EDK Schweizerische Konferenz der kantonalen Erziehungsdirektoren
CDIP Conférence suisse des directeurs cantonaux de l'instruction publique
CDPE Conferenza svizzera dei direttori cantonali della pubblica educazione
CDEP Conferenza svizra dals directurs chantunals da l'educaziun publica

Schlussbericht der HarmoS-Methodologiegruppe

Erich Ramseier
Abteilung Bilddungsplanung und Evaluation der Erziehungsdirektion des Kantons Bern

Urs Moser
Institut für Bildungsevaluation, Universität Zürich

Jean Moreau
Unité de recherche pour le pilotage des systèmes pédagogiques du canton de Vaud

Jean-Philippe Antonietti
Institut de mathématiques appliquées, Université de Lausanne

Bern, Juli 2008

Inhaltsverzeichnis

Die schweizerische Erziehungsdirektorenkonferenz (EDK) hat ein wissenschaftliches Projekt eingesetzt mit dem Auftrag, zu vier Fachbereichen (Schulsprache, Mathematik, Fremdsprachen, Naturwissenschaften) Basisstandards vorzuschlagen. Die Basisstandards sollen sich jeweils auf ein Kompetenzmodell stützen. Zu dieser Entwicklungsarbeit gehört auch eine empirische Validierung dieser Kompetenzmodelle (Validierungsstudie).

Die wissenschaftliche Verantwortung für die Entwicklungsarbeit wurde pro Fachbereich je einem Konsortium übertragen. Mit Mandat vom 1. April 2005 wurde zudem eine Gruppe Methodologie[1] eingesetzt. Diese begleitete die Konsortien während des Projekts bezüglich des methodischen Vorgehens und der technischen Phasen, welche sich auf die Kompetenzmodelle der Konsortien auswirken konnten. Zu den Aufgaben der Gruppe gehörten insbesondere die Konzeption der Validierungsstudie und die Datenauswertungen im Hinblick auf die Skalierung und die Validierung der Kompetenzmodelle. Die Gruppe hatte auch ein Schlussbericht zu verfassen, der die Verfahrensweise und methodologischen Schwierigkeiten bilanziert und Empfehlungen für spätere Etappen enthält.

Dieser hier vorliegende Schlussbericht beschreibt kurz, welche Aufgaben die Gruppe erfüllt hat, wie die Ergebnisse zu beurteilen sind und welche Folgerungen zu ziehen sind. Er dokumentiert die Arbeiten jedoch nicht im Einzelnen und ist somit keine wissenschaftliche Dokumentation der Validierungsstudie.

1. Arbeitsweise

Interne Organisation

Die Methodologiegruppe traf sich 2005 – 2007 jährlich an fünf bis sechs Sitzungen, 2008 an zwei Sitzungen. Erich Ramseier übernahm eine koordinierende Funktion. Daneben fanden zahlreiche Treffen mit den Konsortien und bilaterale Kontakte statt. Für die Skalierung übernahm jedes Mitglied zwei Themengebiete, wie die folgende Zusammenstellung zeigt:

	6. Klasse	*9. Klasse*
Schulsprache	Moser	Moreau
Mathematik	Antonietti	Ramseier
Naturwissenschaften	Moreau	Ramseier
	Lesen, C-Test	*Schreiben, Hören*
Fremdsprachen	Antonietti	Moser

Daneben betreuten Urs Moser, Erich Ramseier und Jean Moreau die Skalierung der Pilotstudien in den 2. Klassen der Schulsprache, Mathematik bzw. Naturwissenschaften.

[1] Jean-Philippe Antonietti, Universität Lausanne; Jean Moreau, URSP Lausanne; Urs Moser, IBE Universität Zürich; Erich Ramseier, Erziehungsdirektion Kanton Bern

Strategiegruppe

Die Methodologie-Gruppe hatte einerseits eine beratende und ausführende Funktion, andernfalls sollte sie auch verbindliche Richtlinien zur wissenschaftlichen Durchführung der Validierungsstudie geben. Um verbindliche Beschlüsse treffen zu können, wurde unter dem Vorsitz von Olivier Maradan eine Strategiegruppe gebildet, der die Konsortiumsleitungen und die Methodologiegruppe angehörten (Sekretariat: Max Mangold). Die gemeinsamen Beschlüsse betrafen nicht nur methodische Fragen, sondern auch die Grundzüge des Kompetenzmodells, die gemeinsame Begrifflichkeit und die Art der Produkte (vgl. Wegleitung). In der Strategiegruppe wurden methodische Spezialthemen, z. B. zum Standard-Setting, vorgestellt (Weiterbildung).

Beurteilung: Die Strategiegruppe wurde zu einem zwar spät eingesetzten, aber zentralen Instrument des Projekts. Sie sorgte für ein Minimum an Gemeinsamkeiten zwischen den Konsortien. Der grosse Zeitdruck stand einer Verbreiterung der gemeinsamen Basis entgegen.

Kontakte mit externen Experten

Die Methodologiegruppe war berechtigt, externe Experten zuzuziehen. Sie traf sich im Sommer 2005 mit Ray Adams, Universität Melbourne/PISA-Leiter, um das Design des Validierungstests und Skalierungsfragen zu klären. Ray Adams wurde auch während der Skalierung mehrfach um eine kurze Beratung per E-Mail gebeten.

Im Rahmen der DACHL-Vereinbarung trafen sich Urs Moser und Erich Ramseier zweimal zu einem Kolloquium und Erfahrungsaustausch mit den Kollegen, die in Deutschland, Österreich und Luxemburg mit ähnlichen Aufgaben betraut sind bzw. diese in einem bereits weit besser definierten Rahmen durchführen (vgl. IQB Berlin).

Beurteilung: Es ist unerlässlich, dass die internationalen Kontakte auf dem Gebiet der Methodologie der Kompetenzmessung weitergeführt und intensiviert werden. Die Schweiz – besonders die Deutschschweiz – hat hier nur eine sehr beschränkte Tradition. Die Treffen mit Wissenschaftlern aus den Nachbarländern machten deutlich, dass in den Nachbarländern weit mehr Ressourcen in die Entwicklung von Kompetenzmodellen und die empirische Validierung eingesetzt werden. Im Gegensatz zur Schweiz wurden in den Nachbarländern feste Organisationsformen zur Entwicklung von Leistungstests eingerichtet.

2. Wegleitung

Die Methodologiegruppe verfasste eine Wegleitung, die Anforderungen an die Kompetenzmodelle und die Validierungsstudie beschreibt. Sie gibt einen guten Überblick über das Projekt – auch wenn sie bzgl. Testdesign, Stichprobe und Testdurchführung nur eine vorletzte Planungsstufe beschreibt. Insbesondere wurden die gemäss Wegleitung noch für 2008 geplanten Erhebungen in den Fremdsprachen und den Naturwissenschaften auf Frühling 2007 vorverschoben und gleichzeitig mit den Erhebungen in Mathematik und Erstsprache durchgeführt.

Im theoretischen Teil hält die Wegleitung die Beschlüsse der Strategiegruppe zum Kompetenzmodell fest. Im Hauptteil werden die Anforderungen an die Aufgaben und die Testvorbereitung beschrieben.

Beurteilung: Die Angaben zur Testvorbereitung und Aufgabenformulierung wurden von den Konsortien gut eingehalten und ermöglichten es, trotz der unterschiedlichen Kulturen in den Fachbereichen zu einer zuverlässigen Kommunikation mit der Methodologiegruppe und zu einem gemeinsamen, auswertbaren Test zu kommen.

3. Pretest

Normalerweise werden nach der Entwicklung und Erprobung der Aufgaben im kleinen Rahmen, evtl. in so genannten kognitiven Labors, und vor dem Einsatz in einem Test alle Aufgaben mindestens einem Pretest unterzogen. Beim gedrängten Zeitplan von HarmoS war dies nicht möglich. Im Jahre 2006 führten die Konsortien auf Drängen der Methodologiegruppe Pretests mit einer beschränkten Zahl bereits entwickelter Aufgaben durch. Die Pretests lieferten einerseits Informationen zu den einbezogenen Aufgaben. Andererseits zeigten sie die Arbeitsschritte auf, die bei einer empirischen Leistungsmessung zu durchlaufen sind, und illustrierten, welche Ergebnisse von der Skalierung erwartet werden können und wie damit umzugehen ist. Der Pretest hatte damit auch eine wichtige Weiterbildungsfunktion.

