Erich Ramseier, Peter Labudde und Marco Adamina

Validierung des Kompetenzmodells HarmoS Naturwissenschaften: Fazite und Defizite

Validation of the HarmoS Science Competency Model: Results and Deficiencies

Zusammenfassung

Im Rahmen eines groß angelegten bildungspolitischen Reformprojekts wurden in der Schweiz in vier Fachbereichen auf Kompetenzmodelle gestützte Bildungsstandards für das 2., 6. und 9. Schuljahr entwickelt. Zum Entwicklungsprozess gehörte bereits eine erste Validierung der Kompetenzmodelle. Der vorliegende Artikel beschreibt kurz das für den Fachbereich „Naturwissenschaften" entwickelte Modell und stellt die darauf bezogene Validierung mittels repräsentativer Stichproben im 6. und 9. Schuljahr dar. Neben dem Design und der Durchführung der Tests werden zentrale Ergebnisse präsentiert: Konstitution einer Kompetenzskala, Kompetenzstruktur, Kompetenzniveaus, Vergleich der naturwissenschaftlichen Kompetenz im 6. und 9. Schuljahr, Festlegung der Bildungsstandards, Kompetenzunterschiede zwischen Teilpopulationen. Der Ertrag und die Lücken dieses ersten Schrittes zu einer Validierung des naturwissenschaftlichen Kompetenzmodells und die notwendigen Folgeschritte werden diskutiert.

Schlüsselwörter: Naturwissenschaften, Kompetenzmodell, Bildungsstandards, Referenzaufgaben, Leistungstests, Validierung.

Abstract

As part of a large-scale education policy reform project in Switzerland, educational standards based on competency models have been developed for grades 2, 6 and 9 in four different departments. A primary validation of these competency models was part of the development process. The following article describes the model that was developed for the science department and displays the related validation using representative student samples from grades 6 and 9. The design and the execution of the tests, as well as key findings involving the following areas are presented: Constitution of a competency scale, competency structure, competency level, comparison of competencies in grades 6 and 9, determining the educational standards and the difference in competencies between sub-populations. The output and the gaps in this primary validation process of the competency model in science education, as well as the necessary subsequent steps are discussed.

Keywords: science, competency model, educational standard, reference tasks, achievement test, validation.

In der Schweiz wurde im Rahmen des bildungspolitischen Großprojekts "Harmonisierung der obligatorischen Schule" (HarmoS) ein Kompetenzmodell für die Naturwissenschaften entwickelt und validiert. Im vorliegenden Artikel beschreiben und analysieren wir den Validierungsprozess und dessen Ergebnisse. Dabei stehen folgende Fragen im Vordergrund:

- Wie wurde die Validierung angelegt?
- Was heißt überhaupt Validieren und inwiefern konnte das Kompetenzmodell validiert werden?
- Wie wurden mit Hilfe eines Validierungstests Bildungsstandards festgelegt?
- Welche Erkenntnisse aber auch Schwierigkeiten und Grenzen ergaben sich bei der Validierung des Kompetenzmodells HarmoS Naturwissenschaften?

Im 1. Kapitel schildern wir kurz die politischen Rahmenbedingungen und stellen das Kompetenzmodell HarmoS vor. Das 2. Kapitel beinhaltet das Design und das Vorgehen bei der Validierung. Im 3. Kapitel stellen wir die Hauptergebnisse der repräsentativen Validierungstests vor, um dann im 4. Kapitel Schlussfolgerungen zu ziehen.

1 Die Rahmenbedingungen und das Kompetenzmodell

1.1 Der politische Rahmen

Die Schweiz zählt 26 Kantone mit 26 unterschiedlichen Schulsystemen. Um diese einander mehr anzunähern, startete die Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK, das Schweizer Pendant zur deutschen Kultusministerkonferenz) vor zehn Jahren das Projekt HarmoS. In diesem geht es zum einen um eine Harmonisierung der Schulstrukturen, z. B. zwei Jahre obligatorischer Kindergarten und darauf aufbauend sechs Jahre Primarschule. Zum anderen werden Bildungsstandards eingeführt und sprachregionale Lehrpläne entwickelt.

Die EDK beauftragte vier Konsortien für jeweils Schulsprache, Fremdsprache, Mathematik und Naturwissenschaften[1], ein Kompetenzmodell für die gesamte obligatorische Schule vom Kindergarten bis zum 9. Schuljahr zu entwickeln und darauf gestützt Bildungsstandards für das Ende des 2., 6. und 9. Schuljahres vorzuschlagen (Labudde & Adamina, 2008).

Modell und Standards gingen in eine breite politische Anhörung (EDK, 2010a, 2010b). Seit ihrer offiziellen Freigabe durch die EDK im Juni 2011 bilden sie einen wichtigen Rahmen für das kommende nationale Bildungsmonitoring und für die Entwicklung der Lehrpläne der Deutschschweiz sowie der französisch- und italienischsprachigen Schweiz (EDK, 2011).

1.2 Auftrag und Rahmen

Die EDK schrieb 2004 die Entwicklung der Kompetenzmodelle und Bildungsstan-

1 In der Schweiz werden – anders als in den Nachbarländern – die Naturwissenschaften in der obligatorischen Schule, d.h. auch in der Sekundarstufe I, als Integrationsfach unterrichtet, im zukünftigen Lehrplan unter dem Titel „Natur und Technik".

dards öffentlich aus und gab bestimmte Rahmenbedingungen vor: Bilden eines Konsortiums aus Fachdidaktikern/-innen, Einbeziehen von Statistikfachleuten und Lehrkräften, sich in den drei Sprachregionen der Schweiz verankern. Sie orientierte sich bei ihrem Auftrag in erster Linie an der Klieme-Expertise (Klieme et al., 2003). Für die Entwicklung der Bildungsstandards bedeutete dies, dass

- vom umfassenden Kompetenzbegriff von Weinert (2001) ausgegangen wird,
- Standards in einem Kompetenzmodell verankert werden, welches das Anforderungsgefüge der gemeinten Kompetenz fundiert beschreibt,
- Basisstandards und nicht Regelstandards formuliert werden,
- das Kompetenzmodell und die Standards bereits in dieser Phase empirisch validiert werden sollen.

Zu den Hauptaufgaben eines Konsortiums gehörten somit: die Ziele für den naturwissenschaftlichen Unterricht klären, ein Kompetenzmodell entwickeln, dieses validieren und allenfalls überarbeiten, Basisstandards vorschlagen und diese durch konkrete Beispiele illustrieren.

Das Konsortium Naturwissenschaften umfasste im Kern 18 Naturwissenschaftsdidaktiker/-innen, welche verschiedene Fächer, Schulstufen und Sprachregionen repräsentierten. Hinzu kamen 2 Statistikfachleute, 30 Lehrpersonen, eine 12-köpfige Begleitgruppe und für die Korrekturen der Tests 40 Hilfskräfte. Während des gesamten Arbeitsprozesses (2005–2008) bestanden regelmäßige Kontakte zu den drei anderen Konsortien, d.h. Schul- und Fremdsprachen sowie Mathematik, und zu einer aus vier Fachleuten bestehenden Methodologiegruppe. Letztere beriet die Konsortien in statistischen Fragen und war für das Design des Validierungstests sowie einen Teil der Testauswertungen verantwortlich.

Das Konsortium nahm im September 2005 seine Arbeit auf, entwickelte 2006 eine erste Version des Kompetenzmodells, überarbeitete es mehrfach aufgrund von Rückmeldungen von Fachdidaktikdozierenden und Lehrpersonen, führte 2007 und im Frühling 2008 verschiedenste Tests durch (Kap. 2) und lieferte Ende 2008 der EDK den wissenschaftlichen Schlussbericht ab (Konsortium, 2008).

1.3 Das Kompetenzmodell

Das Kompetenzmodell besteht aus einer dreidimensionalen Matrix (Abb. 1): 1. Handlungsaspekte, anderswo auch als Kompetenzaspekte oder -bereiche bezeichnet; 2. Themenbereiche (Inhalte), 3. Niveaus (Anforderungsniveaus).

Die 1. Achse umfasst acht Aspekte, die beim jetzigen Entwicklungsstand des Modells als abschließend gelten: 1. Interesse und Neugierde entwickeln, 2. Fragen und untersuchen, 3. Informationen erschließen, 4. Ordnen, strukturieren, modellieren, 5. Einschätzen und beurteilen, 6. Entwickeln und umsetzen, 7. Mitteilen und austauschen, 8. Eigenständig arbeiten, mit anderen zusammenarbeiten. Die Handlungsaspekte bilden für das Konsortium die primäre Achse; für diese – und

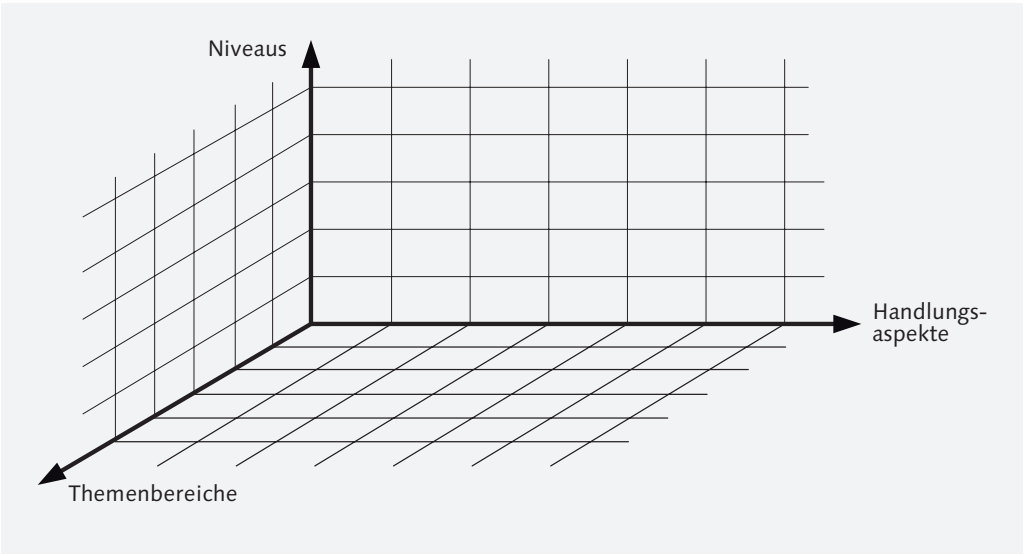


Abb. 1: Das Kompetenzmodell von HarmoS Naturwissenschaften.

nicht für die Themenbereiche – werden die verschiedenen Niveaus beschrieben. Jeder Handlungsaspekt besteht aus drei bis fünf Teilaspekten. Für den Handlungsaspekt „Informationen erschließen" werden zum Beispiel die folgenden fünf Teilaspekte unterschieden: 1. Informationsformen erkennen (Formen, Aufbau und Strukturen von Informationen erkennen, z. B. Aufbau einer Graphik); 2. Informationen zu naturwissenschaftlichen Inhalten frage- und sachbezogen identifizieren und aus Informationsmitteln herauslesen; 3. nach Informationen recherchieren; 4. Informationen umsetzen und für sich erkenntlich, einsichtig und nutzbar machen; 5. Informationen und Informationsquellen einordnen (naturwissenschaftliche Informationen kritisch sichten und die Herkunft von Informationen überprüfen) (Konsortium, 2008, p. 49).

