Z Erziehungswiss (2010) 13:473–505
DOI 10.1007/s11618-010-0130-y

ALLGEMEINER TEIL

ZfE

Modellierung von Lehrerkompetenzen

Nutzung unterschiedlicher IRT-Skalierungen zur Diagnose von Stärken und Schwächen deutscher Referendarinnen und Referendare im internationalen Vergleich

Sigrid Blömeke · Ute Suhl

Zusammenfassung: Auf der Basis der Studie „Mathematics Teaching in the 21st Century (MT21)“, die mit kriteriengeleitet zusammengestellten Stichproben aus Deutschland, Bulgarien, Südkorea, Taiwan, Mexiko und den USA durchgeführt wurde, werden die fachwissenschaftlichen und fachdidaktischen Kompetenzen angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich untersucht. Um Stärken und Schwächen präziser analysieren zu können, wird neben einer traditionellen IRT-Skalierung, die über Einfach-Ladungen die Testperformanz in Mathematik und Mathematikdidaktik abbildet, eine alternative Form der Skalierung durchgeführt. Unter der Annahme, dass die Lösung mathematikdidaktischer Items neben mathematikdidaktischer auch mathematischer Kompetenz bedarf, werden Doppelladungen auf diese beiden Faktoren zugelassen. In einem weiteren Modell erfolgt zudem eine Berücksichtigung der Inhaltsgebiete Arithmetik, Algebra, Funktionen, Geometrie und Stochastik als Erklärungsfaktoren für die Lösung fachbezogener Items. Dieses dritte Modell, das eine hierarchische Struktur von Lehrerkompetenzen annimmt, weist die beste Anpassung an die Daten auf. Nur in diesem Modell werden auch die je spezifischen Schwerpunktsetzungen der an MT21 beteiligten Schul- und Ausbildungssysteme deutlich, und zwar sowohl im Hinblick auf die relative Gewichtung von Lerngelegenheiten in Mathematik und Mathematikdidaktik als auch im Hinblick auf die Lerngelegenheiten in den fünf Inhaltsgebieten.

Schlüsselwörter: Lehrerausbildung · Lehrerkompetenzen · IRT-Skalierung · Internationaler Vergleich · Mathematikunterricht

Online publiziert: 11.08.2010

Univ.-Prof. Dr. S. Blömeke (✉) · Dr. U. Suhl
Philosophische Fakultät IV, Abt. Systematische Didaktik und Unterrichtsforschung,
Humboldt-Universität zu Berlin, Unter den Linden 6,
10099 Berlin, Deutschland
E-Mail: sigrid.bloemeke@staff.hu-berlin.de

Dr. U. Suhl
E-Mail: ute.suhl@staff.hu-berlin.de

Modeling teacher competencies—Using different IRT-scales to diagnose strengths and weaknesses of German teacher trainees in an international comparison

Abstract: This contribution will investigate both the subject-specific and the didactic competencies of trainee teachers for lower secondary education in mathematics on the basis of the study "Mathematics Teaching in the 21st Century (MT21)", which compares criteria-based samples from Germany, Bulgaria, South Korea, Taiwan, Mexico and the USA. In order to analyze their strengths and weaknesses more precisely, the paper considers both traditional IRT-scaling of competencies, which involves a simple loading of test performance in mathematics and didactics of mathematics, and an alternative. Under the assumption that solving items concerning didactics of mathematics requires competencies in didactics of mathematics and mathematics itself, a double-barreled approach was used in the alternative. In a further elaboration of this model, knowledge of arithmetic, algebra, functions, geometry and stochastic was used as a set of further explanatory factors for solving the items. This third model, which presumes a hierarchical structure of teaching competencies, displays the best data fit. Only this model reflects the specialisms and focal points of the education and training systems in the countries participating in MT21 concerning the relative learning opportunities in mathematics and didactics of mathematics and concerning the five subject areas in the field of mathematics.

Keywords: International comparison · IRT-scales · Maths lessons · Teacher competencies · Teacher training

Schulleistungsstudien wie TIMSS und PISA liefern für Deutschland seit mehr als zehn Jahren zuverlässig Auskunft über die Leistungsfähigkeit unserer Schülerinnen und Schüler im internationalen Vergleich. Entsprechende Studien fehlten für den Lehrerbereich lange Zeit fast vollständig. Im vorliegenden Beitrag wird am Beispiel angehender Mathematiklehrkräfte für die Sekundarstufe I am Ende ihrer Ausbildung erstmals eine Analyse der spezifischen Leistungsstärken und -schwächen deutscher Referendare im internationalen Vergleich vorgelegt, und zwar auf der Basis der Untersuchung „Mathematics Teaching in the 21st Century (MT21)". *MT21* wurde in kriteriengeleitet ausgewählten Regionen von sechs Ländern durchgeführt: in Deutschland, Bulgarien, Südkorea, Taiwan, Mexiko und den USA.

Unsere leitende Annahme war, dass sich die Kompetenzen der Lehrkräfte aus diesen sechs Stichproben in ihrem Umfang und ihren Schwerpunkten deutlich unterscheiden. Damit die Ergebnisse möglichst spezifisch und präzise ausfallen und Rückschlüsse auf Wirkungen der Lehrerausbildung zulassen, werden unterschiedliche Skalierungsmodelle verwendet, um neben der herkömmlichen Beschreibung von Testperformanz in den Hauptdimensionen dezidiert Teilkompetenzen der angehenden Mathematiklehrkräfte betrachten zu können. Die zentrale Hypothese ist, dass auf diese Weise *länderspezifische Profile* zu erkennen sein werden. Diese spiegeln vermutlich Schwerpunktsetzungen der jeweiligen Lehrerausbildungs- und Schulsysteme wider.

Mit der Aufstellung und dem Vergleich der unterschiedlichen Skalierungsmodelle soll gleichzeitig ein theoretischer Beitrag zu einer präziseren Modellierung von „Lehrerkompetenzen", vor allem im Hinblick auf deren fachwissenschaftliche und fachdidaktische Komponenten geleistet werden. Die sowohl im deutsch- als auch im englischsprachigen

Raum weit verbreitete Unterscheidung von fachdidaktischem Wissen *(pedagogical content knowledge)* und Fachwissen *(content knowledge)* stellt eine analytisch auf den ersten Blick überzeugende Kategorisierung dar (Shulman 1985; Bromme 1992). In vielen Fachdidaktiken – nicht nur in der Mathematikdidaktik – wird sie aber seit Jahren intensiv diskutiert, insbesondere unter dem Aspekt, ob eine Separierung überhaupt möglich und wenn ja, wie diese genau zu konzeptualisieren sei (Graeber u. Tirosh 2008). In empirischen Studien wird diese Frage besonders virulent, wenn es darum geht zu bestimmen, welche Fähigkeitsdimensionen herangezogen werden, um ein mathematisches oder mathematikdidaktisches Item zu lösen.

Im Folgenden werden zunächst der theoretische Rahmen der Studie *MT21* (1) und die hier verfolgten Forschungsfragen (2) dargelegt, bevor das Untersuchungsdesign beschrieben (3) und auf Details der Skalierung eingegangen wird (4). Im Anschluss wird präsentiert, durch welche Stärken und Schwächen sich die angehenden Lehrkräfte aus den sechs *MT21*-Stichproben im internationalen Vergleich auszeichnen und in welchem Verhältnis diese Ergebnisse zur Gestaltung von Lehrerausbildung und Schule stehen (5). Abschließend werden Schlussfolgerungen für weitere Forschungen und die Lehrerausbildung diskutiert (6).[1]

1 Theoretischer Rahmen

Den Kern des theoretischen Rahmens von *MT21* bildet eine Konzeptualisierung der professionellen Kompetenz, mit der Lehrkräfte berufliche Anforderungen erfolgreich bewältigen können. Im Anschluss an Weinert (1999) wurde diese Kompetenz differenziert in

- kognitive Fähigkeiten und Fertigkeiten (Professionswissen) sowie
- damit verbundene motivationale, volitionale und soziale Bereitschaften und Fähigkeiten, Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll zu nutzen (professionelle Überzeugungen).

Professionelle Kompetenz im Weinertschen Sinne stellt damit ein Konstrukt dar, das kognitive und affektiv-motivationale Dispositionen umfasst. In Tab. 1 ist dokumentiert, welche handlungsrelevanten Situationen und Anforderungen für Mathematiklehrkräfte in den sechs *MT21*-Ländern identifiziert wurden (für weitere Details s. Blömeke et al. 2008). „Unterrichten“ sowie „Beurteilen, Beraten und Fördern“ stellen danach Kernaufgaben von Lehrpersonen dar, die länderübergreifend als unverzichtbar gelten und zugleich als quantitativ messbar erschienen. Hierauf sind daher die Tests in *MT21* ausgerichtet. Im Gegensatz dazu sind „Erziehen“ und „professionelle Ethik“ im internationalen Vergleich unterschiedlich konnotiert und damit nur schwer operationalisierbar. Der in Deutschland bedeutsame Bereich der „Schulentwicklung“ stellt nicht in allen Ländern eine Kernanforderung an Lehrpersonen dar.

Insofern ist darauf hinzuweisen, dass Lehrerkompetenzen über die in *MT21* erfassten Dimensionen hinausgehen. Erziehen, Beraten und Schulentwicklung sowie ein professionelles Ethos sind wichtige Anforderungen an Lehrpersonen und sollen in ihrer Bedeutung durch die Testanlage nicht negiert werden. Eine Reduktion von Lehrerausbildung und Lehrerhandeln auf Unterrichten und Beurteilen würde eine Engführung schulischer Funk-

Tab. 1: Definition beruflicher Anforderungen von Mathematiklehrkräften in *MT21*

Berufliche Aufgaben	Situationen
A: Unterrichten	1. Auswahl/Einordnung von Unterrichtsthemen 2. Unterrichtsplanung
B: Beurteilen, Beraten und Fördern	1. Diagnose von Schülerleistungen 2. Leistungsbeurteilung 3. Beratung von Schüler(inne)n und Eltern 4. Umgehen mit Fehlern, Rückmeldung geben
C: Erziehen	1. Lehrer-Schüler-Beziehung 2. Förderung sozial-moralischer Entwicklung 3. Umgang mit besonderen Risiken 4. Vorbeugung von und Umgang mit Störungen
D: Schulentwicklung	1. Beteiligung an Kooperationen 2. Beteiligung an der Schulevaluation
E: Professionelle Ethik	1. Übernahme professioneller Verantwortung

tionen implizieren, die von *MT21* nicht intendiert ist. Gleichzeitig kann aber festgehalten werden, dass die zentralen Aufgaben von Lehrpersonen damit angemessen erfasst sind.

In Ergänzung zu diesem kompetenzorientierten Zugang zur Testentwicklung, der sich auf die unter A und B genannten beruflichen Situationen bezieht, wurde ein analytischer Ansatz gewählt, um die in diesen Situationen benötigten Teilkompetenzen kognitiver und affektiv-motivationaler Provenienz im *MT21*-Test repräsentieren zu können. Dabei galt es sicherzustellen, dass für eine realistische Abbildung der erforderlichen Teilkompetenzen sowohl die wichtigsten Inhaltsgebiete, in die der Mathematikunterricht unterteilt ist und in denen eine Lehrkraft durchaus unterschiedlich leistungsstark sein kann, als auch die wichtigsten didaktischen Anforderungen an Lehrkräfte zur Bewältigung der oben genannten beruflichen Situationen mit hinreichend vielen Items im Test vertreten sind.

1.1 Inhaltsgebiete des Mathematikunterrichts

Die Inhalte des Mathematikunterrichts wurden in *MT21* in fünf Gebiete unterteilt. Mit Arithmetik, Algebra, Funktionen und Geometrie wurden Inhaltsgebiete berücksichtigt, die traditionell zum Standardrepertoire des Mathematikunterrichts in der Sekundarstufe I gehören. Statistik – in den KMK-Bildungsstandards unter der Leitidee „Daten und Zufall“ zusammengefasst – stellt demgegenüber ein Gebiet dar, dem aufgrund seiner hohen Anwendungsrelevanz in Alltag und Wissenschaft erst neuerdings größeres Gewicht eingeräumt wird (NCTM 2000; KMK 2004).

Im Zuge der Item-Entwicklung galt es im Blick zu haben, dass diese fünf Gebiete hinreichend Items aufweisen. Gleichzeitig hatte eine Orientierung an den in Tab. 1 dokumentierten beruflichen Anforderungen zu erfolgen, um sicherzustellen, dass der Test hinreichend auf den schulischen Alltag ausgerichtet ist. Schließlich erfolgten zwei systematische Einschätzungen der curricularen Validität des *MT21*-Tests in Deutschland, zum ersten im Zuge der Testentwicklung im Hinblick auf die Studienordnungen der deutschen Mathematiklehrerausbildung und zum zweiten nach deren Veröffentlichung im Hinblick auf die Standards für die Lehrerbildung (DMV et al. 2008).

1.1.1 Arithmetik

Die Arithmetik-Aufgaben des *MT21*-Tests erfordern entsprechend beispielsweise die Veranschaulichung von Brüchen als typische Anforderung im Zuge der Unterrichtsplanung, die Ermittlung prozentualer Veränderungen oder die Berechnung der Wahrscheinlichkeit von „1“ als letzter Ziffer quadrierter Zahlen als mathematische Anforderungen im Zuge der Diagnose von Schülerleistungen und Leistungsbeurteilung sowie die Beurteilung der Frage, ob „0,99…“ gleich „1“ ist im Zuge des Gebens von Rückmeldungen zu Schülerfragen. Aus arithmetischer Sicht werden damit grundlegende Aspekte der Zahlentheorie, Brüche und Dezimalzahlen abgedeckt, sodass für Deutschland nach Experteneinschätzung von einer guten Repräsentativität dieses Inhaltsgebietes gesprochen werden kann.

