Blömeke, Sigrid

Vorsicht bei Evaluationen und internationalen Vergleichen. Unterschiedliche Referenzrahmen bedrohen die Validität von Befragungen zur Lehrerausbildung

Zeitschrift für Pädagogik 60 (2014) 1, S. 109-131

D 7484

ZEITSCHRIFT FÜR PÄDAGOGIK

Heft 1 Januar/Februar 2014

■ *Thementeil*

Zukünfte

■ *Allgemeiner Teil*

BELTZ JUVENTA

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 60 – Heft 1
Januar/Februar 2014

Inhaltsverzeichnis

Thementeil: Zukünfte

Allgemeiner Teil

Besprechungen

Dokumentation

Table of Contents

Topic: Futures

Contributions

Allgemeiner Teil

Sigrid Blömeke

Vorsicht bei Evaluationen und internationalen Vergleichen

Unterschiedliche Referenzrahmen bedrohen die Validität von Befragungen zur Lehrerausbildung

Zusammenfassung: Basierend auf der internationalen Vergleichsstudie TEDS-M wird die Validität von Befragungen zur Lehrerausbildung überprüft. Ziel ist zum einen, subjektive Qualitätseinschätzungen zur Ausbildung mit der tatsächlich erreichten Leistung in Zusammenhang zu bringen und so die prädiktive Validität Ersterer zu überprüfen. Zum anderen wird die Gefahr ökologischer Fehlschlüsse nachgewiesen, wenn in Analysen internationaler Daten die falsche Analyseeinheit gewählt wird. In Mehrebenenanalysen mit rund 8 000 angehenden Mathematiklehrkräften der Sekundarstufe I aus 15 Ländern werden vier typische Fragen aus Evaluationsstudien zur Wirksamkeit der Lehrerausbildung mit dem tatsächlich gezeigten mathematischen und mathematikdidaktischen Professionswissen in Beziehung gesetzt. Eine globale Wirksamkeitseinschätzung der Ausbildung weist einen geringen positiven Zusammenhang zum Professionswissen auf. Differenzierten Bewertungen kann dagegen keine prädiktive Validität zugeschrieben werden. Die Ergebnisse mahnen zur Vorsicht, was Schlussfolgerungen zur Lehrerausbildung angeht, die nur auf subjektiven Einschätzungen beruhen. Im internationalen Vergleich müssen zudem die kulturellen Unterschiede zwischen Ländern berücksichtigt werden, da unterschiedliche Referenzrahmen für die Einschätzungen existieren und die erhobenen Konstrukte ihre Bedeutung verändern, wenn sie auf Länderebene aggregiert werden.

Schlagworte: Mehrebenenanalyse, Befragung, Validität, Fehlschluss, Vergleichsstudie

1. Problemaufriss: Befragungen in der Lehrerausbildungsforschung

Studien zur Lehrerausbildung setzen mit wenigen Ausnahmen auf Befragungen von Studierenden, Referendaren und Absolventen. Sie sollen Auskunft zur Wirksamkeit der Ausbildung geben, welche Lerngelegenheiten die angehenden Lehrkräfte hatten und ob sie gut auf ihren Beruf vorbereitet wurden (Oser & Oelkers, 2001; Abs, Döbrich, Vögele & Klieme, 2005). Diesen Studien ist die Annahme gemein, dass die Befragten valide Einschätzungen abgeben können, dass ihren subjektiven Qualitätsaussagen also Vorher-

sagekraft (prädiktive Validität) für die tatsächlich erreichte Leistung am Ende der Ausbildung zukommt.

Inwieweit eine solche Annahme richtig ist, gilt es allerdings empirisch zu prüfen (Messick, 1995). Dabei kann zwischen verschiedenen Formen an Validitätsprüfungen unterschieden werden, beispielsweise klassisch zwischen Inhaltsvalidität, Konstruktvalidität und Kriteriumsvalidität (Lienert & Raatz, 1994) oder ergänzend, in Abhängigkeit von der Reichweite der Ergebnisinterpretation, zusätzlich die Validierung dieser Interpretation und der daraus gezogenen Konsequenzen (Kane, 2006). Angesichts der in Befragungen implizit enthaltenen kausalen Annahme eines Ursache-Wirkungs-Zusammenhangs – Vorhersagekraft von subjektiven Befragungsergebnissen zur Lehrerausbildung für die objektiv erreichte Leistung – kommt einer empirischen Überprüfung dieser Annahme besondere Bedeutung zu. Die prädiktive Validität steht entsprechend im Mittelpunkt dieses Beitrags (zur Prüfung anderer Formen von Validität siehe Abschnitt 3.3).