Die Achse der Themenbereiche enthält deren acht, wobei kein Anspruch auf Vollständigkeit erhoben wird: 1. Planet Erde, 2. Bewegung, Kraft, Energie, 3. Wahrnehmung und Steuerung, 4. Stoffe und Stoffveränderungen, 5. Lebewesen, 6. Lebensräume und Lebensgemeinschaften, 7. Mensch und Gesundheit, 8. Natur, Gesellschaft, Technik. Mit diesen Bereichen werden paradigmatische Inhalte aufgezählt, nicht aber ein Kerncurriculum definiert. Letzteres bleibt den sprachregionalen Lehrplänen vorbehalten.

Auf der 3. Achse werden zu jedem der acht Handlungsaspekte – genauer zu den jeweils drei bis fünf Teilaspekten – für das Ende des 2., 6. und 9. Schuljahres je vier Niveaus I bis IV definiert. Die Differenzierung in vier Niveaus wurde als Vorgabe des Gesamtprojektes übernommen. Die Niveaus für die drei Schuljahre überlappen sich aufgrund der hohen Heterogenität von Schülerleistungen teilweise. So bildet das Niveau IV des 2. Schuljahres zugleich das Niveau I des 6., und die Ni-

veaus III und IV des 6. Schuljahres sind identisch mit den Niveaus I und II des 9. Schuljahres. Bei 8 Handlungsaspekten mit durchschnittlich je 4 Teilaspekten, 3 Schulstufen und je 4 Niveaus ergibt dies unter Berücksichtigung der Überlappungen ungefähr 300 Zellen mit je unterschiedlichen Niveaubeschreibungen. Die Niveaubeschreibungen erfolgten nach zwei Progressionslogiken: eine horizontale Progression bezieht sich darauf, dass weitere – z.T. auch kognitiv anspruchsvollere – Teilaspekte bei zunehmender Ausprägung von Kompetenzen einbezogen werden. Eine vertikale Progression bezieht sich darauf, dass innerhalb von Teilaspekten eine Differenzierung, ein höherer Anspruch der Ausprägung der Kompetenzen und damit eine höhere Komplexität angelegt werden.

2 Design und Vorgehen bei der Validierung

Ein Kompetenzmodell bildet eine theoretische Grundlage für das Lehren und Lernen und insbesondere für das Messen einer Kompetenz. Für seine empirische Validierung ist zentral, dass das Modell in eine entsprechende Kompetenzmessung umgesetzt wird, die ihrerseits validiert wird (Kane, 2006). Dazu werden Tests mit konkreten Aufgaben entwickelt, die das Kompetenzmodell möglichst gut abbilden. Anschließend werden Merkmale der Tests und Fragen untersucht, die sich auf die Struktur und weitere Eigenschaften der so gemessenen Kompetenz beziehen. Dazu gehören die Fragestellungen, die in Kapitel 3 aufgegriffen werden. Aus diesem Programm leiten sich das Design und das Vorgehen bei der Validierung ab.

2.1 Papier-und-Bleistift-Test als Teil der Validierung

Bereits bei der Entwicklung des Kompetenzmodells zeigte sich, dass wichtige Komponenten der Handlungsaspekte nicht ausschließlich in Form von Papier-und-Bleistift-Aufgaben (PB) überprüft werden können. Deshalb wurde bereits frühzeitig nach weiteren Formen gesucht und festgelegt, dass neben Tests in Form von PB-Aufgaben auch Experimentier-, Erkundungs- und Entwicklungsaufgaben eingesetzt werden sollten. Gleichzeitig wurde angestrebt, Testverfahren zu konzipieren, die auf den verschiedenen Schulstufen realisierbar sind (2., 6. und 9. Schuljahr) und die es zudem erlauben, einen Teil der Aufgaben auf jeweils zwei Schulstufen einzusetzen, insbesondere im 6. und 9. Schuljahr. Tabelle 1 gibt einen Überblick über die Tests. Ergänzend zu den Testaufgaben wurden für länger dauernde Aufgabenstellungen, für die Arbeit an außerschulischen Lernorten und für Handlungsaspekte wie „Mitteilen und austauschen“ oder „Eigenständig arbeiten, mit anderen zusammenarbeiten“ Lerngelegenheiten entwickelt und damit Anregungen und konkrete Vorschläge für den naturwissenschaftlichen Unterricht ausgearbeitet. Diese Lerngelegenheiten wurden nicht in die Validierung einbezogen.
Im Rahmen der Experimentiertests wurden Experimente zum eigenständigen Bearbeiten entwickelt und die Aufträge dazu

Tab. 1: Überblick über die durchgeführten Validierungstests

Test	Schuljahr	$N_{Lernende}$	$N_{Klassen}$	Testdurchführung
Angeleiteter Test bestehend aus Papier-und-Bleistift-Aufgaben sowie Experimentieraufgaben	2	593	30	Aug./Sept. 2007 (Beginn 3. Schulj.)
Papier-und-Bleistift-Test (nationaler Validierungstest mit repräsentativer Stichprobe)	6	4124	255	April-Mai 2007
Experimentiertest	6	663	30	April-Mai 2008
Papier-und-Bleistift-Test (nationaler Validierungstest mit repräsentativer Stichprobe)	9	3888	273	April-Mai 2007
Experimentiertest	9	805	44	April-Juni 2008
PISA 2006: Re-Analyse der offiziellen Resultate	9			Juni-Dez. 2008

von den Schülerinnen und Schülern individuell ausgeführt. Dabei wurden in erster Linie der Handlungsaspekt „Fragen und untersuchen" und für einzelne Aufgaben auch „Entwickeln und gestalten" einbezogen.
In den folgenden Ausführungen beschränken wir uns auf den Papier- und Bleistift-Test (PB-Test) im 6. und 9. Schuljahr.

2.2 Stichprobe

Ziel war es, die Tests so anzulegen, dass für die Illustration der angelegten Niveaus, Bereiche und Aspekte im Kompetenzmodell eine ausreichende Zahl von Aufgaben von genügend Personen, nämlich 150 pro Aufgabe und pro Sprachregion, gelöst werden konnten. Damit war auch gesichert, pro Sprachregion genügend Personen zu testen, so dass Leistungsmittelwerte und die Anteile von Schülerinnen und Schüler, die ein bestimmtes Leistungsniveau erreichen, genau geschätzt werden können.

Als Populationen wurde die Schülerschaft des 9. bzw. 6. Schuljahres staatlicher Schulen der deutschen und französischen Schweiz festgelegt und eine Stichprobe im Umfang von 2000 pro Sprachregion angestrebt (vgl. Angaben Tab. 1). Die italienische Schweiz konnte angesichts der eng limitierten Rahmenbedingungen nicht einbezogen werden. Die Stichprobe wurde zweistufig definiert; gewählt wurden zunächst Schulen, darin Klassen. Dabei wurde auf eine repräsentative Berücksichtigung der Schultypen auf der Sekundarstufe I geachtet. In den gewählten Klassen wurden alle Schülerinnen und Schüler getestet. Die Ziehung der Schulen basierte auf den schulstatistischen Daten des Bundesamts für Statistik. Repräsentative Auswertungen verlangten nach einer Gewichtung, da die Stichproben in beiden Landesteilen gleich groß waren, der Anteil der deutschen Schweiz aber knapp drei Viertel der Gesamtpopulation ausmacht (Ramseier & Moreau, 2007; Renaud, 2006).

2.3 Aufgabenkonstruktion

Bei der Aufgabenkonstruktion wurde von den Handlungsaspekten ausgegangen, da diese im Kompetenzmodell die primäre Dimension darstellen. Jede Aufgabe bezieht sich auf eine thematische Situation (inhaltlicher Kontext), zu welcher mehrere Fragen (= Items) gestellt werden. Die Fragen sind so gut als möglich unabhängig voneinander konzipiert und jeweils einem Handlungsaspekt und – durch Experteneinschätzung – einem Niveau zugewiesen. Je Aufgabe werden mehrere Handlungsaspekte überprüft. Bei der Auswahl der inhaltlichen Kontexte wurde eine Balance zwischen lebensweltlichem Bezug und fachlich repräsentativen Konzepten für eine naturwissenschaftliche Grundbildung angestrebt.

Es sind kaum Routineaufgaben, bei denen die Lernenden auf Standardlösungen zurückgreifen können. Die für die Bearbeitung der Aufgaben notwendigen Grundlagen und Bezugspunkte werden in Form von Informationen durch Bilder, Grafiken und Texte aufgebaut. Es wird kein bestimmtes Vorwissen vorausgesetzt. Bisherige Erfahrungen und Konzepte haben jedoch einen Einfluss auf die Kompetenzdisposition für das Bearbeiten der Aufgaben. Eingesetzt wird eine Varietät von Fragenformaten: geschlossen als multiple Choice, offen kurz und lang als Zuordnungen, Ranglisten, Skizzen, Texte, Mind-Map u.a. Zum Beispiel umfasste die Aufgabe „Wie wird das Wetter in den nächsten Tagen sein“ fünf Fragen. Bei der ersten Frage (vgl. Abbildung 2), welche sich auf den Handlungsaspekt „Informationen er-

Wie wird das Wetter in den nächsten Tagen sein?

Im Schulzimmer stellen jede Woche zwei Kinder gemeinsam die Wetterprognosen zusammen. Sie schneiden dazu aus Zeitungen die Meldungen aus und verarbeiten sie. Diese Woche sind auf dem Plakat folgende Angaben:

Angaben	Montag	Dienstag	Mittwoch	Donnerstag	Freitag
Sonnenschein	viel	viel	teilweise	wenig	teilweise
Temperatur Tiefst-/ Höchstwert	14 ° Celsius 28 ° Celsius	15 ° Celsius 30 ° Celsius	16 ° Celsius 32 ° Celsius	17 ° Celsius 23 ° Celsius	16 ° Celsius 26 ° Celsius
Luftfeuchtigkeit	tief	tief	mittel bis sehr hoch	hoch	mittel bis tief

Gib immer zwei Tage an (mit den Abkürzungen Mo, Di, Mi, Do, Fr)!