Beurteilung: Die nur unvollständige Erprobung der Aufgaben in Pretests muss als Notfall-Vorgehen eingestuft werden und darf keineswegs in späteren Tests zum Bildungsmonitoring wiederholt werden. Insbesondere sollten Übersetzungsschwierigkeiten bereits während der Erprobung und nicht erst bei der eigentlichen Validierung aufgedeckt werden können. Der Aufwand für die Entwicklung von Tests in drei Sprachen wurde im HarmoS-Projekt generell unterschätzt.

Trotzdem erwies sich die Erprobung des empirischen Vorgehens und des Zusammenspiels zwischen Konsortien und Methodologiegruppe als sehr wertvoll; ein völliges Auslassen des Pretests und damit dieser Weiterbildung hätte den Erfolg der Validierungsstudie in hohem Masse gefährdet.

4. Design der Validierungsstudie

Die Strategiegruppe legte fest, dass sich die Kompetenzmodelle nicht darauf begrenzen dürfen, was in der Validierungsstudie bereits empirisch überprüfbar ist – inhaltliche Vollständigkeit hat Priorität. Unter dieser Voraussetzung und den zeitlichen und sonstigen Beschränkungen wurde die repräsentative Validierungsstudie auf die Klassen 6 und 9 begrenzt; je nach Fachbereich wurden zudem einzelne Kompetenzbereiche oder -aspekte nicht mit zugehörigen Aufgaben in die Validierungsstudie einbezogen. Daneben wurden in der 2. Klasse für Erstsprache, Mathematik und Naturwissenschaften reduzierte Studien mit nicht-repräsentativen Stichproben durchgeführt. Für die Einstufung der Aufgaben genügt dies weitgehend, solange die Stichprobe die ganze Spannweite individueller Kompetenzen erfasst.

Ziel

Die Validierungsstudie soll primär die Kompetenzmodelle überprüfen (1. Ziel). Dazu muss die Beschreibung der Kompetenz durch Bereiche, Aspekte und Niveaus empirisch abgestützt werden. Dies erfordert, dass genügend Aufgaben erprobt werden und in ihrer Schwierigkeit bekannt sind, so dass damit die Kombinationen von Bereichen, Aspekten und Niveaus erfasst und mit Beispielaufgaben illustriert werden können. Zusätzlich soll die Studie die Verteilung der Schülerschaft auf die Kompetenzniveaus darstellen (2. Ziel). Insbesondere soll angegeben werden können, welcher Anteil der Schülerschaft heute den vorgeschlagenen Basisstandard erfüllt.

Stichprobe und Gewichtung

Das erste Ziel der Studie verlangt den Einbezug so vieler Schülerinnen und Schüler, dass die Aufgabenschwierigkeiten in den Sprachregionen genügend genau geschätzt werden können. Die Wegleitung skizziert auf S. 17f. entsprechende Überlegungen. Unter Einbezug zu erwartender Ausfälle und unter der Voraussetzung, dass die Testpersonen zwei Testhalbtage absolvieren, wurde eine Brutto-Stichprobengrösse von je 6'600 Schülerinnen und Schüler für die 6. und die 9. Klasse geplant. Das zweite Ziel macht es erforderlich, dass die Stichprobe repräsentativ ist. Renaud (2007) beschreibt, wie die zweistufige Stichprobe (zuerst Auswahl von Schulen, dann Auswahl von je 2 Klassen) gebildet wurde.

Beurteilung: Die Stichprobenbildung hat sich bewährt. Die damit gewonnenen Daten erlauben es, Aufgabenschwierigkeiten und Kompetenzverteilungen im Rahmen der erwarteten Genauigkeit zu beschreiben.

Testdesign

Im HarmoS-Tests wurden zahlreiche Aufgaben und ihre Zusammenhänge überprüft, um so ihre Eignung für die Charakterisierung eines Kompetenzmodells abzuklären. Dies erfordert den Einsatz sehr vieler Aufgaben, während eine Testperson in der zur Verfügung stehenden Zeit nur wenige davon lösen kann. In dieser Situation muss eine geeignete Zuordnung von Aufgaben zu Personen und Testzeitpunkten gefunden werden, die eine optimale Auswertung ermöglicht (Matrix-Sampling).

Die vier Konsortien entwickelten die zu bearbeitenden Aufgaben (Hinweise dazu vgl. Wegleitung S. 10 f.) und gruppierten sie in Cluster, die je während 20 oder 30 Minuten bearbeitet werden konnten. Insgesamt entstanden 209 inhaltlich unterschiedliche Cluster, die je in zwei oder drei Sprachen vorliegen mussten. Das Druckverfahren erlaubte es, jeden Cluster in einem eigenen kleinen Testheft zu drucken und das Heft individuell einer Testperson zuzuweisen und anzugeben, zu welchem Zeitpunkt innerhalb der Testdurchführung dieses Heft zu bearbeiten sei.

Das verwendete Testdesign nutzte diese Möglichkeit und sorgte unter Beachtung der Anforderungen der Konsortien für eine möglichst breit streuende und ausgeglichene Aufgabenverteilung. Das Verfahren ist im Anhang ausführlicher beschrieben. Insgesamt wurden so rund 100'000 individuelle Verknüpfungen zwischen Aufgabenclustern, Testpersonen und Testzeitpunkten hergestellt und nur wenige Testpersonen lösten überhaupt die gleiche Kombination von Aufgaben.

Beurteilung: Der Einbezug eines professionellen Druckverfahrens (NZZ-Druckzentrum) ermöglicht die äusserst aufwendige individuelle Zuordnung der Cluster zu den zufällig ausgewählten Schülerinnen und Schülern. Die Umsetzung dieses komplexen Testdesigns sorgte für eine ausgewogene Verteilung der Aufgaben auf die Schülerinnen und Schüler, die der anschliessenden Skalierung der Daten zugute kam. Die Ausgewogenheit gilt allerdings nur zwischen Clustern. Der Einfluss der Reihenfolge der Aufgaben innerhalb eines Clusters bleibt (wie üblich) unkontrolliert.

Schülerfragebogen

Ergänzend zu den Leistungstests und den Informationen aus den Schülerkontrolllisten wurde ein kurzer Schülerfragebogen entwickelt, der Hintergrundinformationen zur kulturellen und sozialen Herkunft sowie zum schulischen Interesse, besonders in den Bereichen Mathematik und Naturwissenschaften, erfasste.

Beurteilung: Die Erfassung der wichtigsten Hintergrundmerkmale, die mit Schulleistungen zusammenhängen, ist für die Validierung von Testaufgaben unerlässlich. Die Informationen wurden bei der Skalierung berücksichtigt; unklar ist, wie weit die Auswertung des Interesses vorangetrieben werden soll. Auch die Auswertung, z. B. nach Migrationshintergrund, wurde nicht ausgeschöpft.

5. Durchführung der Validierungsstudie

Testdurchführung

Nach der gemeinsamen Planung durch das EDK-Sekretariat und die Methodologiegruppe wurde die Datenerhebung wie schon die Übersetzung und Produktion der Testhefte direkt vom EDK-Sekretariat aus durchgeführt. Dies umfasste die Identifikation der ausgewählten Schulen, die Kontaktaufnahme mit den Kantonen und Schulen, den Versand des Testmaterials und der Testinstruktion, die Entgegennahme des ausgefüllten Testmaterials von den Schulen und die Verteilung der ausgefüllten Testhefte auf die Konsortien. Die Testdurchführung ist in einem internen Bericht (Mangold, 2007) näher beschrieben.

Beurteilung: Die Arbeitsteilung zwischen den Konsortien und dem EDK-Sekretariat hat sich bewährt. Für ähnliche Vorhaben, die nicht mehr zu einer Grundlage von HarmoS führen, sondern dem Bildungsmonitoring dienen werden, sind aber andere Strukturen vorzuziehen. Obwohl die Archivierung der Daten im Sommer 2008 abgeschlossen sein wird, fehlt eine Institution, die sich um die professionelle Verwaltung des HarmoS-Testmaterials kümmert. Für die Nutzung des HarmoS-Testmaterials zur Umsetzung der Kompetenzmodelle in den Kantonen wird dies als Nachteil eingestuft.

Die Testaufsicht durch die Lehrpersonen hat zu keinen Problemen geführt, die während der Skalierung aufgedeckt worden wären. Die Eignung dieser Form muss aber für spätere Monitoring-Studien neu geklärt werden.

Datenerfassung

Die Konsortien erhielten die ausgefüllten Testhefte vom EDK-Sekretariat und sorgten für die Datenerfassung. Sie erstellten zu jedem Testheft gemeinsame oder sprachregionale Excel-Dateien mit den codierten Schülerantworten. Bei offen gestellten Aufgaben erforderte dies eine Codierung, die auf genauen Anweisungen beruhen musste, um eine einheitliche Interpretation der Lösungsversuche sicherzustellen.

