27.08.03

A. Helmke und I. Hosenfeld

Vergleichsarbeiten - Standards - Kompetenzstufen: Begriffliche Klärung und Perspektiven für VERA

- ENTWURF -

1. EINFÜHRUNG

Vergleichsarbeiten, wie sie das VERA-Projekt vorsieht, verfolgen simultan mehrere Ziele: von der Qualitätssicherung über die damit verknüpfte angestrebte Verbesserung der Unterrichtsqualität sowie diagnostischer Lehrerkompetenzen, eine fundiertere Schullaufbahnberatung der Eltern bis hin zu positiven Nebenwirkungen wie der beschleunigten Entwicklung von IT-Kompetenzen (wegen der unumgänglichen Nutzung des Internet) und der erleichterten Durchsetzung innovativer Rahmenpläne und Kerncurricula.

Im folgenden geht es nur um das erstgenannte und zentrale Ziel der Vergleichsarbeiten: um die fundierte jährliche Erfassung von Schülerkompetenzen im Bereich der deutschen Sprache und im Fach Mathematik. Angesichts der demnächst zu erwartenden verbindlichen nationalen Bildungsstandards der KMK auch für die Fächer Deutsch und Mathematik in der Grundschule besteht seitens der Bildungspolitik die berechtigte Erwartung, dass sich die Vergleichsarbeiten (obwohl sie als Erhebungszeitpunkt den *Beginn* der 4. Klassenstufe anstreben, während sich die Bildungsstandards auf das *Ende* der 4. Klasse beziehen) an den künftigen Nationalen Standards *orientieren*. Wie eine solche Orientierung aussehen kann, wie Bildungsstandards mit Kompetenzmodellen und -stufen zusammenhängen, und was daraus für die Planung und Auswertungsstrategie des VERA-Projektes folgt, bildet den Gegenstand dieses Textes.

2. BEZUGSNORMEN DER LEISTUNGSBEWERTUNG

Im Interesse einer begrifflichen Klärung und zur Vermeidung nahe liegender Missverständnisse ist es wichtig daran zu erinnern, dass die Bestandsaufnahme - die Feststellung und Sicherung der Wirkungen des Grundschulunterrichts - facettenreich ist, dass sie auf mehreren Ebenen erfolgen kann und dass verschiedene, sich ergänzende Maßstäbe und Interpretationsperspektiven zur Verfügung stehen. Dies sei im Folgenden stichwortartig umrissen.

Hierarchie von Ebenen. Die Bewertung der in Vergleichsarbeiten gezeigten Leistungen kann sich auf die Zuordnung von Testwerten und Kompetenzstufen zu einzelnen Schülern, Klassen, Schulen, Regionen und Bundesländern beziehen.

Verschiedene Gegenstände. Sie ist nicht auf Maße der zentralen Tendenz (wie Durchschnitts- und Mittelwerte, Mediane) beschränkt, sondern Gegenstand der Evaluation wird auch die Bandbreite, Streuung oder Verteilung von Leistungen, z.B. innerhalb einer Klasse, Schule oder im Schulsystem sein.

Scores, Items, Verteilungsinformationen, Kompetenzstufen. Je nach Fragestellung bezieht sich die Beschreibung des Leistungsstandes auf *Testscores* (z.B. für Inhaltsbereiche wie das Rechtschreiben), die sich aus der Bearbeitung meh-

rerer Items zu einem Inhaltsbereich berechnen lassen, oder auf einzelne *Items*. Prinzipbedingt besitzen Informationen, die über mehrere Items hinweg gewonnen wurden, eine höhere Zuverlässigkeit, d.h. zur möglichst exakten Schätzung vorhandener Kompetenzen sollten Testscores herangezogen werden. Für die Lokalisation von Defiziten bzw. die Identifikation von Problembereichen ist es jedoch sinnvoll, die begangenen Fehler itemweise zu analysieren. Außerdem lassen sich Testscores aus verschiedenen Perspektiven Interpretieren: Es können einerseits Aussagen über die bewältigten Anforderungen gemacht werden, d.h. es erfolgt eine Zuordnung zu inhaltlich definierten *Kompetenzstufen*. Andererseits lassen sich Testscores auch in Relation zur Vergleichspopulation betrachten, d.h. es erfolgt eine *Einordnung in die Gesamtverteilung* (z.B. anhand eines Prozentrangplatzes oder anhand der Distanz zum Gesamtmittelwert in Standardabweichungseinheiten).