In vielen Untersuchungen hat sich gezeigt, dass Befragungen in psychologischen und sozialwissenschaftlichen Studien eher geringe prädiktive Validität zukommt (Meyer et al., 2001; Kubinger, 2003; Joint Committee on Standards for Educational Evaluation, 1994). Eine typische Fehlerquelle sind unterschiedliche Referenzrahmen der Befragten, vor deren Hintergrund sie auf die Fragen antworten. In Bezug auf die Lehrerausbildung lässt sich diese potenzielle Fehlerquelle ebenfalls identifizieren. So weisen Studien zum Zusammenhang von fachbezogenen Lehrerkompetenzen und fachbezogenen Schülerleistungen (Baumert et al., 2010) auf die Bedeutung des Fachwissens hin, da vor allem das mathematikdidaktische, aber auch das mathematische Professionswissen mathematische Schülerleistungen signifikant vorhersagt. In Befragungen sehen angehende Lehrkräfte die Notwendigkeit der Fachausbildung dagegen häufig kritisch; sie bewerten den Anteil berufsbezogener Lerngelegenheiten als zu gering und die Wirksamkeit der Fachausbildung eher negativ (Schneider & Bodensohn, 2010). Das Bedürfnis angehender Lehrkräfte nach Handlungssicherheit (Jäger & Milbach, 1994) führt möglicherweise zur Erwartung, in der Ausbildung eher Handlungsroutinen einzuüben statt Fachwissen zu erwerben (Haag & Streber, 2010). Unterschiede in den Erwartungen an die Ausbildung bestehen aufgrund unterschiedlicher Berufswahlmotive eventuell auch zwischen Lehrkräften verschiedener Ausbildungsgänge (Schmidt, Blömeke & Tatto, 2011).

Diese potenziellen Fehlerquellen lassen erwarten, dass das Erfassen subjektiver Einschätzungen zur Lehrerausbildung und die Ausprägung objektiv erreichter Kompetenzen unabhängig voneinander zu betrachten sind. Von einer prädiktiven Validität auszugehen würde erfordern, ihren Zusammenhang empirisch nachzuweisen: Hängen subjektive Einschätzungen von angehenden Lehrkräften tatsächlich mit objektiven Maßen zusammen? Erfordert die Bestimmung von Ausbildungseffektivität mit Blick auf den Kompetenzgewinn nicht auch solche objektiven Maße?

Systematisch untersucht wurde die Validität von Befragungen zur Lehrerausbildung bisher allerdings kaum. König, Kaiser und Felbrich (2012) lieferten eine erste Analyse für zwei Länder, Deutschland und die USA, in der sie das pädagogische Professionswissen angehender Lehrkräfte mit der Einschätzung in Beziehung setzten, inwieweit sich

diese auf ihren Beruf vorbereitet fühlen. Sie fanden maximal schwach positive Zusammenhänge. Dieses Ergebnis lässt sich in die wachsende Zahl an Studien einordnen, die für fächerübergreifende Konstrukte wie kritisches Denken oder interkulturelle Kompetenz kaum valide Vorhersagen der tatsächlichen Leistung durch subjektive Einschätzungen feststellen (Bowman, 2011). Für das *fachbezogene* Lehrerprofessionswissen liegen soweit bekannt überhaupt keine Studien vor, da entsprechende objektive Maße bisher fehlten.

Skepsis gegenüber Befragungsergebnissen ist möglicherweise besonders dann angebracht, wenn es um Vergleiche über Länder hinweg geht (Fischer, 2004). Kulturell unterschiedliche Referenzmaßstäbe oder auch kulturell unterschiedlich ausgeprägte Zustimmungstendenzen können nicht ausgeschlossen werden (van de Vijver & Leung, 1997). So scheint die Tendenz zu sozial erwünschten Antworten in nicht-westlichen Ländern stärker ausgeprägt zu sein als in westlichen (Harzing, 2006). Zugleich tendieren Befragte aus westlichen Ländern offensichtlich stärker zu extremen Einschätzungen als Befragte aus nicht-westlichen Ländern (Takahashi, Ohara, Antonucci & Akiyama, 2002). Die unterschiedlichen Antwortstile liegen vermutlich in kulturell unterschiedlichen Verhaltensnormen begründet (Hofstede, 2001).

Shen und Tam (2008) haben das Problem kulturell unterschiedlicher Referenzmaßstäbe bei subjektiven Indikatoren anhand von TIMSS 1995, 1999 und 2003 durch einen Vergleich mit den objektiv erreichten Schülerleistungen untersucht. Während fachbezogene Einschätzungen der Selbstwirksamkeit bzw. des Leistungsniveaus sowie objektive Testergebnisse unter Beachtung der Mehrebenenstruktur der Daten innerhalb von Ländern in der Regel leicht positiv korrelieren, gilt dies auf der Klassen- oder Schulebene eher nicht. Stellen Länder die Analyseeinheit dar, fällt der Zusammenhang sogar negativ aus, indem sich Schülerinnen und Schüler aus schwächeren Ländern positiver wahrnahmen, als es ihre Testergebnisse ausweisen.

Aus diesem Problem der fehlenden skalaren Äquivalenz subjektiver Einschätzungen war – angeregt durch Klieme und Vieluf (2009) – für die international-vergleichende „Teacher Education and Development Study: Learning to Teach Mathematics (TEDS-M)“[1] der Schluss gezogen worden, ipsative Werte zu berichten (Blömeke, Kaiser & Lehmann, 2010). Diese stellen relative Abweichungen vom jeweiligen Ländermittelwert dar und können so zumindest kulturell unterschiedliche Antwortstile ausgleichen.

