[Eingereichte Fassung; Zitationen erfolgen bitte nach dem Original:
Blömeke, S., Seeber, S., Kaiser, G., Lehmann, R., Schwarz, B., Felbrich, A. & Müller, Ch. (2009). Differentielle Item-Analysen zur Entwicklung professioneller Kompetenz angehender Lehrkräfte während der Lehrerausbildung. In R. Mulder, O. Zlatkin-Troitschanskaia, K. Beck, N. Reinhold & D. Sembill (Hrsg.), Professionalität von Lehrenden – Zum Stand der Forschung. Weinheim: Beck, S. 311-332.]

Blömeke, Sigrid, Dr. phil. habil., Univ.-Professorin, Humboldt-Universität zu Berlin; z.Zt. Michigan State University, East Lansing, MI USA, bloemeke@msu.edu.

Gabriele Kaiser, Dr. phil. habil., Univ.-Professorin, Universität Hamburg, gabriele.kaiser@uni-hamburg.de.

Rainer Lehmann, Dres. Dr. hc., Univ-.Professor, Humboldt-Universität zu Berlin, lehmannr@cms.hu-berlin.de.

Susan Seeber, Dr. phil. habil., Wissenschaftliche Mitarbeiterin, Deutsches Institut für Internationale Pädagogische Forschung Berlin, seeber@bbf.dipf.de.

Björn Schwarz, Wissenschaftlicher Mitarbeiter, Universität Hamburg, schwarz@erzwiss.uni-hamburg.de.

Anja Felbrich, Dr. phil., Wissenschaftliche Mitarbeiterin, Humboldt-Universität zu Berlin, anja.felbrich@staff.hu-berlin.de.

Christiane Müller, Dipl.-Psych., Wissenschaftliche Mitarbeiterin, Humboldt-Universität zu Berlin, christiane.mueller@staff.hu-berlin.de.

Abstract

Das Kohortendesign der Studie *MT21* macht es möglich, die Entwicklung des fachbezogenen Wissens angehender Mathematiklehrer/innen vom Beginn des Studiums bis in das Referendariat hinein zu analysieren. Im Beitrag werden kohortenspezifische differentielle Itemfunktionen untersucht und daraus Schlussfolgerungen für die Gestaltung der Lehrerausbildung gezogen. Anfänger/innen fallen Anforderungen besonders schwer, mit denen sie aus ihrer Schulzeit wenig vertraut sind. Besonders leicht fallen ihnen dagegen schulmathematische Inhalte, die in der Universität nicht mehr aufgegriffen werden, was bei fortgeschrittenen Lehrkräften zu Vergessensprozessen führen kann. Die Ergebnisse können als Beleg für Kleins These der „doppelten Diskontinuität“ gelesen werden.

Sigrid Blömeke, Susan Seeber, Gabriele Kaiser, Björn Schwarz, Rainer Lehmann, Anja Felbrich und Christiane Müller

Differentielle Item-Analysen zur Entwicklung professioneller Kompetenz angehender Lehrkräfte während der Lehrerausbildung

Dank internationaler Vergleichsstudien wie PIRLS, TIMSS und PISA hat der Erkenntnisstand in der Schul- und Unterrichtsforschung in Deutschland in den letzten zehn Jahren große Fortschritte gemacht. Die Lehrerausbildung ist in diesem Zusammenhang allerdings lange ein »blinder Fleck« geblieben (Blömeke 2004). »Mathematics Teaching in the 21st Century (*MT21*)« ist eine der ersten Studien, die die professionelle Kompetenz angehender Lehrerinnen und Lehrer mittels standardisierter Test erfasst.[1] Dies erfolgt zudem anhand einer größeren Fallzahl und im internationalen Vergleich der Länder Bulgarien, Deutschland, Mexiko, Südkorea, Taiwan und USA.

Ziel von *MT21* ist die empirische Erfassung von Effekten der Lehrerausbildung am Beispiel von drei Gruppen angehender Mathematiklehrer/innen der Sekundarstufe I: Anfänger/innen zu Beginn ihres Universitätsstudiums, Studierende am Ende ihres Universitätsstudiums sowie Referendar/innen. Mit Mathematiklehrer/innen nimmt *MT21* eine Personengruppe in den Blick, der für die Vorbereitung der nachwachsenden Generationen auf die Informationsgesellschaft eine zentrale Rolle zukommt (Blum et al. 2004; Bulmahn/Wolff/Klieme 2003; KMK 2003).

1 Ziel des vorliegenden Beitrags

Die *MT21*-Daten machen es möglich, in einem quasi-längsschnittlichen Vergleich die Entwicklung des fachbezogenen Wissens angehender Mathematiklehrer/innen vom Beginn des Studiums bis in das Referendariat hinein zu analysieren. Die Analysen zeigen, dass am Ende der Lehrerausbildung substanziell mehr Wissen vorhanden ist als zu Beginn (Blömeke et al., 2008a). Trotz aller designbedingten Einschränkungen der Aussagekraft dieses Ergebnisses kann es als wichtiger Indikator für die Wirksamkeit der Lehrerausbildung angesehen werden. Die in den untersuchten Dimensionen erreichten Effektstärken erlangen zum Teil beachtliche Ausmaße und deuten eine hohe praktische Relevanz dieser Leistungsunterschiede an.

[1] Eine Förderung erfolgt durch die National Science Foundation, die Alexander von Humboldt-Stiftung, die Humboldt-Universität zu Berlin und die Michigan State University. Die Teilnahme war für Studierende, Referendar/innen, Hochschullehrer/innen und Seminarleiter/innen freiwillig. Ihnen allen möchten wir herzlich für die große Offenheit der Studie gegenüber danken.

Der durchschnittliche Leistungsvorsprung der angehenden Lehrkräfte am Ende der Ausbildung bedeutet allerdings nicht, dass ihnen *jedes* Item leichter fällt als Anfänger/innen oder dass allen Kohorten dieselben Items besonders leicht oder schwer fallen. Im Gegenteil ist festzustellen, dass Anfänger/innen einzelne Items sogar leichter fallen als angehenden Lehrkräften am Ende des Studiums bzw. im Referendariat. Andere Items bereiten Anfänger/innen dagegen überproportional große Lösungsschwierigkeiten.