1. Welches sind die sonnigsten Tage?
2. Welches sind die heißesten Tage?
3. Welches sind die feuchtesten Tage?
4. An welchen Tagen könnte am meisten Regen fallen?

Abb. 2: Beispiel eines PB-Items für die 6. Klasse zum Handlungsaspekt “Informationen erschließen” (Niveau 2) und dem Themenbereich “Planet Erde”.

schließen“ (Teilaspekt 2 Informationen zu naturwissenschaftlichen Inhalten frage- und sachbezogen identifizieren und aus Informationsmitteln herauslesen, vgl. Abschnitt 1.3) mit Niveau 2 bezog, mussten die Schülerinnen und Schüler Informationen in der Tabelle erschließen und herauslesen, für einen Teil ansatzweise auch externes (Vor-)Wissen beiziehen (Zusammenhang zwischen Wetterelementen) und daraus die richtigen Antworten angeben. Bei den weiteren Fragen dieser Aufgabe ging es um bestimmte Wettersituationen in dieser Woche und um die Gegenüberstellung von Prognose und Wettermessung; damit bezogen sich die Fragen auf die zwei Handlungsaspekte „Ordnen, strukturieren, modellieren“ und „Einschätzen und beurteilen“.

Die Aufgabenkonstruktion unterlag zudem formalen Rahmenbedingungen. Für die Bearbeitung der einzelnen Aufgaben standen den Lernenden im Rahmen der Papier-und-Bleistift-Tests zehn Minuten zur Verfügung. Die Bearbeitung musste mit einfachen Mitteln (z. B. auf Papier, mit Farbstiften, ohne Einbezug weiterer Medien wie Ton, Film, Computer) realisierbar sein.

Jede Aufgabe wurde von Fachdidaktikdozierenden entwickelt, mit mehreren Lehrpersonen der entsprechenden Stufe im Sinne einer kommunikativen Validierung besprochen, in mindestens zwei Schulklassen pilotiert und evaluiert (als eine Art kognitives Pretesting), danach überarbeitet und allenfalls nochmals pilotiert. Aus Zeitgründen war es jedoch nicht möglich, einen Pretest mit fundierter statistischer Auswertung durchzuführen.

2.4 Testdesign und Durchführung

Im Papier-Bleistift-Test (PB) wurden die Handlungsaspekte „Informationen erschließen“, „Ordnen, strukturieren, modellieren“ und „Einschätzen und beurteilen“ einbezogen. Es wurden für den PB-Test je 6 bis 7 Aufgaben zu allen acht Themenbereichen für das 6. und 9. Schuljahr zusammengestellt, wobei 16 Aufgaben mit 47 identischen Items auf beiden Stufen eingesetzt wurden. Insgesamt wurden je Stufe 45 Aufgaben entwickelt. Sie enthielten zwischen 54 und 126 Items pro Handlungsaspekt, insgesamt 232 Items für das 6. und 266 Items für das 9. Schuljahr.

Die Aufgaben wurden überwiegend paarweise in 24 Testblöcke aufgeteilt, wobei pro Block jeweils Aufgaben mit unterschiedlicher thematischer Ausrichtung kombiniert wurden. Danach wurden 48 Blockpaare so gebildet, dass jeder Testblock in vier Blockpaaren zusammen mit jeweils thematisch unterschiedlichen Blöcken vorkam. Indem in jedem Blockpaar noch die Reihenfolge der Blöcke variiert wurde, entstanden so 96 unterschiedliche Testhefte. Diese 96 Testhefte wurden gleichmäßig über die Klassen verteilt.

Die Lernenden wurden an einem Schulhalbtag während 90 Minuten in Mathematik und Naturwissenschaften getestet. Sie bearbeiteten eines der 96 naturwissenschaftlichen Testhefte entweder vor oder nach der Pause und komplementär ein Testheft mit Mathematikaufgaben. Ein zweiter Schulhalbtag war dem Test in der Erst- und Fremdsprache gewidmet. Die Durchführung der Tests in den Klassen wurde standardisiert; die Lehrperson lei-

tete den Test auf der Basis einer genauen Anweisung.

2.5 Datenerfassung und -kodierung

Die Korrekturanleitungen für die einzelnen Items wurden von der jeweiligen Autorin bzw. dem jeweiligen Autor der Testaufgaben verfasst und von zwei weiteren Personen überprüft. Alle Korrekturanleitungen des nationalen Validierungstests im 6. und 9. Schuljahr existieren sowohl in deutscher als auch in französischer Sprache.

Bei binären Items – z. B. richtig/falsch-Antworten – wurden die Codes 1 und 0 verwendet, bei offenen Antworten mit Ergebnissen in unterschiedlicher Differenzierung wurde aufgrund eines Kriterienkatalogs ein Codespektrum von 0 bis 2 bzw. 3 angelegt. Die inhaltliche Beschreibung der Codes lag in der Verantwortung der Autorinnen und Autoren der Aufgaben. Diese erfolgte ausgehend von fachwissenschaftlichen und fachdidaktischen Grundlagen, von den Erfahrungen der Pilotierung und in einzelnen Fällen mit Berücksichtigung eines Teils der Ergebnisse. Vor Beginn der Korrektur erfolgte nochmals eine Überarbeitung der Korrekturanleitungen.

Die Korrektur erfolgte in mehreren Schritten: Den Anfang bildete eine ausführliche Schulung der 16 Hilfskräfte, je acht in der deutschen und französischen Schweiz. Um eine genügende Interraterreliabilität sicherzustellen, wurden anschließend je zwei offene Aufgaben aus dem 6. und 9. Schuljahr in der deutschen und französischen Schweiz parallel kodiert. Wenn nötig wurde das Verfahren modifiziert bis ein Kappa von mindestens 0.8 erreicht war. Erst dann erfolgte die eigentliche Kodierung: Bei allen Aufgaben wurde zunächst je ein kleinerer Teil der ca. 300 Antworten in der deutschen und französischen Schweiz parallel kodiert. Anfallende Korrekturprobleme wurden sofort diskutiert, die Korrekturanleitungen entsprechend ergänzt und differenziert. Erst danach wurden alle Antworten einer Aufgabe definitiv korrigiert.

2.6 Rasch-Skalierung

Da die Schülerinnen und Schüler je nach Testheft ganz unterschiedliche Aufgaben lösten, sagt die Anzahl richtig gelöster Aufgaben bzw. Items wenig über ihre Leistung im Test aus. Die Daten wurden deshalb auf der Grundlage des Rasch-Modells skaliert[2]. In diesem Modell wird die Wahrscheinlichkeit, dass jemand ein bestimmtes Item lösen kann, als mathematische Funktion der Differenz zwischen der Fähigkeit der Person und der Schwierigkeit des Items dargestellt. Fähigkeit und Schwierigkeit liegen beide auf einer gemeinsamen, latenten Dimension. Zur Charakterisierung der Schwierigkeit des Items wurde bei HarmoS jener Wert auf der latenten Dimension genommen, bei dem Personen mit genau dieser Fähigkeit eine Wahrscheinlichkeit von 2/3 haben, dieses Item richtig zu lösen – das Item

2 Genauere Angaben zur Skalierung der HarmoS-Tests finden sich in Ramseier (2008).

also mit einiger Sicherheit beherrschen. Zur Charakterisierung mehrstufiger Items, bei denen man 0, 1, 2 oder auch 3 Punkte erreichen konnte, wurden jene Schwellenwerte auf der latenten Dimension verwendet, bei denen Personen mit der entsprechenden Fähigkeit eine Wahrscheinlichkeit von 2/3 haben, bei diesem Item mindestens einen Punkt bzw. mindestens 2 Punkte usw. zu erhalten.

Dank der Einordnung von Personen und Items auf der gleichen latenten Dimension ist es möglich, Abschnitte dieser Dimension durch die Art der dort eingeordneten Items bzw. deren Schwellenwerte zu charakterisieren und damit inhaltlich zu beschreiben, was Personen mit dem jeweiligen Fähigkeits- bzw. Kompetenzwert mit guter Sicherheit können.

Die Fähigkeitswerte der Personen und die Schwierigkeits- bzw. Schwellenparameter der Items müssen aus dem Antwortverhalten der Personen auf die verschiedenen Aufgaben geschätzt werden. Diese Schätzung wurde bei HarmoS mit dem Programm Conquest (Wu, Adams, Wilson & Haldane, 2007) vorgenommen. Sie erfolgte separat für das 6. und das 9. Schuljahr, da primär die Kompetenzbeschreibung auf den einzelnen Klassenstufen überprüft werden musste. Die Skalen wurden wie in PISA so linear transformiert, dass der Mittelwert der Schülerpopulation jeweils 500 und die Standardabweichung 100 beträgt. Das Rasch-Modell ermöglicht es, die beiden so definierten Skalen anhand der in beiden Schuljahren eingesetzten Aufgaben zu verknüpfen.

3 Resultate

3.1 Erfassen der naturwissenschaftlichen Kompetenz im 6. bzw. 9. Schuljahr

Itemselektion

Die Selektion von Aufgaben und Items, die zum anvisierten Kompetenzmodell und der Rasch-Skalierung passen, ist als erstes, wichtiges Resultat der Validierung anzusehen: Dank dieser Aufgabenselektion können das Kompetenzspektrum und insbesondere der Basisstandard anhand empirisch überprüfter Referenzaufgaben illustriert werden.

Von den anfänglich 266 Items des Tests für das 9. Schuljahr zeigten 71 Items eine ungenügende Trennschärfe (Korrelation Item – Gesamtscore < .3); einzelne passten zugleich nicht zum Rasch-Modell (infit < 0.7 oder > 1.3). 18 dieser Items und 26 weitere wiesen große Unterschiede zwischen den Sprachregionen auf (DIF-Parameter > 0.4), d. h. die Schwierigkeit des Items wird pro Sprachregion erheblich anders geschätzt als es dem generellen Unterschied zwischen den Sprachregionen entspricht.

