Rost, Jürgen

Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen

Zeitschrift für Pädagogik 50 (2004) 5, S. 662-678

urn:nbn:de:0111-opus-48346

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 50 – Heft 5
September/Oktober 2004

Inhaltsverzeichnis

Thementeil: Bildungsstandards

Allgemeiner Teil

Besprechungen

Dokumentation

Jürgen Rost

Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen[1]

Zusammenfassung: *Es werden drei Ansätze unterschieden, Kompetenzstufen in einem psychometrischen Modell abzubilden. Im ersten Ansatz werden Kompetenzstufen als Abschnitte auf der zu messenden Leistungsdimension definiert. Diese Segmente können durch die Schwierigkeitsparameter der Items, durch die Verteilung der Personenmesswerte oder durch die Schwierigkeitsparameter der Antwortkategorien definiert sein. Der zweite Ansatz basiert auf einer Charakterisierung der Testaufgaben durch Aufgabenmerkmale bzw. auf der Konstruktion von Tests mittels eines Facettendesigns. Hier werden Schwierigkeits- und Fähigkeitsmodelle unterschieden. Der dritte Ansatz geht davon aus, dass Kompetenzniveaus durch unterschiedliche Lösungsprofile der Testaufgaben definiert sind. Diese Konzeption von Kompetenzstufen kann im Rahmen von Mischverteilungsmodellen abgebildet werden. Es wird argumentiert, dass sich die qualitativen Unterschiede der Schülerleistungen auf verschiedenen Kompetenzstufen am ehesten über Mischverteilungen modellieren lassen.*

Die Debatte um die Bildungsreform der 1970er-Jahre hatte einen starken psychometrischen Akzent. Mit der Kritik an der Notengebung und der schon fast ideologisch geprägten Propagierung des *mastery learning* (Alle schaffen es!) kam das Konzept der kriteriumsorientierten Leistungsmessung auf (Klauer 1987). Schulnoten sollten nicht mehr anhand der Verteilung der Punktwerte in einer Schulklasse vergeben werden, da nach dieser Methode die Noten der Schüler je nach dem Leistungsniveau seiner Klasse sehr unterschiedlich ausfallen können und nicht mehr miteinander vergleichbar sind. Zudem ist ein solches Verfahren der Notenvergabe höchst unfair, bekäme doch derselbe Schüler in einer leistungsstarken Klasse eine schlechtere Note als er in einer leistungsschwachen Klasse bekäme. Die Schulklasse darf nicht mehr die Norm darstellen, an der ein Individuum gemessen wird, sondern es muss ein inhaltliches Kriterium geben, dessen Erreichen oder Nichterreichen über die individuelle Schulnote entscheidet.

Dazu gehörte auch das Konzept des *mastery learning*. Es handelt sich hierbei um die Idee, Schule und Unterricht so zu gestalten, dass alle Schüler eine Chance haben, das vom Lehrer gesetzte Kriterium zu erreichen. Bildung sollte nicht zulasten derer gehen, die ‚abgehängt' werden, weil sie das Kriterium nicht erreichen. In Zusammenhang damit kann auch das Konzept des lernzielorientierten Unterrichts gesehen werden, welches vorsieht, dass sich der Lehrer *vor* dem Unterricht darüber klar wird, welches Lernziel er beim Schüler anstrebt.

Dieses alles erinnert an die derzeitige Diskussion um Bildungsstandards. Tatsächlich ist die Unterscheidung von normorientierter und kriteriumsorientierter Messung von

1 Dieser Aufsatz ist meinem akademischen Lehrer Hans Spada anlässlich seines 60. Geburtstags gewidmet.

fundamentaler Bedeutung für die Testung von Bildungsstandards. Dabei ist normorientiertes Testen das einfachere Konzept und sozusagen das ‚Normale'. Fast alle standardisierten psychologischen und pädagogischen Test- und Fragebogeninstrumente weisen Normtabellen auf, die mit unterschiedlichem Differenzierungsgrad die Interpretation von individuellen Messwerten erlauben. Der PISA-Studie liegen ebenfalls Fragestellungen zugrunde, die mit normorientierten Messwerten beantwortbar sind, geht es doch gerade um die Einordnung nationaler Leistungswerte in eine internationale Referenzverteilung, also in eine Norm (Prenzel u.a. 2001).

Die Erreichung von Bildungsstandards lässt sich jedoch *nicht* mit normorientierten Messungen kontrollieren. Man will ja gerade wissen, wie viel Prozent der Schülerinnen und Schüler einer bestimmten Teilpopulation ein bestimmtes Kriterium, z.B. eine Kompetenzstufe erreicht haben. Prozentränge, Perzentile oder Z-Werte als Maß der Schülerleistung sagen nichts über die Erreichung eines Kriteriums aus. Mit der Formulierung von Bildungsstandards werden Kriterien gesetzt und die eigentliche Schwierigkeit besteht darin, die Kriterien so zu setzen, dass ihre Erreichung empirisch überprüfbar wird.

Man kann sich natürlich auf den Standpunkt stellen, die Aufgaben eines Tests *sind* das Kriterium und die Anzahl der gelösten Aufgaben gibt Auskunft über seine Erreichung. Für dieses Vorgehen ist die (willkürliche) Setzung eines *cut-off scores* erforderlich, z.B. in Form einer mittleren Lösungswahrscheinlichkeit aller Aufgaben. Dies gilt auch, wenn man das früher oft als kriteriumsorientiertes Testmodell bezeichnete Rasch-Modell anwendet. Auch das Rasch-Modell setzt kein Kriterium und sollen die Aufgaben eines Tests oder eine Teilmenge von diesen das Kriterium sein, so fehlt noch die Angabe, um wie viel Logit-Einheiten über dem Mittelpunkt oder dem schwierigsten Item einer Teilmenge von Items eine Person liegen muss, um sagen zu können, sie habe das Kriterium erreicht. Das Rasch-Modell per se ist kein kriteriumsorientiertes Testmodell, es ermöglicht nur die Setzung von Kriterien (Rost 2004).

Diese Kriterien werden an den Aufgabenschwierigkeiten festgemacht und diese wiederum sind konstruktionsbedingt variabel. Man kann ‚denselben' Aufgabeninhalt so in Form einer Testaufgabe bringen, dass diese eher schwer oder eher leicht zu lösen ist. Aussagen über die *relativen* Fähigkeiten der Personen zueinander sind bei Geltung des Rasch-Modells unabhängig von den Schwierigkeiten der verwendeten Testaufgaben. Aussagen über die Erreichung eines Kriteriums sind – natürlich – davon abhängig, welche Aufgaben das Kriterium definieren und wie schwierig sie sind.

Das Problem der Setzung des Kriteriums beim kriteriumsorientierten Testen ist ein Problem der Integration von qualitativer und quantitativer Messung. Angestrebt wird ein qualitativer Messwert der Art: Schüler X hat das Kriterium erreicht, Schülerin Y erfüllt den Bildungsstandard. Zur Verfügung steht ein quantitativer Messwert: das Abschneiden in einem Test. Dieses Problem der Kompatibilität und Integrierbarkeit von Qualität und Quantität durchzieht den gesamten vorliegenden Beitrag.

Die empirische Bildungsforschung war in den 1970er-Jahren in Deutschland jedoch noch nicht von nationalen oder internationalen *assessment*-Studien geprägt, sondern machte die ‚kognitive Wende' mit, die damals fast alle Bereiche der Psychologie erfasste.