Diese kohortenspezifischen Unterschiede in der Itemschwierigkeit sind Ausdruck davon, dass ein theoretisch entwickeltes Testmodell keine perfekte empirische Anpassung aufweisen kann. Auch wenn sich die Modellierungen des fachbezogenen Wissens angehender Mathematiklehrkräfte in *MT21* als sehr gut passend erwiesen haben (Blömeke et al., 2008b), zeigen sich Abweichungen in den Daten. Ursache ist, dass mit dem Test angestrebt wurde, die zu erfassenden Konstrukte möglichst breit abzudecken. Damit ist vorgezeichnet, dass – über zufällige Schätzfehler hinaus – Teile des Tests ggf. nicht von allen Personen gleich häufig gelöst werden, da unterschiedliche Fähigkeiten zum Tragen kommen. Finden sich systematische Abweichungen für ganze Item-Sets bei Gruppen an Personen spricht man von „differentiellen Itemfunktionen“ (Budgell/Namburty/Douglas 1995).

Erkennbar werden solche Disparitäten nicht schon an unterschiedlichen Gruppenmittelwerten, sondern an Verschiebungen der Differenzen von Itemschwierigkeiten zwischen den Gruppen. Anders ausgedrückt: Differentielle Itemfunktionen liegen vor, wenn für ein Item die bedingte Lösungswahrscheinlichkeit, bezogen auf die umfassende, auf allen Items beruhende Fähigkeitsschätzung, zwischen den Gruppen variiert (Baumert/Bos/Lehmann 2000, S. 178ff.). Bezogen auf die vorliegende Studie heißt dies, dass nicht einfach vorausgesetzt werden darf, dass die festgestellten Unterschiede in der professionellen Kompetenz während der Ausbildung gleichförmig stattgefunden hätten. Zwischen Ausbildungsbeginn und -ende können differentielle Lern- oder Vergessensprozesse stattgefunden haben.

Ziel des vorliegenden Beitrags ist es vor diesem Hintergrund zu untersuchen, wodurch sich Items auszeichnen, die Anfänger/innen im Vergleich zu den beiden anderen Gruppen besonders leicht oder besonders schwer fallen, die also kohortenspezifische differentielle Itemfunktionen aufweisen. Auf diese Weise können die fachbezogenen Stärken und Schwächen der drei Kohorten im Detail herausgearbeitet werden, um daraus Schlüsse für die Gestaltung der Lehrerausbildung zu ziehen.

Vermutungen zu den Ursachen von besonderen Stärken und Schwächern von Anfänger/innen bzw. Studierenden am Ende der Universität lassen sich insbesondere aus drei Zugängen ableiten:

1) unter Ausbildungsgesichtspunkten aus besonderen Schwerpunktsetzungen der ersten und zweiten Phase der Lehrerausbildung;
2) unter biographischer Perspektive aus der zeitlichen Nähe bzw. Ferne zur Schulmathematik;
3) unter kohortenspezifischen Gesichtspunkten aus der Entwicklung der bildungspolitischen Debatten zu Schule und Lehrerausbildung „nach PISA“.

Zu 1) Im Zuge der Ausbildung wird das fachbezogenen Wissen gegenüber der Schulmathematik nicht nur auf einem anderen Schwierigkeitsniveau vermittelt, es werden auch zahlreiche neue Themen und mathematische Arbeitsformen angesprochen, mit denen Anfänger/innen nicht vertraut sein können. Während dies vor allem für die erste Phase gilt, macht die zweite Phase mit ungewöhnlichen Lösungsansätzen vertraut. Schüler/innen produzieren in den seltensten Fällen mustergültige Antworten bzw. Aufgabenlösungen. Referendar/innen lernen daher sukzessive, halbrichtige bzw. -falsche Lösungen zu identifizieren und damit umzugehen.

Zu 2) Gleichzeitig wird das in der eigenen Schulzeit erworbene Mathematikwissen während der gesamten Lehrerausbildung als bekannt vorausgesetzt und nicht aktualisiert. Mit zunehmender Ferne zur Schulzeit kann es hier also zu Vergessensprozessen kommen.

Zu 3) Wenn sich Ausbildungsänderungen in der Regel auch nur langfristig durchsetzen, ist doch festzustellen, dass „nach PISA“ eine neue Dynamik in die bildungspolitischen Diskussionen gekommen ist, die zu Akzentverschiebungen in Schule und Lehrerausbildung geführt hat. Beispielsweise wird im schulischen Mathematikunterricht seit einigen Jahren verstärkt Wert auf graphische Interpretationen und Modellierungen gelegt, was insofern zwar von Anfänger/innen, aber weniger von fortgeschrittenen Studierenden und Referendar/innen erlebt worden ist.

2 Theoretischer Rahmen

Den Kern des theoretischen Rahmens von *MT21* bildet eine Konzeptualisierung der professionellen Kompetenz, mit der Lehrerinnen und Lehrer berufliche Anforderungen erfolgreich bewältigen. Im Anschluss an Weinert (1999) wird diese Kompetenz differenziert in die

- bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um berufliche Probleme lösen zu können (Professionswissen),
- die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können (professionelle Überzeugungen).

Professionelle Kompetenz stellt damit ein komplexes multidimensionales Konstrukt dar. In Tabelle 1 ist dokumentiert, welche Situationen und Anforderungen für Mathematiklehrer/innen in den sechs *MT21*-Ländern identifiziert werden konnten (Blömeke, Felbrich & Müller, 2008). »Unterrichten« sowie »Beurteilen und Beraten« stellen international akzeptierte Kernaufgaben von Lehrpersonen dar, die gleichzeitig testbar sind. Auf sie sind daher die Tests in *MT21* ausgerichtet. »Erziehen « und »professionelle Ethik« sind unterschiedlich normativ besetzt und damit nur schwer testbar. »Schulentwicklung« ist nicht in allen Ländern eine Kernanforderung an Lehrer/innen.

Berufliche Aufgaben	Situationen
A: Unterrichten	1. Auswahl/Einordnung von Unterrichtsthemen
	2. Unterrichtsplanung
B: Beurteilen und Beraten	1. Diagnose von Schülerleistungen
	2. Leistungsbeurteilung
	3. Beratung von Schüler/innen 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

Tab. 1: Definition beruflicher Anforderungen von Mathematiklehrerinnen und -lehrern in MT21

Insofern ist darauf hinzuweisen, dass professionelle Kompetenz über die in *MT21* erfassten Dimensionen hinausgeht. Erziehen, Beraten und Schulentwicklung sowie ein professionelles Ethos sind wichtige Anforderungen an Lehrer/innen 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 Funktionen implizieren, die von *MT21* nicht intendiert ist. Gleichzeitig kann aber festgehalten werden, dass das „Kerngeschäft“ (Tenorth 2006) von Lehrpersonen damit angemessen erfasst ist.