Zwischen Methodologie-Gruppe und Konsortien wurde vereinbart, dass für offene Fragen die Übereinstimmung der Codierungen überprüft wurde; in der Regel integriert in die Phase der Codiererausbildung. Als Minimum sollten zu jeder offenen Aufgabe mindestens 30 Antworten von zwei Personen unabhängig voneinander codiert werden. Für die Auswertung stellte die Methodologiegruppe ein Excel-Dokument zur Verfügung, in das die Codes eingetragen werden konnten und das Übereinstimmungsmasse (Kappa, Prozentsatz der Übereinstimmung) berechnete. Die Auswertung der Übereinstimmungskontrolle lag bei den Konsortien. Als Beispiel: In der Mathematik betragen das mittlere lineare Kappa bei 33 zufällig (informell) ausgewählten Items .92 (Spannbreite .77 – 1.0 mit einem Ausreisser bei .64). und die mittlere Übereinstimmung 96.3%.

Beurteilung: Die Datenerfassung lief unter der Regie der Konsortien, so dass die Methodologie-Gruppe wenig dazu sagen kann. Sie scheint weitgehend erfolgreich verlaufen zu sein; die Daten erwiesen sich jedenfalls als auswertbar. Eine besondere Herausforderung stellt allerdings die einheitliche Codierung offener Fragen über die Sprachgrenzen hinweg dar. Diese konnte nicht in allen Fällen erreicht werden. Vor allem die zuverlässige Codierung von Antworten auf offene Aufgabenstellungen (geschriebene Texte) in drei Sprachen verlangt wesentlich mehr Ressourcen, als den Konsortien zur Verfügung gestellt wurden. Die Probleme von Tests in drei Sprachen (Übersetzung, Korrektur der Antworten) spiegelten sich dann auch in den Daten, die auf entsprechende Übersetzungs- und Korrekturschwächen aufmerksam machten. Auf die Sicherstellung der sprachübergreifenden Einheit muss bei späteren Monitoring-Erhebungen deutlich grösseres Gewicht gelegt werden als bei der Validierungsstudie.

Datenaufbereitung

Die Konsortien produzierten mehrere hundert Dateien, die jeweils die Ergebnisse eines Testhefts enthielten. Für die Auswertung mussten sie pro Fachbereich in eine Datenmatrix umgewandelt werden, deren Spalten je ein Item und deren Zeilen je eine Testperson darstellen. Aufgrund des Testdesigns enthält diese Matrix überwiegend leere Zellen (z. B. Mathematik, 9. Klasse zu 92.5%). Diese Umwandlung wurde unter Supervision von Erich Ramseier durch Edi Böni durchgeführt.

In den Naturwissenschaften und den Fremdsprachen wurden Antworten (z. B. Richtig-Falsch-Serien) direkt erfasst. Sie wurden in einer frühen Phase der Umwandlung gemäss Vorgaben der Konsortien in Partial-Credit-Items mit wenigen Bewertungsstufen umgewandelt.

Beurteilung: Die Datenaufbereitung ist zwar aufwendig, gelang aber ohne ernsthafte Probleme.

Gewichtung

Ramseier und Moreau (2007) beschreiben neben der schulinternen Stichprobe die Gewichtung und das Schätzverfahren, die nötig werden um die komplexen Stichproben auszuwerten und die Stichprobenfehler zu schätzen.

Beurteilung: Das Verfahren funktioniert; seine Nutzung in der Studie war aber bescheiden, da nicht viel mehr als der Anteil der Personen in den Kompetenzniveaus gewichtet geschätzt wurde.

6. Skalierung

Modellwahl

Die Skalierung und Auswertung der Testdaten war eine Hauptaufgabe der Methodologiegruppe. Ziel des HarmoS-Tests war es, zahlreiche Aufgaben und ihre Zusammenhänge zu überprüfen, um so ihre Eignung für die Charakterisierung eines Kompetenzmodells abzuklären. Dank der bekannten Schwierigkeit der Aufgaben konnten diese dann benutzt werden, um die Kompetenzniveaus zu illustrieren und so zu einer kriteriumsorientierten Beschreibung der Kompetenz zu kommen.

Der Skalierung wurde das Rasch-Modell zugrunde gelegt (Näheres dazu im Anhang der Wegleitung). Das Rasch-Modell geht von einem bestimmten Zusammenhang zwischen der zu messenden, aber nicht direkt beobachtbaren Kompetenz und der Wahrscheinlichkeit, eine Aufgabe richtig zu lösen, aus. Der Zusammenhang wird durch eine mathematische Funktion beschrieben (Itemcharakteristik-Kurve), die – abgesehen von einer Parallelverschiebung – für jede Aufgabe gleich sein soll.[2]

Für den Einsatz der Item-Response-Theorie waren zwei Gründe ausschlaggebend:

1) Das Rasch-Modell (wie andere Modelle der Item-Response-Theorie) erlaubt es, Aufgaben und Personen auf derselben Skala zu platzieren, die je nach dem als Anforderungs- oder Kompetenzskala angesehen werden kann. Aufgaben werden dort eingeordnet, wo Personen mit der entsprechenden Kompetenz diese Aufgabe mit einer bestimmten Wahrscheinlichkeit (z. B. 50%) lösen können. Damit kann ein Bereich der Kompetenzskala durch die dort platzierten Aufgaben charakterisiert und damit inhaltlich beschrieben werden: Der materielle Zusammenhang zum theoretisch formulierten Kompetenzmodell ist damit überprüfbar.

2) Das Rasch-Modell und andere Modelle der Item-Response-Theorie machen es möglich, die Fülle von Aufgaben in eine kohärente Analyse einzubeziehen, obwohl jede Aufgabe nur von einer kleinen Minderheit der Schülerinnen und Schüler bearbeitet wurde.

[2] Andere Modelle der Item-Response-Theorie würden zusätzliche Variationen dieser Funktion zulassen (vgl. Anhang der Wegleitung). Sie hätten ebenfalls eingesetzt werden können. Das Rasch-Modell wurde vorgezogen, weil es bei der gegebenen relativ kleinen Zahl von Personen pro Aufgabe stabilere Schätzungen ermöglicht und konzeptuell einfach ist.

Beurteilung: Die empirische Validierung von Daten verlangt, dass diese Daten mit Hilfe eines wissenschaftlichen Verfahrens skaliert werden. Die Anwendung dieses Verfahrens setzt verschiedene Bedingungen voraus, die sich direkt auf die Entwicklung von Aufgaben, insbesondere auf das zu verwendende Format, auswirkt. Nicht jede Aufgabe, die in der Schule für das Lernen und Prüfen eingesetzt wird, lässt sich in eine Testaufgabe transformieren. Viele Aufgabenformate haben zudem enorme Folgen auf die Kosten. Es ist für die verantwortlichen und betroffenen Personen (Politik, Verwaltung, Schule, Eltern, Schülerinnen und Schüler, Öffentlichkeit) zentral zu berücksichtigen, dass mit den ausgewählten Aufgabenformaten nur ein Teil von Kompetenzen getestet werden kann und aus Kostengründen häufig grundlegende Kompetenzen wie beispielsweise Sprachproduktion (Sprechen und Schreiben) beim Testen vernachlässigt werden. Im Rahmen der Validierung wurde versucht, so gut wie möglich sämtliche Kompetenzen zu erfassen, also auch produktive Kompetenzen wie Sprechen und Schreiben. Trotzdem decken die empirisch validierten Testaufgaben zum jetzigen Zeitpunkt nur einen Teil der von HarmoS festgelegten Kompetenzen ab.

Itemselektion

Passung zum Rasch-Modell

Zunächst musste die Eignung der Aufgaben überprüft werden. Dies war umso wichtiger als kein vollständiger Pretest durchgeführt wurde. Als erstes Kriterium galt es festzustellen, ob die Items dem Rasch-Modell entsprechen, d.h. ob der empirische Zusammenhang zwischen Kompetenz und Lösungswahrscheinlichkeit genügend gut der theoretisch postulierten Funktion entspricht. Dazu wurde der Infit (weighted mean square) benutzt. Der Infit zeigt für jedes Item, wie viele unerwartete Antworten unter der Annahme des Rasch-Modells beobachtet werden. Der Infit hat ein Erwartungsmass von 1. Ein zu hoher Fit-Wert weist darauf hin, dass die Trennschärfe des Items zu niedrig ist. Ein zu tiefer Fit-Wert weist darauf hin, dass die Trennschärfe zu hoch ist. Die Auswahl der Items richtete sich nach dem Infit, der nicht kleiner als 0.70 und nicht grösser als 1.30 sein durfte (Wright & Linacre, 1994).