Die Mikroanalyse kognitiver Prozesse und Modelle der kognitiven Repräsentation von Wissen waren Themen der empirischen Forschung und der psychometrischen Modellbildung.

Tests sollten dadurch valider gemacht werden, dass man empirisch nachweisen konnte, welche kognitiven Operationen während der Aufgabenbearbeitung beim Respondenten ablaufen und ggf. auch, welche Lerneffekte infolge einer richtigen Lösung stattfanden (Spada 1976). Auch diese Forschung diente indirekt dem Setzen von Kriterien und der Integration von Qualität und Quantität. Natürlich weiß man über das quantitative Testergebnis mehr, wenn man weiß, welche qualitativen kognitiven Prozesse von dem Test erfasst werden. Die Leistungen von Schülerinnen und Schülern mit einem derartigen Test zu erfassen ist auch schon kriteriumsorientiert, wenn auch das Problem des Setzens eines *cut-off* Wertes damit noch nicht gelöst ist.

Das psychometrische Modell, das zur Modellierung der Denk- und Lernprozesse während der Testbearbeitung eingesetzt wurde, war das linear-logistische Testmodell von Fischer (1974), ein Spezialfall des Rasch-Modells. Es zeigte sich dabei auch, dass die strengen Annahmen von Modellen der Leistungsmessung wie dem Rasch-Modell in Konflikt geraten können mit der (mikroanalytischen) Modellierung von Denk- und Lernprozessen. Dies ist nämlich dann der Fall, wenn man annimmt, dass die qualitativ unterschiedlichen Denkprozesse von derselben Person unterschiedlich gut beherrscht werden, sodass ein eindimensionales quantitatives Modell nicht mehr ausreicht, die Personenunterschiede abzubilden.

Die folgende Erörterung unterschiedlicher Arten von Kompetenzmodellen zur Überprüfung von Bildungsstandards knüpft an beide Traditionen an: das kriteriumsorientierte Testen und die Modellierung des Lösungsprozesses. In Abschnitt 1 wird versucht, Kompetenzstufen im Rahmen eindimensionaler Leistungsdiagnostik zu etablieren. Abschnitt 2 behandelt Modelle, in denen die zu erfassende Kompetenz über kognitive Teilkompetenzen modelliert wird, die mit Aufgabenmerkmalen korrespondieren. Abschnitt 3 ist in Hinblick auf qualitative Schülerunterschiede am konsequentesten, indem qualitative Unterschiede im Antwortverhalten eines Tests gekoppelt mit (quantitativen) Niveauunterschieden modelliert werden. Damit werden drei Arten von Kompetenzmodellen unterschieden, die mit jeweils eigenen psychometrischen Modellen verknüpft sind.

1. Modelle eindimensionaler Kompetenzstufen

Der Begriff der Kompetenzmodelle wird in der aktuellen Diskussion um Bildungsstandards zumeist mit dem Begriff der Kompetenzstufen verknüpft. Im einfachsten Fall besteht ein Kompetenzmodell darin, in einer Domäne verschiedene Stufen oder Levels der Leistung zu unterscheiden, die sich in der Elaboriertheit der kognitiven Prozesse auf der jeweiligen Stufe auszeichnen. Der Stufen-, Level- oder auch Niveaubegriff deutet bereits an, dass es sich zwar um qualitativ unterschiedliche Kompetenzausprägungen handelt, die aber auf einer gemeinsamen Kompetenzdimension angeordnet sind.

Die Funktion derartiger Kompetenzstufen ist es, einen Rahmen für die Setzung eines Bildungsstandards zu liefern. Wenn es gelingt, qualitativ unterschiedliche kognitive Anforderungen in Testaufgaben zu erfassen, die mit unterschiedlichen Leistungsniveaus korrespondieren, so wäre die relative Willkür der Festlegung eines quantitativen Standards zumindest eingeschränkt. Das Kriterium wäre dann z.B. nicht mehr ‚80% aller Aufgaben zu lösen', sondern die Aufgaben einer bestimmten Kompetenzstufe zu lösen.

Es können zwar auch innerhalb jeder Kompetenzstufe unterschiedlich schwere Aufgaben konstruiert und damit die Ergebnisse über die Erreichung von Bildungsstandards beeinflusst werden. Wenn es aber tatsächlich gelingt, Aufgaben zu konstruieren, die jeweils von einem bestimmten Leistungsniveau an aufwärts gelöst werden können, so ist die Manipulierbarkeit der Aufgabenschwierigkeiten zumindest in die Grenzen einer Kompetenzstufe eingeschränkt.

Voraussetzung für diese Art der Kriteriensetzung ist der Nachweis der postulierten Schwierigkeitsunterschiede zwischen den Aufgaben. Idealerweise müssten sich die Items der verschiedenen Kompetenzstufen entlang des gemessenen Kontinuums gruppenweise anordnen. D.h., alle Items einer höheren Kompetenzstufe müssen schwerer sein als die Items niedrigerer Kompetenzstufen.

Im Naturwissenschaftstest der PISA-2000 Studie ließen sich die postulierten Kompetenzstufen (Prenzel u.a. 2001) nicht in Form von Schwierigkeitsclustern der Testaufgaben validieren. Es gab eine ganze Reihe von Items, die in einem anderen Schwierigkeitssegment lagen, als es aufgrund ihres Kompetenzlevels zu erwarten gewesen wäre. Auch die theoretische Zuordnung der Items zu den Kompetenzstufen war nicht bei allen Items eindeutig. Als Konsequenz wurde und wird in den Naturwissenschaftstests von PISA 2003 und 2006 eine solche Zuordnung der Items zu Kompetenzstufen nicht mehr vorgenommen.

Eine alternative, ebenfalls empirisch überprüfbare Annahme über Kompetenzstufen betrifft die Verteilung der Personenmesswerte. Wenn die Kompetenzstufen so etwas wie homogene Leistungsniveaus darstellen, so müsste es mehr Personen *innerhalb* jeder Niveaustufe geben als *zwischen* den Stufen. Als Folge würde eine mehrmodale Verteilung resultieren oder zumindest eine Mischverteilung aus soviel Komponenten wie es Kompetenzstufen gibt. Aber auch diese Annahme ist zu streng und unrealistisch. Erfahrungsgemäß sind Leistungsscores, auch oder gerade wenn sie als Summe heterogener Teilleistungen definiert sind, normalverteilt. In den Auswertungen der PISA-Tests wird sogar die Normalverteilung der Messwerte im psychometrischen Modell vorausgesetzt.

Neben den beiden Möglichkeiten, die Kompetenzstufen über die Verteilung der Aufgabenschwierigkeiten oder die Verteilung der Personenparameter zu bestimmen, gibt es den dritten Weg, sie über die von den Schülern produzierten Aufgabenlösungen zu definieren. Dies setzt voraus, dass die Antworten nicht nur dichotom sondern mehrstufig erfasst werden. Gibt ein Schüler eine nur teilweise richtige Lösung zu einer Aufgabe an, ohne die Aufgabe vollständig zu lösen, so kann seine Itemantwort ihn als einer Kompetenzstufe zugehörig ausweisen, auf der nur diese Art von Teillösungen möglich ist. Dabei sind sogar mehr als zwei Abstufungen der Vollständigkeit der Antworten möglich und entsprechenden Kompetenzstufen zuordenbar.