In Ergänzung zu diesem kompetenzorientierten Zugang zur Testung der angehenden Mathematiklehrer/innen wurde ein analytischer Zugang verfolgt, um die Wissens- und Überzeugungsdimensionen strukturell ausdifferenzieren zu können. Im Hinblick auf Professionswissen wird zwischen

- mathematischem,
- mathematikdidaktischem und
- pädagogischem Wissen unterschieden.

Das mathematische Professionswissen repräsentiert das disziplinär-systematische Wissen, das im Unterricht zur Anwendung kommt. Dieses wurde in *MT21* weiter in fünf Inhaltsgebiete analytisch ausdifferenziert. Mit Arithmetik, Algebra, Funktionen und Geometrie erfolgte eine Aufnahme von Gebieten, die zum Standardrepertoire des Mathematikunterrichts in der Sekundarstufe I gehören. Statistik stellt dagegen ein Gebiet dar, dem aufgrund seiner hohen Anwendungsrelevanz in Alltag und Wissenschaft zunehmendes Gewicht eingeräumt wird (NCTM 2000; KMK 2003).

Das mathematikdidaktische Wissen wurde in *MT21* ebenfalls noch einmal ausdifferenziert, und zwar in lehr- und in lernprozessbezogene Anforderungen:

- Lehrbezogene Anforderungen curricularer und unterrichtsplanerischer Art stehen bereits vor Beginn des Unterrichts fest. Planerisch gesehen müssen fachliche Inhalte für die Schülerinnen und Schüler ausgewählt, begründet, angemessen vereinfacht und unter Gebrauch verschiedener Repräsentationen aufbereitet werden (Krauthausen/Scherer 2007; Vollrath 2001). Curricular gesehen handelt es sich um die Frage des Aufbaus mathematischer Kompetenz über die Schuljahre hinweg. Was würde es zum Beispiel für spätere Unterrichtseinheiten bedeuten, wenn man einen klassischen Themenbereich der Schulmathematik in der Sekundarstufe I aus dem Curriculum entfernen würde?
- Lernprozessbezogene Anforderungen während des Unterrichts betreffen das unterrichtliche Handeln von Lehrperson in der unmittelbaren Interaktion mit Schülerinnen und Schülern. Hier geht es darum, deren Antworten – seien sie verbaler oder schriftlicher Art als Reaktion auf Aufgaben oder Fragen – bezüglich kognitiver Niveaus, Komplexität der Struktur sowie eventueller Fehler und Fehlermuster einzuordnen, Rückmeldungen zu geben und angemessen mit Interventionsstrategien darauf zu reagieren (ebd.). Neben der kognitiven Zielorientierung gilt es, Schüler/innen zu motivieren bzw. ihre Motivation aufrecht zu erhalten.

Mit diesem analytischen Zugang greifen wir einerseits die traditionelle Struktur der Lehrerausbildung auf, die die Komponenten Fachwissenschaft, Fachdidaktik bzw. Fachseminar und Erziehungswissenschaft bzw. Hauptseminar umfasst, sodass Lerngelegenheiten in der Lehrerausbildung differenziert mit Lernergebnissen verknüpft werden können. Andererseits berücksichtigen wir die im internationalen Diskurs prominente Ausdifferenzierung des Lehrerwissens in *content knowledge*, *pedagogical content knowledge* und *pedagogical knowledge* (Shulman 1985).

3 Untersuchungsdesign

3.1 Stichprobenziehung

Die deutsche Zielpopulation für *MT21* wurde wie folgt definiert: »Studierende sowie Referendarinnen und Referendare vor dem Zweiten Staatsexamen, die sich im Erhebungszeitraum in einem Ausbildungsgang befinden, mit dem sie eine Lehrbefähigung für das Unterrichtsfach Mathematik in der Sekundarstufe I erwerben«. Aufgrund der Stratifizierung des deutschen Schulsystems und seiner föderalistischen Organisation verbergen sich hinter dieser Definition unterschiedliche Ausbildungsgänge. In *MT21* wurden diese so zusammengefasst, dass zwei Arten an Aussagen möglich wurden:

- zum einen für die Gesamtgruppe an angehenden Sekundarstufen-I-Lehrpersonen. Diese Perspektive betont den einheitlichen Bildungsanspruch der Sekun-

darstufe I, der mit der Vergabe des an allen Schulformen erreichbaren mittleren Schulabschlusses verknüpft ist.

- zum anderen für die beiden großen Gruppen zukünftiger Mathematiklehrer/innen mit einer Lehrbefähigung bis zur Klasse 10 (Grund-, Haupt- und Realschullehrer/innen, GHR) bzw. zukünftiger Mathematiklehrer/innen mit einer Lehrbefähigung für die Klassen 5 bis 13 (Lehrer/innen für das Gymnasium und die Gesamtschule, GyGS). Diese beiden Gruppen durchlaufen unterschiedliche Ausbildungen; mit einer differenzierten Betrachtung wird zudem Rücksicht auf die Stratifizierung des deutschen Schulsystems genommen.

Zur Sicherstellung einer angemessenen Stichprobenqualität wurde in *MT21* eine mehrschrittige kriteriengeleitete Stichprobenziehung durchgeführt. Um die Ausbildung angesichts der Varianz in den Bundesländern auf der Aggregatebene noch adäquat beschreiben zu können, kamen in einem ersten Schritt nur Länder in die Auswahl, in denen die erste Phase der Lehrerausbildung an Universitäten und die zweite Phase an Studienseminaren stattfindet. In einem zweiten Schritt wurden auf der Basis einer Befragung aller Studienabsolvent/innen des Jahres 1997 (Fabian/Minks 2006) Ausbildungsregionen als oberste Ebene der Stichprobenziehung festgelegt, um der Zweiphasigkeit der Lehrerausbildung gerecht zu werden.

Aus dem Bestand an geeigneten Ausbildungsregionen wurden vier Regionen ausgewählt, die verschiedene Landesteile Deutschlands (Norden, Osten und Westen) sowie die Vielfalt an Strukturmerkmalen der Sekundarstufen-I-Ausbildung abbilden. In den Regionen wurden lokale Vollerhebungen durchgeführt, indem die jeweiligen Universitäten und alle umliegenden Studienseminare in die Stichprobe einbezogen wurden. Post hoc wurden die Teilnehmenden in drei Kohorten unterteilt: Studierende im Grundstudium bilden die erste Kohorte der Anfänger/innen in der Lehrerausbildung, Studierende im Hauptstudium bilden eine mittlere Kohorte und Referendar/innen bilden die dritte Kohorte am Ende der Ausbildung.