1.1.2 Algebra

Die Algebra-Aufgaben betreffen zum Beispiel den kumulativen Aufbau curricularer Inhalte im Bereich Algebra als typische Anforderung im Zuge der Auswahl von Unterrichtsthemen, die Veranschaulichung von Äquivalenzrelationen während der Unterrichtsplanung, die Umformung logarithmischer Ausdrücke oder die Bestimmung der Zahl der Lösungen einer gegebenen quadratischen Gleichung zu Diagnosezwecken sowie generalisiertes Wissen über algebraische Gleichungen beim Umgang mit Schülerfehlern. Nach Experteneinschätzung stellt die Item-Auswahl für Deutschland eine angemessene Auswahl aus dem Inhaltsgebiet der Algebra dar, namentlich unter Berücksichtigung der gegebenen Beschränkungen der Testzeit.

1.1.3 Funktionen

Die Aufgaben zum Bereich der Funktionen behandeln die Definition von Stetigkeit und die Interpretation von Graphen als mathematische Anforderungen, eingebettet in vielfältige kontextuelle Probleme und ausgerichtet auf den Umgang mit Schülerfehlern, die Identifikation von Fehlvorstellungen und die Leistungsbeurteilung. Das Gebiet ist aus deutscher Sicht nach Experteneinschätzung nur unzureichend abgedeckt, da der Funktions*begriff* nur implizit vorkommt und graphische Darstellungen vor allem im Vergleich zur Definition dieses Gebietes in den deutschen Standards für die Lehrerbildung (DMV et al. 2008) überrepräsentiert sind.

1.1.4 Geometrie

Die international ausgewählten Inhalte aus der Geometrie werden nach Experteneinschätzung typischerweise auch in Deutschland gelehrt, sodass ihnen eine gute Repräsentativität zugesprochen werden kann. Es geht um die Interpretation von Schülerskizzen zu geometrischen Zusammenhängen, um die Bestimmung von Figuren im Kreis, um Verschiebungen, um das Beweisen von Winkelgrößen in einem Dreieck, um Folgerungen aus der Bestimmung der Mittelsenkrechten in einem Dreieck und um die Diagnose von Schülerfähigkeiten anhand von freigegebenen Aufgaben aus TIMSS 1995 (Harmon et al. 1997).

1.1.5 Statistik

Der Bereich der *Statistik* hat die Analyse und Zusammenfassung von Testergebnissen zu Gruppen und deren graphische Repräsentation sowie die Interpretation des Begriffs „arithmetisches Mittel“ zum Gegenstand. Versteht man Statistik als schulisches Inhaltsgebiet, in dem es vor allem um die Interpretation von Daten geht, so kann man die curriculare Repräsentativität als gegeben ansehen. Das größere Themengebiet der Stochastik, das in den DMV-, GDM- und MNU-Standards für die Lehrerbildung vorgesehen ist (DMV et al. 2008), ist durch die Aufgaben allerdings nicht hinreichend repräsentiert. Im Vergleich zu den anderen Inhaltsgebieten ist die Statistik im Bereich Mathematik durch vergleichweise wenige Items vertreten.

1.2 Didaktische Anforderungen

Knapp zwei Drittel der *MT21*-Items sind direkt auf die Bewältigung der didaktischen Anforderungen ausgerichtet, mit denen Mathematiklehrkräfte laut Tab. 1 in ihrem beruflichen Alltag konfrontiert sind. DMV, GDM und MNU (vgl. DMV et al. 2008) halten diese Anforderungen in ihren Standards für die Lehrerbildung in Deutschland ebenfalls für zentral.

In curricularer und unterrichtsplanerischer Hinsicht – Anforderung A: Unterrichten – wird die Bewältigung beruflicher Probleme verlangt, die sich bereits vor Beginn des Unterrichts stellen. Planerisch gesehen müssen fachliche Inhalte für die Schülerinnen und Schüler ausgewählt, begründet, angemessen vereinfacht und unter Nutzung verschiedener Repräsentationsformen aufbereitet werden (Krauthausen u. Scherer 2007; Vollrath 2001). Curricular gesehen handelt es sich um die Frage des Aufbaus mathematischer Kompetenz im Durchgang oder Aufstieg durch die Schuljahre. Was müsste es zum Beispiel für spätere Unterrichtseinheiten bedeuten, wenn man einen klassischen Themenbereich der Schulmathematik in der Sekundarstufe I aus dem Curriculum entfernte? In den DMV-, GDM- und MNU-Standards (vgl. DMV et al. 2008) werden diese Anforderungen an die Lehrkräfte als „fachbezogene Reflexionskompetenzen“ bezeichnet.

Berufliche Anforderungen während des Unterrichts betreffen das unterrichtliche Handeln von Lehrpersonen in der unmittelbaren Interaktion mit Schülerinnen und Schülern, also Anforderung B: Beurteilen, Beraten und Fördern. Hier geht es darum, deren Antworten – handele es sich nun um verbale oder schriftliche Reaktionen auf Aufgaben oder Fragen – bezüglich ihres kognitiven Niveaus, ihrer Komplexität sowie eventueller Fehler und Fehlermuster einzuordnen, Rückmeldungen zu geben und mit Interventionsstrategien angemessen darauf zu reagieren (Krauthausen u. Scherer 2007; Vollrath 2001). Neben der kognitiven Zielorientierung gilt es zudem, die Lernenden zu motivieren bzw. ihre Motivation aufrecht zu erhalten. Die deutschen Standards für die Mathematiklehrerausbildung formulieren detailreich diese Anforderungen.

Die bisher angesprochenen Bereiche mathematikdidaktischer Kompetenz sind insofern durch die *MT21*-Aufgaben breit repräsentiert. Allerdings ist festzuhalten, dass der in den DMV-, GDM- und MNU-Standards zusätzlich aufgeführte Teilbereich der „mathematikdidaktischen Basiskompetenzen“ nicht in das Design aufgenommen werden konnte. Die *MT21*-Aufgaben sind überwiegend als Reaktion auf einzelne, wohl definierte Situationen

der oben geschilderten Art angelegt, weniger als Herausforderung zu grundsätzlichen Reflexionen über Mathematikunterricht, zur langfristigen Planung von Unterricht oder zum Umgang mit komplexen Situationen.

2 Fragestellungen

Entsprechend dem bisher Dargelegten werden in diesem Beitrag zwei zentrale Fragestellungen verfolgt. Die erste Fragestellung ist empirisch-konzeptueller Art und richtet sich auf die mehrdimensionale Modellierung von Lehrerkompetenzen. Ziel ist, in zwei Schritten ein Modell zu entwickeln, das den dargestellten theoretischen Rahmen und die damit verbundenen Hypothesen zum Verhältnis von mathematischen, mathematikdidaktischen und inhaltsgebundenen Teilkompetenzen präziser widerspiegelt als traditionelle Kompetenzmodelle, die sich auf die Testperformanz beschränken. In einem ersten Schritt wird – geleitet von der folgenden Hypothese – dem Verhältnis von mathematischer und mathematikdidaktischer Kompetenz nachgegangen: Während die Lösung der mathematischen Items allein von der zugrunde liegenden mathematischen Teilkompetenz beeinflusst ist, hängt die Lösung der mathematikdidaktischen Items nicht nur von der mathematikdidaktischen, sondern auch von der mathematischen Teilkompetenz ab (H1). Diese stellt also einen Generalfaktor dar, der die Lösung aller Items beeinflusst. In einem zweiten Schritt werden die Inhaltsgebiete aufgenommen. Dabei wird davon ausgegangen, dass die Lösung sowohl der mathematischen als auch der mathematikdidaktischen Testitems zusätzlich durch inhaltsspezifische Teilkompetenzen beeinflusst ist (H2).

Die zweite Fragestellung ist empirisch-methodischer Natur. Die leitende Annahme hier ist, dass bei einer Modellierung, die die Mehrdimensionalität der Lehrerkompetenzen entsprechend H1 und H2 auf Item-Ebene aufnimmt, stichprobenspezifische Profile zu erkennen sein werden, die Schwerpunktsetzungen der jeweiligen Lehrerausbildungs- und Schulsysteme widerspiegeln, die in traditionellen Modellen überdeckt werden. Dies sollte sowohl für das Verhältnis von Mathematik und Mathematikdidaktik gelten (H3) als auch für die fünf Inhaltsgebiete Arithmetik, Algebra, Funktionen, Geometrie und Statistik (H4).

3 Untersuchungsdesign

3.1 Untersuchungsgruppe

Die *MT21*-Zielgruppe angehender Mathematiklehrkräfte der Sekundarstufe I wurde in einem mehrschrittigen Verfahren kriteriengeleitet definiert. Zunächst wurden die Teilnahmeländer, dann die Ausbildungsinstitutionen und schließlich die Personen bestimmt, die erreicht werden sollten (s. Tab. 2). Insofern ist darauf hinzuweisen, dass es sich bei der *MT21*-Stichprobe nicht um eine Zufallsauswahl handelt, sondern um ein sorgfältig zusammengestelltes *Judgement Sample* (Anderson et al. 2009), das zentralen von Expertinnen und Experten bestimmten Populationsparametern folgt, mit denen die Variation der Lehrerausbildung möglichst umfassend abgebildet wird.

Der Auswahl der sechs Länder Bulgarien, Deutschland, Mexiko, Südkorea, Taiwan und USA lagen Daten aus drei Surveys zugrunde. Die Länder repräsentierten zum Zeit-

Tab. 2: *MT21*-Stichprobe

	Bulgarien	Deutschland	Mexiko	Südkorea	Taiwan	USA
Anzahl Befragte	100	286	149	104	265	223
Anzahl Institutionen	3	4 Universitäten, 22 Seminare	5	4	5	12
Ausbildungsprogramme	SI/II	P/SI; SI/SII	SI	SI/II	SI/II	P/SI; SI; SI/II
Länge (in Jahren)	4	3,5 + 1,5; 4,5 + 2,0	4	4	5	4 oder 5
Auswahlkriterien	Größe	Typ, Größe, Region, Selektivität	Selektivität	Selektivität	Typ, Selektivität	Typ, Größe, Region, Selektivität

P/SI: kombinierte Primar- und Sekundarstufen-I-Ausbildung; SI: reine Sekundarstufen-I-Ausbildung; SI/II: kombinierte Sekundarstufen-I- und II-Ausbildung

punkt der Studie die Haupttypen an Lehrerausbildungssystemen (Eurydice 2004; OECD 2004),[2] sie decken das Spektrum an Schülerleistungen in der Sekundarstufe I ab (Mullis et al. 2008; OECD 2007) und sie stellen überwiegend Länder mit einem hohen Entwicklungsniveau dar, um sozio-ökonomische Verzerrungen gering zu halten (UN 2008). In Bulgarien, Südkorea und überwiegend auch in den USA findet die Sekundarstufen-I-Lehrerausbildung in Vier-Jahres-Programmen an Universitäten statt. In Mexiko handelt es sich um ebenso lange Ausbildungen an Hochschulen, die auf die Lehrerausbildung spezialisiert sind. In Taiwan finden sich beide Typen an Ausbildungsinstitutionen; die Länge der Ausbildung umfasst dabei jeweils fünf Jahre: vier Jahre an den Institutionen und ein Jahr in der Schulpraxis. Die Ausbildung in Deutschland ist zweiphasig und findet zunächst für 3,5 bis 4,5 Jahre an Universitäten bzw. Pädagogischen Hochschulen und dann für 1,5 bis zwei Jahre an staatlichen Studienseminaren statt. Sekundarstufen-I-Lehrkräfte werden entweder in kombinierten Primar- und Sekundarstufen-I-Programmen (Deutschland, USA), in kombinierten Sekundarstufen-I- und -II-Programmen (Bulgarien, Südkorea, Taiwan) oder in spezialisierten Sekundarstufen-I-Programmen ausgebildet (Mexiko, USA).

Innerhalb der sechs Länder wurden anhand von vier Kriterien Institutionen ausgewählt, die die Variationsbreite der Ausbildungssysteme widerspiegeln. Bei den Kriterien handelte es sich um den Ausbildungstyp, die Größe der Institutionen, ihre regionale Lage und ihre Selektivität. Als Indikator für Letztere wurden die für eine Zulassung erforderlichen Schul- oder Testleistungen verwendet. In den beiden großen Teilnahmeländern Deutschland und USA wurden alle vier Kriterien angewandt, in den übrigen vier Ländern nur jene, von denen zu erwarten war, dass sie die Variation am besten abbilden würden.

Innerhalb der Institutionen wurden Vollerhebungen der angehenden Lehrkräfte am Ende der Ausbildung angestrebt. Unterschiedliche Ausschöpfungsquoten wurden durch geeignete Gewichtungsverfahren ausgeglichen, und zwar nach dem Modell prinzipiell gleicher Ziehungswahrscheinlichkeiten (*response homogeneity group*-Modell; Särndal et al. 1997) von Individuen pro Institution und von Institutionen pro Region (Gabler et al. 1994; Kish 1965). Die Gewichte wurden so gewählt, dass die Summe der Gewichte der Anzahl der befragten Personen entspricht (Rosing u. Ross 1992).

