, die mit anderen Mitteln, insbesondere mit Parallelarbeiten, nicht erhältlich sind.
- Das Konzept der *Orientierungsarbeiten (ab 5. Klasse)*, das in den Kantonen der Zentralschweiz realisiert wird (Senn, 2002), umfasst aus den gültigen Lehrplänen entwickelte lernzielorientierte Aufgabensammlungen mit präzisen Angaben zur Bewertung im Hinblick auf verschiedene Zielkriterien. Die Bayerischen „Orientierungsarbeiten in der Grundschule“ (2. und 3. Klassenstufe), die „Diagnosearbeiten“ in Baden-Württemberg sowie die landeseinheitlichen Klassenarbeiten in mehreren anderen Bundesländern (z.B. Saarland) verfolgen ein ähnliches Konzept, nur mit dem Unterschied, dass es sich um obligatorische und flächendeckende Vorhaben handelt. Zwar ist teilweise vorgesehen, die Ergebnisse (in Form von Notenspiegeln oder Punkten pro Aufgaben) zentral rückzumelden; auch hier fehlt jedoch die bei Vergleichsarbeiten vorgesehene Möglichkeit der Evaluation an landesweiten Standards und die Analyse von Trends.
- Sie ähneln *standardisierten Schulleistungstests* (z.B. Hamburger Schulleistungstest, Mietzel & Willenberg, 2000) insofern, als die Aufgaben von Fachleuten entwickelt bzw. ausgewählt

und im Hinblick auf inhaltliche Kriterien (Thematisierung verschiedener Kompetenz- und Wissensbereiche; Abdeckung gültiger Lehrpläne und Curricula) und bestimmte Testgütekriterien (z.B. angemessene Schwierigkeit) geprüft werden. Allerdings sind die Anforderungen an die Testgütekriterien bei Vergleichsarbeiten geringer als bei standardisierten Tests (Arnold, 2001; Heller & Hany, 2001; Lehmann, 2001): *Erstens* ist die Bedingungs- und Durchführungskontrolle geringer, da die Durchführung in der Regel von Lehrkräften und nicht von geschulten Testleitern vorgenommen wird; *zweitens* muss man insbesondere bei sprachlichen Leistungen (außer bei der Rechtschreibung) Abstriche bei der Objektivität machen, *drittens* ist die Vergleichbarkeit über Schulen hinweg in dem Maße eingeschränkt, in dem die schulinterne Auswahl von Aufgaben aus dem Gesamtpool zu unterschiedlich schwierigen Arbeiten führen kann.

- Anders als *Klassenarbeiten*, die sich in der Regel auf einen bestimmten, zuvor durchgenommenen Unterrichtsstoff beziehen, umfassen Vergleichsarbeiten den Stoff des gesamten Schuljahres und beziehen sich gegebenenfalls auch auf entsprechende Vorkenntnisse. Vergleichsarbeiten können, müssen aber nicht als Klassenarbeiten „zählen" – dies ist eine pädagogische und eine bildungspolitische Frage, berührt jedoch nicht das Konzept der Vergleichsarbeiten.
- Von *Lernstandserhebungen* (wie bei IGLU, MARKUS, TIMSS, LAU) unterscheiden sich Vergleichsarbeiten dadurch, dass keine Aussagen über die Leistungen einer ganzen Region (z.B. eines Bundeslandes) beabsichtigt sind und dass sie nicht von umfassenden Lehrer- oder Schülerbefragungen zu unterrichtlichen oder individuellen Bedingungen schulischer Leistungen begleitet werden (vgl. Helmke & Jäger, 2002).
- Ein in der Schweiz verbreitetes Verfahren ist das Selbstevaluations-Verfahren „*Klassen-Cockpit*"[2] (ab 3. Klasse). Es handelt sich um auf den Lehrplan abgestimmte Aufgabensammlungen, die von einem Lehrmittelverlag von den Schulen käuflich erworben werden können. Die Ergebnisse der jeweiligen Klasse werden mit denen einer repräsentativen Stichprobe von 25 Klassen verglichen; zusätzlich zu dem so resultierenden Lernzielprofil enthält die Rückmeldung einen Notenvorschlag. Das rheinland-pfälzische Konzept der Vergleichsarbeiten ist insofern ähnlich, als Vergleiche mit repräsentativen Vergleichsgruppen vorgesehen sind; es unterscheidet sich davon durch die Berücksichtigung des sozioökonomischen Kontextes der Vergleichsstichprobe, die stärkere Partizipation der Schulkollegien, die Vergleiche auch auf Basis einzelner Aufgaben und Fehlermuster sowie durch die pädagogischen Optionen.