Insgesamt nahmen in Deutschland 878 Studierende sowie Referendar/innen aus vier Universitäten und 22 Studienseminaren teil, von denen 849 bereit waren, einen Fragebogen auszufüllen (siehe Tab. 3).

Umfang der realisierten MT21 Stichprobe insgesamt					
849					
Kohorte 1		*Kohorte 2*		*Kohorte 3*	
368		195		286	
Ausbildungsgang		*Ausbildungsgang*		*Ausbildungsgang*	
GHR	GyGS	GHR	GyGS	GHR	GyGS
294	74	143	52	133	153

Tab. 3: Zusammensetzung der Stichprobe nach Kohorten und Ausbildungsgängen

Unter den anwesenden Studierenden und Referendar/innen konnten zum Teil sehr gute Rücklaufquoten erzielt werden (siehe Spalte »Rücklaufquote 1« in Tab. 4). Auch die Rücklaufquote in Bezug auf die für uns erreichbaren Lehrkräfte kann als zufrieden stellend bzw. gut gewertet werden (siehe Spalte „Rücklaufquote 2“). Die Ausschöpfungsquote ist in Bezug auf die wichtige Kohorte 3, die Abschlusskohorte, ebenfalls gut, indem 80% aller formal in der Ausbildung befindlichen Referendar/innen in den teilnehmenden 22 Institutionen erreicht werden konnten. Die niedrigeren Quoten für die Kohorten 1 und 2 entsprechen dem Design.

	Teilnehmer/innen	*Fragebögen*	*Rücklaufquote 1*	*erreichbare Grundgesamtheit*	*Rücklaufquote 2*	*Grundgesamtheit*	*Ausschöpfungsquote*
Gesamt	878	849	97%	1.337	64%	2.761	31%
Kohorte 1	374	368	98%	669	55%	1151	32%
Kohorte 2	215	195	91%	338	58%	1251	16%
Kohorte 3	289	286	99%	330	87%	359	80%

Tab. 4: Rücklauf- und Ausschöpfungsquoten

Die variierenden Ausschöpfungsquoten in der Zusammensetzung der Kohorten wurden nach dem Modell prinzipiell gleicher Ziehungswahrscheinlichkeiten (*response homogeneity group*-Modell; Särndal et al. 1997) von Individuen pro Institution und von Institutionen pro Ausbildungsregion durch geeignete Strukturgewichtungsverfahren schrittweise innerhalb der Ausbildungsgänge ausgeglichen, um die Genauigkeit der Ergebnisse zu verbessern (Gabler/Hoffmeyer-Zlotnik/Krebs 1994; Kish 1965; Lutter 2005). Alle Merkmale des Redressement-Modells korrelieren hoch mit den Testergebnissen, sodass von einer angemessenen Modellgüte ausgegangen werden kann (Schnell 1993).

Die Gewichte wurden so gewählt, dass die Summe der Gewichte der Anzahl der Fälle entspricht (Rekalibrierung des *house-weight* und nicht des *stratum*- bzw. *tot-weight*; Rosing/Ross 1992). Die Gewichtungsfaktoren liegen für die dritte Kohorte wegen der geringen Stichprobenausfälle dicht um 1; die Gewichtungsfaktoren für die ersten beiden Kohorten liegen ebenfalls in einer noch akzeptablen Spannweite, sodass ihnen eine hinreichende Güte bescheinigt werden kann (Gabler/Häder 1997; Gelman/Carlin 2002).

3.2 Untersuchungsinstrumente

Die Entwicklung der Instrumente 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 Mathematiker/innen, Mathematikdidaktiker/innen und Erziehungswissenschaftler/innen aus unterschiedlichen Ländern Item-Entwicklungsworkshops durchgeführt. Auf diese Weise entstand ein umfangreicher Itempool, der mehreren Expertenreviews unterzogen wurde. Die verbleibenden Aufgaben flossen in eine Itempilotierung ein, auf deren Basis die endgültige Auswahl und Zusammenstellung des Leistungstests für die Hauptstudie geschah. In Tabelle 5 ist dokumentiert, in welchem Umfang die Dimensionen fachbezogenen Wissens durch Items repräsentiert sind.

	Mathematik	Mathematikdidaktik	*Insgesamt*
Arithmetik	7	8	15
Algebra	2	14	16
Funktionen	6	10	16
Geometrie	8	7	15
Statistik	3	7	10
Insgesamt	26	46	*72*

Tab. 5: Anzahl der Items in den Dimensionen des MT21-Leistungstests

Die Fabel vom Wettrennen zwischen Hase und Igel ist ein typisches Beispiel für mathematikdidaktische Aufgaben aus dem Inhaltsgebiet Funktionen (siehe Abb. 1). Bei den Items 1 und 2 ist »ja« die korrekte Lösung, bei den Items 3 bis 6 »nein«. Die Einschätzung von Aussage 4 erfordert dabei Präzision, da der in der Abbildung dargestellte Graph und damit die vom Igel zurückgelegte Wegstrecke über das Ziel hinausgeht. Die Einschätzung von Aussage 6 ist schwer, weil die Schlussfolgerung des Schülers neben falschen auch korrekte Elemente enthält.

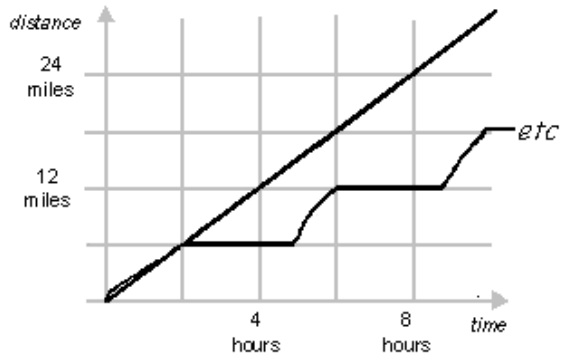
In Abb. 2 ist eine Beispielaufgabe aus dem Inhaltsgebiet der Algebra dokumentiert. In dieser ist die Anzahl der Lösungen einer quadratischen Gleichung zu bestimmen. Option 3 stellt die korrekte Lösung dar. Das Item ist ein Beispiel für die in *MT21* angewendete Idee, Distraktoren nach bekannten konzeptionellen Fehlvorstellungen von Studierenden sowie Referendarinnen und Referendaren zu formulieren. Auf diese Weise sollen Detailanalysen ermöglicht werden, die unterhalb der Ebene „falsch – richtig“ Hinweise auf das Niveau des fachbezogenen Wissens geben, indem den falschen Antworten eine unterschiedliche Qualität zukommt. Im konkreten Fall heißt dies beispielsweise, dass die ersten beiden falschen Optionen den Umgang mit Lösungsstrategien für quadratische Gleichungen testen, während die beiden letzten falschen Optionen den Umgang mit Parametern erfassen.