Die sechs Stichproben sind damit für die in der Auswahl befindlichen Institutionen und Regionen repräsentativ, nicht aber für die beteiligten Länder insgesamt. Eine weitere Adjustierung der Gewichte im Hinblick auf landesspezifische Populationen konnte nicht erfolgen, da entsprechende Angaben fehlen. Insofern ist deutlich darauf hinzuweisen, dass unmittelbare Ländervergleiche auf der Basis einzelner Mittelwerte nicht zulässig sind, da die Qualität der Stichproben nicht hinreichend genau eingeschätzt werden kann. Im Fokus des vorliegenden Beitrags stehen entsprechend nicht die deskriptiven Befunde, sondern die *Struktur*unterschiede, die in den verschiedenen Skalierungsansätzen sichtbar werden und deren Punktschätzer jeweils ggf. vergleichbar verzerrt sind. Insgesamt besteht die *MT21*-Stichprobe angehender Mathematiklehrkräfte der Sekundarstufe I aus 1.127 Personen.

3.2 Untersuchungsinstrumente

Für den fachbezogenen Leistungstest wurde ein rotiertes Testdesign mit zwei Testheften verwendet, um angesichts der beschränkten Erhebungszeit von 90 min eine hinreichend

Tab. 3: Anzahl der Items in den Dimensionen des *MT21*-Leistungstests

	Mathematik	Mathematikdidaktik	Gesamt
Arithmetik	7	15	22
Algebra	10	6	16
Funktionen	7	11	18
Geometrie	8	5	13
Statistik	2	9	11
Gesamt	34	46	80

große Zahl an Items einsetzen zu können. Die Entwicklung des Tests begann mit einer Sichtung vorhandener Studien, um Items zu identifizieren, die sich bereits bewährt hatten. In einem zweiten Schritt wurden anhand der oben dargestellten analytischen Dimensionierung professioneller Kompetenz in den nationalen Projektteams der sechs Teilnahmeländer Items entwickelt. Zudem wurden unter Einbeziehung von Mathematikern, Mathematikdidaktikern und Erziehungswissenschaftlern aus den teilnehmenden Ländern Item-Entwicklungsworkshops durchgeführt. Auf diese Weise entstand ein umfangreicher Itempool, der mehreren Expertenreviews unterzogen wurde. Die verbleibenden Aufgaben wurden in einer Pilotstudie erprobt, auf deren Basis die endgültige Zusammenstellung des Leistungstests für die Hauptuntersuchung erfolgte. In Tab. 3 ist dokumentiert, in welchem Umfang die Teilkompetenzen in der Endfassung repräsentiert sind.[3]

In Abb. 1 ist eine mathematische Beispielaufgabe aus dem Bereich der Algebra dargestellt. In der Aufgabe ist die Anzahl der Lösungen einer quadratischen Gleichung zu bestimmen. Option 3 stellt die korrekte Lösung dar. Das Item ist ein Beispiel für die in *MT21* angewendete Idee, Distraktoren nach bekannten konzeptionellen Fehlvorstellungen angehender Lehrkräfte zu formulieren. Auf diese Weise werden Detailanalysen möglich, die unterhalb der Ebene „falsch – richtig“ Hinweise auf das Niveau des fachbezogenen Wissens geben, indem den falschen Antworten eine unterschiedliche Qualität zukommt.

Das in Abb. 2 dokumentierte Beispiel zur Erfassung mathematikdidaktischen Wissens geht von einer klassischen Aufgabe aus dem Bereich der Geometrie aus und beinhaltet vier Items. Die Aufgabe thematisiert die Angemessenheit von unterschiedlichen Schülerbeweisen und geht in ihrem schülerbezogenen Kern auf eine Aufgabe aus dem „Year 8 Proof Survey“ von Küchemann u. Hoyles (2002) zurück, die in geeigneter Form adaptiert wurde. Die korrekten Lösungen sind für Anna und Bruno jeweils Option 2, für Charlotte

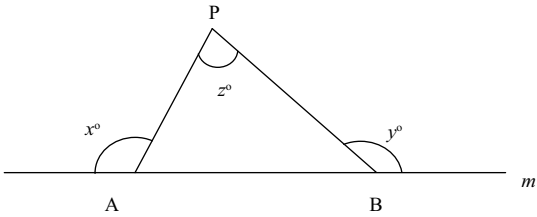
Abb. 1: Mathematische Beispielaufgabe aus *MT21*

B47. Wenn $a > 0$ ist, wie viele verschiedene reelle Lösungen hat dann die Gleichung $x^2 + x - a = 0$?

*Kreuzen Sie **ein** Kästchen an.*

1. 0 ..... ☐
2. 1 ..... ☐
3. 2 ..... ☐
4. Unendlich viele ..... ☐
5. Die Anzahl reeller Lösungen hängt vom Wert von a ab ..... ☐

K4. In der Zeichnung sind A und B zwei feste Punkte auf der Geraden *m*. Punkt P kann bewegt werden, aber bleibt oberhalb von *m* und bleibt verbunden mit A und B.



Anna, Bruno, Charlotte und Daniel diskutieren, ob die folgende Aussage wahr ist:
x° + y° ist gleichgroß wie 180° + z°.

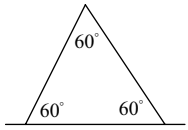
Annas Antwort

Ich habe die Winkel in der Zeichnung gemessen und herausgefunden, dass der Winkel *x* 110° beträgt. Winkel *y* ist 125° und Winkel *z* ist 55°.
110° + 125° = 235°
und 180° + 55° = 235°.

Also sagt Anna, dass es wahr ist.

Brunos Antwort

Ich kann Punkt P so verschieben, dass das Dreieck gleichseitig ist und seine Winkel 60° sind.

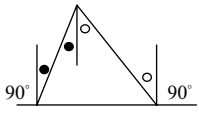


Damit ist *x* 120° und *y* ist 120°.
120° + 120° ist dasselbe wie 180° + 60°.

Also sagt Bruno, dass es wahr ist.

Charlottes Antwort

Ich habe drei parallele Linien gezeichnet, die zur Basis rechtwinklig sind.

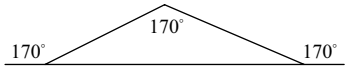


Die beiden mit einem ● markierten Winkel sind gleichgroß und die beiden mit einem ○ markierten sind gleichgroß.
Winkel *x* ist 90° + ● und Winkel *y* ist 90° + ○. Also *x* + *y* ist 180° + ● + ○, was 180° + *z*° ist.

Also sagt Charlotte, dass es wahr ist.

Daniels Antwort

Ich habe an eine Zeichnung gedacht, wo die Winkel *x, y* und *z* alle 170° sind.



In meiner Zeichnung sind *x* + *y* nicht gleich zu 180° + *z*°.

Also sagt Daniel, dass es nicht wahr ist.

*Kreuzen Sie die am **meisten angemessene** Antwort für **jede(n) Schüler(in)** an.*

	Anna	**Bruno**	**Charlotte**	**Daniel**
1. Deine Argumentation enthält einen Fehler. Denke nochmals darüber nach.	☐	☐	☐	☐
2. Es reicht nicht aus, die Aussage an einem Beispiel zu überprüfen. Denke nochmals darüber nach.	☐	☐	☐	☐
3. Ausgezeichnet! Dies ist ein überzeugender Beweis.	☐	☐	☐	☐

Abb. 2: Mathematikdidaktische Beispielaufgabe in *MT21*

Option 3 und für Daniel Option 1. Weitere Item-Beispiele sind in Blömeke et al. (2008) dokumentiert.

Der Umfang und die Inhalte der mathematischen und mathematikdidaktischen Ausbildung der angehenden Lehrkräfte wurden institutionenübergreifend in Form von Dokumentenanalysen der Ausbildungscurricula und Expertenbefragungen durchgeführt. Sie ergeben einen Eindruck davon, welche Schwerpunktsetzungen in den jeweiligen Ausbildungssystemen gesetzt werden (zum Zusammenhang von Lerngelegenheiten in der Lehrerausbildung und Kompetenzerwerb auf Individualebene s. Blömeke et al. 2010).

4 Skalierung

4.1 Modellierung des Verhältnisses von fachwissenschaftlicher und fachdidaktischer Lehrerkompetenz

Datenanalysen in Studien mit standardisierter Testung sind zumeist darauf ausgerichtet, das Kompetenzprofil von Testpersonen über die von ihnen gezeigte Performanz in Untertests abzubilden. Bekanntestes Beispiel hierfür sind Studien zu Schülerleistungen wie die PISA-Studie (Baumert et al. 2001; Prenzel et al. 2004). Die Werte der Testpersonen werden entweder in mehreren eindimensionalen, neuerdings auch gemeinsam in einem mehrdimensionalen Modell probabilistisch skaliert. Unter Performanzgesichtspunkten ist dabei entscheidend, dass jedes Item nur auf eine Dimension lädt, d. h. seine Lösung wird nur auf *eine* Kompetenz zurückgeführt („factorial-simple structure"; McDonald 2000). In den PISA-Skalierungen laden Mathematik-Items beispielsweise lediglich auf die Dimension „mathematische Kompetenz" und nicht auf die Dimension „Lesekompetenz". Die hohe Korrelation von $r = 0{,}77$ (Leutner et al. 2004) zwischen den beiden Konstrukten lässt sich vermutlich teilweise darauf zurückführen, dass Schülerinnen und Schüler auch lesen können müssen, um die Mathematikaufgaben lösen zu können.

Dieselben Skalierungsprinzipen galten für Lehrerstudien wie COACTIV (Brunner et al. 2006) und die ersten nationalen Analysen zu *MT21* (Blömeke et al. 2008). Mathematik-Items luden ausschließlich auf die Dimension „mathematische Kompetenz", mathematikdidaktische Items auf die Dimension „mathematikdidaktische Kompetenz" (s. Abb. 3, links). Auch hier zeigten sich Überschneidungen zwischen den beiden Kompetenzen in hohen latenten Korrelationen um $r = 0{,}80$. Konzeptionell ist dieses Ergebnis unmittelbar einsehbar, setzt die Bewältigung mathematikdidaktischer Anforderungen doch auch die Beherrschung der mathematischen Grundlagen voraus. Die Skalierung der Dimension Mathematikdidaktik bildete insofern die gezeigte Performanz der Lehrkräfte auf den entsprechenden Items als Kombination zweier Teilkompetenzen ab, die zusammen zum Erfolg, heißt zur Lösung dieser Items führten.

So plausibel und notwendig ein solcher deskriptiver Zugang ist, so wenig vermag er es, spezifische Stärken und Schwächen der Lehrkräfte in den zugrunde liegenden Teilkompetenzen sichtbar zu machen, da deren gemeinsame Varianz die Ergebnisse stark prägt. Folgerichtig hatte beispielsweise in *MT21* der starke Zusammenhang von Mathematik und Mathematikdidaktik nur geringe Differenzen in den mathematischen und mathematikdidaktischen Leistungen der angehenden Lehrkräfte erkennen lassen. Um Stärken und

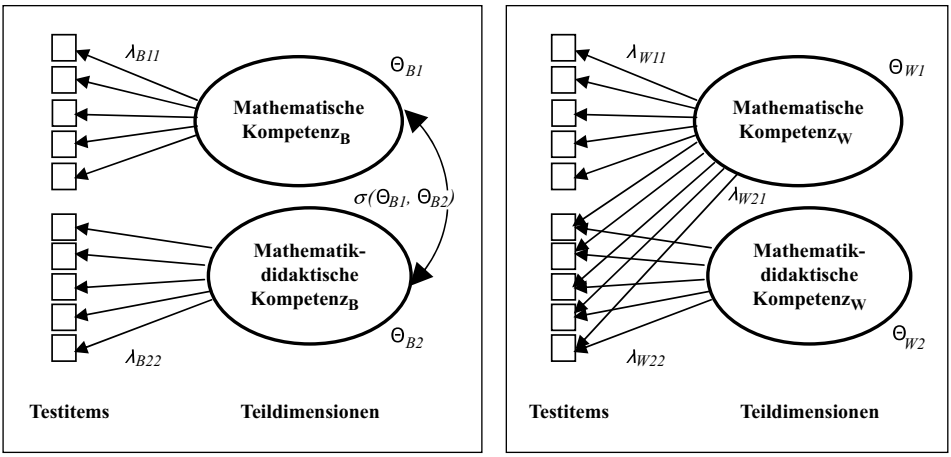


Abb. 3: Between-Item-Mehrdimensionalität (links) und Within-Item-Mehrdimensionalität (rechts; zur Notation s. a. Hartig u. Höhler 2008)

Schwächen präziser analysieren zu können, wird für den vorliegenden Beitrag daher ein zusätzlicher Skalierungsansatz gewählt, der in der empirischen Bildungsforschung bisher eher selten zur Anwendung gelangt ist (Hartig u. Höhler 2008; Koeppen et al. 2008): eine mehrdimensionale probabilistische Skalierung, die für die mathematikdidaktischen Items Doppelladungen zulässt (s. Abb. 3, rechts). Die den mathematischen und den mathematikdidaktischen Items gemeinsame Varianz wird einem Generalfaktor „Mathematik" zugeordnet, sodass die in den Residuen wirksame spezifische Varianz der Mathematikdidaktik besser sichtbar wird. Auf diese Weise wird die Skalierung als diagnostisches Werkzeug genutzt („Multidimensional IRT as a Diagnostic Aid"; vgl. Walker u. Beretvas 2003).