2 Vergleichsarbeiten in Rheinland-Pfalz

Basierend auf einem Beschluss des Landtages 25. April 2002 werden in Rheinland-Pfalz ab 2003 in sämtlichen 4. Klassen der Grundschulen des Landes in Mathematik und später auch in Deutsch Vergleichsarbeiten geschrieben. Das Projekt Vergleichsarbeiten (VERA) umfasst einen Zeitraum von zunächst 5 Jahren. Das Konzept für dieses Projekt wurde von Prof. Dr. Andreas Helmke, gemein-

[2] http://www.klassencockpit.ch/

sam mit Dr. Ingmar Hosenfeld, entwickelt, denen auch die wissenschaftliche Leitung des Projektes, dessen Charakteristika im folgenden kurz skizziert werden sollen, obliegt. Abbildung 1 gibt einen Überblick über die bis 2006 geplanten Komponenten des Projekts.

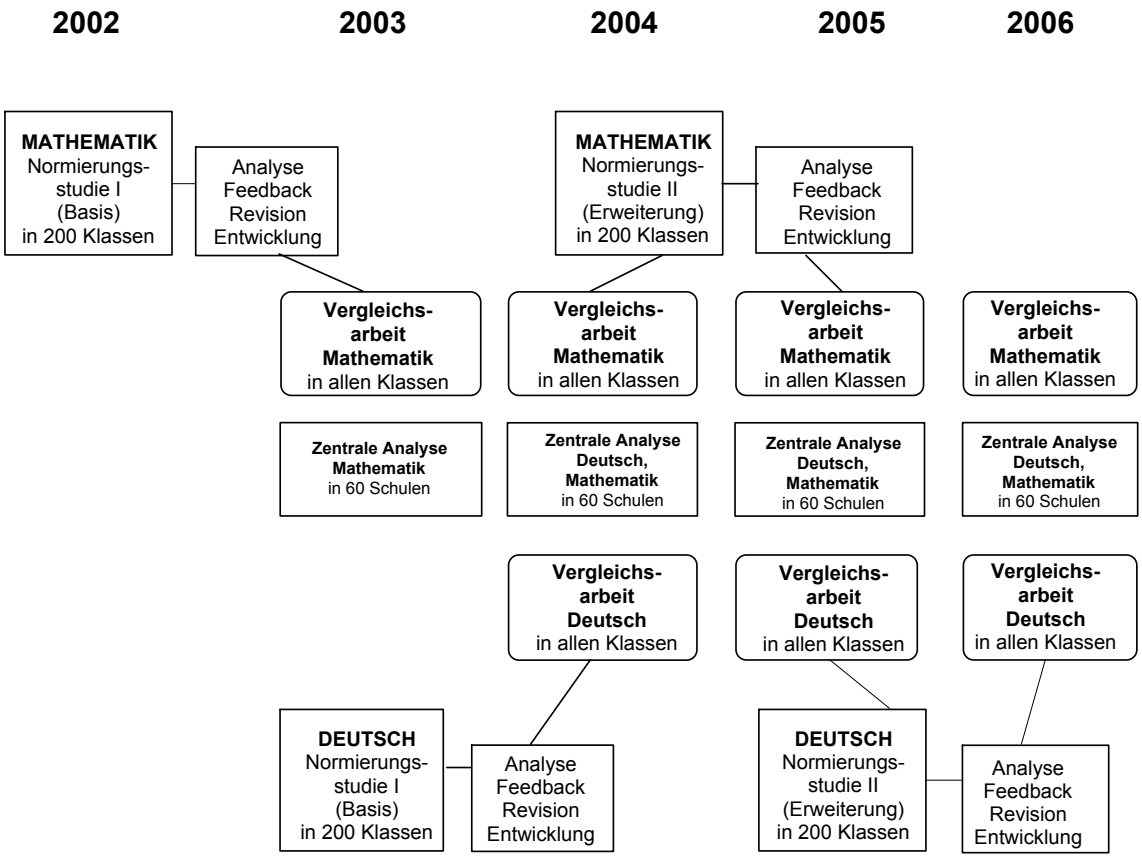

Abbildung 1: Das Design des Projektes VERA (Vergleichsarbeiten) in Rheinland-Pfalz