Eine solche Konzeption von Kompetenzstufen setzt ein Testmodell voraus, das mit ordinalen Itemantworten umgehen kann, also z.B. das ordinale Rasch-Modell (*partial credit* Modell). Im Rahmen dieses Modells werden die Wahrscheinlichkeiten der verschiedenen Antwortkategorien über die Lokation und Abstände der so genannten Schwellen definiert. Eine Schwelle liegt zwischen je zwei benachbarten Antwortkategorien und der Abstand zweier benachbarter Schwellen bestimmt die Wahrscheinlichkeit einer Itemantwort in der zwischen den Schwellen liegenden Kategorie. Auf diese Weise definieren die Schwellen auf dem zu messenden Kontinuum Segmente, in denen bestimmte Antworten wahrscheinlich sind. Diese Segmente könnten mit den postulierten Kompetenzstufen korrespondieren, sodass die Schwellen die Grenzen zwischen den Kompetenzstufen definieren.

Beträgt z.B. die (Schwellen-) Schwierigkeit einer bestimmten Teillösung s_1=-1.2 und die (Schwellen-) Schwierigkeit der vollständigen Lösung s_2=0.5, so definiert das Intervall (-1.2 ; 0.5) den Bereich der mittleren Kompetenzstufe auf dem latenten Kontinuum. Die unterste Kompetenzstufe würde von minus unendlich bis -1.2 reichen, die oberste von 0.5 bis plus unendlich. Das Problem besteht nur darin, dass die vielen Aufgaben eines Tests unterschiedliche Schwellenparameter haben werden und es somit so viele unterschiedliche Segmente für die mittlere Kompetenzstufe gibt, wie es Items mit unterschiedlichen (Schwellen-)Schwierigkeiten gibt. Eine Hilfe bei der Festlegung von Kompetenzstufen bietet dieser Weg wohl nur in speziellen Fällen.

Die drei hier behandelten Wege, Kompetenzstufen in einem psychometrischen Modell abzubilden, lassen sich grafisch veranschaulichen (s. Abbildung 1).

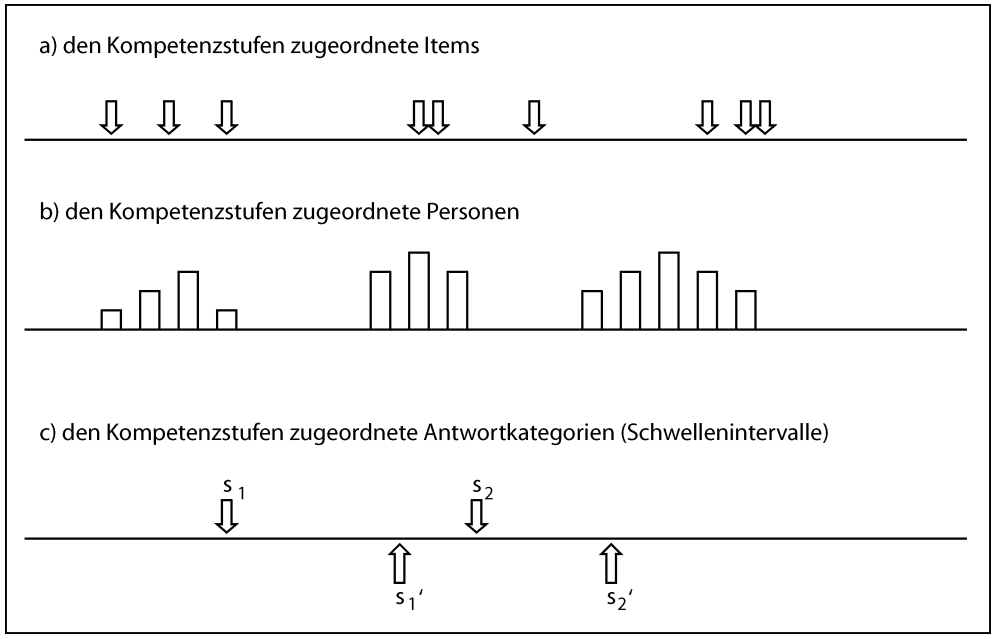

Abb. 1: Drei Modelle für eindimensionale Kompetenzstufen

Sie haben die gemeinsame Eigenschaft, die Kompetenzstufen auf *einer* Dimension zu lokalisieren und sind infolge dessen auch im Rahmen des Rasch-Modells realisierbar. Die Kopplung der Kompetenzstufen an die Aufgabenschwierigkeit, die Personenfähigkeit oder die Schwellenlokation in einem eindimensionalen Modell hat aber gravierende Nachteile. Alle Aufgabenschwierigkeiten und Schwellenlokationen müssen für alle Personen gleich sein, d.h. Personen mit einem unterschiedlichen Kompetenzniveau haben zwar eine unterschiedliche Eigenschaftsausprägung aber es gelten dieselben Schwierigkeits- und Schwellenparameter.

Kompetenzmodelle, deren Kompetenzstufen so geartet sind, dass sich die *relativen* Itemschwierigkeiten für die Stufen unterscheiden, sind in diesem eindimensionalen Rahmen nicht realisierbar. Führt also z.B. eine höhere Kompetenzstufe zu einer kognitiven Verarbeitung aller Items, die auch eine Verschiebung der Aufgabenschwierigkeiten impliziert, so ist das im Rahmen der bisher diskutierten Ansätze nicht abbildbar.

2. Modelle kognitiver Teilkompetenzen

Ein zweiter Weg der Spezifizierung von Kompetenzmodellen besteht darin, Teilkompetenzen an bestimmte Merkmale von Testaufgaben zu knüpfen und die postulierte Korrespondenz (von Aufgabenmerkmalen und Teilkompetenzen) bei der Datenauswertung empirisch zu validieren. Löst ein Schüler Aufgaben, die einen bestimmten kognitiven Prozess erfordern, während er Aufgaben mit einem anderen kognitiven Prozess nicht löst, so hat man eine qualitative Aussage über seine Leistungsstruktur. Man kann Bildungsstandards dann über die Prozesse definieren, die ein Schüler bewältigen muss. Auch dieser Ansatz stellt einen Weg dar, statt eines rein quantitativen Urteils über die Erreichung eines Bildungsstandards die *Art* der zu bewältigenden kognitiven Anforderungen – und somit qualitative Unterscheide – zu berücksichtigen.

Aufgabenmerkmale

Aufgaben	1	2	3	4	5	6	7	8
1	1	0	0	1	1			
2	1	1	0	0	1			
3	0	0	1	1	0	...		
4	...							
5								
6								
				q_{ij}				
k								

Abb. 2: Beispiel einer Itemkomponenten-Matrix

Für eine solche Modellierung des Aufgabenlösungsprozesses stehen in Form von Itemkomponenten-Modellen geeignete psychometrische Hilfsmittel zur Verfügung. Ausgangspunkt stellt eine Itemkomponenten-Matrix dar, die für jede Testaufgabe spezifiziert, welche Aufgabenmerkmale diese Aufgabe aufweist. Diese Aufgabenmerkmale können Lösungsschritte sein, benötigte Wissenselemente oder Repräsentationsmodalitäten.