Beurteilung: Das Infit-Kriterium erwies sich als nicht kritisch und wurde von den meisten Aufgaben erfüllt. In Mathematik, 9. Klasse musste z. B. nur 1 von 269 Aufgaben deshalb ausgeschlossen werden. Die Wahl gerade des Rasch-Modells als Basis der Skalierung hatte somit kaum einengende Wirkungen.

Trennschärfe

Ein zweites Kriterium, das Aufgaben erfüllen sollten, ist die Trennschärfe. Unter der Trennschärfe wird in der klassischen Testtheorie die Korrelation des Items mit dem Gesamtergebnis des Tests verstanden. Der Wertebereich liegt zwischen –1 und 1. Die Trennschärfen der Items sollten nicht kleiner als 0.30 sein. Eine niedrige Korrelation kann daran liegen, dass die Formulierung der Aufgabe missverständlich ist, dass man auch mit einer Fehlüberlegung aufs richtige Resultat kommt, dass fachfremde Aspekte die Lösung erleichtern oder erschweren usw.

Man hat somit gute Gründe, Items mit ungenügender Trennschärfe ($r < 0.3$) auszuscheiden. Eine Ausnahme kann bei extrem leichten (bzw. schweren) Aufgaben gemacht werden. Auch wenn sie konform zum Rasch-Modell sind, ist ihre

Trennschärfe[3] niedrig: Wenn (fast) alle sie lösen, kann (fast) kein Zusammenhang zwischen der Häufigkeit des Lösens und der Kompetenz in dieser Population festgestellt werden. Man kann das Item behalten um zu zeigen, was alle können. Das kann Sinn machen, wenn es nicht um einen effizienten Test geht, sondern um die Illustration der Spannweite einer Kompetenz. Es ist allerdings ein weites Feld von sehr einfachen Aufgaben denkbar und die empirische Überprüfung kann nicht klären, ob das Antwortverhalten bei diesen (fast) immer richtig gelösten Aufgaben überhaupt etwas mit der fraglichen Kompetenz zu tun hat.

Beurteilung: Insgesamt mussten deutlich mehr Aufgaben mangels genügender Trennschärfe ausgeschlossen werden als wegen eines ungenügenden Infits. Der Ausfall ist aber angesichts des fehlenden Pretests nicht übermässig (vgl. Unterlagen der einzelnen Konsortien).

Sprachunterschiede

Ein drittes Kriterium stellt die Messinvarianz dar. Mit ihr ist gemeint, dass der Test für verschiedene Personengruppen dasselbe misst – das Kompetenzmodell dasselbe beschreibt.[4] Das ist z. B. nicht der Fall, wenn Aufgabe 1 in Gruppe A sehr schwierig ist, in Gruppe B dagegen sehr leicht, während es bei Aufgabe 2 gerade umgekehrt ist. Für solche Unterschiede können irgendwelche äusseren Gründe massgeblich sein. Es kann aber auch sein, dass sich die Gruppen A und B Kompetenzen mit unterschiedlicher Struktur aufgebaut haben. Das ist insofern plausibel, als es sich bei Kompetenzen um *gelernte* Fähigkeiten handelt und die Lerngelegenheiten zwischen Gruppen variieren können.

Als „Gruppen" stehen bei HarmoS die Sprachregionen im Vordergrund: Zwischen Sprachregionen können sich Lerngelegenheiten kulturell bedingt unterscheiden und sprachenübergreifende Einheitlichkeit ist bei einem nationalen Projekt der mehrsprachigen Schweiz eine zentrale Frage. Zudem können Übersetzungsprobleme hinter unterschiedlichem Verhalten der Items in den Sprachregionen stecken.

Das Kriterium der sprachlichen Invarianz (differential item functioning) erwies sich als das schwierigste. Ein relativ grosser Teil der Items funktionierte in den drei Sprachen unterschiedlich. Aus diesem Grund entschloss sich die Gruppe Methodologie, bei der Aufgabenauswahl ein grosszügiges und pragmatisch begründetes Kriterium anzuwenden. Ein Item, das in Deutsch, Französisch und Italienisch vorlag, wurde dann als gleich schwierig erachtet, wenn die drei einzeln berechneten Parameter innerhalb 1 Logits lagen beziehungsweise wenn keines der Items mehr als 0.5 Logit vom gemeinsam ermittelten Mittelwert abwich. Lagen die Items nur in zwei Sprachen vor, wurde das Kriterium entsprechend der geringeren Wahrscheinlichkeit von grossen Abweichungen auf 0. 816 Logit beziehungsweise 0.408 Logit angepasst.

Um auch Items verwenden zu können, die dieses Kriterium nicht erfüllten, wurden sie teils „regionalisiert", d. h. das Item wird für jede Region als ein eigen-

[3] „Trennschärfe" wird mit unterschiedlichen Bedeutungen verwendet. Einerseits wird sie in der klassischen Testtheorie wie hier als Korrelation zwischen Item und Test definiert (z. B. Krauth, 1995, S. 266), andererseits wird damit in der Item-Response-Theorie die maximale Steigung der Itemcharakteristik-Kurve bezeichnet (z. B. Rost, 2004, S. 98). Im Rasch-Modell haben alle Items die gleiche maximale Steigung der Itemcharakteristik-Kurve; ihre klassische Trennschärfe variiert je nach Population und ist niedrig, wenn das Item für diese Population ausserordentlich leicht (bzw. schwierig) ist.

[4] Z. B. Embretson und Reise (2000, S. 250).

ständiges Item behandelt, das nur in dieser Region erhoben wurde. Es beeinflusst damit den Vergleich zwischen Regionen nicht mehr, kann aber – unter Berücksichtigung der Region – noch zur Beschreibung der Kompetenz verwendet werden.

Beurteilung: Wie bereits angemerkt, wurde der Aufwand für die Entwicklung von Tests in drei Sprachen deutlich unterschätzt. Der Aufwand für die Entwicklung nationaler Tests ist sehr gross, was in Zukunft besser berücksichtigt werden sollte. Die angewendeten Kriterien für die Entdeckung von Sprachunterschieden sind sehr large und beispielsweise deutlich weniger streng, als die Kriterien, die bei der Entwicklung eines Tests für den internationalen Schulleistungsvergleich PISA angewendet werden. Diese Situation war auch für die Methodologiegruppe unbefriedigend. Das grosszügige Kriterium wurde aufgrund der besonderen Zielsetzung der Studie und des enormen Zeitdrucks eingesetzt und die Itemselektion wurde pragmatisch legitimiert. Ziel war es, nicht allzu viele Items aufgrund von Sprachunterschieden zu verlieren. Ein strenges wissenschaftliches Vorgehen hätte zum Scheitern der nationalen Validierung der Kompetenzmodelle geführt. Das Projekt hätte sprachregional fortgesetzt werden müssen.

Dimensionalität

Die theoretischen Kompetenzmodelle postulieren unterschiedliche Kompetenzaspekte und -bereiche. Zur Validierung kann überprüft werden, ob diesen theoretischen Dimensionen auch Subskalen entsprechen, die voneinander unterschieden werden können und die doch auch so korrelieren, dass sie als Teil einer Kompetenz angesehen werden können. Das Programm „ConQuest" (Wu und Adams, 1998), mit dem skaliert wurde, erlaubt es, solche mehrdimensionale Modelle zu schätzen und mit dem eindimensionalen zu vergleichen. „Conquest" gibt für jedes Modell einen Wert für die Abweichung des Modells von den Daten an. Dieser Wert wird als sogenannte Devianz ausgewiesen. Der Wert ist als eine relative Zahl zu interpretieren. Die Devianz kann folglich nur dann mit der Devianz eines anderen Modells verglichen werden, wenn jedes Mal der gleiche Datensatz verwendet wurde, also gleiche Anzahl Items etc. Je differenzierter ein Modell ist (Anzahl Dimensionen), desto geringer wird in der Regel die Devianz. Wenn also ein mehrdimensionales Modell mit sehr vielen Dimensionen geprüft wird, dann ist die Chance gross, dass die Devianz kleiner wird als beim eindimensionalen Modell. Um zu prüfen, ob ein mehrdimensionales Modell bedeutsam besser zu den Daten passt als ein eindimensionales oder ein anderes mehrdimensionales Modell, wurden die Devianzen anhand eines Chi-Quadrat-Tests und von Informationsmassen (CAIC und BIC) beurteilt. Eine Simulationsstudie von Antonietti und Moreau (2008) weist darauf hin, dass diese Modellprüfung nur unzulänglich gegen die zufällig bessere Anpassung eines mehrdimensionalen Modells mit seinen vermehrten Freiheitsgraden schützt.