J74. Die folgende Aufgabe wurde Schüler(innen) der Sekundarstufe gegeben.

Eine Fabel

Ein Hase und ein Igel beschlossen, einen Marathon (26 Meilen) zu laufen. Der Hase rannte zuversichtlich los, fühlte sich ob seiner hohen Geschwindigkeit überlegen. Er lief 6 Meilen/Stunde während 2 Stunden und entschied sich dann, ein Nickerchen von 3 Stunden einzulegen. Er wollte dieses Muster beibehalten, 2 Stunden laufen, 3 Stunden Pause, bis zum Ende des Rennens. Der Igel behielt bis zum Ende des Rennens das gleichmäßige Tempo von 3 Meilen/Stunde bei und blieb nicht stehen.

Ein Schüler zeichnete zwei Graphen in das untenstehende Koordinatensystem, die mit der Geschwindigkeit der beiden korrespondieren. .



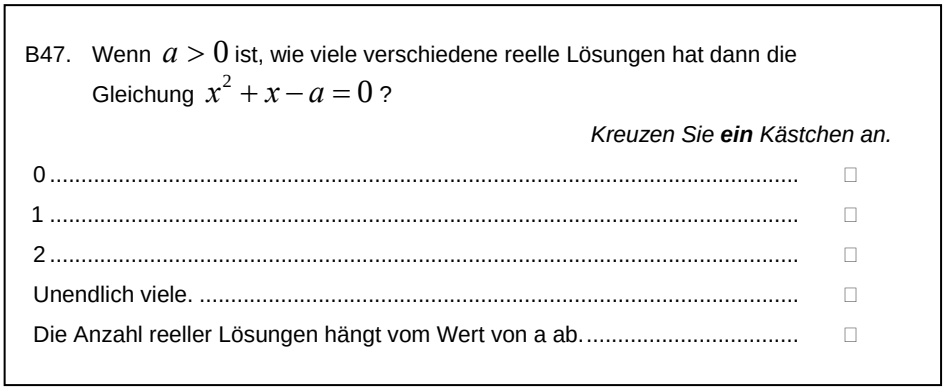
Der Schüler zieht folgende Schlussfolgerung: Der Igel gewinnt das Rennen, weil der Hase zu lange pausiert. Der Igel gewinnt, weil der Igel ein gleichmäßiges Tempo während des ganzen Rennens läuft. (Er läuft langsamer, aber macht keine Pausen.)

Welche der folgenden Aussagen sind für die Antwort des Schülers richtig?

Kreuzen Sie **ein** Kästchen in jeder **Zeile** an.

	Ja	*Nein*	*Bin nicht sicher*
1. Die Geschwindigkeit des Igels kann korrekt aus dem Graphen entnommen werden.			
2. Die Länge der Pausen des Hasen sind korrekt			
3. Die Geschwindigkeit des Hasen kann korrekt aus dem Graphen entnommen werden			
4. Die Strecke des Igels ist korrekt gezeichnet			
5. Die Laufintervalle des Hasen sind korrekt gezeichnet			
6. Die Schlussfolgerung des Schülers folgt aus dem gezeichneten Graphen			

Abb. 1: Mathematikdidaktische Beispielaufgabe in MT21 *für den Inhaltsbereich Funktionen*



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

*Kreuzen Sie **ein** Kästchen an.*

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

Abb. 2: Mathematische Beispielaufgabe aus MT21 für den Inhaltsbereich Algebra

3.3 Datenanalysen

Für den Leistungstest wurde ein rotiertes Testdesign verwendet, um angesichts der beschränkten Erhebungszeit von 90 Minuten eine hinreichend große Zahl an Items einsetzen zu können. Die psychometrischen Eigenschaften des Tests wurden mit Methoden der probabilistischen Testtheorie geprüft, und zwar unter Verwendung des Programmpaketes *ConQuest* (Wu/Adams/Wilson 2006). Über Anker-Items, die in beiden Testheften vertreten waren, gelang auf dieser Basis eine gemeinsame Skalierung aller Personen und Aufgaben.

Das in *ConQuest* implementierte mehrdimensionale *Random Coefficients Multinomial Logit*-Modell kann mehrere latente Fähigkeiten simultan berücksichtigen, womit eine messfehlerfreie Schätzung ihrer Beziehungen möglich wird. Die Lokalisation der Items auf der Fähigkeitsskala wurde über eine 65-prozentige Lösungswahrscheinlichkeit bestimmt; das entspricht einer Erhöhung der Schwierigkeitsparameter um 0,619 Logits. Die Schätzung der Itemparameter erfolgte über *Maximum-Likelihood*-Verfahren, auf deren Basis anschließend die Personenparameter geschätzt wurden. Als solche wurden *Weighted Likelihood Estimators* (WLEs) verwendet (Warm 1989), die optimale Schätzer für individuelle Personenfähigkeiten darstellen (Rost 2004, S. 316). Darüber hinaus wurden *Expected A Posteriori*-Schätzer (*EAP estimates*) berechnet. Aus dem *MT21*-Itemsatz wurden für die vorliegenden Analysen jene Items entfernt, die bei einer der gewünschten Skalierungen keinen zufrieden stellenden Fit, d. h. eine gewichtete Modellabweichung mit Werten < .80 bzw. > 1.20 aufwiesen.

Die Überprüfung der differentiellen Itemfunktionen erfolgte ebenfalls im Rahmen der *Item Response Theory* auf der Basis messfehlerfreier Itemparameter unter

Konstanthalten des mittleren Testergebnisses pro Kohorte (Holland & Wainer, 1993; Klieme/Baumert 2001). Um eine Konfundierung mit den unterschiedlichen Fähigkeitsniveaus sowie den spezifischen Stärken und Schwächen angehender GHR- und GyGS-Lehrkräfte zu vermeiden, wurden die Analysen für die erste Gruppe getrennt durchgeführt. Diese Substichprobe weist zudem zu allen Messzeitpunkten eine hinreichende Fallzahl auf, um stabile Parameterschätzungen zu erhalten.