Eine solche Charakterisierung von Testaufgaben durch ihre Merkmale kann allein schon für die *Konstruktion* der Testaufgaben hilfreich sein, bieten die Aufgabenmerkmale doch ein gewisses Raster, in dessen Rahmen auch der Spielraum für die Aufgabenschwierigkeiten etwas eingeschränkt ist. Was die psychometrische Modellierung solcher Aufgabenmerkmale oder Itemkomponenten anbetrifft, so gibt es bei aller Vielfalt möglicher Modell-Spezifikationen eine fundamentale Unterscheidung: die Konzipierung als *Schwierigkeitsmodell* oder als *Fähigkeitsmodell*.

Die Terminologie stammt aus der Rasch-Messtheorie, die von der Separierbarkeit von Personen- und Aufgabeneinflüssen ausgeht. Jede Person hat einen Messwert (den Personenparameter) θ, der ihre Fähigkeit ausdrückt, und jedes Item hat einen Messwert (den Itemparameter) σ, der seine Schwierigkeit ausdrückt. Die Antwortwahrscheinlichkeit einer Person v bei einem Item i ist dann eine monotone Funktion der Differenz dieser beiden Parameter.

$$(1)\ p_{vi} = f(\theta_v - \sigma_i)$$

Itemkomponenten können entweder zur Modellierung der Aufgabenschwierigkeit oder der Personenfähigkeit herangezogen werden. Das bereits erwähnte linear-logistische Testmodell (LLTM) von Fischer (1974), das übrigens das Obermodell von sehr vielen speziellen Rasch-Modellen darstellt, ist ein reines Schwierigkeitsmodell. Die Designmatrix Q wird verwendet um die *Itemparameter* in additive Bestandteile zu zerlegen.

$$(2)\ p_{vi} = f(\theta_v - \sum_j q_{ij}\eta_j)\,.$$

In dieser Formel ist die Lösungswahrscheinlichkeit einer Person v bei Aufgabe i eine mit den q-Koeffizienten gewichtete Summe von Schwierigkeiten η der Aufgabenmerkmale j. Auch die Mathematik- und Naturwissenschaftsaufgaben aus der PISA 2000 Erhebung wurden über post hoc identifizierte Aufgabenmerkmale dahingehend analysiert, ob sich ihre Schwierigkeiten durch die Itemkomponenten vorhersagen lassen (Neubrand u.a. 2002; Prenzel u.a. 2002). Diese Vorhersagbarkeit war in gewissem Rahmen gegeben, jedoch nicht perfekt (r^2=0,45).

Die Charakterisierung von Testaufgaben durch ihre Merkmale im Rahmen von Schwierigkeitsmodellen kann auch dazu dienen, Kompetenzstufen als *qualitativ* definierte Punkte auf dem gemessenen Kontinuum festzulegen. Beträgt der Schwierigkeitsparameter eines Aufgabenmerkmals z.B. η_j =1,0, so heißt das auch, dass ein Kompetenzniveau, auf dem Schüler Aufgaben mit diesem Merkmal lösen können, um eine Einheit über dem Kompetenzniveau liegt, auf dem Aufgaben mit diesem Merkmal nicht

gelöst werden können. Je nach Anzahl der unterschiedenen Aufgabenmerkmale müssten Kompetenzstufen über einzelne Aufgabenmerkmale oder über Muster verschiedener (auf der jeweiligen Kompetenzstufe beherrschter) Aufgabenmerkmale definiert werden.

Auch diese Art der psychometrischen Modellierung von qualitativ unterschiedlichen kognitiven Prozessen ist eindimensional. Der Personenparameter θ_v bleibt von der Aufgabenstruktur Q unberührt.

Man kann jedoch dieselbe Aufgabenstruktur Q auch im Rahmen einer mehrdimensionalen Modellierung, eines Fähigkeitsmodells, berücksichtigen. Die Annahme, dass jedem Aufgabenmerkmal eine eigene Fähigkeitsdimension entspricht, liegt dem MULTIRA Modell (Rost und Carstensen 2002) zugrunde. Die folgende Modellgleichung zeigt, dass die additive Struktur, im Unterschied zum LLTM (s.o.), jetzt aufseiten der Personenparameter liegt. Tatsächlich handelt es sich nicht um eine *Zerlegung* der Fähigkeitsparameter in Teilfähigkeiten, sondern um eine Vermehrung der Fähigkeitsparameter.

$$(3)\ p_{vi} = f(\sum_j q_{ij}\theta_{vj} - \sigma_i)$$

Dieses Modell ist ebenfalls ein Itemkomponentenmodell (und kein Personenkomponentenmodell), d.h. es verwendet dieselbe Q-Matrix mit Gewichten q_{ij} wie das LLTM. Im Unterschied zu diesem ist es jedoch keine Restriktion des Rasch-Modells sondern eine Verallgemeinerung. Es enthält so viele Personenparameter pro Person, wie es Itemkomponenten gibt und reduziert sich zum eindimensionalen Rasch-Modell, wenn die Q-Matrix ein Spaltenvektor mit Einsen ist. Das LLTM (s. Gleichung (2)) wird zum Rasch-Modell, wenn Q die Einheitsmatrix ist (Einsen in der Hauptdiagonalen und Nullen sonst).

MULTIRA ist ein Fähigkeitsmodell für Aufgabenmerkmale. Allerdings ist es leicht überdimensioniert (im eigentlichen Wortsinn), da es für jedes Aufgabenmerkmal eine eigene Fähigkeit annimmt. Diese Problematik wird auch an einer häufig anzutreffenden Komponentenstruktur, dem vollständigen Facettendesign deutlich. In einem vollständigen Facettendesign werden die Komponenten von 2 Facetten, z.B. dem Inhaltsbereich einer Aufgabe und den zur Lösung erforderlichen kognitiven Prozessen, miteinander gekreuzt, sodass jede Aufgabe genau aus zwei Komponenten besteht, einem Inhalt und einem kognitiven Prozess. Abbildung 3 zeigt die Facettenstruktur des nationalen Naturwissenschaftstests von PISA 2003.

Die Zeilen der Matrix werden durch die inhaltliche Facette strukturiert und stellen 10 Inhaltsbereiche aus Physik, Chemie und Biologie dar, auf die sich die Aufgaben des Feldtests 2003 beziehen. Die Spalten der Matrix sind durch die Facette der kognitiven Teilkompetenzen definiert und umfassen die folgenden Qualitäten kognitiver Prozesse: Bewertungen, divergentes Denken, Umgang mit Grafiken, konvergentes Denken, Nutzung mentaler Modelle, Verbalisierung von Sachverhalten und Umgang mit Zahlen.

Solche Facettentests haben eine typische Struktur der Q-Matrix, die in Abbildung 4 für den einfachen Fall eines 3x3 Facettendesigns wiedergegeben ist.