Klassische und probabilistische Testtheorie (sog. Rasch-Skalierung, Item Response Theory). Traditionell wurde - und wird auch heute noch überwiegend - bei der Bildung von Fragebogen- oder Testskalen die sog. klassische Testtheorie zugrunde gelegt, deren Grundannahme ist, dass sich jeder Messwert aus einem „wahren“ Wert und einem zufälligen Messfehler zusammensetzt, wobei die Messfehler verschiedener Messungen voneinander unabhängig sind. Diese sinnvollen und plausiblen Annahmen spezifizieren, vereinfacht ausgedrückt, wichtige Annahmen über den Zusammenhang mehrerer Items und dem zu messenden Konstrukt (Leistung). Die probabilistische Testtheorie macht ergänzende sinnvolle und plausible Annahmen über den Zusammenhang zwischen dem zu messenden Konstrukt und jedem einzelnen Item, indem die Wahrscheinlichkeit eine Aufgabe zu lösen als Funktion der (latenten) Fähigkeit postuliert bzw. errechnet wird (für eine ausführlichere Darstellung s. Rost, 1996). Ein wesentlicher Vorteil dieser Modelle ist, dass die Aufgabenschwierigkeiten und die Personenfähigkeiten auf der selben Dimension und mit der selben Skalierung abgebildet werden, so dass Aufgabenschwierigkeit und Personenfähigkeit unmittelbar aufeinander bezogen werden können: Ist die Personenfähigkeit höher als die Aufgabenschwierigkeit, so ist es wahrscheinlicher, dass die Person die Aufgabe löst als dass sie die Aufgabe nicht löst. Ist hingegen die Personenfähigkeit niedriger als die Aufgabenschwierigkeit, so ist es wahrscheinlicher, dass die Aufgabe nicht gelöst werden kann. Dabei gilt: je größer der Differenzbetrag zwischen Aufgabenschwierigkeit und Personenfähigkeit, desto höher ist die Wahrscheinlichkeit die Aufgabe zu lösen bzw. an ihr zu scheitern. Aufgrund dieser Eigenschaften ist es möglich, unterschiedliche Bereiche (Zonen, Stufen, Levels) auf dem Leistungskontinuum zu unterscheiden, die eine hierarchische Struktur aufweisen: Wer die Items einer mittleren Schwierigkeitsstufe meistert, der löst mit hoher Wahrscheinlichkeit auch die Items einer darunter liegenden Schwierigkeitsstufe. Dies ist der Kern der Logik des mit TIMSS begonnenen und seit PISA weltweit zum "state of the art" gewordenen "proficiency scaling" im Rahmen von "large scale studies": die Zuordnung von Kompetenzstufen zu verschiedenen Abschnitten auf dem Leistungskontinuum.

Drei Maßstäbe. Die bei Vergleichsarbeiten resultierenden Ergebnisse - z.B. diejenigen einer einzelnen Schule - können an verschiedenen Normen und Kriterien gemessen werden.

a) verteilungsorientiert:

Absolutwerte: Wie gut ist das Leistungsergebnis einer konkreten Schule, verglichen mit dem Durchschnittswert aller Schulen (in diesem Lande, in allen beteiligten Bundesländern)?

Adjustierte Werte: Wie gut ist das Leistungsergebnis dieser Schule, gemessen an den objektiven, von Lehrkräften nicht beinflussbaren Rahmenbedingungen des Lehrens und Lernens (Einzugsgebiet, Klassenzusammensetzung)? Hierbei werden die rohen Werte entsprechend korrigiert (s. Projekt MARKUS; Helmke & Jäger, 2002). Der gleichen Logik folgend kann anstelle einer Adjustierung des Leistungsergebnisses auch ein Vergleich mit den Leistungsergebnissen vergleichbarer Schulen erfolgen (so die PISA-Rückmeldung; Watermann, Stanat, Kunter, Klieme & Baumert, 2003).

b) kriteriumsorientiert:

Früher sprach man von der Orientierung an Lernzielen, in der aktuellen bildungspolitischen Diskussion haben die nationalen Standards diese Rolle übernommen. Standards sind nichts anderes als Kompetenzerwartungen. Die Frage ist hier: Wie viele Schüler/innen zeigen in den Vergleichsarbeiten die Kompetenzen, die von ihnen erwartet werden? Diese scheinbar einfache Frage wird dadurch kompliziert, dass es zur Zeit noch divergierende Vorstellungen darüber gibt, welcher Typ von Standard und welchen Geltungsanspruch / welche Reichweite zugrunde gelegt werden soll:

- *Minimalstandards* beziehen sich auf ein definiertes Minimum an Kompetenzen, die buchstäblich alle Schüler/innen nach 4 Jahren Grundschule erreicht haben müssen, und deren Nichterreichen mit der Erwartung gravierender Schwierigkeiten im weiteren Unterricht und evtl. dem Anspruch auf Förderung gekoppelt ist.

- *Regelstandards*: Kompetenzen, die im "Durchschnitt", "in der Regel" von den Schülern einer Klasse erreicht werden sollen. Die Präzisierung - Bestimmung einer klaren Schwelle oder Grenze - steht seitens der KMK noch aus. In früheren Large-Scale-Studien der IEA hat man hierfür das Konzept der "steering group" herangezogen: Orientierung an einer 80%-Quote, d.h. 4/5 der Klasse (resp. Schule) sollen das gesetzte Ziel erreichen oder übertreffen.

c) verlaufsorientiert

Die dritte und aus pädagogischer Sicht wichtigste Dimension ist die des zeitlichen Verlaufs: Gemessen an einem definierten Startpunkt: Wo und wieviel hat die Schülerschaft einer Schule in einer bestimmten Zeit dazu gelernt, wie hat sich das Profil von Stärken und Schwächen, haben sich Fehlermuster quantitativ und qualitativ geändert? Die Analyse der Veränderungen kann unterschiedliche Schwerpunkte besitzen:

Längsschnitt: Das Leistungsprofil einer Schule zu Beginn der 4. Klasse wird am Ende der 4. Klasse mit teilweise identischen Items wiederholt gemessen, so dass Beschreibungen - und günstigenfalls auch Erklärungen - von Veränderungen möglich sind. Dieses Design liegt z.B. den Follow-Up-Studien der deutschen TIMS-Studie, DESI, PISA2003 zugrunde.