Der vorliegende Beitrag erweitert diesen Forschungsstand, indem die prädiktive Validität fachbezogener Befragungen zur Lehrerausbildung überprüft wird. Subjektiv er-

1 TEDS-M wurde von der International Association for the Evaluation of Educational Achievement (IEA), der US-amerikanischen National Science Foundation (REC 0514431) und den TEDS-M-Teilnahmeländern gefördert. In Deutschland erfolgte eine Förderung durch die Deutsche Forschungsgemeinschaft (BL548/3-1). Alle Darlegungen in diesem Beitrag stammen von der Autorin und spiegeln nicht notwendigerweise die Ansichten der Förderorganisationen wider. Das Problem der fehlenden skalaren Äquivalenz haben Biedermann, Blömeke und Oser bereits häufiger im Zuge von Workshops und Konferenzen thematisiert. Die jeweiligen Diskussionen haben den vorliegenden Beitrag bereichert.

mittelte Einschätzungen mehrerer Indikatoren der fachbezogenen Qualität und Wirksamkeit der Mathematiklehrerausbildung werden mit Testdaten zum mathematischen und mathematikdidaktischen Professionswissen in Beziehung gesetzt. Die Analyse wird für alle 15 TEDS-M-Teilnahmeländer durchgeführt. Sie erfolgt zum einen *intra*-national unter Beachtung der Mehrebenenstruktur der Daten und zum anderen im *inter*nationalen Vergleich auf Länderebene, um zu sehen, inwieweit die von Shen und Tam (2008) gefundenen TIMSS-Probleme auch in TEDS-M bestehen.

2. Theoretischer Rahmen

2.1 Professionelle Kompetenz von Lehrkräften als Kriterium

TEDS-M modelliert die professionelle Kompetenz von Mathematiklehrkräften – wie Weinert (1999) – als latentes Konstrukt, das der Bewältigung von beruflichen Anforderungen zugrunde liegt. Professionelle Kompetenz ist damit auf Performanz im Mathematikunterricht ausgerichtet. In konkreten Situationen wird diese Performanz allerdings – wie in einem klassischen Angebots-Nutzungs-Modell (Fend, 1980; Helmke, 2004) – als von weiteren Faktoren beeinflusst angesehen. Kompetenz und Performanz sind also nicht deckungsgleich. In TEDS-M wurden typische curriculare, planungs- und interaktionsbezogene Anforderungen an die Zielpopulation der Mathematiklehrkräfte in der Sekundarstufe I unter Rückgriff auf etablierte Diskurse der Lehrerforschung definiert (Shulman, 1985; Tatto et al., 2008).

Welche Dispositionen zur erfolgreichen Bewältigung der Anforderungen notwendig sind, wurde aus der Expertiseforschung abgeleitet (Berliner, 2001). Ihren Erkenntnissen zufolge besteht die professionelle Kompetenz von Lehrkräften aus *kognitiven* Leistungsdispositionen in Form von Professionswissen und aus professionsbezogenen *Einstellungen bzw. Wertvorstellungen* (Bromme, 1997; Baumert & Kunter, 2006). Beide Perspektiven wurden durch mehrfache nationale und internationale Expertenreviews validiert (siehe Abb. 1; für Details zum theoretischen Rahmen sowie zentrale Ergebnisse siehe insbesondere den Thementeil der Zeitschrift für Pädagogik 4/2012: Oser & Blömeke, 2012; Biedermann, Brühwiler & Krattenmacher, 2012; Blömeke, Suhl & Döhrmann, 2012; König et al., 2012; Steinmann & Oser, 2012).

Niveau und Varianz der professionellen Lehrerkompetenz sind die zentralen Kriterien, um ein Lehrerausbildungssystem als mehr oder weniger wirksam einzuschätzen (Blömeke, Suhl & Kaiser, 2011). Für den vorliegenden Beitrag wurden das mathematische und das mathematikdidaktische Professionswissen als Kriterien ausgewählt. Den beiden Indikatoren kommt Validierungsstudien zufolge die höchste prädiktive Bedeutung für Schülerleistungen im Mathematikunterricht zu. So zeigen Hill, Rowan und Ball (2005), dass Grundschüler ein signifikant höherer Lernzuwachs in Mathematik gelingt, wenn sie von Lehrkräften mit höherem schulrelevantem Fachwissen und fachdidaktischem Wissen unterrichtet werden. Der Lernzuwachs wird über höhere Instruktionsqualität vermittelt (Hill et al., 2008).