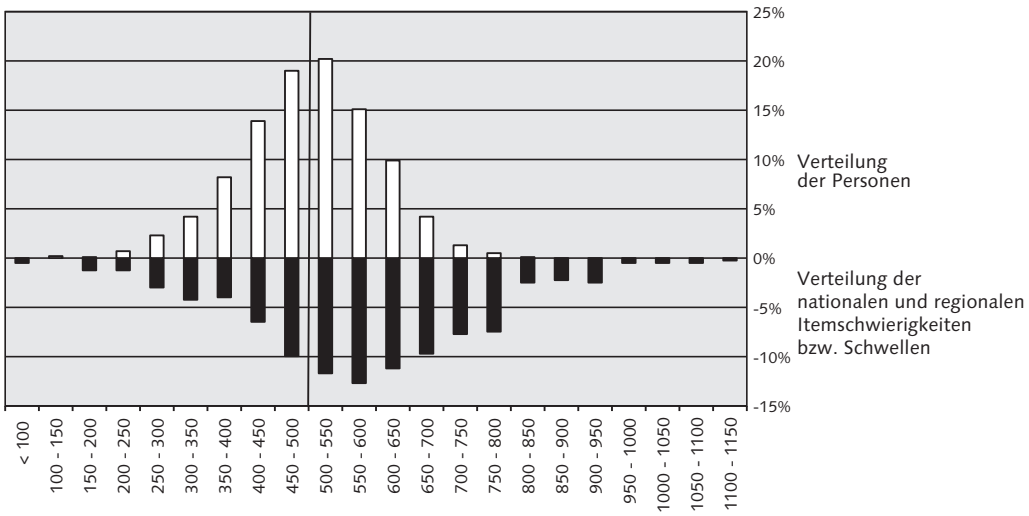
Items mit ungenügender Trennschärfe wurden in der Regel ausgeschieden, zum Teil aber aus wichtigen inhaltlichen Gründen beibehalten. Dies kam primär für sehr leichte oder schwierige Items in Betracht, die zeigen, was fast alle beherrschen oder aber fast niemand in der untersuchten Population kann, und bei denen die Item-Gesamtscore-Korrelation trotz Konformität mit dem Rasch-Modell klein ist.

Items, die sich in den beiden Sprachregionen unterschiedlich verhielten, wurden

nicht ausgeschieden, sondern sie wurden wie zwei Items behandelt, von denen jedes nur in einer Region getestet wurde. Insgesamt wurden so 187 Items als nationale Indikatoren beibehalten und weitere 61 in einer Sprachregion verwendet. Von diesen Items liefern die mehrstufigen gleich zwei oder drei Schwellen, die auf der Fähigkeits- bzw. Schwierigkeitsdimension eingeordnet werden können. Somit stehen nach der Selektion im Test für das 9. Schuljahr 319 national und 83 regional definierte Schwellen zur Charakterisierung dieser latenten Dimension zur Verfügung. Im Test für das 6. Schuljahr wurden von anfänglich 232 Items 184 als nationale beibehalten und 36 wurden als regionale Items eingeführt. Mit dieser Auswahl sind im 6. Schuljahr 335 national und 51 regional definierte Schwellen verfügbar.

Verteilung der Personen und Aufgabenparameter

Die Rasch-Skalierung ermöglicht es, die Verteilung der Personen und der Aufgabenparameter, d. h. der Schwierigkeiten bzw. Schwellenwerte der Items, direkt zu vergleichen. Abbildung 3 zeigt diese beiden Verteilungen. Die Verteilung der Personen im oberen Teil der Abbildung ist auf den Mittelwert 500 und eine Standardabweichung von 100 normiert – diese beiden Werte sagen folglich selbst nichts aus. Die Verteilung weicht nur wenig von einer Normalverteilung ab und ist leicht linksschief ($v = -.20$, $SE = 0.066$): Schülerinnen und Schüler mit sehr niedrigen Kompetenzen sind etwas häufiger als solche mit sehr hohen. Entsprechend liegt der Median mit 504 Punkten leicht über dem Mittelwert. Die Form der Personenverteilung im 6. Schuljahr sieht ähnlich aus.



Senkrechte Linie: Personenmittelwert *M* = 500

Abb. 3: Verteilung der Personen und der Aufgabenparameter im Validierungstest des 9. Schuljahres.

Die Verteilung der Aufgabenparameter im unteren Teil der Abbildung weist einen höheren Mittelwert und eine wesentlich größere Streuung auf als die Personenverteilung. Für die Aufgabenparameter des 9. (6.) Schuljahres sind dies $M = 581$ (624) und $SD = 176$ (194).

Die beiden Verteilungen stimmen nicht ideal überein, wenn es gilt Basisstandards zu beschreiben, da im Bereich von möglichen Standards, die für viele Schülerinnen und Schüler erreichbar sind, nur recht wenige Aufgaben liegen. Dies hängt jedoch damit zusammen, dass die Aufgaben dort lokalisiert sind, wo Personen mit der entsprechenden Fähigkeit sie mit einer Wahrscheinlichkeit von 2/3 lösen können (vgl. Abschnitt 2.6). Für einen Test mit optimalen psychometrischen Eigenschaften für die Beschreibung der Personenwerte ist dagegen maßgeblich, wo Aufgaben mit 50% Wahrscheinlichkeit gelöst werden können. Bei so definierten Itemschwierigkeiten liegen die Mittelwerte der beiden Skalen rund eine Standardabweichung tiefer und stimmen somit recht gut mit den Mittelwerten der Personen überein.

Durch die Normierung der beiden Skalen auf je eine 500/100-Skala geht die Information über die ursprünglich geschätzten Varianzen der beiden Personenverteilungen verloren. Vor der Normierung hat die Skalierung eine Standardabweichung von 0.59 logits für das 6. Schuljahr und von 0.70 für das 9. Schuljahr ergeben. Die größere Streuung im 9. Schuljahr verweist darauf, dass sich die Kompetenzen der Schülerinnen und Schüler im Laufe dieser drei Schuljahre auseinanderentwickeln. Teilweise könnte dies am Schereneffekt liegen, der sich zeigt, wenn Schülerinnen und Schüler durch Selektion unterschiedlichen Schultypen und damit unterschiedlichen Lernumwelten zugeteilt werden (Becker, Lüdtke, Trautwein & Baumert, 2006).

3.2 Kompetenz-Strukturmodell

Das naturwissenschaftliche Kompetenzmodell von HarmoS geht davon aus, dass sich die naturwissenschaftliche Kompetenz nach acht Handlungsaspekten und acht Themenbereichen gliedern lässt. Drei der Handlungsaspekte und alle acht Themenbereiche sind in den Aufgaben des nationalen Validierungstests im 6. bzw. 9. Schuljahr repräsentiert. Grundsätzlich ist es möglich, dass sich die Kompetenz der Schülerinnen und Schüler in jeder dieser 3 mal 8 Kombinationen von Handlungsaspekten und Themenbereichen unterschiedlich entwickelt und ein entsprechendes 24-dimensionales Raschmodell daher ein angemesseneres Bild der Kompetenzstruktur geben würde als das eindimensionale Modell einer einheitlichen naturwissenschaftlichen Kompetenz. Diese Differenzierung sprengt jedoch den Rahmen des empirisch Überprüfbaren. Da im HarmoS-Kompetenzmodell die Niveaus der Handlungsaspekte themenunabhängig beschrieben werden, wird zur Überprüfung der Dimensionalität dem eindimensionalen Modell primär ein Modell gegenübergestellt, das pro Handlungsaspekt eine Leistungsdimension vorsieht, und ergänzend eines, das stattdessen für

Tab. 2: Korrelationen zwischen den Schülerleistungen in Themenbereichen bzw. Handlungsaspekten

Themenbereiche (teils kombiniert)		BS	LE	NM	PE	ST
Bewegung, Energie, Steuerung	BS	-	,70	,60	,72	,71
Lebewesen, Lebensräume	LE	,66	-	,65	,73	,65
Natur, Mensch, Gesellschaft	NM	,60	,69	-	,68	,74
Planet Erde	PE	,63	,73	,73	-	,63
Stoffe	ST	,73	,68	,69	,67	-
Handlungsaspekte		IE	OS	EB		
Informationen erschliessen	IE	-	,83	,84		
Ordnen, strukturieren, modellieren	OS	,94	-	,89		
Einschätzen und beurteilen	EB	,82	,79	-		

oberhalb der Diagonalen jeweils 6. Schuljahr, unterhalb 9. Schuljahr

die Themenbereiche unterschiedliche Leistungsdimensionen vorsieht. Dabei mussten bei letzterem die Themenbereiche „Bewegung, Kraft, Energie" und „Wahrnehmen, Reagieren, Steuern" (in Tabelle 2 mit BS abgekürzt), „Lebewesen" und „Lebensräume und Lebensgemeinschaften" (LE) sowie „Mensch und Gesundheit" und „Natur, Gesellschaft, Technik – Perspektiven" (NM) paarweise zusammengefasst werden, um in beiden Schuljahren stabile Schätzungen zu erreichen.

Die Korrelationen zwischen den drei Handlungsaspekten sind sehr hoch, zwischen .79 und .94 (vgl. Tab. 2). Mit Werten zwischen .60 und .74 liegen die Korrelationen bei den Themenfeldern deutlich tiefer. Beide Male liegen die Werte unter 1 – die Modelle geben die Resultate differenzierter wieder als das eindimensionale Rasch-Modell. Dies zeigt sich auch in den Informationsindices, wonach das Modell nach Themenbereichen am besten, das eindimensionale Modell am schlechtesten abschneidet (CAIC im 9./6. Schuljahr: 106'455/127'048 eindimensional, 106'362/127'018 differenziert nach Handlungsaspekten, 105'849/126'512 differenziert nach Themenbereichen, vgl. Rost, 2004a, p. 344).

Bei der Interpretation dieser Korrelationskoeffizienten ist zu beachten, dass es sich um Schätzungen der wahren Korrelationen zwischen den latenten Dimensionen selbst handelt. Sie sind deshalb höher als die gewohnten Korrelationen zwischen messfehlerbehafteten Variablen. Korrelationen zwischen Handlungsaspekten in der hier gefundenen Höhe sind auch in anderen Untersuchungen zu finden. So korrelieren in der PISA-Erhebung 2006 im 9. Schuljahr der Schweiz die Teilkompetenzen „Erklären", „Erkennen" und „Anwenden" zwischen .90 und .92 (Nidegger, Moreau & Gingins, 2009, p. 106). In der gleichen Untersuchung liegen die Korrelationen zwischen den Wissensbereichen „Wissen über naturwissenschaftliche Un-

tersuchungen und Erklärungen", „Erde und Weltraum", „Lebende Systeme" und „Physikalische Systeme" zwischen .89 und .93 (eigene Berechnung). Anders als bei HarmoS sind hier die Korrelationen zwischen Wissensbereichen bzw. Teilkompetenzen ähnlich hoch.

Im nationalen Naturwissenschaftstest PISA 2003 in Deutschland ist es sogar so, dass die Korrelationen zwischen sechs kognitiven Kompetenzen ($.49 \leq r \leq .89$, kleinste Korrelation zwischen Umgang mit mentalen Modellen und divergentem Denken) teils deutlich kleiner sind als jene zwischen den Fächern Physik, Chemie und Biologie ($.75 \leq r \leq .83$), (Senkbeil, Rost, Carstensen & Walter, 2005, p. 182). Insgesamt scheint es, dass im HarmoS-Test die Differenzierung nach Themenbereichen besonders deutlich ausfällt. Ihnen muss in der weiteren Bearbeitung der HarmoS Bildungsstandards entsprechende Beachtung geschenkt werden (vgl. Abschnitt 4.6).