Intervention: Die längsschnittliche Erhebung des Leistungsprofils zu Beginn und (beispielsweise) am Ende der 4. Klasse wird um eine gezielte Interven-

tion im Verlauf der 4. Klasse ergänzt: (Förderprogramm, Unterrichtsexperiment, etc.). Dies ermöglicht die Beantwortung der Frage, welche Schüler in welchem Ausmaß von solchen Maßnahmen profitieren.

Zeitwandel: Die Leistungen von 4. Klassen werden im jährlichen Abstand immer wieder erhoben, jedes Mal handelt es sich um einen anderen Altersjahrgang (Kohorte). Diese Logik liegt den Zyklen der Studien TIMSS, PIRLS, IGLU zugrunde.

Diese drei Strategien schließen sich nicht wechselseitig aus, sondern lassen sich in vielfältiger Weise miteinander kombinieren: Zum Beispiel könnte es darum gehen, ob sich der Prozentsatz von Schüler/innen, die bei der Vergleichsarbeit 2004 minimale Standards in Mathematik erreichen (*kriteriumsorientiert*), ein Jahr später vergrößert oder verringert (*verlaufsorientiert*), ob dieser Wandel global ist oder ob er in verschiedenen Regionen (oder: in Abhängigkeit von unterschiedlichen Förderkonzepten) eine unterschiedliche Gestalt aufweist (*vergleichsorientiert*), und ob sich bei Inrechnungstellung der vorgefundenen Rahmenbedingungen (wie Einzugsgebiet der Schule, Klassenzusammensetzung) ein anderes Bild ergibt (*adjustiert*).

3. KOMPETENZMODELLE UND KOMPETENZSTUFEN

Es besteht heute völliger Konsens darüber, dass es für die Bewertung des Niveaus von Schülerleistungen und für die Ergebnisrückmeldung an die beteiligten Schulen und Klassen nicht mehr ausreicht, lediglich verteilungsorientierte Aussagen zu machen, und seien diese auch noch so detailliert und präzise. MARKUS dürfte die letzte große Studie gewesen sein, die noch so verfahren hat. Spätestens seit PISA 2000 und IGLU werden heute Aussagen zu inhaltlich definierten Abstufungen von Kompetenzen erwartet. Diese Tendenz wurde durch den Tenor der Expertise zu den Bildungsstandards (Klieme, Avenarius, Blum, Döbrich, Gruber, Prenzel, Reiss, Riquarts, Rost, Tenorth & Vollmer, 2003) noch verstärkt.

Klieme schreibt in der Standards-Expertise:

> "Bildungsstandards ... stützen sich auf Kompetenzmodelle, die in Zusammenarbeit von Pädagogik, Psychologie und Fachdidaktik entwickelt werden müssen. Ein solches Kompetenzmodell unterscheidet *Teildimensionen* innerhalb einer Domäne (also z.B. Rezeption und Produktion von Texten, mündlichen und schriftlichen Sprachgebrauch), und es beschreibt jeweils unterschiedliche *Niveaustufen* auf solchen Dimensionen. Jede Kompetenzstufe ist durch kognitive Prozesse und Handlungen von bestimmter Qualität spezifiziert, die Schüler auf dieser Stufe bewältigen können, nicht aber Schüler auf niedrigeren Stufen ... Vor allem die Fachdidaktik ist gefragt, wenn es festzulegen gilt, welche Anforderungen zumutbar und begründbar sind ... Kompetenzmodelle sollten auch Aussagen darüber machen, in welchen Kontexten, bei welchen Altersstufen und unter welchen Einflüssen sich die einzelnen Kompetenzbereiche entwickeln. Nur so kann von der Schule erwartet werden, dass sie mit geeigneten Maßnahmen zur systematischen Kompetenzentwicklung, zum kumulativen Lernen beiträgt" (S. 15/16).