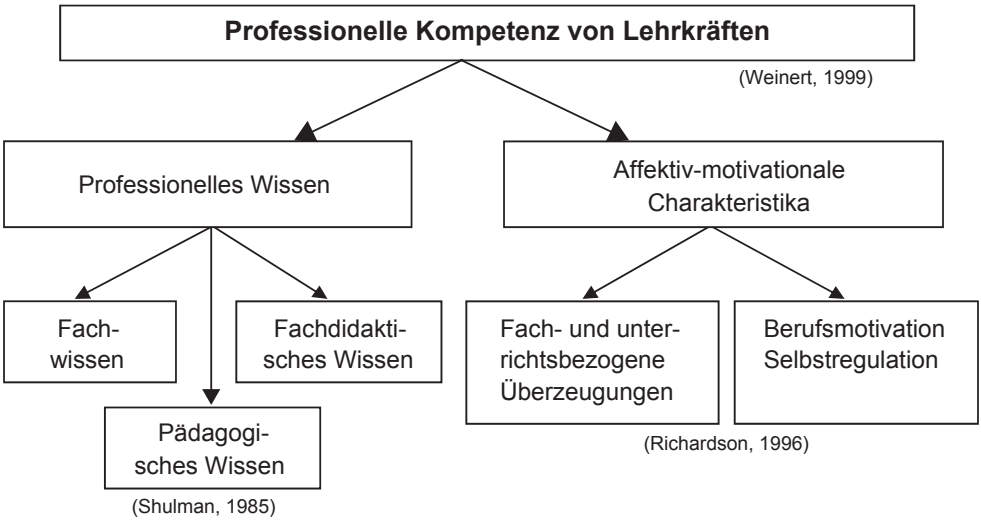

Abb. 1: Analytisches Modell professioneller Lehrerkompetenz

Baumert et al. (2010) gelingt ein vergleichbarer Nachweis für Mathematiklehrkräfte der Sekundarstufe I, die die Zielpopulation des vorliegenden Beitrags darstellen. Es zeigt sich ein signifikanter Zusammenhang zwischen dem mathematikdidaktischen – und mit etwas geringerer Vorhersagekraft auch dem mathematischen – Wissen der Lehrkräfte und den im Laufe eines Schuljahres erreichten mathematischen Schülerleistungen. Der Effekt wird über das Anspruchsniveau, kognitive Aktivierung und Lernerunterstützung vermittelt. Anders, Kunter, Brunner, Krauss und Baumert (2010) replizieren dieses Ergebnis anhand einer Teildimension, der Urteilsgenauigkeit von Mathematiklehrkräften.

2.2 Subjektive Einschätzungen der Lehrerausbildung

Für die Befragung der angehenden Mathematiklehrkräfte zur Qualität und Wirksamkeit ihrer Ausbildung bildeten – wie in den Kompetenztests – typische Anforderungen des Mathematikunterrichts den Bezugspunkt, sodass die Ergebnisse in Beziehung gesetzt werden können. Die subjektiven Einschätzungen wurden mithilfe von vier Indikatoren erhoben, die typischerweise in Evaluationsstudien zur Lehrerausbildung eingesetzt werden (z. B. National Center for Education Statistics, 2003; Ingvarson, Beavis & Kleinhenz, 2007; für einen Überblick, welche weiteren Studien mit entsprechenden Indikatoren arbeiten, siehe König et al., 2012) und die in TEDS-M für den Mathematikunterricht adaptiert wurden.

Die vier Indikatoren beziehen sich auf ein breites Spektrum an Aufgaben von Mathematiklehrkräften der Sekundarstufe I. Zu diesen wurde erstens erfasst, wie häufig die angehenden Lehrkräfte entsprechende berufsvorbereitende Lerngelegenheiten erlebt hatten. Zweitens sollten sie bewerten, wie gut sie sich dadurch auf ihre beruflichen

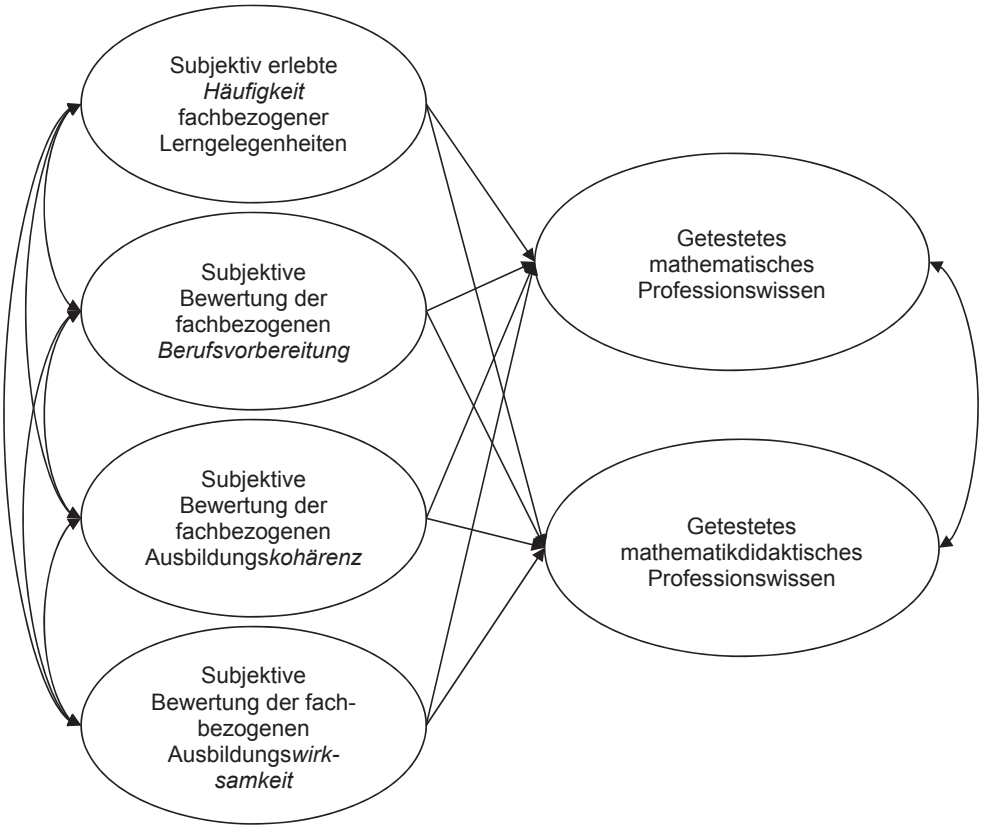

Abb. 2: Modellierung des Zusammenhangs von subjektiven Einschätzungen und getestetem Professionswissen