Zwei charakteristische Merkmale dieses Konzeptes sind, dass erstens ein sehr *großer Aufgabenpool* entwickelt wird und dass zweitens die in einer Vergleichsarbeit tatsächlich einzusetzenden Aufgaben zuvor an einer umfangreichen Stichprobe *normiert* werden. Die Aufgabensammlung für Mathematik umfasst derzeit mehr als 500 Aufgaben; die Arbeiten für Deutsch befinden sich zur Zeit noch im Anfangsstadium. Dieser Aufgabenpool wird von einer Expertengruppe (erfahrene Lehrkräfte, maßgeblich beraten von Fachdidaktikern wie z.B. Prof. J. H. Lorenz für Mathematik und Prof. A. Bremerich-Vos für Deutsch) auf der Basis des aktuellen *Rahmenplans für die Grundschule* erstellt. Alle Mathematikaufgaben werden anhand eines zweidimensionalen Rasters hinsichtlich des angesprochenen mathematischen *Inhaltsbereichs* (z.B. Geometrie, Teilgebiet Symmetrie) und den *Tätigkeitsanforderungen* (z.B. Skizzieren, Schätzen, Argumentieren) klassifiziert. Die Zugrundelegung dieses Raster gewährleistet, dass alle Bereiche des Curriculums in hinreichender Breite und in Kombination mit verschiedenen Anforderungen im Aufgabenpool repräsentiert sind. Alle Aufgaben und die zugehörigen Klassifikationen werden in einer *zentralen Datenbank* verwaltet.

2.1 Pilotierungsstudien

Nach der Entwicklung und Klassifikation der Aufgaben finden – noch vor der eigentlichen Normierung – Pilotstudien in 40 Klassen statt, in denen eine größere Menge von Aufgaben erstmals vorge-

testet werden. Eine wichtige Funktion dieser Pilotierungsstudien besteht unter anderem darin, Informationen über typische Schülerfehler bei den einzelnen Aufgaben zu gewinnen, die z.B. für die Erzeugung informativer Antwortalternativen für Aufgaben im Multiple-Choice-Format benötigt werden.

2.2 Normierungsstudien

Die Basis für den Vergleich der individuellen Ergebnisse wie auch der Ergebnisse ganzer Klassen oder Schulen bilden die Normierungsstudien. Dabei wird an einer *Normierungsstichprobe* von 200 Klassen eine große Anzahl von Aufgaben (in Mathematik gut 150) untersucht und „geeicht", und zwar im jährlichen Wechsel zwischen Mathematik- und Deutschaufgaben. Die Klassifikation anhand des zweidimensionalen Inhalts-Anforderungs-Rasters gewährleistet, dass alle Bereiche des Curriculums und verschiedene Tätigkeitsanforderungen im untersuchten Aufgabensatz repräsentiert sind.

Berücksichtigung des Kontextes. Die Architektur der Normierungsstichprobe sieht vor, dass die Klassen aus Regionen mit unterschiedlichem sozioökonomischen Umfeld stammen („ungünstig", z.B. „sozialer Brennpunkt" / „durchschnittlich" / „günstig"). Die Zuordnung der Klassen zu diesen drei Gruppen erfolgt durch die Schulaufsicht, die für diese Klassifikationsaufgabe die beste Expertise aufweist. Schulen sollen ihre Ergebnisse sowohl mit dem Gesamtdurchschnitt des Landes als auch mit dem Durchschnitt derjenigen Schulen, die ihrer Schule hinsichtlich des Einzugsgebietes am ähnlichsten sind, vergleichen können.

Mantelbogen für Schüler. Angaben zum Alter und Geschlecht der Kinder sowie zu ihrer Herkunftssprache sind erforderlich, um daraus Maße der Klassenzusammensetzung zu entwickeln. Ebenso wichtig ist die Angabe der Zeugnisnoten in Deutsch und Mathematik.

Mantelbogen für Lehrkräfte. Die an der Normierungsstudie teilnehmenden Lehrkräfte erhalten einen Kurzfragebogen, mit dem Ziel, das Verfahren der Normierung (Instruktionen, Gestaltung, Logistik) zu verbessern.

Auswertung. Die Ergebnisse der bearbeiteten Aufgaben sowie der Angaben im „Mantelbogen" (Schülerfragebogen) und im Lehrerfragebogen werden in den Schulen in einen von der Forschungsgruppe entwickelten, von einem Dokumentenscanner lesbaren Bogen eingetragen und an der Universität Landau zentral ausgewertet.

Datenbank. Die Ergebnisse der Normierungsstudien, insbesondere auch über Fehlertypen und -muster, werden in einer Datenbank gespeichert. Diese Ergebnisse stehen bei der Auswahl und vor allem bei der Interpretation der Ergebnisse der Vergleichsarbeiten im Rahmen der Selbstevaluation den Schulen zur Verfügung stehen und können für die pädagogische Arbeit genutzt werden. Dagegen werden schul- oder klassenbezogene Ergebnisse *nicht* an das Ministerium/die Schulaufsicht zurückgemeldet.