	Bew	Div	Gra	Kon	MeM	SaV	Zah
Energieumwandlung							
Bewegungsgesetze							
Wärme							
Elektrizität							
Räuber-Beute-Systeme							
Fortpflanzung							
Atmung und Fotosynthese							
Stärkeumwandlung							
Aggregatzustände							
Teilchenkonzept							

Abb. 3: Die Facettenstruktur des nationalen Naturwissenschaftstests PISA 2003

Aufgabe	Facette 1			Facette 2		
1	1			1		
2	1				1	
3	1					1
4		1		1		
5		1			1	
6		1				1
7			1	1		
8			1		1	
9			1			1

Abb. 4: Designmatrix eines Facettenmodells

Auch bei Facettenmodellen lässt sich wiederum die Unterscheidung von Schwierigkeits- und Fähigkeitsmodellen treffen. Das von Linacre (1989) beschriebene FACETS model ist ein Schwierigkeitsmodell, in dem die Parameterdifferenz des einfachen Rasch-Modells (1) durch einen weiteren Subtrahenden, den Schwierigkeiten δ der Komponenten j der zweiten Facette, erweitert wird.

$$(4)\ p_{vij} = f(\theta_v - \sigma_i - \delta_j)\,.$$

Demgegenüber stellt die Facettenspezifikation von MULTIRA ein typisches Fähigkeitsmodell mit den doppelt indizierten Fähigkeitsparametern θ_{vi} für die Komponenten i der ersten Facette und δ_{vj} für die Komponenten j der zweiten Facette dar.

$$(5)\ p_{vij} = f(\theta_{vi} - \delta_{vj} - \sigma_{ij})\,.$$

Auf die Schwierigkeitsparameter σ der Items wirkt sich die Facettenstruktur nicht aus, hier erhält jedes Item einen eigenen Schwierigkeitsparameter, der doppelt indiziert ist, weil sich jedes Item aus der Kombination zweier Facetten ergibt.

Als ein Hilfsmittel, um Bildungsstandards nicht nur quantitativ zu fixieren, sondern auch die Qualität der Anforderungen deutlich zu machen, gehen Schwierigkeits-Facetten-Modelle nicht über das bereits zum LLTM Gesagte hinaus. Aufgrund ihrer Eindimensionalität werden die qualitativen Unterschiede, die sich aus der Kreuzung der beiden Facetten ergeben, den Erfordernissen eines quantitativen Messmodells untergeordnet.

Anders verhält es sich mit mehrdimensionalen Facettenmodellen. Im Fall des oben dargestellten nationalen PISA Tests (vgl. Abbildung 3) würde ein Fähigkeitsmodell für jede Inhaltskomponente eine Fähigkeitsdimension annehmen, also zehn, und für jede kognitive Operation ebenfalls eine, also sieben Dimensionen.

Im Zuge der Testanalyse des Feldtests und der Hauptuntersuchung von PISA 2003 wurden die Daten des Facettentests auf ihre Dimensionalität geprüft (vgl. Senkbeil u.a. 2004; Rost und Walter 2004). Während die Fähigkeiten der Inhaltsfacette nach den statistischen Kriterien der Modellanpassung entbehrlich sind, kann hinsichtlich der 7 Dimensionen, die den kognitiven Komponenten zugeordnet sind, keine weitere Reduktion der Dimensionalität vorgenommen werden. Abbildung 5 zeigt die zweidimensionale Projektion der sieben kognitiven Teilkompetenzen mittels einer Hauptkomponentenanalyse.

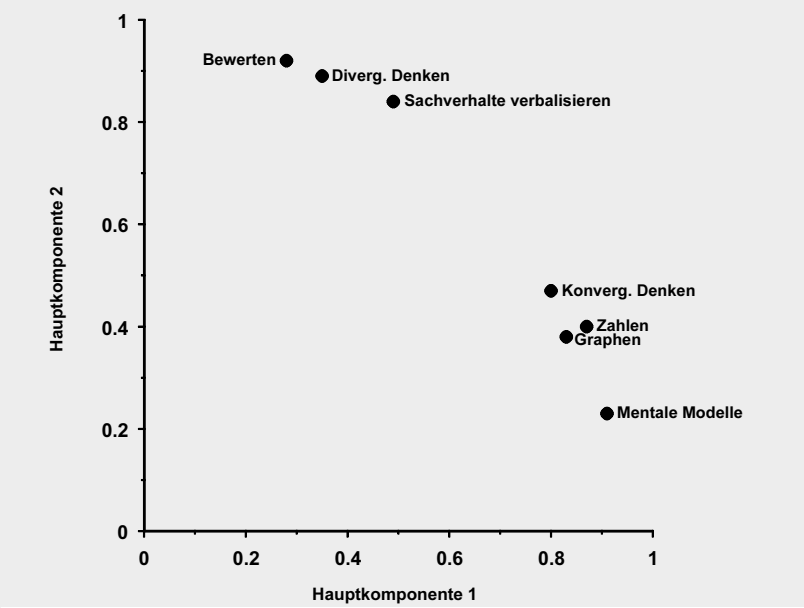

Abb. 5: Die zweidimensionale Struktur der 7 Teilkompetenzen des nationalen Naturwissenschaftstests (Feldtestdaten)

Obwohl dieses Bild eine gut interpretierbare Gruppierung in die ‚härteren' analytischen Kompetenzen ‚Umgang mit Zahlen', ‚Umgang mit Grafiken', ‚mentale Modelle' und ‚konvergentes Denken' einerseits und die ‚weicheren' kreativen Kompetenzen ‚Bewer-

ten‘, ‚divergentes Denken‘ und ‚Sachverhalte verbalisieren‘ aufweist, passt ein entsprechendes zweidimensionales Modell nicht auf die Daten.

In einem solchen mehrdimensionalen Fähigkeitsmodell werden qualitativ unterschiedliche Teilkompetenzen in Form von eigenen quantitativen Variablen erfasst. Diese Art der Synthese von Qualität und Quantität entspricht der Logik des klassischen faktorenanalytischen Modells. Als Rahmen für die Setzung von Bildungsstandards ist das Problem der willkürlichen Festlegung von cut-off scores nach wie vor gegeben, in gewisser Weise vervielfacht auf die Anzahl der Dimensionen.

Der Weg, Bildungsstandards in Form von *Zielprofilen* relevanter Teilkompetenzen zu definieren, ist mit mehrdimensionalen Fähigkeitsmodellen zwar gegeben, über den Verlauf der möglichen unterscheidbaren Profile sagen diese Modelle jedoch nichts aus. Das ist bei dem dritten Typ von Kompetenzmodellen, der im folgenden Kapitel behandelt wird, anders.

3. Niveaubezogene Bearbeitungsmuster

Das Konzept qualitativ unterschiedlicher Kompetenzstufen, die sich auf einer latenten Dimension anordnen, stößt dort an seine Grenzen, wo die Voraussetzungen eindimensionaler Messung verletzt werden. Ist eine bestimmte Testaufgabe auf einer Kompetenzstufe leichter als andere Aufgaben, weil die kognitiven Voraussetzungen zur Lösung der Aufgabe auf dieser Stufe gegeben sind, während auf einer anderen Kompetenzstufe alle Items gleiches Schwierigkeitsniveau haben, so widerspricht das bereits den psychometrischen Voraussetzungen von Messung. Die Konstanz der Itemparameter für alle Personen, bei denen der Test dasselbe messen soll, ist das zentrale Kriterium von Messung.

Misst ein Test in einer Gruppe von Personen etwas anderes, so dürfen seine Itemparameter für diese Gruppe auch andere Werte annehmen. Mehr noch: daran dass die Itemparameter in einer Personengruppe andere Werte annehmen, erkennt man erst, dass der Test in dieser Personengruppe etwas anderes misst. Daraus ergibt sich ein weiteres Konzept für die Messung von Kompetenzstufen: Kompetenzstufen lassen sich definieren als Gruppen von Schülern mit unterschiedlichen Mustern von Itemparametern, verbunden mit unterschiedlichen Niveaus der Testleistung.