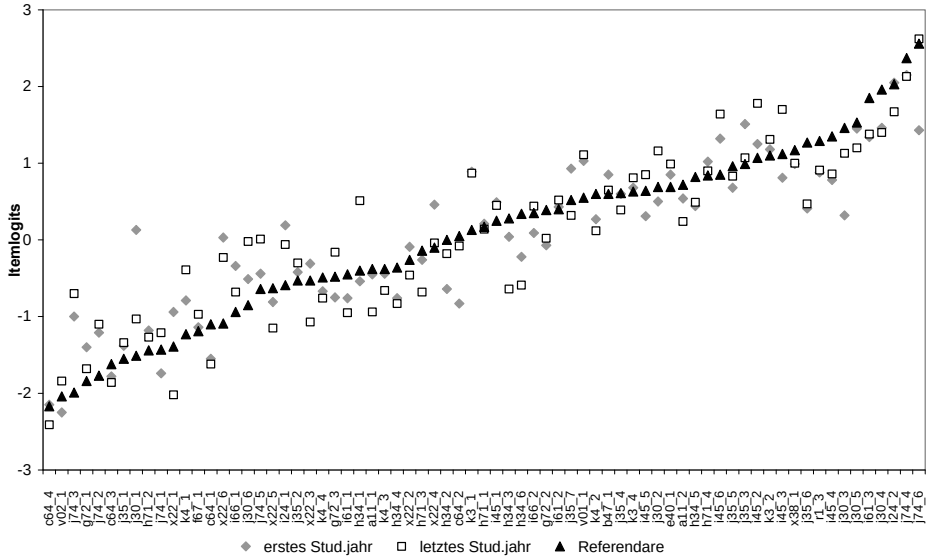


Abb. 3: Schwierigkeitsparameter der MT21-Testitems nach Kohorten getrennt

Die hier zu Grunde gelegte Stichprobe umfasst demnach 570 angehende GHR-Lehrkräfte, von denen sich 294 im Grundstudium, 143 im Hauptstudium und 133 im Referendariat befanden. Vorgegangen wurde dabei in zwei Schritten:

- In exploratorischer Absicht wurde zunächst für jede der drei Kohorten eine separate eindimensionale Skalierung auf der Grundlage des einparametrischen Rasch-Modells mit den 72 fachbezogenen Items des deutschen Tests durchgeführt. Dabei wurde in Kauf genommen, dass die Fallzahlen zum Teil hohe Messfehler zur Folge haben würden. Auf diesem Wege konnten aber – bei jeweils auf 0 zentrierten, also den allgemeinen Kompetenzzuwachs neutralisierenden Itemschwierigkeiten – Testaufgaben identifiziert werden, deren Schwierigkeitsparameter auffällig instabil waren (siehe Abb. 3). Die Differenzen der Logit-Werte können in diesem Zusammenhang unmittelbar als Maß der Effektstärke verwendet werden.
- Sodann wurden die Gruppen mit der in *ConQuest* implementierten Routine zum Umgang mit DIF paarweise miteinander verglichen. In diesem Zusammenhang können die beobachteten Auffälligkeiten auch auf Signifikanz geprüft werden.

Betrug die Abweichung eines kohortenspezifischen Parameters vom kohortenübergreifenden Wert mehr als das Doppelte des geschätzten Messfehlers, wurde auf statistische Signifikanz erkannt, da die Wahrscheinlichkeit einer falschen DIF-Vermutung jenseits solcher Überschreitungen unter $p = 0{,}05$ sinkt (Wu/Adams/Wilson 2006).

3 Ergebnisse

Eine Voraussetzung für die Analyse differentieller Itemfunktionen ist, dass das Rasch-Modell für die gesamte Stichprobe und den vorliegenden Test gilt. Die Passung des eindimensionalen Modells fachbezogenen Wissens zu den Daten kann aufgrund der Fit-Statistiken bestätigt werden. Mit nur zwei Ausnahmen liegt die gewichtete quadrierte Abweichung aller 72 Items im zuvor definierten Rahmen.

Wie schon im Vergleich der Kohorten-Mittelwerte deutlich wurde (Blömeke et al. 2008a), zeigt sich auch in dieser Analyse, dass zu Beginn der Lehrerausbildung substanziell weniger fachbezogenes Wissen vorliegt als am Ende des Universitätsstudiums bzw. im Referendariat. Bei einem Standardfehler von 0,042 trennen 0,329 Logits die Kohorten 1 und 2. Der Unterschied zwischen den Kohorten 2 und 3 wird dann allerdings nicht mehr signifikant.

Auf differentielle Itemfunktionen kann anhand dieses Unterschieds nicht geschlossen werden, da prinzipiell denkbar wäre, dass für Anfänger/innen alle Items im selben Ausmaß schwerer waren als für Studierende am Ende des Universitätsstudiums bzw. im Referendariat. Die Chi-Quadrat-Statistik zur Prüfung der Frage, ob einzelne Items für eine der Kohorten systematisch schwerer oder leichter waren, deutet dann aber bereits an, dass dieses zutrifft: $\chi^2_{[142]} = 274{,}04$, $p < .001$.

Die Spannweite der Differenz in den Itemschwierigkeiten liegt für den Vergleich der Anfänger/innen mit den Studierenden am Ende des Universitätsstudiums bei -1,19 bis 1,16, für den Vergleich der Studierenden am Ende des Studiums mit Referendar/innen bei -0,93 bis 1,29 und für den Vergleich der Anfänger/innen mit den Referendar/innen bei -1,14 bis 1,64 jeweils mit einem Mittelwert von 0 und einer Standardabweichung von 0,5 Logits.

Die Signifikanztests zu den Logit-Differenzen ergeben, dass bei dem ersten Vergleich elf Items und bei dem zweiten Vergleich zusätzlich sechs Items signifikante Abweichungen aufweisen (siehe Tab. 6). Über diese 17 Items mit differentiellen Itemfunktionen hinaus führt die über die Gesamtzeit der Ausbildung festgestellte Logit-Differenz bei drei weiteren Items zu signifikanten Ergebnissen. Insgesamt weisen also 20 Items differentielle Itemfunktionen auf. 11 Items waren für die jeweils vorhergehende Kohorte besonders schwer, neun Items waren für die vorhergehende Kohorte dagegen überproportional viel leichter. Während sich erstere Abweichungen relativ gleichmäßig über die drei Vergleiche verteilen, zeigt sich in Bezug auf letztere Items ein klarer Schwerpunkt auf dem ersten Vergleich.