Aufgaben vorbereitet fühlen. Drittens bewerteten die angehenden Lehrkräfte die Kohärenz der Lerngelegenheiten während der Ausbildung, insbesondere das Verhältnis von Theorie und Praxis – ein als besonders relevant diskutiertes Qualitätsmerkmal (Terhart, 2001). Hinzu trat eine globale Einschätzung der Wirksamkeit der Lehrerausbildung.

Dem bisherigen Forschungsstand zufolge fühlen sich Lehrkräfte subjektiv in vielen Ländern zu wenig und qualitativ unzureichend auf berufliche Anforderungen vorbereitet (Kee, 2012). In Deutschland wird die Lehrerausbildung zudem als praxisfern bzw. fragmentiert kritisiert (Merzyn, 2002). Unseres Wissens liegen bisher keine Studien dazu vor, wie weit solche Befragungsergebnisse mit der objektiv am Ende der Ausbildung vorliegenden Fachkompetenz zusammenhängen. In Abbildung 2 ist das Modell zum Zusammenhang der vier aus den Befragungen stammenden Indikatoren mit dem erworbenen mathematischen und mathematikdidaktischen Wissen dargestellt, das in diesem Beitrag geprüft werden soll. Die Analysen erfolgen für die TEDS-M-Länder zunächst mehrebenenanalytisch unter Beachtung der Clusterstruktur der Daten (als *intra*nationale Analysen bezeichnet). Erwartet werden hier signifikant *positive,* angesichts

der dargelegten potenziellen Fehlerquellen aber *schwache* Zusammenhänge zwischen den subjektiven Indikatoren und den Testergebnissen.

Die *inter*nationalen Analysen erfolgen dann auf Länderebene. Hier stellen also nicht – wie im ersten Schritt – die individuellen Lehrkräfte die Analyseeinheit dar, sondern die typischerweise in Studien wie TIMSS berichteten Mittelwerte der 15 Länder. Analog zu Shen und Tam (2008) werden aufgrund unterschiedlicher kultureller Referenzrahmen der angehenden Lehrkräfte signifikant *negative* Zusammenhänge zwischen ihren subjektiven Einschätzungen der Ausbildung und dem von ihnen erworbenen Professionswissen erwartet.

3. Methode

3.1 Stichprobe

Die Stichprobe der vorliegenden Untersuchung besteht aus allen 15 Ländern, die an TEDS-M teilgenommen haben. Diese decken vier Kontinente ab: Afrika (Botswana), Amerika (Chile, USA), Asien (Georgien, Malaysia, Oman, Philippinen, Singapur, Taiwan, Thailand) und Europa (Deutschland, Norwegen, Polen, Russland, Schweiz). Tabelle 1 gibt einen Überblick über die Ausbildungsstruktur in diesen Ländern.

Insgesamt handelt es sich um 8 185 angehende Sekundarstufen-I-Lehrkräfte, die im letzten Jahr ihrer Ausbildung auf ihr mathematisches und mathematikdidaktisches Wissen hin getestet wurden. Im Mittel sind die Sekundarstufen-I-Lehrkräfte am Ende ihrer Ausbildung 24 Jahre alt (Spannweite: 21 bis 30 Jahre in Georgien bzw. Deutschland). Knapp zwei Drittel der angehenden Mathematiklehrkräfte sind weiblich. Knapp ein Viertel der Väter der Lehrkräfte weist einen tertiären Bildungsabschluss auf. Der Bildungshintergrund der Mütter ist in den meisten Ländern etwas niedriger bzw. entspricht dem der Väter.

Ein mehrstufiges stratifiziertes Samplingdesign sicherte Zufallsziehungen von Ausbildungsinstitutionen und angehenden Lehrkräften mit einer Mathematik-Lehrberechtigung für die Klasse 8. Polen (Ausbildungsinstitutionen mit grundständiger Lehrerausbildung), die Schweiz (Pädagogische Hochschulen in deutschsprachigen Kantonen) und die USA (staatliche Hochschulen) haben eingeschränkt an TEDS-M teilgenommen. In Chile, Georgien, Polen und den USA (kombinierte Rücklaufquote < 75 %) sowie in Norwegen (< 60 %) gelang es nicht vollständig, die Anforderungen der IEA an die Rücklaufquote zu erfüllen. In den USA wurde bei rund einem Fünftel der Stichprobe ein gekürztes Instrument eingesetzt. Für Norwegen wird der kombinierte Wert der vorhandenen Teilstichproben berichtet, um eine bestmögliche Annäherung an das Ergebnis des Landes zu erreichen (für weitere methodische Details siehe Blömeke et al., 2010).