Die hierzu passende Familie psychometrischer Modelle stellen Mischverteilungsmodelle oder latente Klassenmodelle dar. Mischverteilungsmodelle modellieren den Sachverhalt, dass eine beobachtete Verteilung, z.B. die Häufigkeitsverteilung der Antwortmuster eines Tests, eine Mischung mehrerer unterschiedlicher Verteilungen ist. Die Anwendung von Mischverteilungsmodellen zielt darauf ab, die Gesamtpopulation derart zu *entmischen*, dass in den Mischungskomponenten ein bestimmtes Modell passt. Beim Mischverteilungs-Rasch-Modell (mixed Rasch model) soll das Rasch-Modell in jeder Mischungskomponente (latenten Klasse) gelten. Die analysierte Gesamtpopulation wird also in Rasch-skalierbare Teilpopulationen entmischt (Rost 2004).

Bei der Modellierung von Kompetenzstufen kommt hinzu, dass sich die Teilpopulationen im Niveau ihrer Testleistungen unterscheiden müssen, denn sonst würde es sich

nicht um ein Stufenmodell handeln. Zur Illustration wird im Folgenden die entsprechende Analyse eines Ausschnittes des Facettentests aus PISA 2003 gezeigt.

In diesem Test sind für neun naturwissenschaftliche Domänen sieben Items formuliert, die aufgrund ihrer Aufgabenmerkmale jeweils eine der kognitiven Teilkompetenzen ansprechen (s.o.). Im Folgenden werden nur die 7 Aufgaben einer Domäne betrachtet, nämlich ‚Atmung und Photosynthese'. Mit diesen sieben Items wurde das mixed Rasch-Modell mit zwei latenten Klassen gerechnet. Fragen der Modellgeltung und somit auch nach der optimalen Anzahl der latenten Klassen bleiben hier unberücksichtigt.

Unterscheiden sich die Schülerinnen und Schüler nicht nur im Niveau ihrer Leistungen in dieser Domäne, sondern auch darin, dass sich für die unterschiedlichen Leistungsniveaus unterschiedliche Schwierigkeiten ergeben, so erfasst der Test Kompetenzstufen im Sinne des gleichzeitigen Vorliegens qualitativer und quantitativer Leistungsunterschiede. Abbildung 6 zeigt zunächst die quantitativen Leistungsunterschiede zwischen den beiden latenten Klassen. Konkret sind die beiden Verteilungen der Summenscores (d.i. die Anzahl gelöster Testaufgaben) in den latenten Klassen wiedergegeben.

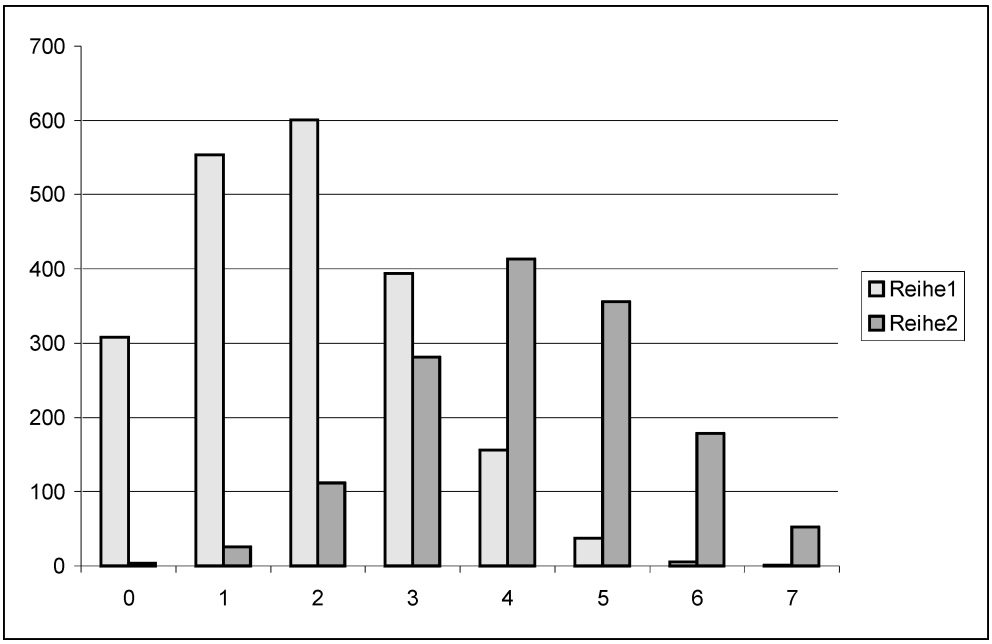

Abb. 6: Die Scoreverteilungen der beiden latenten Klassen für die 7 Items des Inhaltbereiches Atmung

In Klasse 1 (59% der Schüler) haben die meisten Schüler 1 bis 2 der sieben Items gelöst, in der zweiten Klasse (41%) 4 bis 5 Items. Es handelt sich also um zwei Klassen mit einem deutlichen Niveau-Unterschied. Dieser ist verbunden mit einem ebenso deutlichen Unterschied im durchschnittlichen Antwortmuster, welche in Abbildung 7 (S. 674) gezeigt sind.

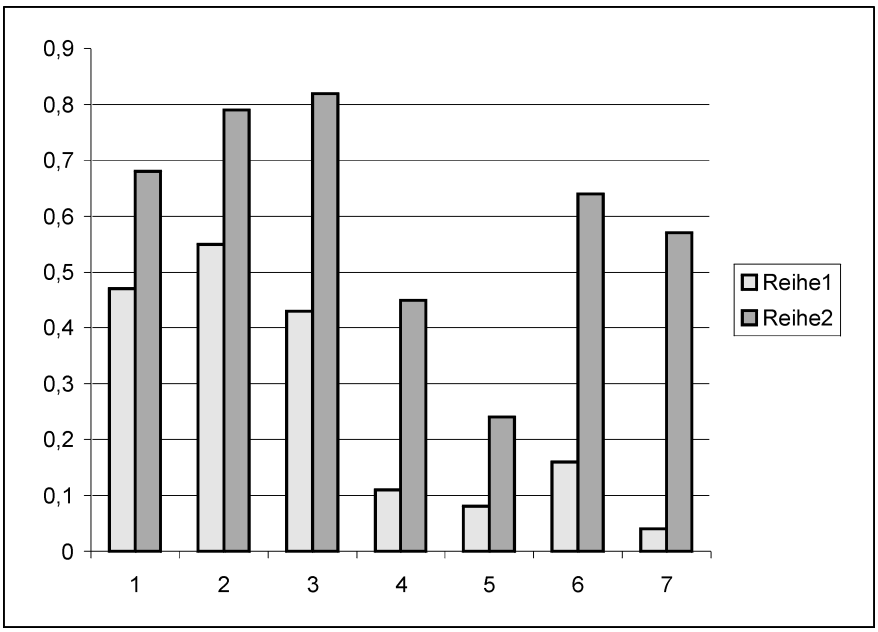

Abb. 7: Die Antwortwahrscheinlichkeiten in den beiden latenten Klassen für die 7 Items des Inhaltsbereiches Atmung

Während die ersten drei Items (divergentes Denken, Bewerten und Sachverhalte verbalisieren) in der leistungsschwachen Klasse noch mittlere Lösungswahrscheinlichkeiten haben, sind die vier Items zu den ‚härteren', analytischen Teilkompetenzen extrem schwierig. Das gleicht sich in der Klasse 2 der leistungsstärkeren Schüler schon fast einander an, wenn man einmal von dem Umgang mit den mentalen Modellen (Item 5) absieht. Die kompetenteren Schüler haben ein anderes Antwortmuster oder Profil der Teilkompetenzen als es die weniger kompetenten Schüler haben.