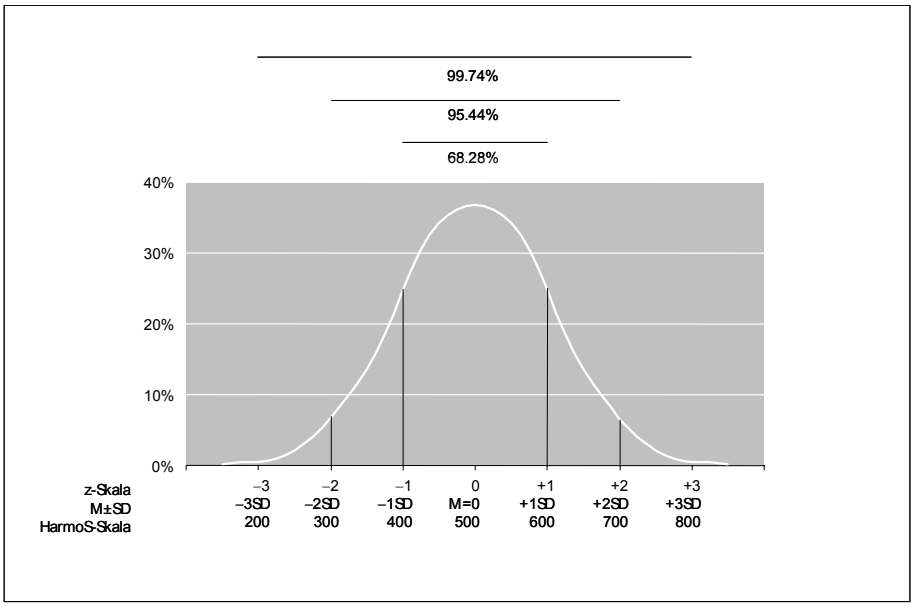
Beurteilung: Die Ergebnisse variieren zwischen den Konsortien. Sie zeigen in der Regel aber, dass Modelle auf der Basis der theoretisch postulierten Dimensionen den Daten besser entsprechen als das eindimensionale. Damit ist allerdings nicht gesagt, dass nicht ganz andere Strukturen noch besser passen würden. Auch die Zuordnung der einzelnen Aufgaben zu den Subdimensionen ist nicht überprüft. Eine gründliche Abklärung der Dimensionalität mit umfassenden Modellvergleichen liess sich aufgrund des Zeitdruckes nicht durchführen.

Skalentransformation

Die Beziehung zwischen Lösungswahrscheinlichkeit, Aufgabenschwierigkeit und Personenfähigkeit im Rasch-Modell geht davon aus, dass die Lösungswahrscheinlichkeit p = 0.5 beträgt, wenn die Aufgabenschwierigkeit und Personenfähigkeit übereinstimmen. Wenn nun die Fähigkeit eines Schülers anhand von Aufgaben beschrieben wird, deren Schwierigkeit seiner Fähigkeit entsprechen, werden somit Aufgaben verwendet, die er mit 50% Wahrscheinlichkeit lösen kann. Diese Lösungswahrscheinlichkeit ist zu gering, als dass die Aufgabe so zur Beschreibung von Bildungsstandards sinnvoll genutzt werden könnte: Eine Lösungschance von 50:50 zeigt nicht das sichere Beherrschen der Aufgabe an. Aus diesem Grund hat sich die Gruppe Methodologie dazu entschlossen, die Schwierigkeitsparameter der Aufgaben so zu transformieren, dass die Lösungswahrscheinlichkeit für Personen mit dieser Fähigkeit p = 0.67 (2/3) beträgt. Diese Transformation verschiebt alle Items in Richtung höherer Kompetenz. Aufgaben, die gerade bei der Schwelle des Bildungsstandards liegen, werden danach von jenen Personen, die den Standard gerade erreichen, mit 67% Wahrscheinlichkeit gelöst.

Im Falle dichotomer Aufgaben handelt es sich bei dieser Transformation um eine Verschiebung aller Itemparameter um den gleichen Wert; sie ändert an der Interpretation wenig. Bei Aufgaben, bei denen je nach Lösungsqualität mehrere Punkte vergeben werden (X = 0,1,2,3…), werden anschauliche Schwellenparameter (z. B. Stelle, wo p(X ≥ 2)) anstelle der direkten Modellparameter benutzt. Die Transformation dieser Schwellenparameter wurde mit einem Algorithmus von J. Ph. Antonietti aus jener der Modellparameter abgeleitet und führt je nach Gesamtkonstellation der Items zu gewissen Verschiebungen unter den Schwellen.

Die Schwierigkeits- und Personenparameter variieren in der originalen Rasch-Metrik um den Nullpunkt und nehmen meist Werte von −3 bis +3 an. Negative Werte sind kein taugliches Mittel zur Beschreibung der Kompetenz von Personen. Um negative Zahlen und Dezimalstellen zu verhindern, entschied sich die Gruppe Methodologie für eine zweite Transformation. Dazu wurden die Schwierigkeits- und Personenparameter jeweils linear so transformiert, dass der Mittelwert der neuen Verteilung dem Punktwert 500 entsprach und die Standardabweichung 100 Punkte betrug.



Eine solche lineare Transformation ändert am Gehalt der Aussagen nichts. Sie erleichtert – wie eine z-Transformation auf den Mittelwert 0 und die Standardabweichung 1 – aber die Einschätzung eines individuellen Wertes im Vergleich zum Mittelwert und zu üblichen Abweichungen und fördert damit normorientierte Aussagen. Dies gilt umso mehr, als die Testwerte angenähert normalverteilt sind.[5]

Rund 68 Prozent der Testergebnisse liegen zwischen dem Mittelwert ± eine Standardabweichung, rund 95 Prozent der Testergebnisse liegen zwischen dem Mittelwert ± zwei Standardabweichungen, also zwischen 300 und 700 Punkten. Die 500/100 – Metrik wurde in Anlehnung an PISA gewählt.

Gesamtbeurteilung der Skalierung: Der Einsatz des Rasch-Modells hat sich bewährt und nicht zu ungebührlichen Einschränkungen innerhalb der in den Test aufgenommenen Aufgaben und der damit repräsentierten Inhalte geführt. Bereits die Bedingungen eines ökonomisch vertretbaren large-scale-assessements setzten aber den möglichen Aufgaben und dadurch den validierbaren Inhalten bestimmte Grenzen. Das Rasch-Modell oder ein anderes Modell der Item-Response-Theorie ist notwendig, um die Daten trotz der Verteilung der vielen Aufgaben auf relativ wenige Personen auswerten zu können. Mit ihm können auch Aufgaben und Personen auf der gleichen Skala eingestuft werden, so dass Abschnitte auf der Kompetenzskala anhand der zugehörigen Aufgaben inhaltlich interpretiert werden können.

Als problematisch haben sich Unterschiede in den Aufgabenschwierigkeiten zwischen den Sprachregionen erwiesen. Diese Thematik sollte anhand der vorhandenen Daten weiter untersucht werden, um zwischen Übersetzungsfragen und unterschiedlichen Kulturen und Lerntraditionen unterscheiden zu können. In kommenden Untersuchungen muss die Aufgabenauswahl sorgfältig auf regionale Repräsentativität abgestimmt werden, falls man sich inhaltlich nicht auf den kleinsten sprachregionalen Nenner beschränken will.

In der gegebenen Zeit konnten nur die allerwichtigsten Kriterien der Modellerfüllung und der Passung der Items überprüft werden. Lückenhaft ist insbesondere die Überprüfung der Dimensionalität. Vor einer Weiterverwendung der Aufgaben müssten weitere Kriterien untersucht werden (vgl. unten).

7. Festlegen der Kompetenzniveaus und Standards

Die Skalierung konstituiert primär eine kontinuierliche Kompetenz- und Anforderungsdimension, auf der neben Personen (mit ihrer Kompetenz) auch Aufgaben (mit ihren Anforderungen) eingeordnet werden. Das Kompetenzmodell spricht dagegen eher von Niveaus, d. h. Abstufungen auf dieser Dimension. Es fragt sich dann, wie diskrete Niveaus definiert und beschrieben werden können und in welcher Beziehung sie zur primär kontinuierlichen Skala der Validierungsstudie stehen. Im einen Extrem (anscheinend eher L1) hat die Skala Priorität und wird a posteriori interpretiert, im andern Extrem sind die Kompetenzniveaus a priori theoretisch formuliert und es wird nach einer Relation zur Skala gesucht (ten-

[5] Die genaue Form hängt vom verwendeten Skalierungsmodell ab. Conquest geht von einer Normalverteilung der latenten Dimension aus; wenn wie hier bei der Schätzung jedoch auch Zusatzmerkmale wie Alter oder Schultyp einbezogen werden, gilt die Normalverteilungsannahme für die entsprechenden Residuen. Zur Logik der Schätzung und der Transformationen vgl. z. B. den technischen Bericht zu PISA, OECD (2003).

denziell eher M). Es gibt zwar Modelle, wie a priori definierte spezifische Aufgabenmerkmale genutzt werden, um Anforderungsniveaus zu charakterisieren; dies gelingt aber erst ansatzweise und setzt eine umfassende Aufgabenklassifikation voraus, was schon aus Zeitgründen nicht realisierbar war. Die Relation zwischen theoretischem Konzept und Skala musste vielmehr durch interpretierende, nicht formalisierte Verfahren hergestellt werden.