	Institutionalisierung	Institutionen	Spezialisierung	Klassen	Sonstiges
Botswana	Pädagogische Hochschulen/Universitäten	7	a	8–10/8–12	3-/4-jährig
Chile	Pädagogische Hochschulen/staatliche/private Universitäten	35	c, d	1–8/5–8	4-/5-jährig; keine staatl. Vorgaben
Deutschland	Universitäten/Studienseminare	16*	b	5–10/5–13	föderale Org.; 4,5-/6,5-jährig
Georgien	Staatliche/private Universitäten	10	a	5–12	3-/5-jährig; keine staatl. Vorgaben
Malaysia	Pädagogische Hochschulen/staatliche/private Universitäten	17	b	7–13	grundständig/konsekutiv; 3-/5-jährig
Norwegen	Pädagogische Hochschulen/Universitäten	45	b, c, d	1–10/8–13	4-jährig; keine staatl. Vorgaben
Oman	Pädagogische Hochschulen/Universitäten	7	a	5–12	4-/6-jährig
Philippinen	Staatliche/private Studienseminare	51	a	7–10	4-jährig; keine staatl. Vorgaben
Polen	Pädagogische Hochschulen/Universitäten	78	a	4–9/4–12	Voll-/Teilzeit; 3-/5-jährig
Russland	Staatliche Universitäten	52	a	5–11	5-jährig; zentral/regional/kommunal organisiert
Schweiz	Pädagogische Hochschulen	15	b	7–9	4,5 jährig; föderale Organisation
Singapur	Universität	1	b	7–10/7–12	2-/5-jährig
Taiwan	Pädagogische Hochschulen/Universitäten	19	a	7–12	4,5-jährig
Thailand	Staatliche Universitäten	44	a	1–12	5-jährig; grundständig/konsekutiv
USA	Staatliche/private Universitäten	50	a	4–8/6–12	grundständig/konsekutiv; 4-/5-jährig

Anmerkungen. * = Anzahl der Bundesländer in der Stichprobe, a = Ein-Fach-Lehrerausbildung (hier: Mathematik), b = Zwei-Fach-Lehrerausbildung oder Ausbildung für mehrere affine Fächer (hier: unter anderem Mathematik), c = Klassenlehrerausbildung mit Mathematik als Schwerpunkt, d = Klassenlehrerausbildung ohne Mathematik als Schwerpunkt. In den meisten Ländern werden mehrere Formen der Sekundarstufen-I-Lehrerausbildung parallel angeboten. Diese sind durch Schrägstriche abgetrennt.

Tab. 1: Ausbildungsgänge und Ausbildungsinstitutionen der Sekundarstufen-I-Lehrerausbildung als differenzielle Lernumwelten

3.2 Instrumente: Erhebung des fachbezogenen Professionswissens

Getestet wurde das mathematische Wissen der Lehrkräfte in Arithmetik, Algebra, Geometrie und Stochastik. Die mit der Bewältigung der Anforderungen verbundenen kognitiven Prozesse wurden in Anlehnung an TIMSS in Wissen abrufen, anwenden und begründen ausdifferenziert. Eine dritte Heuristik für die Itementwicklung stellte das vorab eingeschätzte Schwierigkeitsniveau dar. Mathematikdidaktisches Wissen wurde in Anlehnung an Shulman (1985) sowie Fan und Cheong (2002) in zwei Subdimensionen ausdifferenziert erfasst: curriculares und auf die Planung von Unterricht bezogenes Wissen sowie auf die unterrichtliche Interaktion bezogenes Wissen. Auch hier wurde zwischen drei kognitiven Anforderungen und drei Schwierigkeitsniveaus unterschieden.

Der Test enthielt 76 Mathematik- und 27 Mathematikdidaktik-Items (Multiple-Choice- und offene Formate). Die Testzeit betrug 60 Minuten. Es wurde ein rotiertes Untersuchungsdesign mit drei Testheften eingesetzt, die über Anker-Items miteinander verknüpft waren. Die mathematischen und mathematikdidaktischen Rohdaten wurden in separaten eindimensionalen Modellen rasch-skaliert und auf Mittelwerte von 500 Testpunkten und eine Standardabweichung von 100 transformiert. Die WLE-Reliabilität lag bei guten bis zufriedenstellenden .91 für das mathematische und .70 für das mathematikdidaktische Wissen.

Die Abbildungen 3 bis 5 zeigen beispielhaft drei Items des TEDS-M-Tests. Mit dem ersten Item (Abb. 3) wird mathematisches Professionswissen im Inhaltsbereich Arithmetik erfasst. Es handelt sich um ein Multiple-Choice-Item, in dem Wissen abgerufen werden muss und dessen Schwierigkeit vorab als mittel eingeschätzt wurde. Das Item beschreibt eine Problemstellung, die mit Hilfe des kombinatorischen Modells einer zufälligen Auswahl von k aus n Elementen ohne Wiederholung und ohne Berücksichtigung der Reihenfolge gelöst werden kann. Die Anzahl der Möglichkeiten lässt sich berechnen, durch Schlussfolgern ermitteln oder abzählen. Die richtige Lösung ist C und wurde in Deutschland von 37 % sowie im internationalen Mittel von 34 % der Probandinnen und Probanden gewählt (Spannweite: 9 % in Georgien bis 92 % in Taiwan).

In einer Klasse sind 10 Schüler. Einmal werden 2 Schüler der Klasse zufällig ausgewählt, ein anderes Mal 8 Schüler. Welche der folgenden Aussagen ist richtig?