3.3 Kompetenzniveaus

Im HarmoS Kompetenzmodell werden verbal im 6. bzw. 9. Schuljahr je vier diskrete und gestufte Kompetenzniveaus beschrieben. Die Rasch-Skalierung im Validierungstest führt dagegen primär zu kontinuierlichen Kompetenzdimensionen. Die üblichste Art, Niveaus und Kontinuum zu verknüpfen besteht darin, den Niveaus Abschnitte auf der Skala zuzuordnen – auch wenn dies weder unproblematisch noch die einzige Möglichkeit ist (Rost, 2004b; Schecker & Parchmann, 2006).

Bei HarmoS ordneten Mitglieder des Konsortiums gleich zu Beginn den einzelnen Items das erwartete Kompetenz- bzw. Schwierigkeitsniveau zu, wobei sie sich auf eine erste Version verbaler Niveaubeschreibungen stützten. Diese Zuordnung entsprach aber lediglich einer subjektiven Einschätzung und basierte nicht auf einem theoretischen Modell schwierigkeitsgenerierender Faktoren wie etwa bei Kauertz (2007) oder Bernholt (2010). Sie diente primär als Orientierung bei der Konstruktion des Validierungstests und sollte zu einer angemessenen Streuung der Aufgabenschwierigkeit im Test verhelfen. In einem gewissen Sinne wurden so dennoch a priori Kompetenzniveaus konstituiert.

Wie Abbildung 4 zeigt, überlappen sich Items, die unterschiedlichen Niveaus zugeordnet wurden, nach ihrer im Test ermittelten Schwierigkeit erheblich. Die Korrelation zwischen erwartetem Niveau und empirischer Schwierigkeit ist mit $r = .48$ nicht beeindruckend ($r = .47$ im 6. Schuljahr). Dennoch fallen die Mittelwerte der empirischen Schwierigkeitswerte je nach erwartetem Niveau deutlich unterschiedlich aus ($M = 498$ für Niveau 1, 632 für Niveau 2 und 747 für Niveau 3). Diese Ergebnisse zeigen, wie wichtig es ist, dass Kompetenzniveaus nicht nur verbal umschrieben und mit Aufgaben illustriert werden, von denen Fachleute bloß annehmen, sie würden diese Niveaus widerspiegeln. Die Kenntnis der empirischen Aufgabenschwierigkeit ist für realistische Angaben unerlässlich. Im Projekt wurde dies für die Reformulierung der Kompetenzniveaus genutzt. Eine direkte

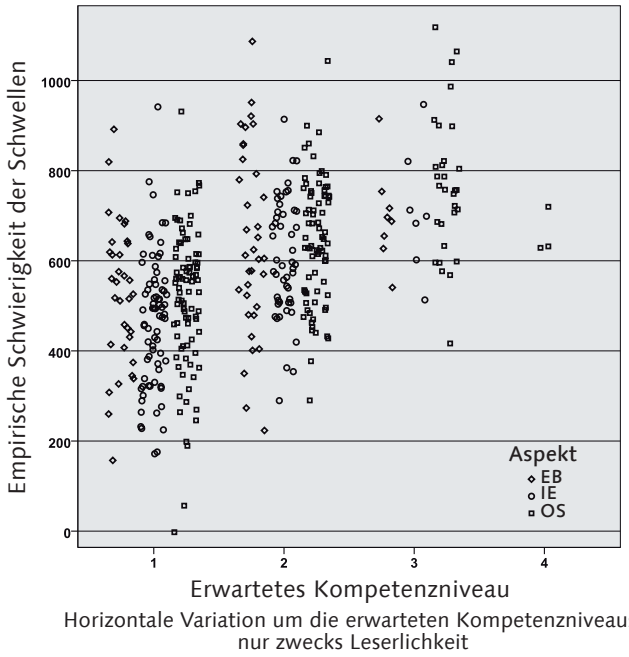


Abb. 4: Empirische Aufgabenschwierigkeit (Schwellen) nach erwartetem Niveau und Handlungsaspekten im 9. Schuljahr

Umsetzung der Niveaubeschreibungen in Abschnitte auf der Schwierigkeitsdimension konnte unter dem bestehenden Zeitdruck nicht vorgenommen werden.

3.4 Vergleich zwischen 6. und 9. Schuljahr

Das Kompetenzmodell beschreibt die naturwissenschaftliche Kompetenz für das 2. bis 9. Schuljahr in einer einheitlichen Struktur. Dazu gehört auch, dass grundsätzlich von einer Überlappung von Kompetenzniveaus zwischen den Schuljahren ausgegangen wird. Um diese Überlappung zu überprüfen, wurden 47 Items in die Tests beider Schuljahre aufgenommen so dass nun die Kompetenzverteilungen im 6. und 9. Schuljahr verglichen werden können.

Die Itemschwierigkeiten dieser gemeinsamen Items wurden sowohl bei der Skalierung im 6. als auch im 9. Schuljahr geschätzt. Der Unterschied zwischen den beiden Mittelwerten dieser Itemschwierigkeiten zeigt gerade, um wie viel sich die Nullpunkte der beiden Skalen unterscheiden. Durch Addition dieses Unterschieds lassen sich dann gemäß dem Rasch-Modell Werte von der einen Skala auf die andere umrechnen. Vorausgesetzt ist, dass die Skalen dasselbe messen. Ein Indiz dafür ist, dass die Streuungen dieser Itemschwierigkeiten um ihren jeweiligen Mittelwert gleich sind (je $SD = 0.77$ logits).

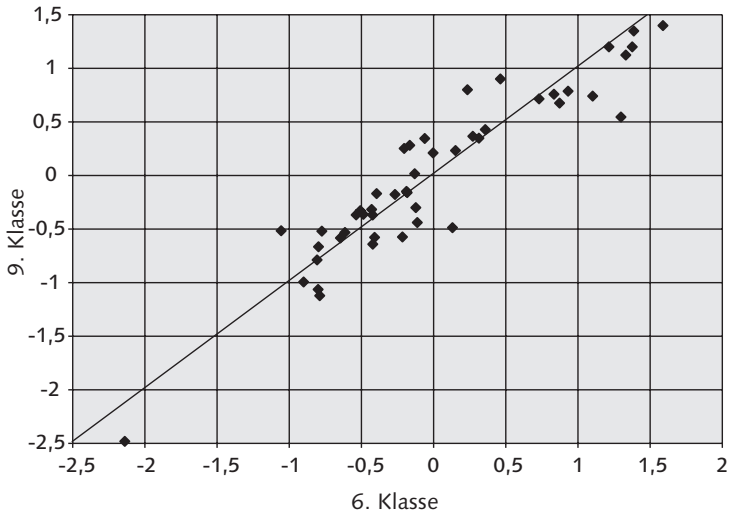


Abb. 5: Relative Itemschwierigkeit der gemeinsamen Items im Test des 6. und 9. Schuljahres.

Zudem zeigt Abbildung 5, dass sich die relativen Schwierigkeiten (Itemschwierigkeit minus mittlere Schwierigkeit aller gemeinsamen Items in logits) zwischen den Schuljahren nur mäßig und ohne nennenswerten systematischen Trend unterscheiden.

Unter Berücksichtigung der unterschiedlichen Nullpunkte der beiden Skalen zeigt sich, dass der Populationsmittelwert des 9. Schuljahres um 0.39 logits ($SE = 0.042$) über jenem des 6. Schuljahres liegt. In der auf 500/100 normierten Skala des 9. Schuljahres entspricht dies einem Zuwachs[3] von 56 Punkten, auf der Skala des 6. Schuljahres von 67 Punkten.[4]

Der ermittelte Zuwachs ist in verschiedener Hinsicht klein. Zunächst ist er gemessen am damit verknüpften Gewinn an Aufgabenbewältigung bescheiden: Wenn ein Schüler bzw. eine Schülerin des 6. Schuljahres mit gerade durchschnittlichen Kompetenzen eine bestimmte Aufgabe mit einer Wahrscheinlichkeit von 50 % lösen kann, dann ist die entsprechende Wahrscheinlichkeit im 9. Schuljahr 60 %, also nicht massiv höher. Wie bescheiden der Unterschied zwischen den Schuljahren ausfällt, zeigt sich auch daran, dass er über die drei Jahre hinweg weniger als halb so groß ist wie zwischen den Schultypen mit Grundansprüchen bzw. hohen Ansprüchen innerhalb des 9. Schuljahres (vgl. Tabelle 3).

Der bescheidene Kompetenzzuwachs ist auch kaum verträglich mit der Vorstellung, dass sich nur zwei der vier Kompetenzniveaus des 6. bzw. 9. Schuljahres überdecken (vgl. Abschnitt 1.3). Falls man bei dieser theoretischen Vorgabe von üb-

3 Insofern als keine Gründe für erhebliche sonstige Populationsunterschiede vorliegen, kann die Mittelwertsdifferenz weitgehend als Zuwachs interpretiert werden.

4 Alternativ wurden die Daten des 6. und 9. Schuljahres unter Einbezug der Link-Items gemeinsam skaliert. Diese Analyse bestätigt die hier gefundenen Unterschiede, insbesondere auch die im 6. Schuljahr kleinere Varianz der Personenverteilung (vgl. Abschnitt 3.1).

lichen Niveaubreiten ausgeht, z. B. 0.8 logits beim PISA Naturwissenschaftstest 2006 (OECD, 2009, pp. 246, 293), muss man in Kauf nehmen, dass die Population im 9. Schuljahr mit großer Mehrheit nur die untersten beiden Kompetenzniveaus erreicht, falls sich jene des 6. Schuljahres symmetrisch über die vier dort festgelegten Niveaus verteilt. Andernfalls kann man die Kompetenzniveaus so schmal definieren, dass sich beide Populationen symmetrisch über die Niveaus verteilen. Dann wären die beiden sich überlappenden Niveaus aber sehr schmal und inhaltlich kaum aussagekräftig.