Empirisch sind die beiden Modelle – unter der festgelegten Restriktion gleicher Faktorladungen (s. unten) – mit Einfachladungen und Doppelladungen gleichwertig, da sie dieselben Zusammenhänge zwischen den Items implizieren. Das neue Modell nutzt lediglich eine andere Form der Parametrisierung. Konzeptionell steht hinter diesem Zugang die Annahme, dass jede Mathematiklehrkraft über eine generelle mathematische Kompetenz verfügt, die für die Lösung aller fachbezogenen Test-Items erforderlich ist, dass sie aber *zusätzlich* über eine spezifische mathematikdidaktische Kompetenz verfügt, die nur bei der Lösung der mathematikdidaktischen Items relevant wird. Ein solcher Ansatz kann als „within-item multidimensionality" (Adams et al. 1997) bezeichnet werden, da die Lösung der Mathematikdidaktik-Items von mehreren Dimensionen simultan beeinflusst wird, während bei einem traditionellen Vorgehen (Adams et al. 1997: „between-item multidimensionality") die Überlappung zwischen Mathematik und Mathematikdidaktik nur in den latenten Korrelationen sichtbar wird. In der Tradition der Strukturgleichungsmodellierung werden vergleichbare Ansätze als „bi-factor model" (Holzinger u. Swineford 1937) oder „nested-factor model" (Mulaik u. Quartetti 2000) bezeichnet.

Hartig u. Höhler (2008) haben die Aussagekraft eines solchen ergänzenden Vorgehens für das Fremdsprachenlernen am Beispiel der DESI-Daten demonstriert. Der Englisch-Leseverständnis-Test und der Englisch-Hörverständnis-Test, als unabhängige Tests

konzipiert, wurden zunächst traditionell über mehrdimensionale Modelle mit Einfach-Ladungen skaliert (Nold u. Rossa 2008; Nold et al. 2008). Eine spätere Modellierung mit Englisch-Lesekompetenz als Generalfaktor und Hörverständnis als zusätzlichem zweitem Faktor für diesen spezifischen Testteil ermöglichte dann eine präzise Analyse der speziellen Fähigkeiten im Bereich Hörverständnis. Erst diese konnte auf relative Stärken von Jungen aufmerksam machen, die zuvor von den generellen Stärken der Mädchen überdeckt worden waren.

Die Ladungen der Items sowohl im *MT21*-Within- als auch im -Between-Modell wurden frei geschätzt. Dabei wurde zusätzlich festgelegt, dass die Ladungen für die Mathematik-Items einerseits als auch für die Mathematididaktk-Items andererseits pro Dimension gleich sein müssen (s. Abb. 3), um das Modell parametrisch so einfach wie möglich zu halten (Stout 2007). Damit ist zum einen eine Entsprechung zur Denklogik von Raschmodellen gewährleistet, indem jedem Item dasselbe Gewicht zukommt. Zum anderen geht man den von Rost (2004, S. 370 f.) kritisch diskutierten Problemen aus dem Weg, die mit einem Freisetzen der Ladungen verbunden wären.

Damit der spezifische Mathematikdidaktik-Faktor nur die Residualvarianz repräsentiert, wurde die Korrelation zwischen Mathematik und Mathematikdidaktik im Within-Modell auf $r = 0{,}00$ festgesetzt, d. h. die beiden Dimensionen stehen im Vektorraum orthogonal zueinander. Alle Schätzungen wurden mit *MPlus* in der Version 5.2 mit Combination-Add-on durchgeführt (Muthén u. Muthén 2008). Das in *MPlus* implementierte zwei-parametrische mehrdimensionale logistische Item-Response-Modell kann mehrere latente Kompetenzen simultan berücksichtigen. Um das Cluster-Sampling berücksichtigen zu können, wurde ein robuster Maximum-Likelihood-Schätzer (MLR) und für die Schätzung der Standardfehler ein Sandwich-Schätzer eingesetzt. Die Varianz der beiden Dimensionen wurde auf 1 festgesetzt.

Im *MT21*-Within-Modell wurden die Dimensionen für die vorliegende Studie als additiv-kompensatorisch betrachtet (Reckase u. McKinley 1991), d. h. niedrige Ausprägungen in einer können durch höhere Ausprägungen in einer anderen Dimension kompensiert werden. Dies ergibt sich weniger aus konzeptionellen Überlegungen, unter denen ein multiplikatives Verhältnis nach Möglichkeit geprüft werden sollte, sondern aus der Anlage des Tests, die einen Kompromiss zwischen den teilnehmenden Ländern darstellt. Der Test enthält Items, die spezifisch die mathematische Teilkompetenz erfassen, und Items, zu deren Lösung mathematische und mathematikdidaktische Teilkompetenzen zusammenfließen müssen. Um ein multiplikatives Modell aufzustellen, würden darüber hinaus Items benötigt, für deren Lösung mathematikdidaktische, aber keine mathematische Kompetenz benötigt würde. Aus deutscher Sicht haben wir uns im Zuge der Instrumentwicklung stark hierfür eingesetzt und Beispiele geliefert (Identifizierung mathematischer Denkstile, Probleme von Schülerinnen und Schülern mit Migrationshintergrund beim Umgang mit Mathematikaufgaben, Indikatoren für Rechts-Links-Diskriminationsschwäche), um uns die Möglichkeit der Prüfung offen zu halten, ohne uns allerdings angesichts des damaligen Standes der Theorieentwicklung im englischsprachigen Raum durchsetzen zu können.

Für den Umgang mit fehlenden Werten wurde die *Full Information Maximum Likelihood*-Methode verwendet. Dieses Verfahren setzt voraus, dass es sich bei fehlenden Werten um zufällig fehlende Werte handelt (Little u. Rubin 1987), was für designbedingt

fehlende Werten angesichts der zufälligen Verteilung der beiden Testhefte uneingeschränkt angenommen werden kann. Insgesamt waren in beiden Modellen 83 Item-Parameter zu schätzen: 80 frei geschätzte Item-Schwierigkeiten τ_i sowie zwei frei geschätzte Faktorladungen λ und die latente Korrelation zwischen Mathematik und Mathematikdidaktik im Between-Modell bzw. drei frei geschätzte Faktorladungen im Within-Modell. Bei einer Stichprobengröße von 1.127 Personen ist so eine hinreichende Personenzahl pro Parameter gewährleistet.

Als Personenparameter wurden *Expected A Posteriori* (EAP)-Schätzer berechnet. Dieses zweischrittige Vorgehen bei den deskriptiven Stichprobenvergleichen anstelle einer direkten Schätzung der Unterschiede im Rahmen eines latenten Modells geht auf das spezifische Erkenntnisinteresse des vorliegenden Beitrags zurück, das die Interaktion von Personen und Items in den Vordergrund stellt. Entsprechend der Logik von Strukturgleichungsmodellen, in der MPlus von Muthén u. Muthén (2008) entwickelt wurde, können hier bei latenten Zusammenhängen nur die Varianzen der Residuen der latenten Variablen auf 1 gesetzt werden, wenn Prädiktoren (in diesem Falle die Länderzugehörigkeit) eingefügt werden. Um die Faktorladungen angemessen vergleichen zu können, ohne dass diese von den eingeführten Prädiktoren abhängen, müssen jedoch die Varianzen der latenten Variablen auf 1 gesetzt werden. In Bezug auf die Schätzung der Mittelwerte ist dieses Vorgehen insofern eher unproblematisch, als es eher konservativ ist und Unterschiede damit eher unterschätzt werden.

Im Interesse einer plausibleren Interpretation der Personenwerte wurden die EAP konventionell auf einen Mittelwert von 500 für die 1.127 Lehrkräfte mit einer Standardabweichung von 100 transformiert. Die Standardfehler der Gruppen-Mittelwerte wurden unter Berücksichtigung der Gewichte auf dem üblichen Weg geschätzt (Quadratwurzel aus der Varianz eines Merkmals im Verhältnis zur Stichprobengröße). Auf die Anwendung eines komplexeren Verfahrens (*Balanced Repeated Replication*, *Jackknifing* oder *Bootstrapping*) wurde wegen des großen Aufwandes verzichtet. Damit geht allerdings die Gefahr einher, dass die Standardfehler eher unterschätzt werden.

4.2 Einbezug von Inhaltsgebieten

In einem Folgeschritt führen wir die Inhaltsgebiete Arithmetik, Algebra, Funktionen, Geometrie und Statistik als zusätzliche Erklärungsfaktoren ein und um detailliert Einsicht in entsprechende Stärken und Schwächen der verschiedenen Stichproben zu erhalten. Damit wird nicht nur wiederholt an uns herangetragenen Forderungen aus der Mathematikdidaktik nachgekommen, die Inhaltsabhängigkeit mathematischer und mathematikdidaktischer Kompetenzen stärker zu berücksichtigen, sondern das Vorgehen ist in Bezug auf den vorliegenden Test auch insofern von Vorteil, als die Inhaltsgebiete im Test ungleichgewichtig vertreten sind. Rein mathematische Aufgaben erfordern beispielsweise häufiger algebraische als statistische Kenntnisse, während die mathematikdidaktischen Aufgaben beispielsweise häufiger arithmetische als geometrische Kenntnisse erfordern (s. oben, Tab. 3). Nicht ausgeschlossen werden kann auch, dass einzelne Teilkompetenzen mit Item-Schwierigkeiten konfundiert sind, etwa im Falle von Arithmetik und Algebra (Reckase u. McKinley 1991).

Der reinen Within-Modellierungslogik folgend würden die Inhaltsgebiete neben die Mathematikdidaktik als spezifische Faktoren zu treten haben, die orthogonal zueinander und zur mathematischen Kompetenz als Generalfaktor stünden. Eine solche Modellierung bringt allerdings Interpretationsprobleme mit sich:[4] Auf der einen Seite wäre nach der Bedeutung der fünf latenten mathematischen Inhaltsdimensionen zu fragen, wenn zugleich eine latente Dimension „Mathematische Kompetenz" geschätzt wird, die orthogonal zu den Inhaltsgebieten steht. Dieses Problem tritt auf, selbst wenn Korrelationen zugelassen werden. Auf der anderen Seite erscheint insbesondere eine Interpretation jener unkorrelierten latenten Variablen schwierig, auf die dreifache Ladungen spezifiziert werden.

Wir haben uns daher für eine Kombination von Between- und Within-Ansatz in Form eines Second-Order-Modells entschieden (de la Torre u. Song 2009). Auf der ersten Ebene werden dabei die Vorteile des Between-Modells genutzt: leicht interpretierbare Einfach-Ladungen, während auf der zweiten Ebene die Vorteile des Within-Modells genutzt werden: Spezifizierung der Mehrdimensionalität. Dabei ist es durchaus möglich, mehrere Generalfaktoren zuzulassen – in unserem Falle die mathematische und die mathematikdidaktische Kompetenz einer Lehrkraft. Eine solche hierarchische Anordnung von Kompetenzen mit Generalfaktoren auf der höheren Ebene und spezifischen Kompetenzen – in diesem Falle den Inhaltsgebieten – auf der darunter liegenden Ebene ist prinzipiell nicht neu (Carroll 1993; Gustafsson u. Snow 1997). Sie ist allerdings bisher noch nicht auf Lehrerkompetenzen angewendet worden. In anderen Gebieten hat sich gezeigt, dass der Präzisionsgewinn durch die Anwendung eines solchen Modells gegenüber konventionellen IRT-Skalierungen dann besonders groß ist, wenn – wie im vorliegenden Fall – relativ kurze Subtests für die spezifischen Teilkompetenzen und relativ hohe Korrelationen zwischen diesen vorliegen (de la Torre u. Song 2009). Inhaltlich beinhaltet ein Second-Order-Modell die Annahme, dass die Lösung unserer Test-Items stark inhaltsspezifisch geprägt ist. Die mathematische bzw. mathematikdidaktische Kompetenz hat jeweils keinen *direkten* Effekt. Sie erklären im Sinne einer Grundlage aber, warum die Inhaltsgebiete positiv korrelieren.

Mit dieser komplexen Modellierung wird über das von Hartig u. Höhler (2008) gewählte Vorgehen hinausgegangen. Mathematisch unterscheiden sich das reine Within-Modell und das kombinierte Between- und Within-Modell nicht stark voneinander, sondern sie lassen sich ebenso leicht ineinander überführen wie beispielsweise in ein Testlet-Modell (Yung et al. 1999; Wang u. Wilson 2005). Das Second-Order-Modell (vgl. Abb. 4) ist aber besser inhaltlich zu interpretieren.

Aufgrund des exponentiellen Wachstums der erforderlichen Zahl von Iterationsschritten bei zunehmender Anzahl von Dimensionen erfolgte die numerische Integration im Falle des komplexen Modells mit den Inhaltsgebieten als zusätzlicher Dimension nach dem Monte-Carlo-Algorithmus mit 300 Integrationspunkten pro Dimension. Darauf hinzuweisen ist, dass die Modellierungen angesichts der relativ geringen Stichprobengrößen mit der gesamten *MT21*-Stichprobe zugleich erfolgen. Offen bleiben muss insofern, ob das Messmodell über die verschiedenen Länder hinweg invariant ist.