2.3 Vergleichsarbeiten

Planung. Diese werden – beginnend mit Mathematik (ab Herbst 2003), später zusätzlich mit Deutsch (ab Herbst 2004) – in jährlichem Turnus in allen 4. Klassen der Grundschulen in Rheinland-Pfalz durchgeführt. Für die Durchführung der Vergleichsarbeiten in den Schulen und für die

innerschulische Auswertung sind die Schulen selbst verantwortlich. Es wird jedoch empfohlen, dass Lehrkräfte nicht die Vergleichsarbeiten der eigenen Klassen korrigieren. Zur schulinternen Vorbereitung des Auswahlprozesses und der Durchführung sollen die Schulen (bzw. Fachgruppen innerhalb von Schulen) Gelegenheit zur inhaltlichen und organisatorischen Vorbereitung der Vergleichsarbeit haben.

Aktive Beteiligung des Kollegiums. Ein Eckpfeiler des Vergleichsarbeiten-Konzeptes ist sein partizipativer Charakter: die aktive Einbindung der Schulen (des Kollegiums, der Fachgruppe) in die *Auswahl der Aufgaben*: Jeweils die Hälfte der Aufgaben wird zentral vorgegeben, die andere Hälfte wird - auf der Basis bestimmter Kriterien, die gewährleisten, dass die verschiedenen Aufgaben- und Kompetenzbereiche abgedeckt sind - von den Schulen gewählt. Außerdem haben die Schulen die Möglichkeit, selbst *neue Aufgaben zu entwickeln* und für die jeweils folgende Normierungsstichprobe vorzuschlagen.

Transparenz und Qualität der Durchführung. Organisation, Durchführung und Auswertung der Vergleichsarbeiten werden auf der Grundlage von noch durch die Forschungsgruppe zu entwickelnden detaillierten Handreichungen und Instruktionen von den Schulen selbst geleistet. Die Schulaufsicht begleitet die Schulen und stellt die ordnungsgemäße Durchführung sicher. Zusätzlich erfolgt eine *formative Evaluation*, bei der in ausgewählten Klassen (Basis: Freiwilligkeit und Vorankündigung) die Durchführung beobachtet wird. Ziel ist die Optimierung des gesamten Verfahrens.

Das Konzept der Zentralstichprobe. Um über die Bestandsaufnahme der Leistungen auf der Ebene von Schulen und Klassen hinaus zentrale Trends beschreiben zu können, ist eine *Zentralstichprobe* vorgesehen. Das heißt: Zu jedem Durchführungstermin der Vergleichsarbeiten werden 60 zufällig ausgewählte Schulen (anders als die konstant bleibende Normierungsstichprobe wechselt die Zentralstichprobe von Durchführung zu Durchführung) untersucht, deren Ergebnisse eingeschickt und dann *zentral* (d.h. nicht schulspezifisch) analysiert werden. Schwerpunkt dieser Auswertung ist der Vergleich zwischen den vom Ministerium vorgegebenen und den von den Schulen selbst gewählten Aufgaben. Zugleich fließen die Ergebnisse dieser Schulen in die Weiterentwicklung der Aufgabensammlung ein.

Kein Schulranking. Ein Ranking von Schulen ist weder vorgesehen noch ist es - infolge der schulspezifischen Wahlmöglichkeiten - möglich. Anders dagegen auf *Klassenebene innerhalb von Schulen*: Da Wahlmöglichkeiten nur auf der Ebene der Schulen existieren und Parallelklassen immer identische Aufgabensätze bearbeiten, sind innerschulische Vergleiche und darauf basierende pädagogische und fachdidaktische Diskussionen nicht nur möglich, sondern ausdrücklich erwünscht.

3 Nutzung der Ergebnisse der Vergleichsarbeiten

3.1 Beratung der Eltern

Die Eltern werden von der Schule in geeigneter Weise über die individuellen Ergebnisse der Vergleichsarbeiten ihrer Kinder informiert. Eltern können auf diese Weise fundierter beraten werden, was die Schullaufbahn ihrer Kinder anbelangt: Die Vergleichsarbeit ist – neben den Noten und der Schullaufbahnempfehlung der Lehrkraft – eine wichtige *objektive Zusatzinformation*. Sie bietet für

jeden Schüler Vergleichsinformationen zum Leistungsstand auf der Klassen-, Schul- und Landesebene und kann so helfen, Über- wie Unterschätzungen des Leistungsniveaus der Kinder in Deutsch und Mathematik zu vermeiden. Zu welchen Verzerrungen es bei der Notengebung gelegentlich kommt, weiß man seit Ingenkamps Pionierarbeiten (Ingenkamp, 1971), und die neueren Evaluationsstudien, etwa LAU in Hamburg (Lehmann, Gänsfuß & Peek, 1999) haben dies erneut gezeigt. Insbesondere der Vergleich auf Landesebene stellt einzigartige Informationen bereit, die sehr viel weitergehende Analysemöglichkeiten erlauben als dies auf der Ebene der Einzelschule möglich ist. Als *alleinige* Entscheidungsgrundlage für die Grundschulempfehlung ist das Ergebnis einer Vergleichsarbeit aus methodischen und inhaltlichen Gründen jedoch unzureichend; es kann und soll lediglich *ergänzenden* Charakter haben.