Item für die vorhergehende Kohorte signifikant ...	Univ.-Anfang – Univ.-Ende	Univ.-Ende – Referendariat	Univ.-Anfang – Referendariat	Insgesamt
... schwerer	4	4	3	**11**
... leichter	7	2	0	**9**
Insgesamt	**11**	**6**	**3**	**20**

Tab. 6: Zahl der Items mit differentiellen Itemfunktionen

Diese Verteilung hängt mit der Art der Lern- bzw. Vergessensprozesse zusammen, die differentielle Itemfunktionen bewirken können. Analysiert man die Items im Einzelnen lassen sich alle Abweichungen mit den eingangs angeführten Begründungen erklären.

Zunächst zu jenen Items, die für die jeweils vorhergehende Kohorte noch schwerer waren, als im Mittel zu erwarten gewesen wäre. Im Fall der beiden Vergleiche, in die Anfänger/innen involviert sind, handelt es sich um Items, die entweder Inhalte, Itemformate oder kognitive Anforderungen enthalten, mit denen sie aus ihrer Schulzeit weniger vertraut sind. Solche Inhalte sind beispielsweise komplexe Zahlen oder abschnittweise definierte, stückweise lineare Funktionen. Es können aber auch Fehler aufgrund einer schulbedingt stärkeren algorithmischen Prägung gemacht werden oder weil Anfänger/innen mit der mathematischen Fachterminologie nicht hinreichend vertraut sind und so eher Fehlvorstellungen unterliegen. Im Falle der Vergleiche, in die Referendar/innen involviert sind, haben Letztere einen Vorteil im Umgang mit Schülerfehlern, Item-Formaten, die Bewertungen verlangen und textförmigen Beschreibungen mathematischer Sachverhalte, da sie mit diesen durch ihren Unterricht stärker vertraut sind. Dies gilt insbesondere für Items, die schülernah in Alltagssprache dargestellt sind, während an der Universität besonderer Wert auf mathematische Exaktheit gelegt wird.

Bei jenen Items, die für die jeweils vorhergehende Kohorte signifikant leichter gewesen sind, handelt es sich in Bezug auf den Vergleich der Anfänger/innen mit angehenden Lehrkräften am Ende des Studiums zum einen um Inhalte, Item-Formate oder kognitive Anforderungen, die typisch für Schulmathematik sind, die in der Universität aber nicht mehr aufgegriffen werden. Dies gilt zum Beispiel für graphische Interpretationen von Funktionen. Je weiter die Kohorten davon entfernt sind, umso eher ist von Vergessensprozessen auszugehen. Zum anderen handelt es sich um Inhalte oder kognitive Anforderungen, die erst in den letzten Jahren verstärkt Einzug in die Schule gehalten haben, sodass frühere Jahrgänge an Studierenden sie nicht in demselben Ausmaß erlebt haben wie heutige Studienanfänger/innen. Dies gilt zum Beispiel für Beschreibende Statistik oder Modellieren. Hier profitiert also die „jüngere“ Generation von Änderungen in der schulischen Mathematikausbildung. In Bezug auf den Vergleich von Studierenden und Referendar/innen ergeben sich Vorteile für Erstere in klassisch universitären Themengebieten. Wenn die-

se Veranstaltungen gerade gehört werden bzw. kürzlich gehört wurden, können die Items häufiger richtig beantwortet werden.

4 Zusammenfassung, Diskussion und Ausblick

Im Rahmen der Studie „Mathematics Teaching in the 21st Century (*MT21*)“ wurde erstmals in Deutschland das fachbezogene Wissen angehender Lehrerinnen und Lehrer im internationalen Vergleich getestet. Ziel der 6-Länder-Studie war die empirische Erfassung von Effekten der Lehrerausbildung am Beispiel der Ausbildung von angehenden Mathematiklehrkräften der Sekundarstufe I. In *MT21* wurde eine Konzeptualisierung professioneller Kompetenz entwickelt, die einer international-vergleichenden empirischen Erfassung zugänglich und gleichzeitig anschlussfähig an die deutsche Diskussion ist.

Das Kohortendesign der Studie ermöglicht vertiefte Analysen zu Entwicklungsprozessen während der Lehrerausbildung. Diese weisen auf einen deutlichen Leistungsvorsprung von Referendar/innen und Studierenden am Ende der Universitätsausbildung gegenüber Studienanfänger/innen hin. Dieses aus der zentralen *MT21*-Publikation (Blömeke/Kaiser/Lehmann 2008) bekannte Ergebnis konnte im vorliegenden Beitrag bestätigt werden.

Darüber hinaus zeigen sich differentielle Itemfunktionen. Elf Items waren für die jeweils vorhergehenden Kohorten systematisch noch schwerer, als dies aufgrund der mittleren Leistungsunterschiede zu erwarten gewesen wäre. Die entsprechenden Unterschiede in der Schwierigkeit lassen sich fast vollständig auf Spezifika der ersten bzw. zweiten Phase der Lehrerausbildung zurückführen, weisen also auf besondere Stärken dieser hin.

Gleichzeitig waren neun Items für die vorhergehenden Kohorten signifikant leichter. Für die Lehrerausbildung lassen sich insbesondere aus den zugrunde liegenden Interpretationen dieser differentiellen Itemfunktionen als Vergessensprozesse – von der Schulzeit in die Universitätsausbildung hinein und dann noch einmal von dieser in die berufliche Praxis hinein – wichtige Schlussfolgerungen ziehen. Felix Klein (1933, S. 1) hat bereits Anfang des letzten Jahrhunderts auf das Problem der doppelten Diskontinuität hingewiesen und Lösungsansätze formuliert:

> „Der junge Student sieht sich am Beginn seines Studiums vor Probleme gestellt, die ihn in keinem Punkte mehr an die Dinge erinnern, mit denen er sich auf der Schule beschäftigt hat; natürlich vergisst er daher alle diese Sachen rasch und gründlich. Tritt er aber nach Absolvierung des Studiums ins Lehramt über, so soll er plötzlich eben diese herkömmliche Elementarmathematik schulmäßig unterrichten; da er diese Aufgabe kaum selbstständig mit seiner Hochschulmathematik in Zusammenhang bringen kann, so wird er in den meisten Fällen recht bald die althergebrachte Unterrichtstradition aufnehmen, und das Hochschulstudium bleibt ihm nur eine mehr oder minder angenehme Erinnerung, die auf seinen Unterricht keinen Einfluss hat.

> Diese doppelte Diskontinuität, die gewiss weder der Schule noch der Universität jemals Nutzen gebracht hat, bemüht man sich neuerdings endlich aus der Welt zu schaffen, einmal indem man den Unterrichtsstoff der Schulen mit neuen, der modernen Entwicklung der Wissenschaft und der allgemeinen Kultur angepassten Ideen zu durchtränken sucht [...], andererseits aber durch geeignete Berücksichtigung der Bedürfnisse der Lehrer im Universitätsunterricht."