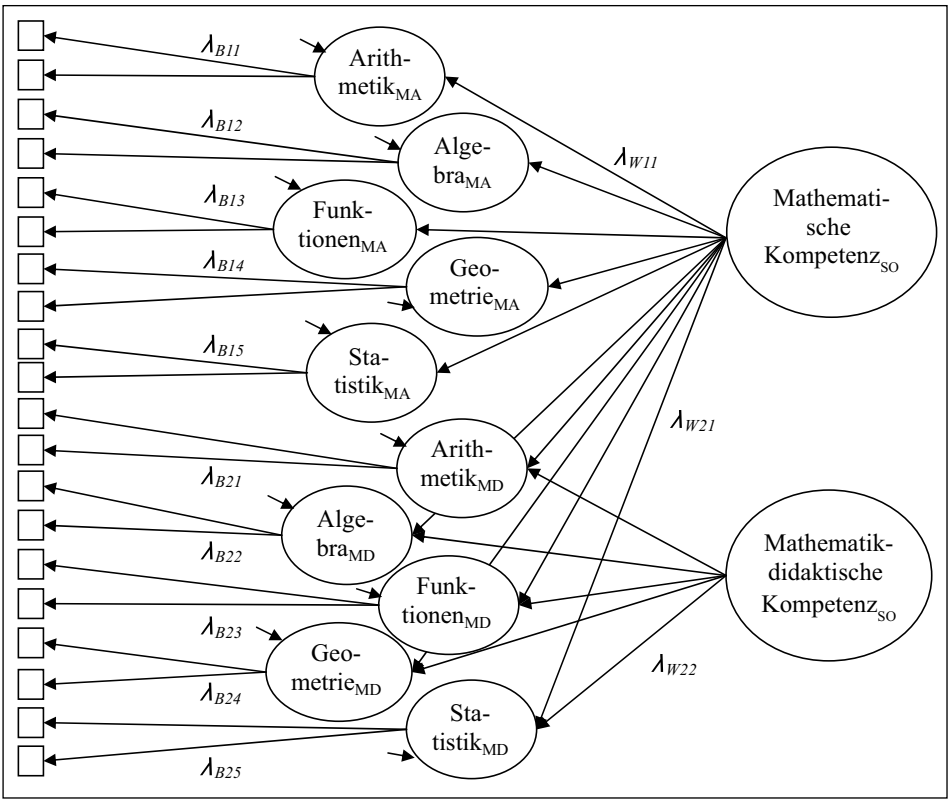


Abb. 4: Kombination von Within-Item- und Between-Item-Mehrdimensionalität professioneller Kompetenzen von Lehrpersonen unter Berücksichtigung inhaltlicher Profile in einem Second-Order-Modell

5 Ergebnisse

5.1 Modellanpassungen und Faktorladungen

Zunächst soll anhand der Anpassung der verschiedenen Modelle an die *MT21*-Daten gezeigt werden, dass multidimensionale Modelle Vorteile gegenüber einer eindimensionalen Skalierung bringen und dass eine Between- und Within-Modellierung unter den vorgenommenen Restriktionen bezüglich der Faktorladungen mathematisch äquivalent sind.

Die beiden zweidimensionalen Modelle weisen eine signifikant bessere Anpassung an die Daten auf als das eindimensionale Modell (s. Tab. 4), während sie im Vergleich untereinander dieselbe Abweichung aufweisen. Die latente Korrelation zwischen mathematischer und mathematikdidaktischer Kompetenz am Ende der Lehrerausbildung ist im internationalen Datensatz zwar geringer als im deutschen Datensatz allein, aber mit $r = 0{,}67$ noch immer sehr hoch.

Tab. 4: Modellanpassungen der verschiedenen Modelle im Vergleich zum eindimensionalen Modell ($n = 1.127$)

Modell	Log-likelihood (LL)	Scaling correction factor (SCF)	Parameter	AIC	BIC_{adj}
Eindimensional	−36.150	1,16	82	72.464	72.615
Mathematik-Mathematikdidaktik (Between)	−35.991	1,14	83	72.148	72.301
Mathematik-Mathematikdidaktik (Within)	−35.991	1,14	83	72.148	72.301
Second-Order-Modell (Between und Within kombiniert)	−35.441	1,18	93	71.068	71.239

AIC: Informationskriterium nach Akaike; BIC_{adj}: adjustiertes Informationskriterium nach Bayes

Tab. 5: Standardisierte Faktorladungen für das eindimensionale und die zweidimensionalen Modelle ($n = 1.127$)

Modell	Faktorladungen Mathematik-Items	Faktorladungen Mathematikdidaktik-Items	
Eindimensional	0,55 (0,01)***		0,26 (0,01)***
Mathematik-Mathematikdidaktik (Between)	0,55 (0,01)*** (λ_{B11} in Abb. 3)		0,33 (0,01)*** (λ_{B22} in Abb. 3)
Mathematik-Mathematikdidaktik (Within)	0,55 (0,01)*** (λ_{W11} in Abb. 3)	0,22 (0,01)***	0,24 (0,01)***
		$Mathematik_W$ (λ_{W21} in Abb. 3)	$Mathematikdidaktik_W$ (λ_{W22} in Abb. 3)

***$p < 0{,}001$

Das Modell, das als hierarchisch angelegtes Second-Order-Modell zusätzlich die Inhaltsgebiete berücksichtigt, weist nicht nur gegenüber dem eindimensionalen, sondern auch gegenüber den zweidimensionalen Modellen eine nochmals verbesserte Anpassung auf. Dies gilt neben der Log-Likelihood und dem AIC insbesondere auch für das adjustierte Bayesiansche Informationskriterium, das die Modellanpassung in Relation zur Zahl zu schätzender Parameter setzt. Ein Chi-Quadrat-Differenz-Test beruhend auf der Log-likelihood und dem *Scaling Correction Factor* unter Berücksichtigung der Zahl der Parameter ergibt zudem eine hochsignifikant geringere Abweichung des Second-Order-Modells verglichen mit den beiden anderen mehrdimensionalen Modellen ($TRd = 728{,}48$).

Die Faktorladungen zeigen die relative Bedeutung der jeweiligen Teilkompetenzen für die Lösung der Items in den verschiedenen Modellen an. Sowohl im eindimensionalen als auch in den beiden zweidimensionalen Modellen kommt der mathematischen Kompetenz für die Lösung der Mathematik-Items substanzielle Bedeutung zu (s. Tab. 5). Deren Varianz wird durch nur einen Faktor also in hohem Maße aufgeklärt.

Im eindimensionalen und im Between-Modell laden die Mathematikdidaktik-Items dagegen zwar hochsignifikant, aber deutlich geringer auf die angenommene mathematik-

didaktische Kompetenz. Dabei fallen die Ladungen im Between-Modell signifikant höher aus als im eindimensionalen Modell. Nur im Within-Modell wird allerdings die spezifische Bedeutung der beiden Einflussfaktoren deutlich. Die Ladung auf den Mathematikdidaktik-Faktor fällt geringer aus als im Between-Modell. Dafür zeigen sich substanzielle Ladungen auf den Mathematikfaktor. Alle Ladungen sind zudem hochsignifikant, was erneut auf die Angemessenheit dieser Ausdifferenzierung verweist.

Im Hinblick auf die Faktorladungen ist die Hinzunahme der Inhaltsgebiete vor allem im Bereich der mathematischen Kompetenz von hoher Bedeutung (s. Tab. 6). Alle Items laden hochsignifikant auf die jeweiligen Inhaltsgebiete. Für die arithmetischen, algebraischen und funktionenbezogenen Items kann nun ein deutlich höherer Varianzanteil erklärt werden. Lediglich für die Statistik-Items sinkt – angesichts der geringen Itemzahl im *MT21*-Test kaum erstaunlich – die Faktorladung. Die inhaltsbezogenen Teilkompetenzen laden dann wiederum alle hochsignifikant auf den Generalfaktor mathematische Kompetenz, der einen Großteil der Varianz erklärt. Diese Struktur stützt einmal mehr die Konzeption einer hierarchischen Struktur mathematischer Lehrerkompetenzen mit einem Generalfaktor und spezifischen inhaltsbezogenen Teilkompetenzen.

In Bezug auf die Mathematikdidaktik-Items zeigen sich auf den ersten Blick mit Ausnahme des Bereichs Funktionen geringere direkte Ladungen auf die Inhaltsgebiete. Hierfür lassen sich zwei Erklärungsansätze formulieren: Die geringeren Ladungen können zum einen auf eine geringere Inhaltsabhängigkeit der Itemlösung hinweisen. Zum anderen kann es sich aber auch um einen Hinweis auf eine geringere Reliabilität der über die Mathematikdidaktik-Items definierten Inhaltsgebiete handeln. Bei diesen Items werden jeweils die mathematischen Inhalte mit verschiedenen didaktischen Aufgaben kombiniert. Dies führt möglicherweise dazu, dass die Items der so definierten Inhaltsgebiete im Vergleich zu den über die reinen Mathematik-Items definierten Inhaltsgebieten heterogener ausfallen, was sich dann wiederum in einer geringeren Reliabilität – insbesondere bei den Inhaltsgebieten Arithmetik und Algebra – niederschlagen würde.

Entscheidender ist in diesem Teil des Modells aber – vor allem im Vergleich zu den Ergebnissen der oben dokumentierten Between- und Within-Modelle –, dass die Varianz der inhaltsbezogenen mathematikdidaktischen Teilkompetenzen dann zu einem sehr hohen Maße aufgeklärt wird, und zwar laden diese signifikant sowohl auf den mathematischen als auch auf den mathematikdidaktischen Generalfaktor. Dabei ist das Einflussgewicht des Ersteren deutlich höher als das des Letzteren. Diese Struktur verweist zum einen auf die Bedeutung eines eigenständigen mathematikdidaktischen Generalfaktors und zum anderen noch einmal auf die Bedeutung der inhaltlichen Ausdifferenzierung der mathematischen Kompetenz.

5.2 Mathematische, mathematikdidaktische und inhaltsspezifische Kompetenzen angehender Lehrkräfte

Im nächsten Schritt werden zunächst die Ergebnisse zur mathematischen Kompetenz angehender Lehrkräfte der Sekundarstufe I im internationalen Vergleich der sechs *MT21*-Stichproben berichtet. Diese stimmen für die beiden zweidimensionalen Between- und Within-Modelle naturgemäß überein (s. a. die entsprechende Korrelation der Parameterschätzungen in Tab. 7). Für das Second-Order-Modell fallen sie erwartungsgemäß

Tab. 6: Standardisierte Faktorladungen für das kombinierte Between- und Within-Modell mit den Generalfaktoren mathematische und mathematikdidaktische Kompetenz sowie spezifischen Faktoren für fünf Inhaltsgebiete ($n = 1.127$)

Mathematische Kompetenz 0,79*** (λ_{W11} in Abb. 4)					Mathematikdidaktische Kompetenz				
					$Mathematik_{SO}$ 0,63*** (λ_{W21} in Abb. 4)		$Mathematikdidaktik_{SO}$ 0,28*** (λ_{W22} in Abb. 4)		
ARI_{MA} (λ_{B11})	ALG_{MA} (λ_{B12})	FUN_{MA} (λ_{B13})	GEO_{MA} (λ_{B14})	STA_{MA} (λ_{B15})	ARI_{MD} (λ_{B21})	ALG_{MD} (λ_{B22})	FUN_{MD} (λ_{B23})	GEO_{MD} (λ_{B24})	STA_{MD} (λ_{B25})
0,75 (0,02)***	0,60 (0,02)***	0,71 (0,02)***	0,53 (0,03)***	0,36 (0,06)***	0,28 (0,02)***	0,19 (0,04)***	0,56 (0,02)***	0,37 (0,03)***	0,43 (0,02)***

***$p < 0{,}001$

ARI: Arithmetik, ALG: Algebra, FUN: Funktionen, GEO: Geometrie, STA: Statistik

Tab. 7: Korrelation der Parameterschätzungen für mathematische Kompetenz aus den Between- und Within-Modellen sowie dem Second-Order-Modell

	Mathematische Kompetenz_{B}	Mathematische Kompetenz_{W}
Mathematische Kompetenz_{W}	1.00***	
Mathematische Kompetenz_{SO}	0,98***	0,98***

***$p < 0{,}001$

Tab. 8: Korrelation der Parameterschätzungen für mathematikdidaktische Kompetenz aus den Between- und Within-Modellen sowie dem Second-Order-Modell

	Mathematikdidaktische Kompetenz_{B}	Mathematikdidaktische Kompetenz_{W}
Mathematikdidaktische Kompetenz_{W}	0,69***	
Mathematikdidaktische Kompetenz_{SO}	0,60***	0,89***

***$p < 0{,}001$

ebenfalls fast deckungsgleich aus. Diese Ähnlichkeit des Konstrukts „Mathematische Kompetenz" aus dem Second-Order-Modell mit den beiden anderen Modellen spiegelt sich auch in den hohen Korrelationen wider (s. Tab. 7).

Anschließend wird auf die Ergebnisse zur mathematikdidaktischen Kompetenz eingegangen, und zwar zunächst aus dem Between-Modell, das die unmittelbare Testperformanz der angehenden Lehrkräfte bei den Mathematikdidaktik-Items dokumentiert, bevor unter Kontrolle der mathematischen Kompetenz auf das spezifische mathematikdidaktische Profil der Lehrkräfte aus dem Within-Modell eingegangen wird. Die in Tab. 8 dokumentierte Korrelation der beiden Parameterschätzungen macht erwartungsgemäß deutlich, dass die beiden Konstrukte zwar signifikant positiv korrelieren, dass es sich aber um unterschiedliche Konstrukte handelt. Das mathematikdidaktische Ergebnis aus dem Second-Order-Modell spiegelt wie erwartet weitgehend das Within-Ergebnis wider. Als einziger Unterschied lässt sich ein weiter eingeschränkter Wertebereich erkennen, also weniger herausragende Stärken und Schwächen, wie es bereits für das Within- im Vergleich zum Between-Modell gilt, was auf die durch die Hinzunahme der Inhaltsbereiche nochmals verringerte Residualvarianz zu erklären ist.