3.2 Standard- und Qualitätssicherung

Es gibt eine Vielfalt von Möglichkeiten, die Ergebnisse der Vergleichsarbeiten für Zwecke der Qualitätssicherung, der Standardsicherung und für eine Bestandsaufnahme zu nutzen.

- Vergleich der Ergebnisse der Schule / der Klasse mit den Ergebnissen der *Normierungsstudie*, basierend auf den in dieser Schule eingesetzten Aufgaben: entweder mit dem Durchschnitt aller Grundschulen (Mittelwert) oder mit derjenigen Gruppe von Schulen, deren Einzugsgebiet dem eigenen am meisten ähnelt. Diese Vergleiche dienen der Standortbestimmung.
- Vergleich mit vorgegebenen Kriterien (insbesondere den in Entwicklung begriffenen *Bildungsstandards*, die in den künftigen Aufgabenpool einfließen werden). Insofern sind die Vergleichsarbeiten als wichtige Vorarbeiten auf dem Weg zu künftigen verbindlichen Bildungsstandards (Klieme, 2002) zu betrachten.
- Vergleich der Ergebnisse *paralleler Klassen* hinsichtlich Gesamtleistung, Profil der Stärken und Schwächen, Leistungsunterschiede innerhalb der Klasse, Fehlermustern, Vorkommen extrem schlechter und exzellenter Leistungen. Diese Daten sind wichtiger „Rohstoff“ für originär pädagogische und didaktische Überlegungen.
- Vergleiche der Ergebnisse der Zentralstichprobe im zeitlichen Verlauf. Dies erlaubt - ähnlich wie beim PISA- und TIMSS-Zyklus - die Schätzung von allgemeinen Trends im Hinblick auf die Leistungsfähigkeit von Grundschülern in Mathematik und Deutsch.

Internationale Verortung. Durch Verwendung von Aufgaben, die in internationalen Vergleichsstudien eingesetzt wurden und nach deren Publikation freigegeben werden, lassen sich auch internationale Vergleiche ziehen. So sind in den Aufgabenpool „Mathematik“ des VerA-Projektes auch zahlreiche bereits freigegebene Aufgaben der TIMSS/I-Studie (Grundschule) der IEA eingeflossen. Sobald die IEA die Daten der IGLU-Studie freigibt (dies erfolgt meist bereits ein Jahr nach der Publikation der Ergebnisse) können auch freigegebene IGLU-Aufgaben verwendet werden.

3.3 Pädagogische Interventionen und Verbesserung der Unterrichtsqualität

Vergleiche über die Zeit (Trends): Da die Vergleichsarbeiten jährlich wiederholt werden, ergeben sich – auf Schulebene wie auf der des Landes ("Zentralstichprobe") – die gleichen Möglichkeiten, die auf der Makro-Ebene auch beim PISA-Zyklus gegeben sind: die Feststellung von Trends. Zum

Beispiel: Haben unterrichtszentrierte Schulprogramme, Maßnahmen der Sprachförderung oder Unterrichtsentwicklungsprogramme „gegriffen“, und lässt sich der Erfolg an einer Verbesserung der relativen Position und des absoluten Niveaus auch empirisch belegen?
Sofern die praktische Durchführung und Auswertung bei den Schulen liegt, erfordert dies den Einsatz identischer Aufgaben, so dass diese Prüfung nicht im Rahmen des regulären Vergleichsarbeiten-Zyklus erfolgen kann[3]. Es wird empfohlen, die Vergleichsarbeit zu einem späteren Zeitpunkt im Schuljahr zu *wiederholen*, so dass Veränderungen im Hinblick auf einen kriterialen Maßstab verfügbar werden.