Kreuzen Sie nur ein Kästchen an.

A.	Es gibt mehr Möglichkeiten, 2 Schüler aus der Klasse zu wählen als 8.	☐
B.	Es gibt mehr Möglichkeiten, 8 Schüler aus der Klasse zu wählen als 2.	☐
C.	Die Anzahl der Möglichkeiten, 2 Schüler aus der Klasse zu wählen ist genauso groß wie die Anzahl der Möglichkeiten 8 zu wählen.	☐
D.	Es ist nicht möglich zu entscheiden, für welche Auswahl mehr Möglichkeiten existieren.	☐

Abb. 3: Beispiel-Item zur Erfassung arithmetischen Professionswissens

Welche der folgenden Sachverhalte können durch eine Exponentialfunktion beschrieben werden?

Kreuzen Sie ein Kästchen pro Zeile an.

		Ja	Nein
A.	Die Höhe *h* eines Balls *t* Sekunden nachdem er in die Luft geworfen wurde.	☐	☐
C.	Der Wert *W* eines Autos nach *t* Jahren, wenn die Wertminderung *d* % pro Jahr beträgt.	☐	☐

Abb. 4: Beispiel-Items zur Erfassung algebraischen Professionswissens

Mit der Aufgabe in Abb. 4 wird mathematisches Professionswissen im Inhaltsbereich Algebra erfasst. Es handelt sich um zwei Multiple-Choice-Items, in denen Wissen angewendet werden muss und deren Schwierigkeit vorab als gering bzw. mittel eingeschätzt wurde. Zum Lösen der Aufgabe muss die Exponentialfunktion als funktionaler Zusammenhang zweier Größen bekannt sein und die angehenden Mathematiklehrkräfte der Sekundarstufe I müssen außerdem in der Lage sein, diese Funktion als mathematisches Modell auf außermathematische Sachverhalte zu übertragen bzw. zu erkennen, welche der gegebenen außermathematischen Situationen sich mithilfe eines exponentiellen Zusammenhangs erklären lässt. Dies gilt für das Item C, was in Deutschland von 72 % sowie im internationalen Mittel von 60 % der Probandinnen und Probanden erkannt wurde (Spannweite: 28 % in Chile bis 95 % in Taiwan). Bei Item A handelt es sich dagegen um eine quadratische Funktion. In Deutschland lösten dieses Item rund 60 %, im internationalen Mittel rund 40 % korrekt (Spannweite: knapp 20 % in Chile bis rund 66 % in Russland bzw. Singapur).

Mit dem letzten Beispiel-Item (Abb. 5) wird in Teil b mathematikdidaktisches Professionswissen erfasst, das auf die unterrichtliche Interaktion bezogen ist. Es handelt sich um eine Kurzantwort-Aufgabe, deren Schwierigkeit vorab als gering eingeschätzt wurde. Auf den ersten Blick erscheinen die beiden algebraischen Problemstellungen sehr ähnlich, die zu analysieren sind. In beiden Fällen ist die Gesamtanzahl der Objekte ebenso bekannt wie die Verhältnisse der Objektmengen, die jedes Kind besitzt. Aus diesen Angaben soll die Menge der Objekte jedes einzelnen Kindes ermittelt werden. Im ersten Fall besitzt Peter 132 Murmeln, David 22 und Jonathan 44. Im zweiten Fall besitzt Anna 132€, Philipp 22€ und Lukas 44€.

Nur gut ein Drittel der angehenden Lehrkräfte (38 %) war im internationalen Mittel in der Lage, einen Grund zu benennen, warum die zweite Aufgabe schwerer ist als die erste (Spannweite: 7 % in Georgien bis 71 % in der Schweiz). In Deutschland gelang dies 58 %. Im ersten Fall beziehen sich beide Verhältnisangaben auf David und können die Murmelanzahlen von Peter und Jonathan durch Vielfache der Murmelanzahl von David ausgedrückt werden. Im zweiten Fall müssen die gesuchten Größen indirekt aus den Angaben über Anna bestimmt werden. Damit sind Überlegungen zum Verhält-

Die folgenden Aufgaben stammen aus einem Mathematikschulbuch für die Sekundarstufe I.

1. Peter, David and Jonathan spielen mit Murmeln. Zusammen haben sie 198 Murmeln. Peter hat 6-mal so viele Murmeln wie David und Jonathan hat 2-mal so viele Murmeln wie David. Wie viele Murmeln hat jeder der Jungen?

2. Die drei Kinder Anna, Philipp und Lukas besitzen zusammen 198 €. Anna hat 6-mal so viel Geld wie Philipp und 3-mal so viel wie Lukas. Wie viele Euro hat jedes Kind?
   (a) Lösen Sie beide Aufgaben.
   (b) Üblicherweise bereitet die zweite Aufgabe Schüler(inne)n der Sekundarstufe I größere Probleme als die erste. Nennen Sie einen Grund, der für den unterschiedlichen Schwierigkeitsgrad verantwortlich sein könnte.

Abb. 5: Beispiel-Item zur Erfassung mathematikdidaktischen Wissens (b)

nis der Anzahl der Euros von Philipp und Lukas nötig, wobei Überlegungen zu Verhältnissen auf einem höheren Schwierigkeitsniveau liegen als der einfache Vergleich von Größen.