Abschließend werden die inhaltsspezifischen Profile aus dem Second-Order-Modell berichtet, und zwar der Übersichtlichkeit halber zunächst im Detail bezogen auf die mathematische Kompetenz und dann zusammenfassend für die mathematikdidaktische Kompetenz.

Betont sei noch einmal, dass es sich bei den *MT21*-Stichproben nicht um repräsentative, sondern um kriteriengeleitet zusammengestellte Gruppen an Mathematiklehrkräften für die sechs Teilnahmeländer handelt. Dennoch sollte die sorgfältige Auswahl der Ausbildungsinstitutionen ein angemessenes Abbild sicherstellen.

5.2.1 Ergebnisse zur mathematischen Kompetenz im internationalen Vergleich

Die mathematische Kompetenz der Stichproben aus Südkorea und Taiwan liegt am Ende der Ausbildung deutlich über der Kompetenz der Stichproben aus den übrigen vier Län-

dern (s. Tab. 9). Drei leistungsschwächere Gruppen an Lehrkräften aus Bulgarien, USA und Mexiko liegen dagegen deutlich unter dem internationalen Mittelwert der 1.127 Lehrkräfte. Die Abstände sind dabei so deutlich, dass selbst eine Berücksichtigung der Stichprobenstruktur bei der Schätzung der Standardfehler vermutlich zu keiner anderen Schlussfolgerung führen würde. Die mathematische Kompetenz der deutschen Stichprobe liegt zwischen diesen beiden Gruppen, sie unterscheidet sich nicht vom internationalen Mittelwert. Die Unterschiede zwischen den sechs *MT21*-Teilnahmeregionen sind ausweislich einer Varianzanalyse statistisch hoch signifikant: $F = 192{,}89$; $p < 0{,}001$. Sie erklären 43 % der Varianz in den Testleistungen, was als großer Effekt eingeordnet werden kann.

Die Ergebnisse der Gruppen an der Spitze – die Teilnahmeregionen aus Südkorea und Taiwan – sowie am Ende der Leistungsskala – die Teilnahmeregionen aus Mexiko – zeichnen sich jeweils durch eine hohe Homogenität aus. In Bezug auf die Stichproben aus Deutschland und den USA fallen dagegen die hohe Standardabweichungen und vor allem die enormen Spannweiten zwischen dem schwächsten und dem besten Ergebnis auf, die fast an die Spannweite für die gesamte Stichprobe heranreichen.

Blickt man, um den Ursachen für diese breite Streuung näher zu kommen, auf die Merkmale jener angehenden Lehrkräfte, deren mathematische Kompetenz um mindestens eine Standardabweichung unter bzw. über den nationalen Mittelwerten liegt, zeigt sich ein je charakteristisches Profil: Angehende deutsche Mathematiklehrkräfte mit einem besonders schwachen Ergebnis gehören überwiegend einem kombinierten Primar- und Sekundarstufen-I-Ausbildungsgang an, der zu einem Lehramt in den Grund-, Haupt- und Realschulen führt, während jene mit einem sehr guten Ergebnis überwiegend einem kombinierten Sekundarstufen-I- und -II-Lehramt angehören und damit ein Gymnasial- bzw. Gesamtschullehramt anstreben. Entsprechende ausbildungsgangspezifische Differenzen lassen sich auch für die US-Gruppe finden, wo die Sekundarstufen-I-Ausbildung ebenfalls entweder kombiniert mit einer Primarstufen- oder einer Sekundarstufen-II-Ausbildung stattfindet.

5.2.2 Ergebnisse zur mathematikdidaktischen Kompetenz im internationalen Vergleich

Im Hinblick auf mathematikdidaktische Kompetenz werden basierend auf ihrer Testperformanz die stärksten Leistungen erneut von den südkoreanischen und taiwanesischen Stichproben erbracht, die auch signifikant über dem internationalen Mittelwert und den Leistungen aller übrigen *MT21*-Länder liegen (s. Tab. 10). Allerdings ist ihr Vorsprung geringer als im Falle der mathematischen Kompetenz. Die Lehrkräfte aus den deutschen Teilnahmeregionen liegen mit ihren Testleistungen zusammen mit der Stichprobe aus den USA um den internationalen Mittelwert. Die amerikanische Gruppe erbringt also relativ zu den übrigen Stichproben bessere Leistungen in Mathematikdidaktik als in Mathematik. Signifikant darunter liegen die angehenden Mathematiklehrkräfte aus den mexikanischen und – mit noch einmal deutlichem Abstand dahinter – aus den bulgarischen Teilnahmeregionen. Letztere Gruppe weist damit in Relation zu den anderen fünf *MT21*-Teilnahmeregionen sehr viel schlechtere mathematikdidaktische als mathematische Leistungen auf. Erneut sind die Unterschiede zwischen den sechs Stichproben statistisch hochsignifikant:

Tab. 9: Mathematische Kompetenz (θ_{B1} bzw. θ_{W1}) im internationalen Vergleich

Stichproben	M	S.E.	95 % CI	SD	min-max	Spann-weite	Abweichung
Südkorea	588	5,0	578–598	51	467–725	257	▲
Taiwan	577	3,6	570–584	58	381–725	343	▲
Deutschland	489	5,3	488–500	90	173–679	506	–
Bulgarien	454	9,4	435–472	94	173–650	477	▼
USA	446	4,2	438–455	67	231–736	505	▼
Mexiko	429	4,4	420–438	53	284–588	305	▼
Gesamtstichprobe	*500*	*2,8*	*494–506*	*93*	*173–736*	*563*	

M: Mittelwert, S.E.: Standardfehler des Mittelwertes, CI: Konfidenzintervall, SD: Standardabweichung, min – max: Minimum – Maximum, Abweichung: vom Mittelwert der sechs Stichproben

Tab. 10: Mathematikdidaktische Kompetenz im internationalen Vergleich (Between-Modell, d. h. basierend auf der Testperformanz; θ_{B2})

Stichprobe	M	S.E.	95 % CI	SD	min-max	Spann-weite	Abwei-chung
Südkorea	563	5,0	553–573	51	400–682	282	▲
Taiwan	557	3,6	550–564	59	346–685	339	▲
Deutschland	501	4,8	491–510	82	131–672	542	–
USA	487	4,4	478–495	66	297–686	389	–
Mexiko	430	4,8	420–439	59	276–587	311	▼
Bulgarien	415	10,2	395–435	102	131–643	512	▼
Gesamtstichprobe	*500*	*2,6*	*495–505*	*86*	*131–686*	*556*	

M: Mittelwert, S.E.: Standardfehler des Mittelwertes, CI: Konfidenzintervall, SD: Standardabweichung, min – max: Minimum – Maximum, Abweichung: vom Mittelwert der sechs Stichproben

$F = 134{,}27$; $p < 0{,}001$. Die Varianzaufklärung liegt mit 35 % etwas unter der für mathematische Kompetenz.

Für die Stichproben aus Südkorea, Taiwan und Mexiko lässt sich auch für die mathematikdidaktische Kompetenz eine hohe Homogenität der Ergebnisse feststellen. Die Streuung der deutschen Testleistungen liegt dagegen erneut im oberen Bereich, im Falle der mathematikdidaktischen Kompetenz ähnlich hoch wie die der Lehrkräfte aus den bulgarischen Teilnahmeregionen. Ausbildungsgangsspezifisch zeigt sich ein ähnliches Profil wie in Bezug auf mathematische Kompetenz: Die Gruppe an Referendarinnen und Referendaren mit besonders schwachen Ergebnissen wird dominiert von GHR-Lehrkräften, während jene mit einem sehr guten Ergebnis überwiegend ein Gymnasial- bzw. Gesamtschullehramt anstreben. Auffällig ist im Falle der mathematikdidaktischen Kompetenz die deutlich größere Homogenität der US-Ergebnisse.

Blickt man auf die relativen mathematikdidaktischen Stärken und Schwächen der angehenden Lehrkräfte, d. h. auf ihre Testleistungen unter Kontrolle mathematischer Kompetenz (s. Tab. 11), ergibt sich ein deutlich anderes Bild als zuvor. Die sich in der

Tab. 11: Mathematikdidaktische Stärken und Schwächen im internationalen Vergleich (Within-Modell, d. h. unter Kontrolle mathematischer Kompetenz; θ_{W2})

Stichprobe	M	S.E.	95 % CI	SD	min-max	Spann-weite	Abwei-chung
USA	529	4,2	520–537	62	349–675	326	▲
Deutschland	511	4,3	502–519	73	265–693	428	–
Taiwan	507	3,7	500–514	60	309–653	344	–
Südkorea	505	5,7	494–517	58	294–635	341	–
Mexiko	469	4,9	460–479	60	316–643	327	▼
Bulgarien	427	8,5	411–444	85	215–618	403	▼
Gesamtstichprobe	*500*	*2,1*	*496–504*	*72*	*215–693*	*477*	

M: Mittelwert, S.E.: Standardfehler des Mittelwertes, CI: Konfidenzintervall, SD: Standardabweichung, min – max: Minimum – Maximum, Abweichung: vom Mittelwert der sechs Stichproben

zuvor verwendeten Skalierung andeutenden relativ besseren mathematikdidaktischen Testleistungen der amerikanischen und mexikanischen Teilnahmeregionen treten in diesem Modell weit deutlicher hervor. Die US-Stichprobe weist Stärken in der Mathematikdidaktik auf, die signifikant über denen der übrigen *MT21*-Teilnahmeregionen liegen. Unter Kontrolle der mathematischen Kompetenz liegen die Ergebnisse der südkoreanischen und taiwanesischen Gruppen in Mathematikdidaktik nur noch um den internationalen Mittelwert und damit auf gleicher Höhe mit denen der deutschen Stichprobe. Das spezifisch mathematikdidaktische Profil der angehenden Lehrkräfte aus den mexikanischen Teilnahmeregionen liegt zwar signifikant unter diesen Ergebnissen, allerdings ist der Abstand deutlich geringer, als wenn keine Kontrolle der mathematischen Kompetenz erfolgt. Erneut sind die Unterschiede zwischen den sechs Stichproben statistisch hochsignifikant: $F = 43{,}87$; $p < 0{,}001$. Die Varianzaufklärung ist mit 15 % allerdings geringer als zuvor.

Die Bedeutsamkeit der Within-Modellierung mit ihrem Blick auf Mathematikdidaktik als spezifischer Teilkompetenz unter Kontrolle einer generellen mathematischen Lehrerkompetenz wird deutlich, wenn die Schwerpunktsetzungen der Lehrerausbildungen in den einzelnen Ländern betrachtet werden.

In den USA liegt der mathematikdidaktische Ausbildungsanteil für die Mehrheit angehender Lehrkräfte, und zwar für diejenigen, die einen Bachelor of Education erwerben, über dem mathematischen Anteil, der außerordentlich gering ist (Schmidt et al., im Druck). Zudem werden die fachbezogenen Studienanteile fast ausschließlich von Mathematikdidaktikern an den *Schools of Education* in einer integrierten Form gestaltet. Das heißt, auch hier bestehen umfangreiche Lerngelegenheiten, die unmittelbar auf schulische Anforderungen ausgerichtet sind. Rein fachwissenschaftliche Lehrveranstaltungen an den Fakultäten für Mathematik sind dagegen selten und gelten in der Regel nur für einen Teil der kleinen Gruppe an Sekundarstufenlehrkräften, die zunächst einen polyvalenten Bachelor in Mathematik erworben haben und sich dann entscheiden, in den Lehrerberuf zu gehen.

In Mexiko ist Mathematik als eigenständiger Anteil überhaupt nicht in der Lehrerausbildung für die Sekundarstufe I vertreten, sondern dieser wird in die Mathematik-

didaktik-Veranstaltungen integriert. Der Mathematikdidaktik kommt ein Drittel der Ausbildungszeit zu. Dies ist der im Vergleich aller sechs *MT21*-Stichproben mit Abstand höchste Anteil und dürfte die relative Stärke dieser Gruppe erklären.

Im Unterschied dazu ist die bulgarische Mathematiklehrerausbildung stark fachwissenschaftlich ausgerichtet (vgl. Schmidt et al., im Druck). Angehende Lehrkräfte müssen allein 98 Semesterwochenstunden und damit knapp die Hälfte der vierjährigen Ausbildungszeit in Mathematik belegen. Die 36 Semesterwochenstunden umfangreiche Fachdidaktik-Ausbildung hat ebenfalls eine starke fachwissenschaftliche Prägung, indem deutlich mehr als die Hälfte das Erlernen der Schulmathematik vom höheren Standpunkt zum Gegenstand hat. Von den fachdidaktischen Pflichtveranstaltungen sind letztlich nur acht Semesterwochenstunden im engeren Sinne didaktisch ausgerichtet.

Aus diesen Gegensätzen – relativ starke fachdidaktische Prägung der Lehrerausbildung in Mexiko und den USA sowie relativ starke fachwissenschaftliche Prägung der Ausbildung in Bulgarien – ergibt sich auch die Erklärung für das Profil der *MT21*-Stichproben aus Deutschland, Südkorea und Taiwan. In allen drei Ländern hat die fachdidaktische Ausbildung in Relation zur fachwissenschaftlichen einen deutlich geringeren Stellenwert als in Mexiko und den USA, aber einen höheren Stellenwert als in Bulgarien. Speziell für Deutschland mit seinem in sehr unterschiedliche Ausbildungsgänge zerfallenden System muss an dieser Stelle allerdings dezidiert festgehalten werden, dass diese Aussage nur auf aggregiertem Niveau gilt. Blickt man eine Ebene tiefer wird deutlich, dass sich der substanzielle fachwissenschaftliche Anteil vor allem aus der Gymnasiallehrer- und der substanzielle Anteil an fachdidaktischer Ausbildung vor allem aus der GHR-Ausbildung speist.