Förderung. Vor allem der Gesichtspunkt der individuellen Förderung spricht für eine Wiederholungsmessung: Die erste Welle der Kompetenzmessung kann die Grundlage für gezielte Förderung darstellen, und mit einer Wiederholungsmessung kann empirisch belegt werden, ob, bei wem und in welchem Ausmaß diese Förderung effektiv war. Diese Sichtweise entspricht dem Umdenken in der Bildungspolitik (Lange, 2001; Terhart, Czerwenka, Ehrlich, Jordan & Schmidt, 1994; Terhart, 2002): Weg von der Methode *irgendwelcher* Verbesserungen des Inputs in der Hoffnung, es werde sich schon *irgendwo irgendwann* eine Verbesserung einstellen – hin zu einer empirisch orientierten, wirkungszentrierten Sichtweise. Wenn eine Förderungsmaßnahme nachweislich und nachhaltig war, dann – und nur dann – war sie erfolgreich! Wenn nicht, muss man den Ursachen nachgehen und ggf. andere Maßnahmen probieren. Außerdem stellt eine Wiederholungsmessung eine Abkehr vom aus verschiedenen Gründen unzulänglichen statischen Prinzip der Einpunktmessung dar. Eine optionale Wiederholungsmessung kann also eine solide Datenbasis sein, um den tatsächlichen Erfolg von Förderungsmaßnahmen empirisch zu belegen.

3.4 Fachdidaktische Impulse und Diskussionen

Die Ergebnisse können Informationen über klassen- oder schulspezifische *Fehlermuster* (z.B. Häufung bestimmter Fehlertypen) liefern und damit Anlass für didaktische Diskussionen, veränderten Unterricht oder auch für eine kritische Analyse der verwendeten Lehrwerke sein: Die Lehrkräfte können die (falschen) Antworten in ihrer eigenen Klasse mit den in den Normierungsstudie gewonnen Antworten auf der Landesebene vergleichen und so die Frage beantworten: Wo haben die Schüler meiner Klassen Stärken, wo Schwächen? Welche Typen von Fehlern (z.B. Flüchtigkeitsfehler, systematische Fehler) kommen bei mir gar nicht, anders, oder häufiger vor als im Landesdurchschnitt? Woran könnte das liegen – am Lehrwerk, am Unterricht? Wo muss gefördert, wiederholt, neu erklärt werden, und wo zeigen sich Stärken, auf denen man aufbauen kann? Die Reflexion über Fehlertypen und -ursachen stellt einen wichtigen didaktischen Ansatz zur Verbesserung der *Aufgabenkultur* und *Fehlerkultur* an Schulen dar (Kötter, Struchholz, Niegemann & Auffenfeld, 1986; Weinert, 1999; Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (BLK), 1997).

---

[3] Dagegen erlaubt die von der Forschungsgruppe verwendete Methodologie (Stichwort: „Rasch-Modell“) solche Schätzungen auch auf der Ebene vergleichbarer (also nicht identischer) Aufgaben – dieses Verfahren dürfte für den autonomen Einsatz in Schulen jedoch zu komplex sein.

3.5 Erfassung und Verbesserung diagnostischer Kompetenzen

Maßnahmen der Förderung und Individualisierung setzen ein Mindestmaß an diagnostischer Kompetenz voraus. Das Thema der diagnostischen Kompetenzen von Lehrkräften wird in der Ausbildung so gut wie ignoriert, und die Forschung hat gezeigt, dass es hier erhebliche Defizite gibt (Schrader & Helmke, 1987; Helmke, 2003; Artelt, Stanat, Schneider & Schiefele, 2001; Schrader, 1997; Schrader & Helmke, 2001; Hosenfeld, Helmke & Schrader, 2002). Basierend auf den für die Vergleichsarbeiten ausgewählten Aufgaben lassen sich diagnostische Urteilsleistungen zu den folgenden Sachverhalten überprüfen und weiterentwickeln:

Unterschiede zwischen Aufgaben. Dazu ist eine Vorhersage nötig, welche Aufgaben für die Klasse insgesamt am schwersten und welche leicht sind. Dies erfordert eine Facette der Diagnosekompetenz, die eine stärker *fachdidaktische Komponente* hat. Hier muss überlegt werden, *warum* eine Aufgabe schwer ist, und in dieses Urteil gehen neben einer Rückschau auf die mutmaßlichen Vorkenntnisse und unterrichtlichen Lerngelegenheiten originär didaktische Überlegungen ein. Eine Mathematikaufgabe beispielsweise mag schwer sein, weil sie a) nicht authentisch ist, b) eine komplexe Mathematisierung erfordert, c) mehrere Schritte erfordert, d) ohne verfügbare Routinen zu bearbeiten ist, e) in einen komplexen Text eingebettet ist etc. Im schulinternen Austausch über solche Ergebnisse kann ein Potenzial für die Verbesserung des Unterrichts liegen. Ergebnisse von Vergleichsarbeiten können auf diese Weise auch Rohstoff für didaktisch orientierte schulinterne Lehrerfortbildung sein.

Interindividuelle Unterschiede. Lehrer können *vor* der Vergleichsarbeit 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 wird die Forschungsgruppe gezielte Hilfestellungen und Anregungen geben, so dass die diagnostische Sensibilität wirkungsvoll trainiert werden kann.