Der Zuwachs über die drei Schuljahre ist auch im Vergleich zu den Ergebnissen anderer Studien klein. So wird in PISA (OECD, 2007a, p. 55) der Zuwachs in den naturwissenschaftlichen Kompetenzen in einem Jahr auf 38 Punkte geschätzt, bei einer gleichartigen 500/100-Normierung der Skala. In TIMSS betrug der Unterschied zwischen der Stichprobe im 6. bzw. 7. Schuljahr der Schweiz 38 Punkte (Beaton et al., 1996, p. 29). Hochgerechnet auf drei Jahre gäbe das erheblich mehr als im HarmoS-Test. Auch in einem echten Längsschnitt wurde der Zuwachs in einem Jahr auf rund 1/3 einer Standardabweichung

Tab. 3: Kompetenzunterschiede[1] nach Region, Schultyp und Geschlecht in der Gesamtskala und in Teildimensionen im 9. Schuljahr

	Region: Differenz deutsche - französische Schweiz			**Schultyp:** Differenz hohe Ansprüche - Grundansprüche			**Geschlecht:** Differenz Frauen - Männer		
	Differenz	***SE***	***p*-Wert**	**Differenz**	***SE***	***p*-Wert**	**Differenz**	***SE***	***p*-Wert**
Gesamtskala	17	7,0	,013	138	8,3	<.001	-12	5,4	,027
Themenbereiche (teils kombiniert)									
Bewegung, Energie, Steuerung	12	7,5	,103	126	9,7	<.001	-32	6,7	<.001
Lebewesen, Lebensräume	10	10,5	,318	188	10,8	<.001	-5	7,6	,551
Natur, Mensch, Gesellschaft	19	9,4	,042	152	10,5	<.001	8	6,9	,257
Planet Erde	15	6,5	,023	120	7,6	<.001	-12	5,5	,025
Stoffe	29	6,8	<.001	125	8,9	<.001	-22	6,4	,001
Handlungsaspekte									
Informationen erschliessen	22	8,0	,006	151	9,2	<.001	-19	5,7	,001
Ordnen, strukturieren, modellieren	12	6,9	,093	136	8,5	<.001	-18	5,8	,002
Einschätzen und beurteilen	27	7,2	,000	122	9,1	<.001	4	5,4	,485

[1] Punktdifferenzen basierend auf der $M = 500 / SD = 100$-Normierung der Gesamtskala
Berechnung basierend auf plausiblen Werten und designgerechter Fehlerschätzung

geschätzt, diesmal allerdings für Mathematik im 7. bis 8. Schuljahr in Deutschland (Becker et al., 2006, p. 237).
Zum relativ geringen Zuwachs im HarmoS-Validierungstest dürfte beitragen, dass hier – gerade im Gegensatz zu TIMSS – eine sehr allgemeine Scientific Literacy getestet wird. Zudem wird nach gegenwärtigem Lehrplan (noch) nicht entsprechend dem HarmoS-Kompetenzmodell unterrichtet.

3.5 Kompetenzen in Teilpopulationen

Im HarmoS Validierungstest steckt ein beträchtliches deskriptives Potenzial, das bisher kaum genutzt wurde. Auch hier können nur beispielhaft einige Kompetenzunterschiede für das 9. Schuljahr dargestellt werden.
Der HarmoS-Validierungstest war bei der Aufgabenselektion auf einen fairen Vergleich zwischen der deutschen und französischen Sprachregion ausgerichtet. In der Gesamtskala der naturwissenschaftlichen Kompetenz zeigt sich nun, dass in der deutschen Schweiz leicht höhere Leistungen erreicht werden als in der französischen Schweiz – ein Unterschied, der statistisch signifikant ausfällt (5 %-Niveau). Dieser Unterschied zieht sich mit Variationen durch alle Handlungsaspekte und Themenbereiche und ist beim Handlungsaspekt „Einschätzen und beurteilen" sowie beim Themenbereich „Stoffe" am ausgeprägtesten.
Die Sekundarstufe I ist in der Schweiz in den 26 Kantonen sehr unterschiedlich organisiert. Die Spannweite reicht von integriert-differenzierten Gesamtschulen als einzigem Angebot bis zu viergliedrigen Systemen. Dies führt zu einer Vielfalt von Schultypen mit unterschiedlichen Leistungsanforderungen. Statt diese Vielfalt darstellen zu wollen, werden hier nur die Leistungen jener Schultypen verglichen, die lediglich Grundanforderungen stellen oder aber hohe, i. A. gymnasiale Ansprüche stellen. Die beiden polaren Typen repräsentieren gemäß unserer Einteilung 29 bzw. 27 Prozent der Schülerpopulation. Die Ergebnisse der übrigen Schultypen liegen immer dazwischen und zeigen weniger Profil.
Wie zu erwarten, übertrifft der Kompetenzmittelwert im Schultyp mit hohen Anforderungen in der Gesamtskala jenen im Schultyp mit Grundanforderungen ganz erheblich (Tab. 3). Interessanter ist, dass dieser Unterschied im Handlungsaspekt „Informationen erschließen" signifikant größer ist als bei „Einschätzen und beurteilen".
Zwischen den Themenbereichen variieren die Unterschiede zwischen den Schultypen noch ausgeprägter. Dies geht soweit, dass der größte Unterschied (beim zusammengefassten Bereich „Lebewesen, Lebensräume") um die Hälfte größer ist als bei „Planet Erde" und 188 Punkte ausmacht. Der Unterschied beim Spitzenreiter ist auch noch signifikant größer als beim Bereich „Natur, Mensch, Gesellschaft", der den zweitgrößten Unterschied zwischen Schultypen aufweist.
Nach Geschlecht zeigt sich ein leichter aber signifikanter Kompetenzvorsprung der jungen Männer in der Gesamtskala. Es ist bemerkenswert, dass dieser Vorsprung

beim Aspekt „Einschätzen und beurteilen“ ganz entfällt, was signifikant zu den andern beiden Aspekten kontrastiert. Unter den Bereichen fällt der Vorsprung der Männer bei „Bewegung, Energie, Steuerung“ signifikant größer aus als in der Gesamtskala; beim kombinierten Bereich „Natur, Mensch, Gesellschaft“ deutet sich dagegen eher ein Rückstand an.

Die gefundenen Unterschiede bestätigen was bereits anderswo festgestellt wurde. So schneidet die deutsche Schweiz auch bei PISA 2006 besser ab als die französische (Nidegger et al., 2009) und die jungen Männer verzeichnen auch bei PISA 2006 einen leichten Vorsprung vor den jungen Frauen (6 Punkte, vgl. OECD, 2007b, p. 27). Dieser ist beim Bereich „Physikalische Systeme“ besonders ausgeprägt (32 Punkte, a.a.O., p. 52), er ist auch bei „Erde und Weltraum“ zu finden (26, a.a.O., p. 48) und entfällt bei „Lebende Systeme“ (a.a.O., p. 50).[5]

Die Konsistenz der HarmoS-Ergebnisse mit PISA kann man als Hinweis auf die Validität der HarmoS-Skala ansehen. Die je nach Themenbereich unterschiedlichen Kompetenzdifferenzen zwischen den Schultypen könnten ein interessanter Hinweis auf unterschiedliche Gewichtungen der Themen im Unterricht sein.

3.6 Basisstandards

Die Basisstandards wurden in einem iterativen Prozess zwischen normativen Setzungen aufgrund des erarbeiteten Kompetenzmodells und Lehrplänen einerseits und den Ergebnissen der Tests andererseits festgelegt: Entwicklung einer ersten Version des Kompetenzmodells mit Niveaubeschreibungen, Überarbeitung des Modells und der Beschreibungen aufgrund der Rückmeldungen von Lehrpersonen und Experten/-innen, Durchführen von Tests, weitere Überarbeitung des Modells und der Beschreibungen aufgrund der Testresultate, Vorschlagen der Basisstandards (meist entsprechend Niveau I oder II), Veranschaulichen der Standards durch typische Aufgaben im Sinne von Ankerbeispielen (Referenzaufgaben). Dabei wurden die aus den Tests bekannten Lösungshäufigkeiten der Referenzaufgaben benutzt, um realistische Standards zu formulieren. So liegen die Lösungshäufigkeiten jener Referenzaufgaben, die aus den beiden repräsentativen Validierungstests übernommen wurden, zwischen 65 und 89 Prozent.

Als Resultat dieses Prozesses hat das Konsortium der EDK verbal umschriebene und durch Referenzaufgaben charakterisierte Basisstandards vorgeschlagen. Nachdem die EDK sie aufgrund von internen Entscheiden und Rückmeldungen aus breiten interessierten Kreisen mehrfach überarbeitet hat, hat sie die Basisstandards 2011 freigegeben und damit politisch in Kraft gesetzt. Für den Handlungsaspekt „Informationen erschließen“ wird zum Beispiel am Ende des 6. Schuljahres eine der Grundkompetenzen (Basisstandards) als „Can-do“ wie folgt formuliert: **„aus aufbereiteten Informationen** (z. B. aus Lehrmitteln, aus Jugendsachbüchern, aus dem Internet) **Angaben und Sachverhalte**

5 Die Aspekte bei HarmoS lassen sich weniger gut bestimmten Teilkompetenzen von PISA zuordnen (Nidegger et al., 2009, p. 104).

herauslesen und in selber gewählten Formen zusammenstellen sowie Informationen lesen und mit eigenen Worten, Sachbegriffen und Darstellungsformen beschreiben und wiedergeben (z. B. Angaben aus einer Grafik in einer Tabelle zusammenstellen, einfaches Schema z. B. zu Energieumwandlungen zeichnen, kurzer Sachtext zu einem Phänomen wie «Farben beim Regenbogen» entwerfen, Mind-Map z. B. zum Lebensraum Wald darstellen oder Erklärung zum Stichwort «Puls» mit Text und Skizze zusammenstellen)" (EDK, 2011, p. 27, Hervorhebung im Original).

Anhand des repräsentativen Validierungstests lässt sich bestimmen, welcher Anteil der Schülerschaft den Basisstandard erfüllt. Dazu muss zu diesem Test ein Schwellenwert festgelegt werden, ab dem der Basisstandard als erreicht angesehen wird. Dieser Schwellenwert kann nicht stringent aus der verbalen Beschreibung des Standards abgeleitet werden; vielmehr muss er interpretiert und umgesetzt werden, womit unvermeidlich auch subjektiv geprägte Beurteilungen ins Spiel kommen. In Ländern mit langer Testtradition wurden aber zahlreiche Verfahren entwickelt, die absichern, dass dieser Prozess explizit, transparent und nachprüfbar erfolgt (Cizek, 2001; Nichols, Twing, Mueller & O'Malley, 2010; Zieky & Perie, 2006). Ein solches Verfahren konnte unter dem Zeitdruck der politischen Vorgaben in HarmoS jedoch nicht realisiert werden. Es kann auch hier nicht vollwertig nachgeholt werden. Es ist aber möglich, in einer Annäherung zu bestimmen, wie oft die Basisstandards im Validierungstest erreicht werden, indem die Referenzaufgaben zur Operationalisierung des Schwellenwerts genutzt werden.