5.2.3 Ergebnisse zu den inhaltsspezifischen Stärken und Schwächen im internationalen Vergleich

Unter einer inhaltsbezogenen Perspektive lassen sich mit den Ergebnissen aus der Second-Order-Skalierung länderspezifische Stärken und Schwächen in Arithmetik, Algebra, Funktionen, Geometrie und Statistik ausmachen (s. Tab. 12). Vorab sei festgehalten, dass in Statistik von den sechs Stichproben vergleichsweise homogene Leistungen erzielt werden, während sich die Ergebnisse im Gebiet Funktionen am stärksten unterscheiden.

Die angehenden Lehrkräfte aus den südkoreanischen Teilnahmeregionen verfügen über besondere Stärken in Arithmetik, Algebra und Geometrie, wo ihre Ergebnisse ausweislich des Standardfehlers noch einmal signifikant über denen der taiwanesischen Stichprobe und bis zu 1,5 Standardabweichungen über dem Mittelwert aller Lehrkräfte liegen. Diese großen Abstände sind bemerkenswert. Für die angehenden Lehrkräfte in der Stichprobe aus Taiwan lässt sich eine ähnliche Stärke im Vergleich zur gesamten *MT21*-Stichprobe in Funktionen feststellen, wenn der Abstand zu den Kolleginnen und Kollegen aus Südkorea auch nicht so groß ist, dass der Unterschied signifikant wird.

Die angehenden Mathematiklehrkräfte aus den deutschen Teilnahmeregionen zeigen leichte Stärken in Arithmetik, ohne dass sie damit allerdings signifikant über dem Mittelwert aller Lehrkräfte liegen. Signifikante Schwächen zeigen sie in Algebra. In diesem Inhaltsgebiet liegt die Leistung der deutschen Stichprobe doch deutlich unter dem Mittelwert aller Lehrkräfte. Bulgarische Lehrkräfte weisen eine relative Stärke im Gebiet

Tab. 12: Inhaltsbezogene mathematische Stärken und Schwächen angehender Lehrkräfte im internationalen Vergleich (Second-Order-Modell mit mathematischer und mathematikdidaktischer Kompetenz als Generalfaktoren)

Stichprobe	Arithmetik			Algebra			Funktionen			Geometrie			Statistik		
	M	S.E.	Abweichung	M	S.E.	Abweichung	M	S.E.	Abweichung	M	S.E.	Abweichung	M	S.E.	Abweichung
Südkorea	631	8,0	▲	648	9,0	▲	612	8,5	▲	632	8,0	▲	616	6,6	▲
Taiwan	604	5,5	▲	613	6,2	▲	620	5,8	▲	598	6,3	▲	599	4,8	▲
Deutschland	511	7,7	–	476	8,6	–	498	8,5	–	496	7,3	–	496	6,7	–
USA	434	7,7	▼	440	7,8	▼	416	7,2	▼	446	6,6	▼	451	5,8	▼
Bulgarien	419	13,8	▼	430	13,9	▼	457	14,8	▼	436	14,5	▼	419	13,8	▼
Mexiko	389	6,7	▼	416	7,7	▼	389	8,2	▼	413	6,5	▼	405	6,5	▼
Gesamtstichprobe	*502*	*4,1*		*506*	*4,3*		*504*	*4,4*		*506*	*3,9*		*503*	*3,7*	

M: Mittelwert, S.E.: Standardfehler des Mittelwertes, CI: Konfidenzintervall, SD: Standardabweichung, min – max: Minimum – Maximum, Abweichung: vom Mittelwert der sechs Stichproben

Funktionen auf, wo sie nur eine halbe Standardabweichung unter dem Mittelwert bleiben. Besonders deutlich bleiben sie hinter diesem in zwei Inhaltsgebieten zurück, und zwar in Arithmetik und Statistik. Ein nahezu spiegelbildliches Profil hierzu weisen die angehenden Lehrkräfte der amerikanischen Stichprobe auf, die in Statistik nur eine halbe Standardabweichung unter dem Mittelwert aller Lehrkräfte liegen, dafür aber in Funktionen eine besondere Schwäche zeigen.

Angehende Mathematiklehrkräfte aus Mexiko sind die Stichprobe, die in allen Inhaltsgebieten die größten Schwächen zeigen. In Arithmetik und Funktionen fallen diese besonders deutlich aus.

Auch diese Profile spiegeln kulturelle Schwerpunktsetzungen, und zwar in diesem Fall nicht nur der Lehrerausbildung, sondern auch der Schulsysteme.

Die befragten südkoreanischen Lehrkräfte haben im Laufe ihrer Ausbildung deutlich mehr fortgeschrittene Inhalte in Algebra, insgesamt deutlich mehr Inhalte in Geometrie und etwas mehr Inhalte in Arithmetik belegt als jene aus Taiwan (vgl. Schmidt et al., im Druck). Für Arithmetik lässt sich auch im Schulcurriculum und in den Schulbüchern ein deutlich stärkerer Akzent in Südkorea als in anderen Ländern feststellen (Schmidt et al. 1997). Umgekehrt haben die Befragten aus Taiwan während ihrer Lehrerausbildung mehr Inhalte im Bereich Funktionen als jene aus Südkorea belegt, wobei allerdings für beide Stichproben ein Deckeneffekt festgestellt werden muss. Über schulische Schwerpunktsetzungen in Taiwan liegen kaum Informationen vor, da das Land nicht an der TIMSS-Curriculumstudie teilgenommen hat.

Unter den umfangreichen verpflichtenden fachwissenschaftlichen Lerngelegenheiten der bulgarischen Lehrerausbildung finden sich nur vier SWS in Arithmetik und sechs SWS in Statistik, dafür aber 30 SWS in Funktionen (Schmidt et al., im Druck). Einen im internationalen Vergleich starken Fokus auf Funktionen bei gleichzeitiger Unterrepräsentation arithmetischer und vor allem statistischer Inhalte weisen auch das Schulcurriculum und die verwendeten Schulbücher in Bulgarien auf (Schmidt et al., im Druck; Schmidt et al. 1997).

In Bezug auf Deutschland ist festzuhalten, dass Arithmetik sowohl für angehende GHR- als auch für angehende Gymnasiallehrkräfte ein zentrales Inhaltsgebiet ist (Blömeke et al. 2008). Algebra ist dagegen nur für Letztere ein bedeutender Bereich (Tietze et al. 1997). Dieses Profil entspricht Schwerpunktsetzungen im deutschen Schulcurriculum und in den Schulbüchern (Schmidt et al. 1997).

Wahrscheinlichkeitsrechnung und Statistik ist in vielen westlichen Ländern in den letzten Jahren eine wachsende Bedeutung für das Schulcurriculum zugeschrieben worden. Die USA haben hier eine führende Rolle eingenommen (Wu u. Dianzhou 2006). Vor allem Funktionen sind hier dagegen ein Inhaltsgebiet, das deutlich weniger stark in Schulcurricula gefordert wird als in anderen Ländern (Schmidt et al. 1997). Entsprechend gehört „Probability and Data“ in der Mathematiklehrerausbildung zum Standard, während dies für Funktionen keinesfalls gilt (Schmidt et al., im Druck).

In Mexiko ist an den *Normal Schools* als Spezialuniversitäten für Lehrerausbildung keine eigenständige fachbezogene Lehrerausbildung im engeren Sinne zu finden (vgl. ebd.). Dies erklärt, warum die Stichprobe generell in keinem Inhaltsgebiet eine sichtbare Stärke aufweist. Erkennbare schulische Schwerpunktsetzungen vor allem in Geo-

metrie können sich daher vermutlich nur in dem geringen Umfang niederschlagen, wie ihn Tab. 12 ausweist.

In Bezug auf die mathematikdidaktischen Inhaltsgebiete werden die länderspezifischen Profile ebenfalls sichtbar, allerdings fallen sie aufgrund der eingeschränkten Wertebereiche weniger stark aus. Was auffällt, sind die durchgängig geringeren Abstände der US-Stichprobe zum Mittelwert aller Lehrkräfte. Zwar liegen sie in jedem mathematikdidaktischen Inhaltsgebiet signifikant unter diesem, die praktische Bedeutsamkeit ist mit meist nur einer Viertel Standardabweichung oder weniger aber sehr viel geringer als in Bezug auf die mathematischen Inhaltsgebiete. Zudem liegen die Ergebnisse in vier der fünf Inhaltsgebiete signifikant über denen der Lehrkräfte aus Bulgarien und Mexiko. Dies verweist auf die Bedeutsamkeit, zwischen mathematischen und mathematikdidaktischen Inhaltsgebieten zu unterscheiden.

6 Zusammenfassung, Diskussion und Folgerungen

Ziel des vorliegenden Beitrags war, zum einen erstmals auf Testdaten beruhende international-vergleichende Ergebnisse zu mathematischen, mathematikdidaktischen und inhaltsbezogenen Kompetenzen angehender Lehrkräfte darzulegen sowie zum anderen zu zeigen, wie sich unterschiedliche IRT-Skalierungen als Werkzeug zur Diagnose von Stärken und Schwächen in Teilkompetenzen nutzen lassen. Die konzeptionelle Überlappung von Mathematik und Mathematikdidaktik führt bei der ausschließlichen Dokumentation von Testperformanz dazu, dass spezifisch mathematikdidaktische Stärken und Schwächen leicht übersehen werden. Vergleichbares gilt für inhaltsbezogene Stärken und Schwächen zum Beispiel in Geometrie oder Funktionen.

In zwei Schritten ist in diesem Beitrag eine mehrdimensionale Modellierung von Lehrerkompetenzen gelungen, die das Verhältnis von mathematischen, mathematikdidaktischen und inhaltsbezogenen Kompetenzen ausweislich der Modellanpassungen und Faktorladungen präziser widerspiegelt als traditionelle Kompetenzmodelle. In Übereinstimmung mit unserer Ausgangshypothese H2 wurde deutlich, dass die Lösung der mathematischen Items sowohl von einer generellen mathematischen Kompetenz als auch von spezifischen inhaltsbezogenen Teilkompetenzen in Arithmetik, Algebra, Funktionen, Geometrie und Statistik beeinflusst ist. In Übereinstimmung mit Hypothese H1 zeigte sich, dass die Lösung der mathematikdidaktischen Items von einer generellen mathematikdidaktischen, darüber hinaus aber auch von der generellen mathematischen Kompetenz sowie von inhaltsspezifischen Teilkompetenzen abhängt.

Die Ergänzung der traditionellen mehrdimensionalen Kompetenzmodellierung als Between-Multidimensionalität durch die Modellierung einer Within-Multidimensionalität, die den mehrdimensionalen Charakter der Lehrerkompetenzen direkt auf der Ebene der Items bzw. im Second-Order-Modell auf der Ebene der Inhaltsgebiete aufnimmt, indem Mehrfachladungen zugelassen werden, machte die Herausarbeitung stichprobenspezifischer Profile möglich. Diese spiegeln Schwerpunktsetzungen der jeweiligen Lehrerausbildungs- und Schulsysteme wider, die in traditionellen Modellen überdeckt werden. Entsprechend unserer Ausgangshypothesen gilt dies sowohl für das Verhältnis

von Mathematik und Mathematikdidaktik (H3) als auch für die fünf Inhaltsgebiete Arithmetik, Algebra, Funktionen, Geometrie und Statistik (H4).

Basierend auf der Within-Skalierung wurde deutlich, dass die *MT21*-Stichprobe aus den USA über besondere mathematikdidaktische Stärken verfügt. Relativ zum Abschneiden der übrigen Stichproben gesehen, gilt dies auch für die Lehrkräfte aus den mexikanischen Teilnahmeregionen. Diese Ergebnisse bilden die Schwerpunktsetzungen der jeweiligen Lehrerausbildungen in der Mathematikdidaktik weit besser ab als Between-Skalierungen, die nur Einfach-Ladungen zulassen. Gleichzeitig relativierten sich die mathematikdidaktischen Ergebnisse der Stichproben aus Südkorea und Taiwan deutlich.

Vergleichbares lässt sich in Bezug auf inhaltsbezogene Profile feststellen. Die relativen Stärken der befragten Lehrkräfte aus den USA in Statistik, der deutschen Lehrkräfte in Arithmetik, der bulgarischen und taiwanesischen Befragten in Funktionen sowie der südkoreanischen Lehrkräfte in Arithmetik, Algebra und Geometrie spiegeln Schwerpunktsetzungen nicht nur der Lehrerausbildungen, sondern auch der jeweiligen Schulsysteme. Umgekehrt gilt dies für die Schwächen der befragten Deutschen in Algebra, der US-Lehrkräfte in Funktionen sowie der Bulgaren in Arithmetik und Statistik.

Die Ergebnisse des vorliegenden Beitrags sollten allerdings nicht dahingehend missverstanden werden, dass Within-Modellierungen Between-Modellierungen zukünftig ersetzen sollten. Sie stellen eher eine sinnvolle Ergänzung dar, indem sie der Frage nachgehen, was sichtbar wird, wenn die Leistungen im Detail betrachtet werden: „For researchers interested in the specific abilities contributing to the overall competence to solve specific test items, the within-item model can yield more interesting information than the between-item model, which is simply a descriptive measurement model. The within-item model is a more elaborated model of the interaction between the person and the test item. […] Performance in complex tasks is decomposed into more basic abilities, providing a more detailed picture of the competence assessed.“ (Hartig u. Höhler 2008, S. 93) Unter Testperformanzgesichtspunkten ist dagegen festzuhalten, dass es die *MT21*-Stichproben aus Südkorea und Taiwan sind, die ein besonders hohes Niveau auf dem Generalfaktor Mathematik aufweisen und dementsprechend mehr Items in allen Inhaltsgebieten und auch der Mathematikdidaktik richtig lösen.