Aber selbst wenn Kompetenzmodelle zur Verfügung stünden: Von solchen Modellen zur Entwicklung konkreter Testitems ist es ein weiter Weg:

> "Entsprechende Testaufgaben können allerdings nicht einfach aus den Kompetenzbeschreibungen "abgeleitet" werden. Sie müssen generiert und auf ihre Validität hin geprüft werden" (S. 16)

Das von Klieme skizzierte Lastenheft für die Entwicklung wissenschaftlich fundierter Kompetenzmodelle ist umfassend. Stellt man in Rechnung, dass neben den Beiträgen der Fachdidaktik, Pädagogik, Kognitions- und Entwicklungspsychologie insbesondere auch methodische Expertise unabdingbar erscheint, dann wird klar: Es handelt sich bei der Entwicklung fachspezifischer Kompetenzmodelle um ein umfangreiches interdisziplinäres Forschungsprogramm, bei dem wir zurzeit erst am Anfang stehen. Angesichts dringlicher und unabweisbarer Forderungen der Bildungspolitik plädieren wir für ein mehrgleisiges Verfahren: Zum einen muss natürlich die Entwicklung von Kompetenzmodellen vorangetrieben werden, hier stimmen wir Klieme vollkommen zu. Zum anderen ist es jedoch nötig und möglich, auch dann zu vorläufigen Aussagen zu Kompetenzstufen zu gelangen, wenn noch kein umfassendes und bewährtes Kompetenzmodell entwickelt ist. Für diese Strategie spricht auch, dass die Entwicklung von Testaufgaben auf der Basis von Standards (Kompetenzerwartungen) und Kompetenzmodellen keine Einbahnstraße ist, sondern ein Kreisprozess:

a) Kompetenzmodelle steuern und erleichtern die Konstruktion von Aufgaben, z.B. durch die gezielte Variation schwierigkeitsbestimmender Merkmale. Die damit verbundenen Annahmen werden empirisch getestet.

b) Die Ergebnisse dieser Tests - bei VERA z.B. die Ergebnisse von Normierungs- und Zentralstichproben - haben jedoch möglicherweise massive Rückwirkungen, und zwar in mehrfacher Weise: Zum einen tragen sie zur Präzisierung der theoretischen Annahmen zur Struktur der zugrunde liegenden Fähigkeiten bei. Zum anderen können die Ergebnisse von Vergleichsstudien Hinweise darauf geben, wie realitätsangemessen und fair bestimmte Kompetenzerwartungen sind. Standards und damit verbundene Kompetenzerwartungen sind keine Dogmen, die ausschließlich am grünen Tisch produziert werden, sondern müssen sich selbst auf den empirischen Prüfstand stellen lassen.

Diesen Kreisprozess kann grafisch wie folgt repräsentiert werden:

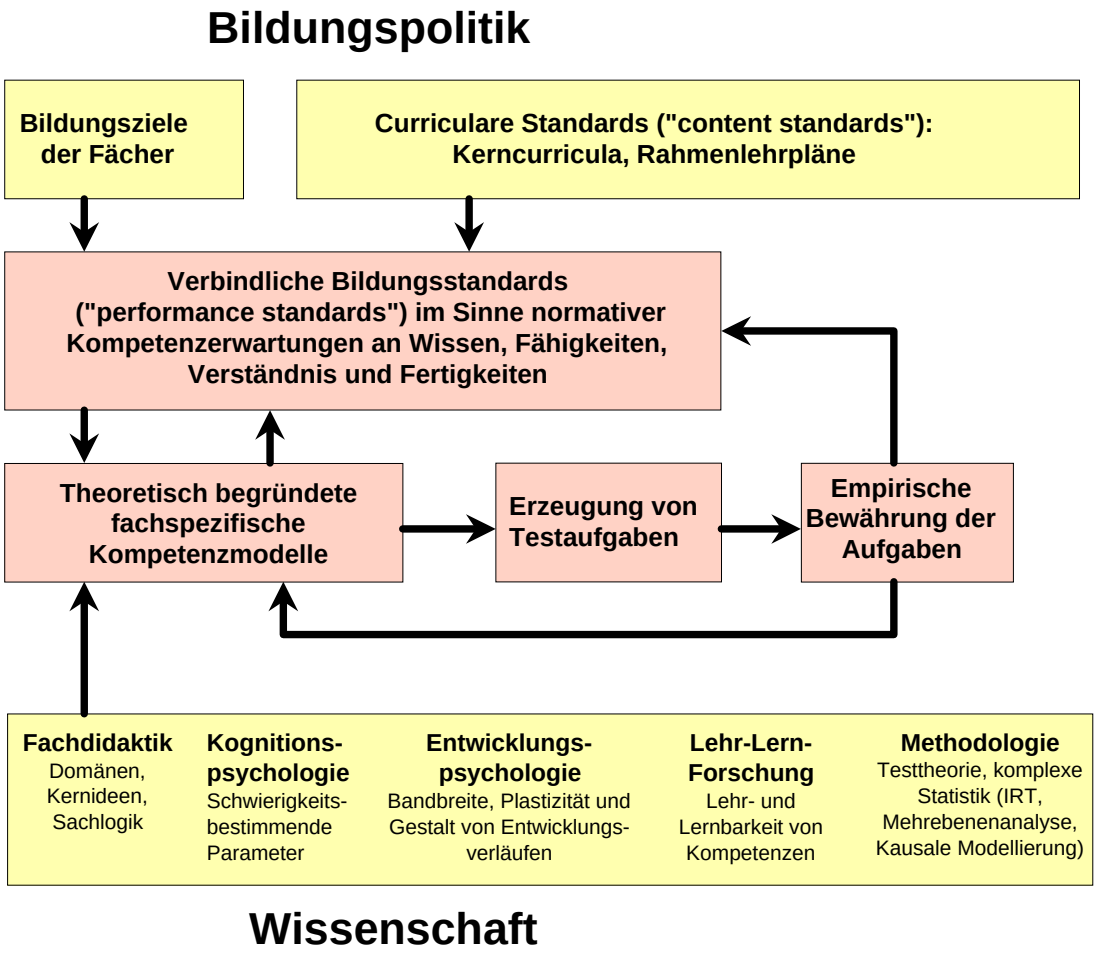

Abbildung 1: Ein Rahmenmodell zur Veranschaulichung des Kreisprozesses von Kompetenzmodellen und empirischen Tests

Über diesen grundlegenden Konsens hinaus gibt es jedoch zur Zeit eine große Ungewissheit darüber, a) auf welchem Wege man zu Kompetenzmodellen gelangt und b) wie man auf dieser Basis Niveaustufen definiert. Warum?

Notwendigkeit aktiver Konstruktionsleistungen. Entgegen verbreiteten Vorstellungen gibt es für die Entwicklung von Kompetenzmodellen und die darauf basierende Differenzierung von Niveaustufen - anders als für die "Rasch-Skalierung", die der Bildung von Stufen voraus geht - keine Software, kein "Programm". Kompetenzstufen sind nicht irgendwo "vorhanden", um nur entdeckt oder gefunden zu werden, sondern sie werden konstruiert. Die sachgerechte und methodisch akzeptable Durchführung der damit verbunden Prüf- und Auswertungsschritte ist ein schwieriger und langwieriger Prozess - wie unsere eigenen mehrjährigen Erfahrungen im Projekt DESI der KMK zeigen.

Mehrere Wege zu Kompetenzstufen. Gelegentlich wird propagiert und suggeriert, es gebe nur einen richtigen gangbaren Weg zu Kompetenzmodellen und -stufen, verbunden mit monopolistischen Bestrebungen. Dies ist keineswegs der Fall, wie ein Blick auf die in den Projekten TIMSS-II, TIMSS-III, PISA 2000 und DESI verwendeten sehr unterschiedlichen Prozeduren zeigt (vgl. Klieme, 2003).

- Bei *TIMSS-II* (Baumert, Lehmann, Lehrke, Schmitz, Clausen, Hosenfeld, Köller & Neubrand, 1997) wurden die Schwellen zwischen den Kompetenzstufen nachträglich („post hoc“) anhand der skalierten Fähigkeits-/Schwierigkeitsdimension gesetzt. Anschließend wurde die inhaltliche Interpretation und Benennung der Stufen vorgenommen, indem aus der „holistischen“ Zusammenschau *aller* Aufgaben der entsprechenden Schwierigkeitszone der Kern gemeinsamer Anforderungen beschrieben wurde. (Dieser Prozess kann in etwa analog zur Interpretation eines Faktors in einer Faktorenanalyse verstanden werden, wobei jedoch alle auf den Faktor ladenden Items miteinbezogen werden). Dieses Verfahren ist unmittelbar plausibel, weist aber auch Schwächen auf: Es besteht die Gefahr eines atheoretischen Vorgehens und der Beliebigkeit oder Subjektivität bei der Benennung der Stufen.

- Bei *TIMSS-III* (Baumert, Bos & Lehmann, 2000b, 2000a) wurde prinzipiell der gleiche Weg beschritten, d.h. auch hier wurden die Schwellen gesetzt und die Interpretation aus dem „Inhalt“ entsprechender Items extrahiert. Der entscheidende Unterschied im gewählten Vorgehen war, dass zur Interpretation nicht alle Aufgaben des Bereichs herangezogen wurden, sondern nur diejenigen, deren Lösungswahrscheinlichkeiten zwischen den Stufen deutlich variierten. (Im Analogon zur Faktorenanalyse: Die Faktorinterpretation findet nur anhand von Markieritems statt, die erstens sehr hohe Ladungen auf den jeweiligen Faktor und zweitens keine Nebenladungen auf weiteren Faktoren besitzen).

- Bei *PISA* 2000 wurden die Schwellen ebenfalls gesetzt, die Interpretation erfolgte jedoch auf der Grundlage von systematischen Analysen der Anforderungsmerkmale (inhaltliche Zuordnung + schwierigkeitsbestimmende Faktoren), die noch vor dem Vorliegen empirischer Daten von Experten eingeschätzt wurden.

Allen drei Studien ist gemein, dass die Kompetenzstufen erst *nach* dem Vorliegen der empirischen Daten gebildet wurden, da die Itementwicklung nicht auf der Basis eines a priori gegeben Kompetenzmodells erfolgte. Die so erzielten Ergebnisse können als robust bezeichnet werden, die Unterschiede zwischen den drei Verfahren erbringen jedoch nicht unmittelbar und per se Qualitätsunterschiede im Ergebnis sondern zielen eher auf eine einfachere Durchführung.