Beide Leistungsgruppen sind nicht durch einen *cut-off* der Scoreverteilung definiert. Ihre Scoreverteilungen überlappen sich, sodass z.B. sehr viele Schüler aus beiden Kompetenzstufen den Score 3 haben. Dieser Score wird aber auf beiden Kompetenzstufen mit sehr unterschiedlichen Antwortmustern erreicht. Abbildung 8 (S. 675) zeigt die Lösungswahrscheinlichkeiten, die Schüler mit dem Score 3 haben, in Abhängigkeit von ihrer Kompetenzstufe. Dieses Ergebnis illustriert, dass eine Aufteilung der Schüler nach qualitativen Gesichtspunkten (Antwortmuster) nicht nur ein *ergänzender* Aspekt der Messung von Kompetenzstufen ist, sondern dass die allein an quantitativen Gesichtpunkten (Summenscore) orientierte Auswertung dadurch auch relativiert oder gar falsifiziert werden kann. Nun würde man allein an sieben Aufgaben eines inhaltlichen Bereiches keine zuverlässige Diagnose individueller Kompetenzstufen festmachen. Man würde vielmehr erwarten, dass die Kompetenzstufen der Schüler über die Inhaltsbereiche hinweg korrespondieren. Diese Ergebnisse sollen jedoch der Gesamtanalyse der Daten vorbehalten bleiben.

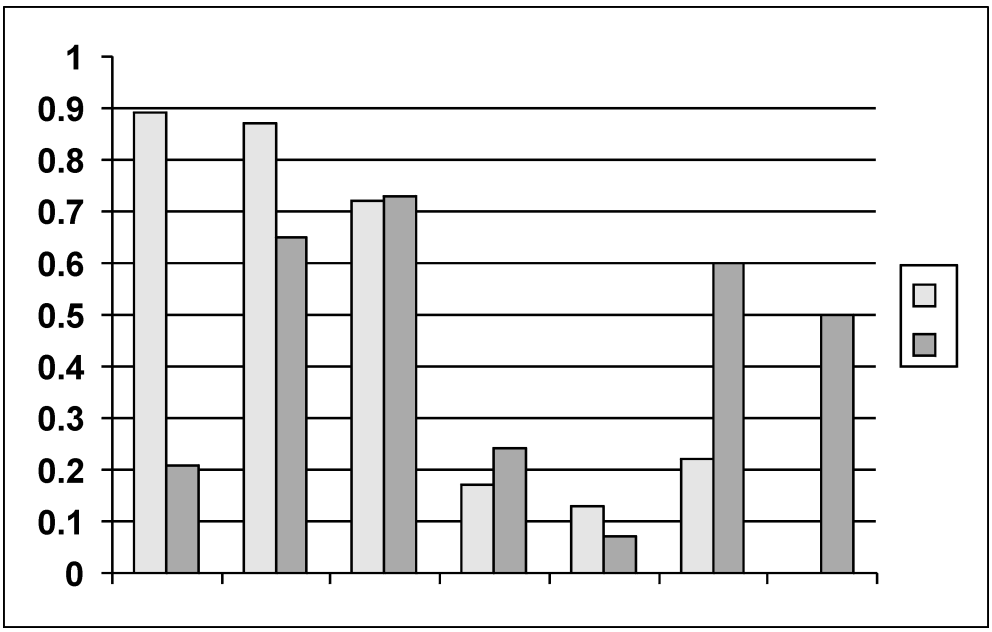

Abb. 8: Die Lösungswahrscheinlichkeiten von Schülern mit Score 3 in der leistungsschwachen Klasse (helle Balken) und der leistungsstarken Klasse (dunkle Balken)

Mischverteilungsmodelle im Allgemeinen und das mixed Rasch-Modell im Besonderen sind geeignet, die notwendige qualitative Charakterisierung von Kompetenzstufen und ihre Anordnung auf einer latenten Dimension widerspruchsfrei zu verbinden. Die Anordnung von Kompetenzstufen auf einer Rasch-Skala (Abschnitt 1) ist nur um den Preis machbar, dass die relativen Itemschwierigkeiten auf allen Kompetenzstufen gleich bleiben. Itemkomponenten-Modelle als Schwierigkeitsmodelle (Abschnitt 2) unterliegen einem vergleichbaren Dilemma, da alle mit den Aufgabenmerkmalen verknüpften kognitiven Prozesse mit derselben Fähigkeit eines Schülers durchgeführt werden. Die mehrdimensionale Erweiterung dieses Ansatzes löst zwar das Problem der gleichzeitigen Berücksichtigung von Quantität (latente Dimensionen) und Qualität (Art der Dimensionen) auf sehr elegante Weise, doch wird das Problem der Setzung von Kriterien (Bildungsstandards) nicht gelöst, sondern eher vervielfacht.

Auch beim mixed Rasch-Modell werden a priori keine Kriterien gesetzt, was auch nicht notwendig ist, da über die Erreichung eines Kriteriums die Zugehörigkeit zu den latenten Klassen entscheidet. Deren Größe wird auch nicht vorgegeben, sondern stellt ein Ergebnis der Datenanalyse dar. Ebenso werden die Zielprofile der erwarteten latenten Klassen in der Regel nicht vorgegeben, was allerdings technisch über die Restriktion der Modellparameter möglich wäre. Art, Anzahl und Größe der Mischungskomponenten stellen empirische Ergebnisse dar, die sich übrigens inferenzstatistisch absichern lassen, und stellen somit minimale Anforderungen an die präexperimentelle Hypothesen-

bildung. Allerdings steht und fällt die Identifizierung qualitativer Kompetenzstufen mit der Auswahl geeigneter Indikatoren, sprich mit der Formulierung geeigneter Items.

Bezogen auf die Formulierung von Bildungsstandards setzt dieses Kompetenzmodell voraus, dass sehr gute Vorstellungen dazu bestehen müssen, welche kognitiven Prozesse typisch für die Erreichung eines Bildungsstandards sind und welche Prozesse auch von Schülern niedrigerer Kompetenzstufen bewältigt werden können. Über *cut-off* Werte auf irgendwelchen Dimensionen braucht man sich zunächst keine Gedanken zu machen, obwohl das Modell quantitative Leistungsunterschiede auf den einzelnen Kompetenzstufen berücksichtigt. Im Extremfall kann ein Schüler, der einen Bildungsstandard erfüllt, einen niedrigeren Score aufweisen, als ein Schüler, der den Standard nicht erfüllt (wenn er nämlich gerade die schwereren, für die höhere Kompetenzstufe typischen Items löst). Das ist die Konsequenz daraus, dass man mit einem Mischverteilungsmodell der qualitativen Beschreibung des Leistungsprofils (durch die latenten Klassen) die gleiche Priorität wie der quantitativen Messung einräumt (durch Geltung eines Messmodells innerhalb der Klassen).

4. Schlussfolgerungen

Es wurden drei Arten von Kompetenzmodellen vorgestellt und hinsichtlich ihrer Tauglichkeit diskutiert, die Erfassung von Bildungsstandards psychometrisch zu fundieren. Man kann sie verkürzt als Modelle für Kompetenzstufen, Kompetenzdimensionen und Kompetenzmuster bezeichnen.