Lernfortschritte. Wird die empfohlene Mehrfacherhebung vorgenommen, dann ergibt sich nicht nur die Möglichkeit einer Diagnose des Lernfortschritts (s.o.), sondern auch die Möglichkeit einer *Diagnose der Diagnosefähigkeit*, d.h., ob sich die diagnostische "Trefferquote" verändert hat und wenn ja, in welcher Hinsicht.

3.6 Unterstützung der Implementation des Grundschulrahmenplans

In Rheinland-Pfalz wurde vor kurzem ein radikal geänderter, modernen pädagogischen und mathematikdidaktischen Forderungen entsprechender Rahmenplan für Mathematik in der Grundschule verabschiedet (Ministerium für Bildung, Frauen und Jugend Rheinland-Pfalz, 2002); entsprechende neue Rahmenpläne für andere Fächer werden folgen. Was üblicherweise gegen Testuntersuchungen eingewendet wird („teaching on the test“), kann hier im Gegenteil eine erwünschte und sehr positive Wirkung entfalten: die Akzeptanz der Inhalte des Rahmenplans, die sich in den Aufgaben der Vergleichsarbeiten widerspiegeln.

4 Formative Evaluation des Gesamtvorhabens

Die Lehrkräfte der Normierungsstudie erhalten regelmäßig einen kurzen Lehrerfragebogen. Ziel dieser flankierenden Befragung ist einerseits das Abstellen möglicher Schwachstellen der *Normierungsstudie* und der *Vergleichsarbeiten* selbst: Durchführung, Instruktionen, Ergebnisdarstellung, die Beurteilung des Nutzens für die Unterrichtsentwicklung und Vorschläge für Verbesserungen etc. Darüber hinaus werden die Unterstützungssysteme und die von der Forschungsgruppe entwickelten Handreichungen und Rückmeldungen beurteilt. Die Ergebnisse dieser formativen Evaluation werden sowohl bei den jeweils folgenden Normierungsstudien als auch bei der Vergleichsarbeit des jeweils kommenden Jahres berücksichtigt. Zudem haben alle Lehrkräfte die Möglichkeit, für die jeweils nächste Welle der Vergleichsarbeiten eigene begründete Aufgabenvorschläge zu machen, die (nach einer Vorauswahl durch die Expertengruppe) Bestandteil des Aufgabenpools und der folgenden Normierungsstudie II (Erweiterungen) werden können.

Zu allen pädagogischen Optionen erarbeitet die Forschungsgruppe wissenschaftlich begründete, praktikable und hinreichend erprobte (das heißt: vorgetestete) Verfahrensvorschläge und leistet Unterstützung (z.B. anhand von Beispielen, die passwortgeschützt ins Internet gestellt werden).

5 Perspektiven

Das bisher skizzierte Gesamtkonzept des VERA-Projektes ist aus zwei Gründen nur vorläufiger Natur:

(1) Es hat einen stark partizipativen Charakter, einschließlich der wichtigen und ernstgenommenen Komponente der regelmäßigen *formativen Evaluation*, so dass sich schon deshalb im Laufe der Zeit vielfältige Änderungen ergeben werden.

(2) Die KMK-Beschlüsse zur Entwicklung bundesweit *orientierender Bildungsstandards auch für die Grundschule* hat dazu geführt, dass es für dieses Projekt bereits jetzt eine ernsthafte Kooperationsabsicht mit Nordrhein-Westfalen gibt; mehrere andere Bundesländer haben ihr Interesse an dem hier vorgestellten Konzept bekundet. Dies kann und wird erhebliche Synergieeffekte haben. Aber es heißt natürlich auch, dass begründete Änderungs- oder Erweiterungsvorschläge der Partnerländer ernst genommen werden und mit Modifikationen des Konzeptes verbunden sein können.

Literatur

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.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K. J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69-140). Opladen: Leske + Budrich.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K. J. & Weiß, M. (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (BLK). (1997). *Gutachten zur Vorbereitung des Programms "Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts"* (Heft 60 der BLK-Reihe "Materialien zur Bildungsplanung und Forschungsförderung"). Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

Ditton, H., Arnoldt, B. & Bornemann, E. (2002). Entwicklung und Implementation eines extern unterstützenden Systems der Qualitätssicherung an Schulen - QuaSSu. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen. Zeitschrift für Pädagogik. 45. Beiheft* (S. 374-389). Weinheim: Beltz.

Heller, K. A. & Hany, E. A. (2001). Standardisierte Schulleistungsmessungen. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 87-102). Weinheim: Beltz.

Helmke, A. (2003). *Unterrichtsqualität: Erfassung, Bewertung, Verbesserung*. Velber: Kallmeyersche Verlagsbuchhandlung.