Die Basisstandards zu den im Validierungstest erfassten Handlungsaspekten werden im Schlussbericht (Konsortium, 2008) mit je 11 Referenzaufgaben für das 6. und 9. Schuljahr illustriert. Man kann jede dieser Aufgaben als Vorschlag ansehen, den Basisstandard so zu definieren, dass Schülerinnen und Schüler, die diesen Standard gerade erfüllen, die betreffende Aufgabe mit einer Wahrscheinlichkeit von 2/3 lösen können. Dann sind die Schwierigkeitsparameter der elf Referenzaufgaben jeweils ein Vorschlag zur Festlegung des Schwellenwerts des Basisstandards. Der Mittelwert dieser elf Parameter kann als beste Annäherung an den Basis-Schwellenwert im Validierungstest angesehen werden. So erhält man für den Basisstandard im 6. Schuljahr einen Schwellenwert von 447 Punkten, im 9. Schuljahr von 426 Punkten. Diese Schwellenwerte werden im 6. Schuljahr von 28 %, im 9. von 22 % der Schülerinnen und Schüler nicht erreicht.

Die Basisstandards für das 6. und 9. Schuljahr wurden hier unabhängig voneinander festgelegt. Dass der Anteil der Schülerschaft, der den Basisstandard erfüllt, im 9. Schuljahr höher ist, ergibt sich aus diesen Festlegungen und kann nicht als Erfolg des Unterrichts interpretiert werden.

Will man wie HarmoS (EDK, 2010a, p. 2), dass praktisch alle Schülerinnen und Schüler die Basisstandards erfüllen, so ist man gemäß dieser provisorischen Festlegung des Standards noch weit vom Ziel entfernt. Zwar kann man sich erhoffen, dass mehr Schülerinnen und Schü-

ler den Basisstandard erfüllen, wenn der Unterricht später besser auf die Anforderungen der naturwissenschaftlichen Standards und des zugehörigen Kompetenzmodells ausgerichtet ist. Es ist aber davon auszugehen, dass es ganz erhebliche Anstrengungen braucht, damit ein wesentlich größerer Anteil der Schülerinnen und Schüler das vorgegebene Kriterium erfüllt.

4 Fazite und Defizite

Welche Schlussfolgerungen lassen sich ziehen und welche Antworten ergeben sich zu den eingangs gestellten Fragen? Die folgenden Fazite und Defizite beziehen sich in der Reihenfolge auf die Unterkapitel 3.1 bis 3.6. Vorausgehend ist als erstes positives Fazit festzuhalten, dass sich die grundsätzliche Orientierung des Harmos-Projekts an der Klieme-Expertise bewährt hat: Bereits während der Formulierung der Basisstandards und des Kompetenzmodells konnte mit dessen Validierung begonnen werden, wie das u. a. die Teilnehmenden der Schwerpunkttagung „Kompetenzmodelle und Bildungsstandards: Aufgaben für die naturwissenschaftsdidaktische Forschung“ (Labudde et al., 2009) gefordert hatten. Dies führt zu einer einfacheren Situation als in Deutschland, wo Basis- bzw. Mindeststandards noch immer erst Diskussionsgegenstand sind (GFD, 2009) und ein differenziertes Kompetenzmodell und seine Validierung auf im Voraus formulierte Bildungsstandards bezogen werden müssen (Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010).

4.1 Empirisch verankerte Beispiele

Im 6. und 9. Schuljahr wurden 90 Aufgaben mit insgesamt 498 Items eingesetzt. Jedes Item wurde von mindestens 300 Schülerinnen und Schülern beantwortet, wovon jeweils 150 aus der deutschen bzw. französischen Schweiz stammten. Im 6. Schuljahr ist es das erste Mal, im 9. Schuljahr nach PISA das zweite Mal, dass man in der Schweiz auf nationaler Ebene aus naturwissenschaftsdidaktischer Perspektive die Leistungen von Kindern und Jugendlichen repräsentativ erfasst hat. Die Aufgaben und Items wurden auf ihre Eignung für einen Leistungstest und auf ihre Fairness bezüglich der Verwendung in unterschiedlichen Sprachregionen überprüft und ihr Schwierigkeitsgrad bzw. ihre Einordnung in eine Kompetenzskala wurde ermittelt. Damit kann man sich erstmals auf empirisch verankerte Beispiele abstützen, sei es beim Festlegen und Illustrieren von Bildungsstandards, sei es in der zukünftigen Aus- und Weiterbildung von Lehrpersonen. An zahlreichen von uns gehaltenen Vorträgen, Seminaren und Kursen zeigten Studierende, Lehrkräfte und Fachdidaktikdozierende größtes Interesse an den Referenzaufgaben und deren Resultaten.

4.2 Kompetenz-Strukturmodell

Die Analyse der Testergebnisse nach Themenbereichen hat gezeigt, dass die Leistungen in den verschiedenen Themenbereichen nur in mittlerem Ausmaß korrelieren. Denkbar ist, dass die Auswahl

der thematischen Kontexte und damit der Bezug zum vorhandenen Vorwissen der Schülerinnen und Schüler einen nicht unbedeutenden Einfluss haben. Dies konnte allerdings im Rahmen der Tests nicht genauer überprüft werden. Streng genommen muss damit ein Kompetenzmodell, das die Leistungen über die verschiedenen Themenbereiche hinweg in einer Leistungsdimension zusammenfasst, in Frage gestellt werden. Es ist allerdings nicht unüblich, dass parallel ein- und mehrdimensionale Rasch-Modelle verwendet werden (Ramseier, 2008). Zumindest müssen aber als Konsequenz die einzelnen Themenbereiche in künftigen Tests zur Überprüfung der Bildungsstandards unbedingt ausgewogen vertreten sein, damit repräsentative Ergebnisse erreicht werden können.

Demgegenüber fallen die Korrelationen zwischen den im PB-Test berücksichtigten Handlungsaspekten deutlich höher aus (vgl. Tabelle 2). Dass im HarmoS-Modell primär nach Handlungsaspekten differenziert wird, kann somit zwar fachdidaktisch, aber nicht psychometrisch begründet werden.

Für die Überprüfung der Kompetenzstruktur wurde ein eingeschränktes Modell verwendet. So wurde jedes Item genau einem Kompetenzaspekt und einem Themenbereich zugeordnet, obwohl vielfach mehrere Aspekte relevant sein dürften. Dass diese vereinfachte Zuteilung der Items zu nur einem Kompetenzaspekt nicht optimal ist, könnte gerade auch für die hohe Korrelation zwischen den Aspekten verantwortlich sein. Auch wurde die erwartete Dimensionalität nur konfirmatorisch überprüft. Es kann nicht ausgeschlossen werden, dass andere Dimensionen das Antwortverhalten differenzierter beschreiben.

4.3 Kompetenzniveaus

Bereits bei der Entwicklung des Kompetenzmodells zeigten sich Schwierigkeiten, auf die Teilaspekte der Handlungsaspekte bezogen innerhalb eines Zyklus vier wirklich differente Anspruchsniveaus zu formulieren und dies auch sprachlich zum Ausdruck zu bringen. Bei der Konzeption von Aufgaben ergaben sich zudem Schwierigkeiten, Items präzise auf bestimmte Anspruchsniveaus hin zu formulieren.

Die geringe Korrelation zwischen a priori erwartetem Niveau, wie dies bei der Entwicklung der Items festgelegt wurde, und der berechneten empirischen Schwierigkeit eines Items aufgrund der Testergebnisse zeigt, dass die Einschätzung von Aufgabenschwierigkeiten äußerst diffizil ist. Umso wichtiger ist die empirische Verankerung von Referenzaufgaben.

Über die a priori Einschätzung hinaus konnte im gegebenen Zeitrahmen kein systematischer Bezug zwischen den verbalen Niveaubeschreibungen im Kompetenzmodell und der empirisch bestimmten Kompetenzdimension und ihrer möglichen Stufung ausgearbeitet werden. Das ist umso bedauerlicher, als das Raschmodell dank der Zuordnung von Personen und Aufgaben auf der gleichen Dimension die Grundlage dafür liefert, die Merkmale der für die einzelnen Skalenbereiche typischen Aufgaben zu identifizieren und

sie mit den Niveaubeschreibungen des Kompetenzmodells zu konfrontieren. Dies und die theoretische Absicherung solcher Niveaus sind dringende Aufgaben bei der Weiterentwicklung des Kompetenzmodells – Aufgaben, die auch für die auf das Kompetenzmodell gestützte Aufgabenentwicklung wichtig sind.

4.4 Zuwachs vom 6. zum 9. Schuljahr

Dank 47 gemeinsamen Items im Test für das 6. bzw. 9. Schuljahr kann der Kompetenzzuwachs zwischen diesen beiden Schuljahren geschätzt werden. Mit knapp 2/3 einer Standardabweichung fällt der Zuwachs im Vergleich zur Streuung innerhalb der Schulstufen bescheiden aus und entspricht nicht einer zu erwartenden Progression in drei Schuljahren. Das macht es schwierig, wie theoretisch vorgesehen ein Gesamtmodell zu konstruieren, in dem sich zwei der vier Kompetenzniveaus des 6. und 9. Schuljahres überlappen, und gleichzeitig diese Niveaus inhaltlich deutlich abgestuft zu formulieren. Es wird eher ein Gesamtmodell zu präzisieren sein, in dem drei der Niveaus einander zugeordnet werden. Zu beachten ist ferner, dass hier nur Kompetenzverteilungen in unterschiedlichen Stichproben verglichen wurden. Aussagen über individuelle Kompetenzentwicklungen sind somit nicht möglich.

4.5 Teilpopulationen

Die im Validierungstest gefundenen Unterschiede zwischen den Sprachregionen und die Unterschiede zwischen den Geschlechtern stimmen gut mit den Ergebnissen von PISA 2006 überein. Die Unterschiede zwischen den Schultypen entsprechen den Erwartungen. Diese Ergebnisse können als Belege für die Konstruktvalidität der Tests und damit des Kompetenzmodells angesehen werden. Ein weiterer Beleg dafür erbrachten Nidegger et al. (2009), die die Rohdaten des Naturwissenschaftstests von PISA 2006 nach den Dimensionen des hier beschriebenen HarmoS-Kompetenzmodells analysierten und die Ergebnisse mit jenen verglichen, die auf der originalen PISA-Struktur beruhen.

4.6 Basisstandards

In einem iterativen Prozess zwischen normativen Setzungen und Reformulierungen des Kompetenzmodells einerseits und der Berücksichtigung der Testergebnisse andererseits wurden für das 2., 6. und 9. Schuljahr verbal umschriebene Basisstandards definiert und mit Referenzaufgaben illustriert. Diese Vorgehensweise ist in dem Sinne pragmatisch, als die verschiedenen Protagonisten einbezogen wurden und die Standards nicht nur definiert, sondern auch mittels empirisch gesicherter Referenzaufgaben veranschaulicht werden. Als Manko muss vermerkt werden, dass die Basisstandards aus zeitlichen Gründen nicht nach einem expliziten Verfahren in

Basis-Schwellenwerte in den Validierungstests umgesetzt werden konnten und dass für die in den Tests nicht berücksichtigten Handlungsaspekte – welche allerdings sehr schwierig und aufwändig zum Testen sind – Basisstandards formuliert wurden, welche normativen Setzungen mit Bezug zu Erfahrungen von Fachdidaktikern und Schulpraktikern entsprechen. Damit bleibt trotz der Leistungsmessung in repräsentativen Stichproben des 6. und 9. Schuljahres offen, welcher Anteil der Schülerschaft heute die Basisstandards erfüllt.