- Bei *DESI* erfolgte bereits die Aufgabenentwicklung anhand von Kompetenzmodellen. Dabei zeigte sich jedoch, dass zwischen den theoretischen Vorstellungen der Sprachdidaktiker einerseits und den Erfordernissen der Psychometrie andererseits eine große Kluft bestand. Sehr viel Zeit, viele Diskussionen und empirische Pilotierungen waren nötig, um eine gemeinsame Sprache zu finden und um fachspezifische Kompetenzmodelle zu Teilleistungen im Bereich der Beherrschung der deutschen und englischen Sprache zu entwickeln.

Qualitativ unterschiedliche Stufen vs. Kontinuum? Eines der Probleme der Kompetenzmodelle ist das der Niveau-"Stufe". Mit dem Konzept "Stufe" verbindet man im Deutschen allgemeinen die Vorstellung von qualitativ unterschiedlichen Zuständen. In der Entwicklungspsychologie spricht man beispielsweise von Stufen nur dann, wenn auf grund neuartiger Einsichten und (z.B. hirnphysiologisch bedingter) kognitiver Leistungsmöglichkeiten plötzlich Vorgänge und

Konzepte verstanden werden, die vorher unverständlich waren (z.B. dass ein Gegenstand auch dann weiter existiert, wenn er der Wahrnehmung entzogen ist - Leistung der Objektpermanenz, siehe Piaget). Bei vielen kognitiven Leistungen ist dieser Stufencharakter allerdings nicht geprüft - und in manchen Fällen, wie etwa beim Wortschatz, ist die Annahme von qualitativen Sprüngen, von diskreten Stufen oder Plateaus ohnehin nicht plausibel. Wenn man von "Abstufungen" oder „Zonen“ statt von "Stufen" der Kompetenz spräche (wie man "level of competence" ebenfalls übersetzen könnte), wäre das aus unserer Sicht angemessener. Dafür spricht auch die Tatsache, dass die Modellbildung nicht so strikt ist, wie es das Wort „Stufe“ vermuten lässt: es handelt sich ja nicht um eine deterministisches sondern um ein probabilistisches Modell. Dies bedeutet, dass Schülerinnen und Schüler, die einer bestimmten Kompetenzstufe zugeordnet werden, die entsprechenden Aufgaben mit „hinreichender Sicherheit“ lösen können, wobei im Regelfall eine Lösungswahrscheinlichkeit von .65 als Schwellenwert definiert wird. Wir plädieren daher hier für eine entspanntere und liberale Sichtweise: Wo es inhaltlich überzeugend und empirisch belegbar ist, ist das Stufenkonzept hilfreich. Wo es sich um kontinuierlich verteilte Merkmale handelt, lassen sich gleichwohl Abschnitte oder Zonen unterschiedlicher Ausprägung definieren, ohne dass damit die Fiktion von Stufen verbunden wäre.

Konsens Fachdidaktik - Psychologie? Im Gegensatz zur Didaktik der Mathematik und der Naturwissenschaften ist man in der Fachdidaktik des Deutschen und des Englischen - nicht nur in Deutschland - noch weit von einer entfalteten Diskussion über Kompetenzmodelle, geschweige denn von einem Konsens darüber, entfernt. Dies liegt teilweise an der Empiriefremdheit dieser Fachdidaktiken, aber auch an inhaltlichen Divergenzen und daran, dass verschiedene Fachsprachen die Kommunikation erschweren.

4. VERGLEICHSARBEITEN VS. STANDARDS-TESTUNG

Neben den angesprochenen theoretischen und methodischen Punkten ist ein weiterer Aspekt zu beachten: der Unterschied zwischen den geplanten Vergleichsarbeiten und den künftigen Überprüfungen der Standards. Gelegentlich wird beides gleich gesetzt - dies wäre jedoch aus mehreren Gründen fehlerhaft:

Anfang vs. Ende der 4. Klassenstufe. Die KMK-Standards beziehen sich ganz klar auf Abschlüsse und auf Gelenkstellen des Bildungswesens, im Bereich der Grundschule also auf das, was Grundschüler/innen nach 4 Jahren Grundschule wissen, können und verstehen sollen. Demgegenüber werden die Vergleichsarbeiten bei VERA eher gegen Anfang der vierten Klasse stattfinden. Zwischen beiden Zeitpunkten liegt ein knappes Schuljahr, und in dem aus neurobiologischer, lernpsychologischer und unterrichtlicher Sicht viel passiert. Vergleichsarbeiten können daher nicht beanspruchen, den Grad der Erreichung "der Standards" (der Grundschule) zu *erfassen*. Sie können sich jedoch auf mehrfache Weise an diesen Standards *orientieren* - dazu unten mehr.

Flächendeckende Gesamterhebung vs. Stichprobenuntersuchung. Wegen der für VERA charakteristischen Verknüpfung von Evaluation und Unterrichtsentwicklung ist das Konzept der Vergleichsarbeiten per definitionem auf die Durchführung in der Fläche gerichtet. Dagegen wäre es für die Überprüfung der erreichten Standards Zeit- und Ressourcenverschwendung, eine Totalerhebung durchzuführen; hier reichen repräsentative Stichproben vollkommen aus.