Bei Kompetenzstufenmodellen können die drei Fälle unterschieden werden, dass die Items, die Personen oder (bei mehrstufigen Itemantworten) die Schwellen die Lokation der Stufen auf dem latenten Kontinuum definieren. Obwohl alle drei Methoden einen eleganten Ausweg aus dem Problem der (willkürlichen) Setzung von cut-off Werten darstellen, wird dieser Ansatz letztlich nicht den postulierten qualitativen Unterschieden zwischen den Kompetenzstufen gerecht.

Mehrdimensionale Kompetenzmodelle wurden aus dem alten Ansatz der Itemkomponenten- (Aufgabenmerkmals-) Modelle hergeleitet, die als Schwierigkeitsmodelle ebenfalls vor dem Problem stehen, dass qualitative Unterschiede zwischen kognitiven Prozessen nicht mit den Erfordernissen eindimensionaler Messung in Konflikt stehen dürfen. Mehrdimensionale Itemkomponenten-Modelle ermöglichen es demgegenüber, Bildungsstandards als zu erreichendes Qualifikationsprofil zu definieren, d.h. das Muster der von einem Schüler erreichten Messwerte mit einem Zielprofil zu vergleichen. Die Identifizierung der Zielprofile wird jedoch nicht durch die zugehörigen psychometrischen Modelle unterstützt. Insofern müsste das Zielprofil wieder durch cut-off Werte auf allen beteiligten Dimensionen festgelegt werden, wobei sich Probleme des Umgangs mit dem Messfehler ergeben können.

Die Möglichkeit, Kompetenzen als *Muster* von Teilkompetenzen zu definieren, bieten so genannte Mischverteilungsmodelle und hier insbesondere das mixed Rasch-Modell. Es teilt gleichzeitig die Personen in Gruppen mit gleichem Kompetenzprofil ein

und misst innerhalb dieser Gruppen die quantitative Kompetenzausprägung. Es ist das einzige der hier besprochenen Modelle, das der qualitativen Unterscheidung von Personengruppen den Vorrang vor der Quantifizierung der individuellen Leistung gibt.

Geht man davon aus, dass die Entscheidung über die Erreichung eines Bildungsstandards eine kategoriale und somit eine qualitative Diagnose ist, so spricht das dafür mit einem psychometrischen Ansatz zu arbeiten, in dem nicht nur quantitative sondern auch qualitative Personenunterschiede abgebildet werden. Dabei ist die Methodologie der Mischverteilungsmodelle noch nicht gänzlich ausgereift, um auf alle Datenstrukturen, die im Rahmen der Messung von Bildungsstandards anfallen, angewendet werden zu können.

So können insbesondere große Itemmengen und unvollständige Testdesigns noch zu Problemen der Identifikation der latenten Klassen führen. Auch die Berücksichtigung starker präexperimenteller Annahmen über die Profile der Kompetenzstufen in Form von Parameterrestriktionen wirft noch Entwicklungsbedarf auf. Dagegen ist die Identifizierung der latenten Klassen unter Vorgabe von erwarteten Klassengrößen standardmäßig möglich. So lässt sich z.B. das Leistungsprofil derjenigen 20% der getesteten Schüler ermitteln, die am stärksten vom Profil der 80%-Majorität abweichen.

Die Formulierung von Bildungsstandards sollte die technischen und das heißt hier die psychometrischen Möglichkeiten der Verarbeitung entsprechender Daten berücksichtigen. Neue Entwicklungen zu Möglichkeiten der kriteriumsorientierten Leistungsmessung und neue Erfahrungen mit Item-response Modellen zur Modellierung kognitiver Prozesse sollten für die Entwicklung eines geeigneten Formates der Formulierung und empirischen Kontrolle von Bildungsstandards genutzt werden. Die Messung quantitativer Leistungsvariablen und die Identifizierung qualitativer Personenunterschiede schließen sich nicht mehr aus und müssen auch nicht auf unterschiedliche Daten rekurrieren. Wenn sich Bildung nicht quantitativ messen lässt, weil die Messobjekte ‚bunt' gemischt sind, so kann man nur versuchen, sie so zu entmischen, dass sie messbar werden.

Literatur

Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.) (2003) unter Beteiligung folgender Autoren: Klieme, E./Avenarius, H./Blum, W./Döbrich, P./Prenzel, M./Reiss, K./Riquarts, K./Rost, J. u.a: Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Bonn: BMBF.

Fischer, G.H. (1974). Einführung in die Theorie psychologischer Tests. Bern: Huber.

Klauer, K.J. (1987). Kriteriumsorientierte Tests. Göttingen: Hogrefe.

Linacre, J.M. (1989) Many-faceted Rasch measurement. Chicago: MESA Press.

Neubrand, M./Klieme, E./Lüdtke, O./Neubrand, J. (2002): Kompetenzstufen und Schwierigkeitsmodelle für den PISA-Test zur mathematischen Grundbildung . In: Unterrichtswissenschaft 30, S. 100-119.

Prenzel, M./Rost, J./Senkbeil, M./Häußler, P./Klopp, A. (2001): Naturwissenschaftliche Grundbildung. Testkonzeption und Ergebnisse. In: Deutsches PISA-Konsortium (Hrsg.): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich, S. 192-250.

Prenzel, M./Häußler, P./Rost, J./Senkbeil, M. (2002): Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? In: Unterrichtswissenschaft 30, S. 120-135.

Rost, J./Carstensen, C.H. (2002): Multidimensional Rasch Measurement via Item Component Models and Faceted Designs. In: Applied Psychological Measurement 26, S. 42-56.

Rost, J./Carstensen, C.H./Bieber, G./Neubrand, M./Prenzel, M. (2003): Naturwissenschaftliche Teilkompetenzen im Ländervergleich. In Deutsches PISA-Konsortium (Hrsg.): PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Opladen: Leske + Budrich, S. 109-129.

Rost, J./Walter, O. (2004): Multimethod Item Response Theory. In: Eid, M./Diener, E. (Eds.):Handbook of Psychological Measurement: A Multimethod Perspective. American Psychological Association (i. Druck).

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

Senkbeil, M./Rost, J./Carstensen, C.H./Walter, O. (2004): Der nationale Naturwissenschaftstest PISA 2003. Entwicklung und empirische Überprüfung eines zweidimensionalen Facettendesigns. Eingereicht bei Zeitschrift für empirische Pädagogik.

Spada, H. (1976): Modelle des Denkens und des Lernens. Bern: Huber.

Abstract: *Three different approaches of defining levels of competence in a psychometric framework are distinguished. The first approach defines levels of competence as intervals on the latent continuum to be measured. These intervals may be defined by the difficulties of the items, by the distribution of the person scores, or by the difficulty parameters of the response categories. The second approach is based on a characterization of the tasks by means of a set of task properties or a facet design. In this context, difficulty and ability models are to be distinguished. The third approach assumes that levels of competence are related to different patterns or profiles of responses on the tasks. Such concepts of levels of competence can be formalized within the framework of mixture distribution models. It is argued that qualitative differences of students that belong to different levels of competence are best modelled by means of mixture models, e.g. the mixed Rasch model.*

Anschrift des Autors:
Prof. Dr. Jürgen Rost, Leibniz-Institut für die Pädagogik der Naturwissenschaften IPN,
Abteilung Pädagogisch-psychologische Methodenlehre, Olshausenstr. 62, 24098 Kiel,
E-Mail: rost@ipn.uni-kiel.de.