Helmke, A. & Jäger, R. S. (Hrsg.). (2002). *Die Studie MARKUS - Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext*. Landau: Verlag Empirische Pädagogik.

Hosenfeld, I., Helmke, A. & Schrader, F.-W. (2002). Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen. Zeitschrift für Pädagogik, 45. Beiheft* (S. 65-82). Weinheim: Beltz.

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

Klieme, E. (2002). Bildungsstandards als Beitrag zur Qualitätsentwicklung im Schulsystem. *DIPF informiert, 3*, 2-6.

Kötter, L., Struchholz, H., Niegemann, H. M. & Auffenfeld, A. (1986). Fehleranalytische Verfahren bei pädagogischen Diagnosen - Ansätze, Probleme, Perspektiven. In H. Petillon, J. Wagner & B. Wolf (Hrsg.), *Schülergerechte Diagnose. Theoretische und empirische Beiträge zur Pädagogischen Diagnostik. Festschrift zum 60. Geburtstag von Karlheinz Ingenkamp* (S. 89-114). Weinheim: Beltz.

Lange, H. (2001). Die bildungspolitische Bedeutung von Schulleistungsvergleichen. In G. Kaiser, N. Knoche, D. Lind & W. Zillmer (Hrsg.), *Leistungsvergleiche im Mathematikunterricht* (S. 1-28). Hildesheim: Franzbecker.

Lehmann, R. H. (2001). Messung von Schulleistungen im Primar- und Sekundarbereich. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 131-142). Weinheim: Beltz.

Lehmann, R. H., Gänsfuß, R. & Peek, R. (1999). Ergebnisse der Erhebung von Aspekten der Lernausgangslage und der Lernentwicklung - Klasse 7. *Hamburg macht Schule, 6*, S. 27-29.

Lehmann, R. H., Peek, R. & Gänsfuß, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen. Bericht über die Untersuchung im September 1996*. Hamburg: Behörde für Schule, Jugend und Berufsausbildung, Amt für Schule.

Mietzel, G. & Willenberg, H. (2000). *Hamburger Schulleistungstest für 4. und 5. Klassen*. Hogrefe: Göttingen.

Ministerium für Bildung, Frauen und Jugend Rheinland-Pfalz (Hrsg.). (2002). *Weiterentwicklung der Grundschule. Rahmenplan Grundschule. Allgemeine Grundlegung Teilrahmenplanung Mathematik*. Grünstadt: Sommer.

Ministerium für Schule und Weiterbildung, Wissenschaft und Forschung des Landes Nordrhein-Westfalen (Hrsg.). (2000). *Qualitätsentwicklung und Qualitätssicherung. Aufgabenbeispiele Klasse 7: Deutsch* (Materialien Schulentwicklung). Düsseldorf: Ministerium für Schule, Wissenschaft und Forschung des Landes Nordrhein-Westfalen.

Orth, G. (2002). Vergleichsarbeiten. In H.-G. Rolff & J. Schmidt (Hrsg.), *Schulaufsicht und Schulleitung in Deutschland*. Neuwied: Luchterhand.

Peek, R. (2001). Die Bedeutung vergleichender Schulleistungsmessungen für die Qualitätskontrolle und Qualitätsentwicklung von Schulen und Schulsystemen. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 323-336). Weinheim: Beltz.

Schrader, F.-W. (1997). Lern- und Leistungsdiagnostik im Unterricht. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (Enzyklopädie der Psychologie, Pädagogische Psychologie, Vol. 3, S. 659-699). Göttingen: Hogrefe.

Schrader, F.-W. & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik, 1*, 27-52.

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

Schrader, F.-W. & Helmke, A. (in Druck). Evaluation - und was danach? Ergebnisse der Rezeptionsstudie WALZER bei Schulleitungspersonen. *Schweizerische Zeitschrift für Bildungswissenschaften*.

Senn, W. (2002). *Beurteilen sprachlicher Kompetenzen mit Hilfe der Orientierungsarbeiten der Bildungsplanung Zentralschweiz*. Davos: Bildungsplanung Zentralschweiz.

Terhart, E. (2002). *Nach PISA*. Hamburg: Europäische Verlagsanstalt.

Terhart, E., Czerwenka, K., Ehrlich, K., Jordan, F. & Schmidt, H. J. (1994). Berufsbiographien von Lehrern und Lehrerinnen.

Thürmann, E. (2002). Unbekanntes Land. Wie Bildungsstandards das Lernen verbessern sollen. *forum schule, 3*, 22-25.

Weinert, F. E. (1999). Aus Fehlern lernen und Fehler vermeiden lernen. In W. Althof (Hrsg.), *Fehlerwelten. Vom Fehlermachen und Lernen aus Fehlern* (S. 101-109). Opladen: Leske + Budrich.