Wenn primär von der kontinuierlichen Skala ausgegangen wird, können die Intervallbreiten mehr oder weniger gleich gross gewählt werden. In der Regel betragen die Intervallbreiten dabei rund 1 Einheit (logit). Dieses Vorgehen hat den Vorteil, dass die Ergebnisse relativ ähnlich interpretiert werden können: Wenn ein Schüler einem Niveau zugeordnet wird, dann gilt bei dieser Breite als Faustregel, dass die Zuteilung eines Schülers zu einem Niveau bedeutet, dass er im Durchschnitt mindestens 50 Prozent der Aufgaben des Niveaus richtig lösen kann. Wenn man sich mehr an theoretischen Merkmalen orientiert, ist eher damit zu rechnen, dass die Intervallbreiten variieren.

Unter dem gegebenen Zeitdruck haben die Konsortien die Niveaus weitgehend ohne Einbezug der Methodologiegruppe festgelegt. Sie konnten sich dabei an der in der Wegleitung formulierten Zielgrösse von 4 Niveaus orientieren. Die Niveaudefinitionen unterscheiden sich in der Folge (etwa in der Breite auf der Skala) erheblich.

Auch die Basisstandards wurden von den Konsortien ohne engen Einbezug der Methodologiegruppe festgelegt. Die Methodologiegruppe hat zwar gängige Methoden vorgestellt, mit denen die Bestimmung eines Grenzwerts systematisiert wird. Ihre Verwendung würde nichts daran ändern, dass Standards immer eine Setzung und Einschätzung sind. Sie würden aber den Setzungsprozess transparent machen und den Einbezug unterschiedlicher Personengruppen (Stakeholders) systematisieren.

Der Anteil der Schülerschaft, der die vorgeschlagenen Standards erfüllt, konnte anhand der Validierungsstudie bestimmt werden und variiert zwischen den Konsortien (vgl. deren Berichte). Die Wegleitung postuliert auf S. 9 klar: „Basisstandards sind in erster Linie inhaltlich zu begründen. Sie sollen aber auch realistisch sein." Glaubt man einer Publikation von de Pietro, Müller und Wirthner (2008, S. 44, 46), so wurde in der Erstsprache im Gegensatz eine Orientierung an der statistischen Norm umgesetzt, die gar auf einer EDK-Vorgabe von 20% beruhen soll.

Beurteilung: Kompetenzniveaus und Basisstandards wurden von den Konsortien unter grossem Zeitdruck weitgehend selbständig festgelegt. Die Methodologiegruppe kann entsprechend wenig dazu sagen. Der allerdings aufwendige Einsatz bekannter formaler Methoden könnte die Niveaus und Standards besser absichern. Bei einer primär kontinuierlichen Kompetenzdimension bleiben diskrete Niveaus aber blosse Zusätze zur Charakterisierung von Kompetenzbereichen und markieren keine echten Stufen.

Das Vorgehen scheint zwischen Konsortien variiert zu haben und evtl. teilweise der ursprünglichen Konzeption zu widersprechen. Angesichts der Bedeutung, die Niveaus und Standards (im Gegensatz zur primär validierten kontinuierlichen Kompetenz) erhalten, ist die Methodik und Transparenz der Festlegungen insgesamt nicht befriedigend. Wichtig ist, dass die Kommunikation – vor allem im wissenschaftlichen Kontext – transparent erfolgt und auch die Politik sich der limitierten Aussagekraft der Standards bewusst ist. Die Konsortien sollten sich bei einer so wichtigen Angelegenheit noch einmal ein Jahr Zeit nehmen können um

die Standardsetzung in Ruhe mit Vertretern aus Wissenschaft und Praxis sowie mit weiteren Stakeholdern zu diskutieren. Die bisherige Standardsetzung dürfte den Regeln eines expliziten, wissenschaftlich gestützten Verfahrens nicht genügend entsprochen haben.

8. Archivierung

Am Ende der Arbeit mit den Konsortien werden die der Auswertung zugrunde liegenden Daten auf einer DVD zusammengefasst und der EDK übergeben. Sie stehen damit für weitere Auswertungen und die Verknüpfung mit kommenden Monitoring-Untersuchungen zur Verfügung.

9. Folgerungen

Beurteilungen, die Folgerungen enthalten oder nahe legen, sind bereits bei den einzelnen Abschnitten des Berichts angebracht worden. Hier folgen einige generelle Aussagen.

Stand der Validierungsstudie

Die durchgeführte Validierungsstudie hat wichtige Ergebnisse erbracht. Schon nur, dass die Kompetenzmodelle von Anfang an mit Aufgaben illustriert werden können, deren faktische Schwierigkeit bekannt ist und nicht nur vermutet wird, rechtfertigt den Aufwand. Allerdings muss auch festgestellt werden, dass die Validierung ein langfristiger Prozess ist und im bisherigen Rahmen erst ein erster Schritt gemacht werden konnte. Zudem ist die Validierungsstudie mit über 10'000 repräsentativ ausgewählten Getesteten und einer Fülle von eingesetzten Testaufgaben ein bedeutsames Unternehmen der empirischen Bildungsforschung in der Schweiz, das aber bisher nur beschränkt ausgewertet werden konnte. Aus der Sicht der Methodologiegruppe wurde eher eine Grundlage für die wissenschaftliche Validierung der Kompetenzmodelle geschaffen als dass die Modelle bereits hätten validiert werden können. Der Start ist gelungen, das Ziel noch nicht erreicht, sofern in der Schweiz die Kompetenzmodelle im wissenschaftlichen Sinne empirisch validiert und die Standards nach wissenschaftlichen Verfahren bestimmt werden sollen.

Weiterführung

Im Hinblick auf ein valides künftiges Bildungsmonitoring der Volksschule sollte der Validierungsprozess weitergeführt werden. Manche Forschungsfragen müssen aus einer kombinierten fachbezogenen und methodischen Perspektive definiert werden. Einige Fragen können auch aus rein methodologischer Sicht genannt werden. Dazu gehören die folgenden Themen, die danach unterschieden werden, ob sie anhand der Auswertung der vorliegenden Daten untersucht werden können (weitere Auswertungen) oder ob in kleinerem Rahmen neue Datenerhebungen nötig sind (Anschlussarbeiten).

Weitere Auswertungen

a) Wie zeigen sich unterschiedliche Curricula in den Testergebnissen? Analyse der Skalen, Subskalen und Einzelitems (differential item functioning DIF) nach Schultyp und teilweise Kanton.

b) Wie zeigt sich der Migrationshintergrund bzw. die Fremdsprachigkeit in den Testergebnissen (DIF, Subskalen im Vergleich)?

c) Weitere Modelltests, z. B. DIF nach guten/schwachen Schülern

d) Vollständigere Bestimmung der Fehler der verwendeten Schätzwerte; auch bessere Anpassung der empirischen Bestimmung des Populationsanteils, der den Basisstandard erfüllt, an die theoretische Vorgabe (Bspl. Mathematik: gegenwärtige Berechnung anhand der Gesamtskala; verbal wird aber die Erfüllung für jede Subskala verlangt)

e) Dimensionalität untersuchen: Mehrfachzuordnung der Aufgaben zu Skalen (Im Gegensatz zur bisherigen Analyse dürften viele Aufgaben mehrere Kompetenzaspekte ansprechen); explorative Dimensionsanalyse

f) Relation der Kompetenzen zwischen den einzelnen Fachbereichen; z. B. auch Analyse der Mathematik-Kompetenz unter Kontrolle der Lesekompetenz. Welcher Anteil der Schülerschaft scheitert in allen vier Fächern bzw. in mindestens einem Fach an den Bildungsstandards?

g) Voraussage der Itemschwierigkeit aus sonstigen Aufgabenmerkmalen: Präzisierung der Beschreibung von Kompetenzniveaus

h) Zentrales Ergebnis der Validierung ist die Kenntnis der Schwierigkeit der Aufgaben. Diese wurde jedoch unter den einschränkenden Bedingungen der Studie ermittelt. So sind die Aufgaben immer nur in einem Testheft enthalten, das sie in einer bestimmten Reihenfolge präsentiert. Die Auswertung zeigt, dass in einigen Testheften die letzten Aufgaben oft nicht gelöst wurden. Solche fehlenden Angaben wurden bei der Schätzung der Aufgabenschwierigkeit ausgeschlossen. Es ist aber wahrscheinlich, dass schwächere Schüler besonders oft in Zeitnot kamen und Aufgaben ausliessen. Damit dürfte die Schwierigkeit der Aufgaben unterschätzt worden sein. Die Bedeutsamkeit dieses Effekts kann analysiert werden: Zusammenhang zwischen der Häufigkeit des Auslassens von Aufgaben, der Fähigkeit der entsprechenden Personen und der Position der Aufgabe.

i) Motivation ist ein Aspekt der Kompetenz. Im Schülerfragebogen wurden zwar Angaben zur Motivation erfasst. Sie aber noch besser ausgewertet und mit den Kompetenzmodellen in Verbindung gebracht werden.