Nicht eindeutig interpretiert werden konnten die durchgehenden Schwächen der mexikanischen Stichprobe in allen Inhaltsgebieten, obwohl Schwerpunktsetzungen in Geometrie und Statistik in Schule und Lehrerausbildung zu erkennen sind. Möglicherweise ist das generelle mathematische Kompetenzniveau zu gering, damit diese sichtbar werden können. Hier stellen sich weitere Forschungsaufgaben, denen im Rahmen der internationalen IEA-Studie zur Mathematiklehrerausbildung TEDS-M nachgegangen werden kann, an der eine Reihe an Ländern teilnimmt, die keine fachliche Ausbildung in Mathematik vorsehen, sodass ein relativ schwaches Abschneiden zu erwarten ist.

Etwas aus dem Rahmen fällt auch das Mathematik-Ergebnis für Bulgarien. Die Schwäche dieser Stichprobe im Bereich Mathematikdidaktik sowie das inhaltsbezogene Profil lassen sich angesichts der konkreten Lehrangebote sowie der Schwerpunktsetzungen in Schule und Lehrerausbildung gut nachvollziehen. Angesichts der hohen Gesamtzahl an fachbezogenen Veranstaltungen hätte man aber eine höhere generelle mathematische Kompetenz erwartet. Um die Ursachen hierfür zu ergründen, muss vermutlich auf Merkmale außerhalb der Ausbildung geblickt werden. Bulgarien weist wie viele osteuropäische

Länder eine starke mathematisch-naturwissenschaftliche Tradition auf. Der Lehrerberuf war zwar nicht gut bezahlt, aber hoch angesehen und bot zudem lebenslange Sicherheit. Die ökonomischen Krisen seit Ende des Kalten Krieges haben hier zu Erosionen geführt. So ist die Besoldung noch einmal deutlich zurückgegangen und das Ansehen schulischer Bildung hat stark nachgelassen. Viele Schulen mit mathematisch-naturwissenschaftlichem Profil wurden aufgelöst. Die Auswirkungen hiervon waren bereits im ständigen Absinken der bulgarischen Ergebnisse in den TIMS- und PISA-Studien der letzten fünfzehn Jahre festzustellen. *MT21* deutet darauf hin, dass dieser Trend nun möglicherweise auch in der Lehrausbildung angekommen ist. An TEDS-M nehmen weitere osteuropäische Länder teil, sodass es interessant sein wird zu untersuchen, inwieweit sich entsprechende Ergebnisse auch für diese finden lassen.

Die deutsche Stichprobe erreicht unabhängig von der Form der Skalierung fast immer den internationalen Mittelwert. Dies entspricht den im Vergleich zu den übrigen Stichproben im mittleren Umfang gebotenen Lerngelegenheiten. Wie zuvor bereits angesprochen, kommt dieses Mittel in Deutschland allerdings nur zustande, wenn man Besonderheiten von Ausbildungsgängen unberücksichtigt lässt. Die Ausdifferenzierung unserer Lehrerausbildung in Ausbildungsgänge für Grund-, Haupt- und Realschulen einerseits sowie Gymnasial- und Gesamtschullehrkräfte andererseits führt zu einem entweder stark fachwissenschaftlichen oder einem stark fachdidaktischen Akzent. Zudem spielen algebraische und funktionenbezogene Inhalte in letzterer Ausbildung eine deutlich stärkere Rolle als in Ersterer. Hier stellt sich mit den größeren Stichproben der TEDS-M-Studie dringend die Aufgabe zu untersuchen, ob sich ausbildungsgangspezifische Unterschiede finden lassen.

Generell stellt sich für kommende Studien – und TEDS-M stellt hier angesichts des Aufschwungs der empirischen Lehrerforschung nur die erste Möglichkeit dar – die Aufgabe, Fragen nachzugehen, die angesichts des Charakters der *MT21*-Stichprobe und des *MT21*-Tests offen gelassen werden mussten. So muss mit größeren Fallzahlen geprüft werden, inwieweit die vorgestellten Modelle über die Länder hinweg invariant sind. Mit repräsentativen Stichproben sind die deskriptiven Ergebnisse zu replizieren, bevor hieraus Schlussfolgerungen gezogen werden können.

In konzeptioneller Hinsicht stellt sich die Aufgabe, einen Mathematikdidaktik-Test zu entwickeln, der Aufgaben enthält, die ohne mathematische Kompetenz gelöst werden können. Beispiele dafür wurden eingangs aufgelistet. Erst dann wird die Aufstellung und Überprüfung eines multiplikativen, nicht-kompensatorischen Modells möglich, das intuitiv für das Verhältnis von mathematischer und mathematikdidaktischer Kompetenz plausibler ist als ein additiv-kompensatorisches Modell. Ein solcher Test sollte dann auch ausgewogener die Inhaltsgebiete berücksichtigen und diese breiter abdecken. Im *MT21*-Test ist vor allem die Statistik mit zu wenigen Items vertreten, und das Gebiet der Funktionen ist zu eng definiert.

Abschließend sei festgehalten, dass die vorliegenden Ergebnisse nicht nur für den Bereich der Modellierung von Lehrerkompetenzen darauf hinweisen, dass stärker unterschiedliche Wege in der Skalierung gegangen werden sollten, auch wenn die Ergebnisse dann schwerer zu vermitteln sind. Da sie aber unterschiedliche Schlussfolgerungen zulassen, können sie vor zu schnellen Schlüssen schützen.

Anmerkungen

1 *MT21* wurde von der *National Science Foundation* (REC-0231886) und der Alexander-von-Humboldt-Stiftung gefördert. Die hier geäußerten Thesen und Interpretationen sind die der Autorinnen und repräsentieren nicht die Meinungen der Stiftungen. Die Autorinnen danken Jan-Eric Gustafsson, Johannes Hartig, Richard T. Houang, Rainer Lehmann, Gabriele Kaiser und William H. Schmidt für wertvolle Anmerkungen und Hinweise zu früheren Versionen dieses Beitrags. Verbleibende Missverständnisse und Irrtümer gehen selbstverständlich ausschließlich zulasten der Autorinnen.

2 Die Datenerhebung für *MT21* fand 2006 statt. Seither ist im Zuge des sogenannten „Bologna-Prozesses" vor allem die Lehrerausbildung in Deutschland in einigen Bundesländern Veränderungen unterzogen worden. Diese sind in der Stichprobe nicht abgebildet, da die deutschen Lehrkräfte ihre Ausbildung noch vollständig im grundständigen System erhalten hatten.

3 Für den vorliegenden Beitrag konnte auf den vollen Itempool des *MT21*-Tests zurückgegriffen werden. Für die Analysen des nationalen Datensatzes (Blömeke et al. 2008) hatten aufgrund der geringeren Stichprobengröße acht Items ausgeschlossen werden müssen.

4 Wir danken den anonymen Gutachterinnen und Gutachtern, die auf dieses Problem aufmerksam gemacht und uns insofern zu einem Neuansatz bewogen haben.

Literatur

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2009). *Statistics for business and economics* (10. Aufl.). Boston: South-Western College.

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

Blömeke, S., Kaiser, G., & Lehmann, R. (Hrsg.). (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*. Münster: Waxmann.

Blömeke, S., Suhl, U., Kaiser, G., Felbrich, A., Schmotz, C., & Lehmann, R. (2010). Lerngelegenheiten und Kompetenzerwerb angehender Mathematiklehrkräfte im internationalen Vergleich. *Unterrichtswissenschaft, 38*(1), 29–50.

Bromme, R. (1992). *Der Lehrer als Experte: zur Psychologie des professionellen Wissens*. Bern: Huber.

Brunner, M., Kunter, M., Krauss, S., Baumert, J., Blum, W., Dubberke, T. et al. (2006). Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildung? *Zeitschrift für Erziehungswissenschaft, 9,* 521–544.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

DMV, GDM, & MNU (2008). *Standards für die Lehrerbildung im Fach Mathematik. Empfehlungen von DMV, GDM, MNU.* http://www.math.uni-sb.de/ag/lambert/LAHLAR/Standards-LehrerbildungMathematik.pdf. Zugegriffen: 13. Aug. 2009.

Eurydice (2004). *Der Lehrerberuf in Europa: Profil, Tendenzen und Anliegen, Bericht IV: Die Attraktivität des Lehrerberufs im 21. Jahrhundert, Allgemein bildender Sekundarbereich I.* Brüssel: Eurydice.

Gabler, S., Hoffmeyer-Zlotnik, J. H. P., & Krebs, D. (Hrsg.). (1994). *Gewichtung in der Umfragepraxis*. Opladen: Westdeutscher Verlag.

Graeber, A., & Tirosh, D. (2008). Pedagogical content knowledge: Useful concept or elusive notion. In P. Sullivan & T. Woods (Hrsg.), *Knowledge and beliefs in mathematics teaching and teaching development. The international handbook of mathematics teacher education* (Vol. 1, S. 117–132). Rotterdam: Sense Publisher.

Gustafsson, J. E., & Snow, R. E. (1997). Ability profiles. In R. F. Dillon (Hrsg.), *Handbook on testing* (S. 107–135). Westport: Greenwood Press.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez, E. J., & Orpwood, G. (1997). *Performance assessment in IEA's third international mathematics and science study (TIMSS)*. Chestnut Hill: TIMSS International Study Center, Boston College.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie, 216*(2), 89–101.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41–54.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

KMK 2004 = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss (Jahrgangsstufe 10)*. München: Wolters Kluwer.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in research on competence modeling and assessment. *Zeitschrift für Psychologie, 216*(2), 60–72.

Krauthausen, G., & Scherer, P. (2007). *Einführung in die Mathematikdidaktik*. München: Elsevier.

Küchemann, D., & Hoyles, C. (2002). *Technical report for the longitudinal proof project: Year 8 survey 2000*. London: University of London, Institute of Education.

Leutner, D., Klieme, E., Meyer, K., & Wirth, J. (2004). Problemlösen. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 147–175). Münster: Waxmann.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99–114.

Mulaik, S. A., & Quartetti, D. A. (2000). First or higher order general factors? *Structural Equation Modeling, 4,* 193–211.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report. Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Muthén, B., & Muthén, L. (2008). MPlus Version 5.21. Base Program and Combination Add-On (32-bit). Software.

NCTM 2000 = National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics*. Reston: NCTM.

Nold, G., & Rossa, H. (2008). Sprechen Englisch. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff et al. (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 170–179). Weinheim: Beltz.

Nold, G., Rossa, H., & Chatzivassiliadou, K. (2008). Leseverstehen Englisch. In E. Klieme, W. Eichler, A. Helmke, R. H. Lehmann, G. Nold, H.-G. Rolff et al. (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 130–138). Weinheim: Beltz.

OECD (2004). *Education at a glance. OECD indicators 2004*. Paris: OECD.

OECD (2007). *PISA 2006. Science competencies for tomorrow's world* (2 Vols.). Paris: OECD.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (Hrsg.). (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs.* Münster: Waxmann.

Reckase, M., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Mesaurement, 14,* 361–373.

Rosing, M. J., & Ross, K. N. (1992). Sampling and administration. In J. P. Keeves (Hrsg.), *The IEA technical handbook* (S. 51–90). The Hague: IEA.

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

Särndal, C.-E., Swensson, B., & Wretman, J. (1997). *Model assisted survey sampling.* New York: Springer.

Schmidt, W. H., McKnight, C. C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics*. Dordrecht: Kluwer.

Schmidt, W. H., Blömeke, S., & Tatto, M. T. (im Druck). *Teacher preparation from an international perspective.* New York: Teacher College Press.

Shulman, L. S. (1985). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Wittrock (Hrsg.), *Handbook of research on teaching* (3. Aufl., S. 3–36). New York: Macmillan.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*(4), 313–324.

Tietze, U.-P., Klika, M., & Wolpers, H. (1997). *Mathematikunterricht in der Sekundarstufe II* (Bd. 1). Braunschweig: Vieweg.

Torre, J. de la, & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*(8), 620–639.

UN 2008 = United Nations (2008). Human development index. http://hdr.undp.org/en/statistics/. Zugegriffen: 18. Juni 2009.

Vollrath, H.-J. (2001). *Grundlagen des Mathematikunterrichts in der Sekundarstufe*. Heidelberg: Spektrum.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Mesaurement, 40*(3), 255–275.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149.

Weinert, F. E. (1999). *Konzepte der Kompetenz. Gutachten zum OECD-Projekt „Definition and Selection of Competencies: Theoretical and Conceptual Foundations (DeSeCo)“*. Neuchatel, Schweiz: Bundesamt für Statistik.

Wu, M., & Dianzhou, Z. (2006). An overview of the mathematics curricula in the west and east – discussions on the findings of the Chongquing paper. In F. Leung., K. D. Graf, & F. Lopaz-Rea (Hrsg.), *Mathematics education in different cultural traditions – a comprehensive study of east Asia and the West* (S. 181–193). New York: Springer.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order model and the hierarchical factor model. *Psychometrika, 64,* 113–128.