Die Ankündigung von Felix Klein wird allerdings immer noch nur eher singulär umgesetzt, von einer breiten Umsetzung seiner Ideen sind die deutschen Universitäten weit entfernt. Mathematiklehramtsstudierende erleben zu Beginn ihrer universitären Studien noch immer einen Bruch zur Schulmathematik, die sie nach ihrer universitären Ausbildung dann wiederum unterrichten sollen.

Für die weitere Forschung stellt sich aufbauend auf den hier vorgestellten Analysen eine Reihe von Aufgaben. In methodischer Hinsicht ist darauf hinzuweisen, dass es sich einerseits um eine Gelegenheitsstichprobe und andererseits um eine Querschnittsstudie handelte. Bevor weit reichende Konsequenzen für die Bildungspraxis gezogen werden, sind daher zunächst einmal Replikationen der Ergebnisse anhand repräsentativer Stichproben und echten Längsschnittstudien notwendig. TEDS-M (siehe dazu den Beitrag von Blömeke, Kaiser et al. in diesem Band) bietet dazu eine erste Gelegenheit.

Zudem zeigt sich die Notwendigkeit, Studien wie die hier vorgestellte in der Berufseingangsphase fortzusetzen bzw. neu zu initiieren. Empirische Studien zu den ersten Jahren der beruflichen Praxis, die auf Testdaten beruhen, existieren so gut wie nicht. Erst dann ließe sich aber klären, welche langfristigen Wirkungen der Ausbildung zugeschrieben werden können und ob sich ähnliche Lern- und Vergessensprozessen, wie sie sich hier andeuten, feststellen lassen.

Literatur

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

Blömeke, S. (2004): Empirische Befunde zur Wirksamkeit der Lehrerbildung. In: Blömeke, S./Reinhold, P./Tulodziecki, G./Wildt, J. (Hrsg.): Handbuch Lehrerbildung. Bad Heilbrunn/Braunschweig: Klinkhardt/Westermann, S. 59–91.

Blömeke, S./Felbrich, A./Müller, Ch. (2008): Theoretischer Rahmen und Untersuchungsdesign. In: Blömeke, S./Kaiser, G./Lehmann, R. (Hrsg.): Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematik-Studierender und -referendare – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung. Münster: Waxmann, S. 15–48.

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./Kaiser, G./Schwarz, B./Seeber, S./Lehmann, R./Felbrich, A./Müller, Ch. (2008a): Entwicklung des fachbezogenen Wissens in der Lehrerausbildung. In: Blömeke, S./Kaiser, G./Lehmann, R. (Hrsg.): Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematik-Studierender und -referendare – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung. Münster: Waxmann, S. 135–170.

Blömeke, S./Lehmann, R./Seeber, S./Schwarz, B./Kaiser, G./Felbrich, A./Müller, Ch. (2008b): Niveau- und institutionenbezogene Modellierungen des fachbezogenen Wissens. In: Blömeke, S./Kaiser, G./Lehmann, R. (Hrsg.): Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematik-Studierender und -referendare – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung. Münster: Waxmann, S. 105–134.

Blum, W./Neubrand, M./Ehmke, T./Senkbeil, M./Jordan, A./Ulfig, F./Carstensen, C.H. (2004): Mathematische Kompetenz. In: Prenzel, M./Baumert, J./Blum, W./Lehmann, R./Leutner, D./Neubrand, M. et al. (Hrsg.): PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann, S. 47–92.

Budgell, G. R./Namburty, S. R./Douglas, A. Q. (1995): Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19, S. 309–321.

Bulmahn, E./Wolff , K./Klieme, E. (2003): Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Berlin: BMBF.

Camilli, G./Shepard, L. A. (1994): Methods for identifying biased test items. Bd. 4. Thousand Oaks: Sage.

Fabian, G./Minks, K.-H. (2006): Dokumentation des Scientific Use Files „HIS-Absolventenpanel 1997“. Hannover: Hochschul-Informations-System.

Gabler, S./Häder, S. (1997): Wirkung von Gewichtungen bei Face-to-Face und Telefonstichproben. Eurobarometerexperiment 1994. In: Gabler, S./Hoffmeyer-Zlotnik, J. H. P. (Hrsg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag, S. 221–245.

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

Gelman, A./Carlin**,** J. B. (2002): Poststratification and Weighting Adjustments. In: Groves, R. M./Dillman D. A./Eltinge, J. L./Little, R. J. A. (Hrsg.): Survey Nonresponse. New York: Wiley, S. 288–302.

Holland, P. W./Wainer, H. (1993): Differential Item Functioning. Mahwah, NJ: Lawrence Erlbaum.

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

Klein, F. (1933): Elementarmathematik vom höheren Standpunkte aus. Erster Band. Berlin, Springer.

Klieme, E./Baumert, J. (2001): Identifying national cultures of mathematics education. Analysis of cognitive demands and differential item functioning in TIMSS. European Journal of Psychology of Education, 16, S. 385–402.

[KMK] Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2003): Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Bonn: KMK.

Krauthausen, G./Scherer, P. (2007): Einführung in die Mathematikdidaktik. Heidelberg: Elsevier, 3. Aufl.

Lutter, M. (2005): Gewichtungsverfahren in der empirischen Sozialforschung. Resultate Monte-Carlo-simulierter Redressment-Prozeduren. Duisburg-Essen: Universität (Diplomarbeit).

[NCTM] National Council of Teachers of Mathematics (2000): Principles and Standards for School Mathematics. Reston, VA: NCTM.

Rosing, M. J./Ross, K. N. (1992): Sampling and administration. In: Keeves, J. P. (Hrsg.): The IEA Technical Handbook. The Hague: IEA, S. 51–90.

Rost, J. (2004): Lehrbuch Testtheorie – Testkonstruktion. Bern: Hans Huber, 2. Aufl.

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

Schnell, R. (1993): Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität“ und Gewichtungsverfahren. Zeitschrift für Soziologie, 22, S. 16–32.

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

Tenorth, H.-E. (2006): Professionalität im Lehrerberuf. Ratlosigkeit der Theorie, gelingende Praxis. Zeitschrift für Erziehungswissenschaft, 9, S. 580–597.

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

Warm, T. A. (1989): Weigthed Likelihood Estimation of Ability in Item Response Models. Psychometrika, 54, S. 427–450.

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

Wu, M./Adams, R. J./Wilson, M. R. (2006): ConQuest. Generalized item response modelling software. Melbourne: ACER.