Um dennoch näherungsweise etwas über das Ausmaß des Erreichens des Basisstandards aussagen zu können, kann man in heuristischer Weise definieren, dass Lernende, welche den Standard gerade erfüllen, entsprechende Referenzaufgaben mit einer Wahrscheinlichkeit von 2/3 lösen können. Gemäß dieser Definition hätten in den Validierungstests nur 72 % (78 %) der Schülerinnen und Schüler des 6. (9.) Schuljahres die Basisstandards erreicht. Sollte ein späteres Bildungsmonitoring ähnliche Werte liefern, ließe sich kaum argumentieren, die Schweizer Kinder würden die gesetzten Basisstandards in befriedigendem Ausmaß erfüllen. Insbesondere zeigen die Erfahrungen im Projekt HarmoS, dass für die transparente Festlegung und Operationalisierung von Basisstandards nach den gängigen Verfahren (z. B. Cizek, 2001) im künftigen Bildungsmonitoring genügend Zeit eingeplant werden muss.

4.7 Bilanz

Die Validierung eines Kompetenzmodells ist ein langwieriger, umfassender und kaum abschließbarer Prozess. Im HarmoS-Projekt konnten davon erste Schritte geleistet werden. Dazu sind auch Erfahrungen zu zählen, die im Verlaufe der in der Praxis breit abgestützten Bearbeitung des Modells und seiner Konkretisierung und Illustration in Aufgaben sowie mit den auf die Erfassung komplexer Kompetenzen gerichteten Experimentiertests gewonnen wurden. Die hier beschriebenen repräsentativen Validierungstests ermöglichten es auf das Kompetenzmodell bezogene Aufgaben bzw. Items dahingehend zu überprüfen, ob sie sich in eine Dimension naturwissenschaftlicher Kompetenz einordnen lassen und über die kulturellen und didaktisch-unterrichtlichen Grenzen der Sprachregionen hinweg verwendbar sind. Damit konnten insbesondere die Basisstandards mit empirisch verankerten Aufgaben von bekannter Schwierigkeit illustriert werden. Weiter konnten die Dimensionalität des Kompetenzmodells überprüft und der Kompetenzunterschied zwischen 6. und 9. Schuljahr sowie die Überlagerung der beiden Kompetenzverteilungen untersucht werden.

Nach diesen ersten Schritten bleibt noch manches zu tun. Neben den bereits aufgeführten Defiziten, insbesondere dem fehlenden Bezug zwischen theoretischen Kompetenzniveaus und der aufgabengestützten inhaltlichen Beschreibung von Kompetenzniveaus, lassen sich folgende weitere Lücken feststellen:

- In die Tests wurden nur fünf, in die repräsentativ angelegten Papier-und-Bleistift-Tests nur drei der acht Handlungsaspekte einbezogen.
- Der Kompetenzentwicklungsprozess wurde nicht erfasst, die Tests erfolgten als Momentaufnahmen in verschiedenen Schuljahren.
- Lehrpersonen und weitere Experten/-innen erhielten zwar häufig die Gelegenheit, Modell, Standards und Aufgaben zu kommentieren bzw. arbeiteten bei deren Entwicklung mit, jedoch ersetzt dies nicht eine stringente kommunikative Validierung.
- Die Schülerinnen und Schüler, welche in den Tests mitarbeiteten, wurden nicht auf der Grundlage dieses Modells ausgebildet; es zeigte sich insbesondere, dass sie zu einzelnen Handlungsaspekten und Themenbereichen noch wenig oder sogar keine schulischen Lernerfahrungen haben.

Im Weiteren lässt sich festhalten, dass das Potenzial der vorhandenen Daten noch nicht ausgeschöpft wurde: einerseits ließen sich die Ergebnisse, z. B. einzelner Aufgaben oder Items, fachdidaktisch analysieren, andererseits würden die Daten interessante Analysen aus der Perspektive der Bildungsforschung ermöglichen, z. B. Zusammenhänge zwischen sozioökonomischem Hintergrund, Interessen und Leistungen.
Die erwähnten Defizite dürfen nicht darüber hinwegtäuschen, dass trotz gewisser konzeptioneller und testtheoretischer Defizite das Kompetenzmodell, die Beschreibungen der Standards und insbesondere auch die zahlreichen Referenzaufgaben ein solides Fundament bilden: einerseits für die anstehenden politischen Entscheidungen und die Entwicklung des neuen Lehrplans, andererseits für die Entwicklung des Unterrichts hin zu einem vermehrt kompetenzorientierten Lernen und Lehren.

Literatur

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzales, E. J., Smith, T. A. & Kelly, D. L. (1996). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Becker, M., Lüdtke, O., Trautwein, U. & Baumert, J. (2006). Leistungszuwachs in Mathematik. Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? *Zeitschrift für Pädagogische Psychologie, 20,* 233–242.

Bernholt, S. (2010). *Kompetenzmodellierung in der Chemie (Dissertation).* Oldenburg: Carl von Ossietzky Universität.

Cizek, G. J. (Ed.) (2001). *Setting Performance Standards. Concepts, Methods, and Perspectives.* Mahwah, NJ: Lawrence Erlbaum.

EDK (2010a). *Basisstandards für die Naturwissenschaften: Unterlagen für den Anhörungsprozess.* Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK). www.edk.ch => HarmoS (6. Mai 2011).

EDK (2010b). *DAS KANN ICH. Gemeinsame Grundkompetenzen für unsere Schülerinnen und Schüler: Schweizerische Bildungsstandards für vier Fachbereiche.* Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK). www.edk.ch => HarmoS (6. Mai 2011).

EDK (2011). *Grundkompetenzen für die Naturwissenschaften. Nationale Bildungsstandards. Freigegeben von der EDK-Plenarversammlung am 16. Juni 2011.* Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK). www.edk.ch -> HarmoS -> Nationale Bildungsziele (22.November 2011).

GFD (Gesellschaft für Fachdidaktik) (2009). Mindeststandards am Ende der Pflichtschulzeit. *Zeitschrift für Didaktik der Naturwissenschaften, 15,* 371–377.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4. ed., pp. 17–64). Westport, CT: American Council on Education.

Kauertz, A. (2007). *Schwierigkeitserzeugende Merkmale physikalischer Testaufgaben (Dissertation).* Essen: Universität Duisburg-Essen.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E., & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften, 16,* 135–153.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards: eine Expertise.* Bonn: Bundesministerium für Bildung und Forschung.

Konsortium (2008). *HarmoS Naturwissenschaften+: Kompetenzmodell und Vorschläge für Bildungsstandards (Wissenschaftlicher Schlussbericht).* Bern: PHBern.

Labudde, P. & Adamina, M. (2008). HarmoS Naturwissenschaften: Impulse für den naturwissenschaftlichen Unterricht von morgen. *Beiträge zur Lehrerbildung, 26,* 351–360.

Labudde, P., Duit, R., Fickermann, D., Fischer, H., Harms, U., Mikelskis, H. et al. (2009). Schwerpunkttagung ‹Kompetenzmodelle und Bildungsstandards: Aufgaben für die naturwissenschaftsdidaktische Forschung›. *Zeitschrift für Didaktik der Naturwissenschaften, 15,* 343–370.

Nichols, P., Twing, J., Mueller, C. D. & O'Malley, K. (2010). Standard-Setting Methods as Measurement Processes. *Educational Measurement: Issues and Practice, 29*(1), 14–24.

Nidegger, C., Moreau, J. & Gingins, F. (2009). Kompetenzen der Schülerinnen und Schüler in den Naturwissenschaften: Erkenntnisse aus PISA und HarmoS. In Bundesamt für Statistik (Ed.), PISA 2006: *Analysen zum Kompetenzbereich Naturwissenschaften. Rolle des Unterrichts, Determinanten der Berufswahl, Vergleich von Kompetenzmodellen* (pp. 92–120). Neuchâtel: Bundesamt für Statistik (BFS).

OECD (2007a). PISA 2006. *Science Competencies for Tomorrow's World. Volume 1 – Analysis.* Paris: OECD.

OECD (2007b). *PISA 2006. Volume 2: Data/Données.* Paris: OECD.

OECD (2009). *PISA 2006. Technical Report.* Paris: OECD.

Ramseier, E. (2008). Validation of competence models for developing education standards: Methodological choices and their consequences. *Mesure et évaluation en éducation, 31*(2), 35–53.

Ramseier, E. & Moreau, J. (2007). *Stichprobendesign und Gewichtung in der HarmoS-Validierungsstudie.* Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (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. (2004a). *Lehrbuch Testtheorie, Testkonstruktion* (4. ed.). Bern: Huber.

Rost, J. (2004b). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik, 50,* 662–678.

Schecker, H. & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften, 12,* 45–66.

Senkbeil, M., Rost, J., Carstensen, C. H. & Walter, O. (2005). Der nationale Naturwissenschaftstest PISA 2003. Entwicklung und empirische Überprüfung eines zweidimensionalen Facettendesigns. *Empirische Pädagogik, 19,* 166–189.

Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and Selecting Key Competencies* (pp. 45–65). Göttingen: Hogrefe & Huber

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER ConQuest Version 2. Generalised item response modelling software.* Melbourne: ACER Press.

Zieky, M. & Perie, M. (2006). *A Primer on Setting Cut Scores on Tests of Educational Achievement.* Princeton, NJ: Educational Testing Service.

Kontakt
Prof. Dr. Erich Ramseier
Mitglied der
EDK-Methodologiegruppe HarmoS
Pädagogische Hochschule PHBern
Zentrum für Forschung und Entwicklung
Fabrikstraße 2
CH-3012 Bern
erich.ramseier@phbern.ch

Autoreninformation
Prof. Dr. Marco Adamina
Ko-Leiter HarmoS Naturwissenschaften
Pädagogische Hochschule PHBern
Institut für Vorschulstufe und Primarstufe
Brückenstraße 73
CH-3005 Bern
marco.adamina@phbern.ch

Prof. Dr. Peter Labudde
Ko-Leiter Konsortium HarmoS
Naturwissenschaften, Zentrum
Naturwissenschafts- und Technikdidaktik
Pädagogische Hochschule FHNW
Riehenstraße 154
CH-4058 Basel
peter.labudde@fhnw.ch