Schulweise identische Aufgabensätze vs. Aufgabenrotation. Die Überprüfung von Standards wird sich - wie dies auch PISA, IGLU und TIMSS getan haben und wie es DESI tun wird - der bewährten Methode der Aufgabenrotation bedienen ("multi-matrix-sampling"). Auf diese Weise ist die Bandbreite der überprüfbaren Leistungen um ein vielfaches höher - dagegen sind die Schüler innerhalb einer Klasse und sind die Klassen innerhalb einer Schule dann nicht mehr unmittelbar vergleichbar. Dies ist für den Zweck der Standards-Überprüfung im Sinne des system monitoring unerheblich. Dagegen erfordert es das Ziel des VERA-Projektes (Anregung innerschulischer didaktischer und diagnostischer Diskussion und Vergleiche), pro Schule unbedingt die jeweils gleichen Aufgabensätze zu bearbeiten.

5. STRATEGIEN DER ENTWICKLUNG VON KOMPETENZSTUFEN BEI VERA

Als Konsequenzen aus dem bisher Dargestellten ergibt sich für das Vorgehen von VERA folgendes Arbeitsprogramm:

Statistische Überprüfung der Fähigkeitsstrukturen in den Fächern Mathematik und Deutsch. Wie stark überlappen sich die Fähigkeiten unterschiedlicher Teilbereiche (in Mathematik z.B. in Geometrie, Arithmetik, Sachrechnen oder in der Muttersprache in den Bereichen Schreiben, Leseverstehen und Sprachreflexion)? Lohnt es sich, oder ist es sogar erforderlich, anstelle eines globalen Kompetenzmodells "Beherrschung der deutschen Sprache" unterschiedliche Teilkompetenzen zu erfassen? Solche Prüfungen der Dimensionalität von Fähigkeiten werden üblicherweise mit dem statistischen Werkzeug der konfirmatorischen Faktorenanalyse durchgeführt (so auch bei PISA und IGLU), für das inzwischen vielfältige Software (LISREL, EQS, AMOS, MPLUS) zur Verfügung steht. Allerdings setzen diese Programme relativ vollständige Datensätze voraus, die bei der Normierung (Aufgabenrotation) und der Zentralstichprobe (verschiedene Aufgabensätze) naturgemäß nicht gegeben ist. Für die zehn zentral vorgegebenen Aufgaben stellt diese Technologie das Mittel der Wahl für die Analyse der Zentralstichprobendaten dar. Die von uns bevorzugte, alternative Strategie ist die simultane Rasch-Skalierung (mit CONQUEST) mehrerer Dimensionen, wobei die Korrelation zwischen den latenten Dimensionen als Maß der Ein- oder Mehrdimensionalität der verschiedenen Teilfähigkeiten berechnet werden kann.

Statistische Überprüfung der Rasch-Homogenität der Leistungsdaten. Ziel ist die Entwicklung von homogenen Skalen, die die statistischen Voraussetzungen des Rasch-Modells erfüllen. Hier werden wir pragmatische Entscheidungen treffen, d.h. auch bei kleineren unmaßgeblichen Verletzungen von Voraussetzungen für die Rasch-Skalierung werden wir diesen Weg so weit gehen müssen wie möglich, auch wenn dies aus der Sicht von Puristen & Dogmatikern angreifbar sein mag. Nur so lassen sich überhaupt Kompetenzstufen entwickeln. Diese sind nicht nur für eine inhaltlich fundierte Berichterstattung auf der Ebene des Systems nötig, sondern auch für die Klassifikation von Schulen, Klassen und Schülern (Verteilung auf verschiedene Kompetenzstufen auf allen drei Ebenen).

Fortlaufende Ergänzung des Itempools (Normierung). Die fortlaufende „Renovierung“ des Itempools dient verschiedenen Zwecken. Neben der Vermeidung unerwünschter Lern- oder Trainingseffekte und der Berücksichtigung der Item-

vorschläge aus den Fachkonferenzen ist das Ziel insbesondere eine fortlaufende Verbesserung der Passung zwischen Nationalen Standards und zuordbaren Aufgaben in den Vergleichsarbeiten.

Experten-Ratings der Ausprägung schwierigkeitsbestimmender Merkmale. Dazu liegen zurzeit folgende Daten vor:

- Für Mathematik: Aufgabenkontext, Länge des Aufgabentextes, Sprachliche Komplexität, Menge erforderlicher Lösungsschritte sowie Anzahl der zu berücksichtigenden Parameter, jeweils in drei Stufungen.
- Für Deutsch ergeben sie je nach Bereich unterschiedliche Ratings, die Texte werden z.B. nach relativer Kürze, Gliederung/Ordnung, Motivierungsqualität, syntaktischer Komplexität und Vertrautheit mit dem Wortschatz beurteilt (ebenfalls in je drei Stufungen).