Anschlussarbeiten

a) Die Bedeutsamkeit fehlender Angaben kann in einer Zusatzstudie in einer neuen Stichprobe und mit anderer Aufgabenreihenfolge wesentlich genauer untersucht werden.

b) Die fixe Abfolge der Aufgaben innerhalb der Testhefte hat nicht nur die Häufung unbearbeiteter Aufgaben am Ende eines Testhefts zur Folge. Um im Testheft einen sinnvollen Arbeitskontext zu konstituieren, beziehen sich viele Aufgaben auch auf gemeinsame oder verwandte Inhalte. Das führt jedoch zu einer Abhängigkeit zwischen den Aufgaben, aus der Lösung der einen kann z. B. für die Lösung der nächsten etwas gelernt

werden. Solche Abhängigkeiten widersprechen aber den Voraussetzungen des verwendeten Skalierungsmodells. Dass sie wahrscheinlich sind, wird u.a. dadurch belegt, dass eine versehentlich in zwei Testheften gestellte Mathematik-Aufgabe je nach Heft unterschiedlich oft korrekt gelöst wurde. Auch hier könnte eine Zusatzstudie mit variierender Abfolge der Aufgaben Klärung bringen.

c) Teilweise wurden Aufgabenstellungen und Themen aus der Validierung ausgeschlossen, weil sie im gegebenen Rahmen nicht empirisch umgesetzt werden konnten. In neuen Studien kann die Validierung auf solche Fragestellungen ausgeweitet werden. Damit kann der Gefahr entgegengetreten werden, dass Standards und Tests zu einer unangemessenen Verengung des Lehrstoffs führen. Das Konsortium Naturwissenschaften hat solche authentische Tests bereits in die erste Validierungsphase einbezogen.

d) Analyse von Entwicklungsprozessen über die Klassenstufen hinweg.

Diese Arbeiten sind wichtig, damit klar ist, ob und wie die Aufgaben als einzelne Einheiten in einer kommenden Monitoring-Studie verwendet werden können, um die Verbindung zu den heute definierten Standards herzustellen. Sie führen auch zu allgemeinen Erkenntnissen, die beim Aufbau des Monitorings notwendig sein werden. Das Praktische daran ist, dass sie mit relativ kleinen Stichproben durchgeführt werden können, die zwar das ganze Leistungsspektrum abdecken sollten, aber nicht repräsentativ sein müssen.

Organisatorischer Rahmen

Die Validierungsstudie wurde unter massiven Einschränkungen und mit erheblichen Auslassungen durchgeführt. Zu den Begrenzungen gehören das Fehlen von Pretests, der zu enge Zeitrahmen für fast alle Arbeitsschritte, die Arbeit der Methodologiegruppe im blossen Nebenjob und die beschränkte Interaktion und fehlende Integration zwischen Fachbereichen und Methodologie.

Auch wenn die Validierungsstudie dank des grossen Einsatzes vieler Beteiligter zu einem teilweisen Erfolg geführt hat, wäre es falsch, diese Mangelsituation unbesehen als Grundlage für die Planung künftiger Bildungsmonitoring-Studien zu nehmen und damit zu perpetuieren.

Bildungsmonitoring und Verhältnis Fachdidaktik – Psychometrie

In der Validierungsstudie war eine deutliche Trennung und Hierarchie zwischen fachlich-inhaltlicher und methodischer Arbeit gegeben: Die Fachkonsortien waren verantwortlich, die Methodologiegruppe gab via Strategiegruppe zwar Richtlinien, war aber letztlich beratend und ausführend tätig.

Die starke Trennung führte generell dazu, dass empirische Ergebnisse in der inhaltlichen Auswertung und Interpretation nicht umfassend genutzt wurden und bestehende Verfahrensweisen für die Definition der Niveaus und Standards nicht eingesetzt wurden.

Die Validierungsstudie befand sich in einer besonderen Situation: Es galt primär, ein theoretisches Konstrukt – das Kompetenzmodell – aufzubauen; empirische Ergebnisse hatten nur eine stützende Funktion. Den inhaltlich-fachlichen Aspekten kommt daher zu Recht eine sehr grosse Bedeutung zu. In kommenden Monitoring-Studien wird die Situation umgekehrt sein: Auf der Basis bestehender Kompetenzmodelle geht es darum, die Kompetenzverteilung in Regionen, Kanto-

nen zuverlässig zu erfassen und Merkmale der Schulen und Personen einzufangen, die eine Interpretation von Unterschieden unterstützen. Zudem muss sichergestellt werden, dass in späteren Studien zuverlässig Trends aufgezeigt werden können. All dies stellt hohe methodische Ansprüche.

Sicher werden auch in den kommenden Bildungsmonitoring-Studien fachlich-inhaltliche Gesichtspunkte wichtig sein. Die Priorität muss aber tendenziell eher der umfassend verstandenen methodisch-analytischen Sichtweise gegeben werden – schon nur um für das Monitoring der einzelnen Fächer einen einheitlichen Bezugsrahmen zu haben.

Bildungsmonitoring und Forschungsstrukturen

Die schweizerische Bildungsforschung leidet darunter, dass zu einem Thema kurzfristige Initiativen und Projekte lanciert und Strukturen und Kompetenzen aufgebaut werden. Nach Abschluss der Projekte zerfallen diese Strukturen und damit die Kompetenzen meist wieder. Sie fehlen dann, wenn das Thema erneut aufgegriffen werden muss.

Genau dies droht wieder im Anschluss an HarmoS – und dies obwohl absehbar ist, dass die in HarmoS aufgebauten Kompetenzen schon bald für das Bildungsmonitoring benötigt werden. Um dies zu verhindern, sollten fachbereichsbezogene Entwicklungsschwerpunkte unterstützt werden. Angesichts der gegenwärtig schmalen Basis in der Schweiz muss die methodologische Expertise bzgl. Kompetenzmessung erst recht gefördert und institutionell abgesichert werden. Unsere Nachbarländer Deutschland und Österreich haben auf die bei ihnen ähnliche Mangelsituation reagiert und Strukturen geschaffen und teils auch Forschungsprogramme lanciert. Im Gegensatz zu diesen Ländern hat die Schweiz die Entwicklung von Bildungsstandards von Anfang an auf empirische Grundlagen abgestützt. Diese Pionierleistung muss nun in nachhaltige Formen gebracht werden, wenn die Schweiz nicht den Vorteil ihres innovativen Ansatzes verlieren und in Rückstand geraten will.

Ausblick in die Zukunft

Der Erfolg der Umsetzung der Bildungsstandards hängt von verschieden Faktoren ab, die im Folgenden kurz diskutiert werden:

- Die bisherigen Ergebnisse ermöglichen es, in Zukunft die Qualitätssicherung durch Leistungsmessung an Kompetenzmodellen auszurichten, die fachdidaktisch begründet und zum Teil empirisch validiert worden sind.
- Die vorliegenden Daten ermöglichen es, die Validierung der Kompetenzmodelle weiterzuführen und legen die Grundlage, in Zukunft die Kompetenzmodelle empirisch nach allen Regeln der Kunst zu validieren und die begonnenen Arbeiten sinnvoll abzuschliessen.
- Bildungsstandards werden in zweierlei Hinsicht umgesetzt. Erstens soll ihre Einhaltung im kommenden Bildungsmonitoring regelmässig überprüft werden. Die EDK hat es in der Hand, diese Monitoring-Tests so anzulegen, dass sie den Kompetenzmodellen entsprechen und auch deren Validierung weiterführen.
- Zweitens dienen die Bildungsstandards der individuellen Standortbestimmung der Schülerinnen und Schüler. Die Entwicklung entsprechender Grundlagen ist den sprachregionalen EDK's überlassen. Wie sie diese Aufgabe lösen, ist offen. Sicher werden einige Institutionen von sich aus Tests

auf den Markt bringen. Die Tatsache, dass HarmoS bzw. die EDK den Kantonen und Schulen keine Instrumente, sondern nur Modelle zur Verfügung stellt, birgt die Gefahr, dass es zu einer unkontrollierten Dynamik im Bereich der Testentwicklung kommt. Entscheidend für eine Harmonisierung über Standards sind streng genommen nicht die Kompetenzmodelle, sondern die wissenschaftlichen Skalen, die eine einheitliche Metrik zur Beschreibung und Beurteilung von Schulleistungen ermöglichen.