Die Validität dieser Merkmale (Zusammenhänge mit den realen Schwierigkeiten) kann für Mathematik anhand der Daten der Zentralstichprobe 2003 überprüft werden. Aufgrund weiterer theoretischer Überlegungen, gemeinsam mit den Fachdidaktikern, und ihrer empirischen Bewährung, wird der Satz dieser schwierigkeitsbestimmender Merkmale zunehmend präzisiert. Er dient auch als Werkzeug für die Entwicklung künftiger Items. Dies soll (u.a.) die Anschlussfähigkeit des VERA-Vorgehens an die von Klieme empfohlene Methode sichern, von Kompetenzmodellen auszugehen ("Kompetenzstufen sind ein zentrales Hilfsmittel für die Konstruktion von Aufgaben ... mit unterschiedlichem Schwierigkeitsniveau", Standards-Expertise, S. 17)

Die schwierigkeitsbestimmenden Merkmale spielen in VERA auch bei der Bildung von Kompetenzstufen eine wichtige Rolle (s.o. die Strategie bei PISA 2000). Die Berücksichtigung derjenigen Merkmale, die nachweislich mit der empirischen Aufgabenschwierigkeit zusammenhängen, erlaubt eine inhaltlich klarere Interpretation der Kompetenzstufen, da z.B. zwischen rein inhaltlichen und formal-gestalterischen Anforderungen der Aufgaben unterschieden werden kann. Fallen beispielsweise viele Aufgaben zur Addition und Subtraktion im Zahlenraum bis 1000 in eine Kompetenzstufe, so liegt es nahe diese Stufe entsprechend zu definieren. Dies erscheint jedoch auf den ersten Blick nicht angemessen, wenn auch eine größere Zahl von Aufgaben dieses Inhalts der nächst höheren Stufe zuzuordnen sind. Weiß man jedoch, dass diese Aufgaben z.B. alle einen langen Aufgabentext besitzen und das die Aufgabentextlänge nachweislich mit der Aufgabenschwierigkeit in Verbindung steht, dann ist die ursprüngliche Definition der Kompetenzstufe durchaus plausibel. Wir versuchen also quasi die Aufgabenschwierigkeit in verschiedene Anforderungen zu zerlegen und so zu exakteren Vorstellungen zu gelangen, welche Kompetenzen im Detail beherrscht werden.

Experten-Rating der VERA-Items im Hinblick auf die Standards. Alle Mathematik- und Deutsch-Aufgaben von VERA werden von Experten darauf hin beurteilt, wie relevant sie für die einzelnen Standards sind. Umgekehrt wird auch für alle Standards beurteilt, ob und welche Entsprechungen diese bei den VERA-Aufgaben haben. Diese beiden Prozeduren liefern einen Teil der empirischen Basis, die nötig ist, um anschlussfähig zu sein.

Follow-up-Studie: Anfang und Ende der 4.Klassenstufe. Wenn man dem Anspruch gerecht werden will, eine Orientierung an den Standards zu ermögli-

chen, dann benötigen wir neben dem Expertenrating der Relevanz von VERA-Items für die Standards ab 2004 unbedingt eine solide empirische Basis, um etwas darüber aussagen zu können, welche Relevanz die VERA-Items nachweislich für das Leistungsprofil am Ende der 4. Klasse haben. Prädiktoren wären also VERA-Items am Anfang der 4. Klasse; am Ende der 4. Klasse würden diese Items wiederholt und ergänzt (a) um die originalen (laut Schindler erheblich komplexeren und deshalb für VERA gar nicht geeigneten) Items, und falls möglich (b) um IGLU-Items (bzw. KESS-, LAU-, ELEMENT-Items). Damit ließe sich zum einen der Prognosewert von VERA für die Erreichung der Standards empirisch belegen, zum anderen ergeben sich Hinweise auf die Instruktionssensitivität der VERA-Items: Wo zeigen sich Steigerungen, Stagnationen, Verluste?

6. LITERATUR

Baumert, J., Bos, W. & Lehmann, R. H. (Hrsg.). (2000a). *TIMSS/III - Dritte Internationale Mathematik - und Naturwissenschaftsstudie-Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe, Vol. 2). Opladen: Leske + Budrich.

Baumert, J., Bos, W. & Lehmann, R. H. (Hrsg.). (2000b). *TIMSS/III - Dritte Internationale Mathematik - und Naturwissenschaftsstudie-Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit, Vol. 1). Opladen: Leske + Budrich.

Baumert, J., Lehmann, R. H., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O. & Neubrand, J. (Hrsg.). (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde*. Opladen: Leske + Budrich.

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

Klieme, E. (2003). *Definition von Kompetenzstufen – Diskussionsbeitrag zur Arbeitsstrategie im DESI-Projekt* (Präsentation). Frankfurt.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Frankfurt a. M.: DIPF.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.

Watermann, R., Stanat, P., Kunter, M., Klieme, E. & Baumert, J. (2003). Möglichkeiten und Grenzen der Nutzung von Schulrückmeldungen im Rahmen von Schulleistungsuntersuchungen: Das Disseminationskonzept von PISA 2000. In Deutsches PISA - Konsortium (Hrsg.), *PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 393-410). Opladen: Leske + Budrich.