Die Politik tut gut daran, klare Regeln über die Verwendung des Begriffs «HarmoS-Test» festzulegen. Andernfalls besteht die Gefahr eines Wildwuchses von Labeln. Jede Institution, die einen Test auf den Markt bringt, wird von sich aus behaupten, dass dieser Test HarmoS-kompatibel sei. Klassencockpit und Stellwerk werden auf einmal auch als HarmoS-Tests verkauft. Streng genommen müsste der Begriff «HarmoS-Test» aber an die nationale Skala gebunden sein und dadurch geschützt werden. Wenn der Begriff «HarmoS-Test» oder «HarmoS-kompatibel» ohne (empirisch festgelegte) Bedingungen verwendet werden darf, dann muss das ganze Vorhaben aus einer methodologischen Perspektive als unnötig eingestuft werden. Die Gruppe Methodologie schätzt dieses Szenario als wahrscheinlich ein, weil damit die Kosten für aufwändige HarmoS-Tests eingespart werden können.

- Als wichtige Bedingung für die Verwendung des Labels „HarmoS-kompatibel" sollte von Testanbietern und -durchführenden verlangt werden, dass die von ihnen verwendeten Testaufgaben und die anonymisierten Testergebnisse (Basisdaten, nicht nur Skalenwerte) für wissenschaftliche Analysen zur Verfügung stehen. Damit können die Qualität der Tests kontrolliert und die methodische Weiterentwicklung sichergestellt werden.
- Tests für die individuelle Standortbestimmung und für das Bildungsmonitoring können nicht als zwei völlig getrennte Gegenstände angesehen werden: Es muss beide Male um die gleichen Kompetenzmodelle, Standards und nationalen Skalen gehen, wenn man Chaos im Bildungssystem vermeiden will. Um dies sicherzustellen, müssen teilweise gemeinsame Aufgaben verwendet werden. Eine gute Koordination der beiden Umsetzungsformen kann zudem zu ganz beträchtlichen Synergien führen.

Generelle Einschätzung

Es ist zu hoffen, dass die bisherigen Arbeiten für die Validierung der Kompetenzmodelle realistisch eingeschätzt werden und mit den provisorischen Ergebnissen ein redlicher Umgang gepflegt wird. Alle Beteiligten, insbesondere das EDK-Sekretariat, haben einen grossen Effort geleistet. Mit grossem Einsatz liess sich bereits vieles erreichen. Es ist aber wichtig, dass die Arbeiten zur Konsolidierung und Validierung der Kompetenzmodelle und Bildungsstandards mit hohem wissenschaftlichem Engagement weitergeführt werden.

10. Anhang

Literatur

Antonietti, J.-P., & Moreau, J. (2008). *Perspectives méthodologiques dans le cadre d'un projet d'harmonisation de la scolarité en Suisse.* Paper presented at the conference of the Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE), Genève.

De Pietro, J.-F., Müller, R., & Wirthner, M. (2007). HarmoS-L1: Vers des standards de base pour la langue de scolarisation. – In Richtung Basisstandards im Bereich der Schulsprache. *Babylonia* (4), 40-52.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Mahwah, NJ: Lawrence Erlbaum.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Psychologie Verlags Union.

Mangold; M. (2007). *Durchführung der empirischen Phase. Projekt HarmoS.* Interner Bericht. Bern: EDK

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4. ed., pp. 681-699). Westport, CT: American Council on Education.

Ramseier, E. & Moreau, J. (2007). *Stichprobendesign und Gewichtung in der HarmoS-Validierungsstudie.* Arbeitspapier. Bern: EDK

Renaud, A. (2006). *Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles*. Neuchâtel: Office fédéral de la statistique.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest. Generalised Item Response Modelling Software*. Melbourne: Australian Council for Educational Research.

HarmoS Test Design

The four consortia were responsible for developing the tasks and for most aspects of the test design like content coverage or choice of item formats, thereby supported by a guideline of the methodological group. They assigned each task to one cluster, each cluster taking 20 or 30 minutes of testing time. In total, they produced 209 clusters which had to be administered in German, French, and mostly in Italian (cf. table 1).

Since mathematics and science were tested in one session and languages in the other, these sessions could be designed separately. Both testing sessions consisted of two positions before and two after a break. In each position, students could work on one cluster. Since each student only worked on 8 clusters (matrix design), a main challenge was to distribute and combine clusters in a way that all tasks of a subject could be scaled on a common scale, correlations between content areas of competence and between aspects of competence could be estimated, and position effects were controlled at least between clusters.

Table 1: Number of Clusters/Booklets by Subject, Grade and Language Region

Subject	Grade 6				Grade 9			
	German	French	Italian	Total	German	French	Italian	Total
Duration: 20 minutes								
Mathematics	26	26	26	78	34	34	34	102
Science	24	24		48	24	24		48
First language	15	15	15	45	15	15	15	45
Foreign languages	8	8		16	16	16		32
Duration: 30 minutes								
First language	19	19	7	45	20	20	8	48
Foreign languages	4	4		8	8	8		16
Total	96	96	48	240	117	117	57	291
	Grade 6 and 9 together							
All Clusters	213	213	105	531				
Clusters with different content	209	209	105					

Since mathematics and science were tested in one session and languages in the other, these sessions could be designed separately. Both testing sessions consisted of two positions before and two after a break. In each position, students could work on one cluster. Since each student only worked on 8 clusters (matrix design), a main challenge was to distribute and combine clusters in a way that all tasks of a subject could be scaled on a common scale, correlations between content areas of competence and between aspects of competence could be estimated, and position effects were controlled at least between clusters.

The printing arrangement allowed for assigning booklets individually to persons and positions. Therefore, a balanced design was not created by assigning certain combinations of four clusters systematically to a limited number of test booklets (usual multi-matrix-design). Instead, each cluster was assigned to a single short booklet and these booklets were randomly assigned to persons and positions, thereby aiming for a number of principles. For mathematics, grade 9, where each cluster included tasks from one content area of competence, these principles are:

- Each cluster is tested equally often in total and in each position

- Each cluster is equally often combined with each cluster from other content areas in pairs of clusters
- Pairs of clusters are evenly distributed across the class and student sample

The assignment of the 34 mathematics clusters had to be combined with the assignment of the 24 science clusters. Pairs of mathematics and science clusters were combined randomly. To each cluster, the same number of students was assigned; therefore most students worked on a pair of clusters from each of the two subjects but some students worked on two pairs of mathematics clusters.

To distribute clusters to students and positions, a list of all possible cluster combinations which meet the mentioned design principles was constructed by programming it. Copies of this list were combined to create a series longer than the number of students. The cluster combinations within each list were randomly sorted. Finally, the series was matched to the stratified list of students.

As a result of this procedure, the mentioned principles are only attained with some minor random variation. Each student got a set of test materials with a specific combination of four booklets per session. If one only considers position 1 and 2 in the German part of Switzerland, already 791 different combinations of mathematics clusters entered analysis in grade 9 – each combination only 1 or 2 times. If one considers all 4 positions of the mathematics and science test, this augments to 2079 different cluster combinations – only 10 of them dealt with by two students. Given such diversity, possible interactions between clusters are optimally controlled and tasks are linked with each other in many different ways. Despite some random variation, the design is almost equal to a full balanced incomplete block design which is considered as unrealistically complex (Mazzeo, Lazer & Zieky, 2006, p. 685). This is helpful, since in the mathematics test of grade 9, 92.5% of all information is missing by design.

To generate the cluster combinations for the other three subjects, a similar approach was applied. For the languages, the situation was somewhat more complex: For first language some tasks refer to an acoustic stimulus presented to the whole class. The clusters of these tasks were administered at the beginning of the testing session and had to be assigned on class level. For foreign languages, clusters of English, German, and French as foreign language had to be handled. In total, about 100’000 individual links between clusters, positions and persons were produced. Erich Ramseier, BiEv, ED Bern