3.3 Instrumente: Erhebung der subjektiven Einschätzungen

Zur Erfassung der subjektiv eingeschätzten Häufigkeit, mit der die angehenden Lehrkräfte auf ihren Beruf vorbereitet wurden, wurden 26 fachbezogene Anforderungen an eine Mathematiklehrkraft der Sekundarstufe I definiert. Für diese war auf vierstufigen Skalen von „nie“ bis „oft“ anzugeben, wie häufig in der Lehramtsausbildung Gelegenheit bestanden hatte, sie zu erlernen. Ein Beispiel ist „an das Vorwissen in Mathematik und die vorhandenen intellektuellen Fähigkeiten von Schüler(innen) anknüpfen“. Die Reliabilität dieser Skala (Cronbachs Alpha) lag zwischen .88 (Botswana) und .96 (Malaysia).

Wie gut sich die angehenden Mathematiklehrkräfte auf ihren Beruf vorbereitet fühlen, wurde – wie in König et al. (2012) – mithilfe von 13 Anforderungen erfasst, die vierstufig von „überhaupt nicht“ bis „in großem Maße“ einzuschätzen waren. Ein Beispiel ist, wie gut sie sich darauf vorbereitet fühlen, „angemessene Lernziele für die Schüler(innen) in Mathematik zu entwickeln“. Cronbachs Alpha lag hier zwischen .87 (Oman) und .92 (Chile, Malaysia, USA).

Die Bewertung der Mathematiklehrerausbildung im Hinblick auf deren Kohärenz wurde mit einer Skala erhoben, die sechs Items umfasste. Auf sechsstufigen Likertskalen von „stimme überhaupt nicht zu“ bis „stimme völlig zu“ war beispielsweise anzugeben, inwieweit die Fachausbilder im Referendariat „Ihre Erfahrungen, die Sie in Ihrer Lehramtsausbildung gesammelt haben, (schätzen)“. Cronbachs Alpha lag zwischen .83 (Oman) und .95 (Philippinen).

Mit einem Einzelitem wurde schließlich die globale Einschätzung der Ausbildungs*wirksamkeit* erfasst. Auf die Frage „Wie wirkungsvoll war Ihre Lehramtsausbildung insgesamt, um Sie auf den Beruf des Mathematiklehrers (der Mathematiklehrerin) vor-

zubereiten?“ war eine vierstufige Bewertung von „sehr wirkungslos“ bis „sehr wirkungsvoll“ abzugeben.

Die Daten aus den Befragungen wurden rasch-skaliert und linear transformiert (Tatto et al., 2012). Der Mittelwert 10 zeigt eine neutrale Position an, sodass – basierend auf den Informationen über den Grad an Zustimmung als eine Form der Item-Schwierigkeit (Rasch, 1980) – höhere Werte Zustimmung und geringere Werte Ablehnung signalisieren. Der empirische Mittelwert der θ-Skala wurde mit Hilfe der Test-Charakteristik-Kurve identifiziert (Rost, 2004). Diese Form der Skalierung weist den Vorteil auf, dass das Skalenniveau durch die Ansiedlung der Parameterschätzungen auf einem Kontinuum Intervall-Skalenniveau aufweist (Rasch, 1980), was für Zusammenhangsanalysen günstiger ist. Zudem können im hier verwendeten *Partial-Credit*-Modell die Schwellen zwischen den einzelnen Kategorien der Likertskalen präziser modelliert werden als in einem klassischen Score, der gleiche Abstände unterstellt (Rost, 2004).

Die Rasch-Skalierung stellt durch die Ansiedlung von Item- und Personenparametern auf einer Skala und das zweistufige Vorgehen bei den Schätzungen dieser auch einen effizienten Weg dar, mit fehlenden Werten umzugehen. Im ersten Schritt der Kalibrierung können nicht erreichte Items in Übereinstimmung mit der TIMSS-Praxis als *missings* und im zweiten Schritt als „falsch“ codiert werden, um möglichst präzise Schätzungen zu erhalten (Ludlow & O’Leary, 1999). Die Standardabweichung der Rohdaten wurde beibehalten, sodass die Effektstärken mit jenen der Einzel-Items vergleichbar sind.

In Ergänzung zur oben berichteten Alpha-Reliabilität wurde die psychometrische Qualität der Skalen, insbesondere ihre nationale und internationale Passung an die Daten sowie ihre Messinvarianz über die TEDS-M-Teilnahmeländer hinweg, mithilfe konfirmatorischer Faktorenanalysen mit gleich gewichteten Länderstichproben und unter Berücksichtigung der genesteten Datenstruktur geprüft. In diesem Zusammenhang wurde im Rahmen von Modellvergleichen auch die strukturelle Validität geprüft (Messick, 1995). Hierzu gehörte, ob sich die erhobenen Dimensionen trennen lassen und ob sie stabil über die verschiedenen Untersuchungspopulationen existieren. Die nationale Konstruktvalidität der Skalen, die verlangt, dass Testwerte in ihrer Bedeutung eindeutig und in einer vorab definierten theoretischen Perspektive interpretierbar sind (Borsboom, Mellenbergh & van Heerden, 2004), wurde in allen Teilnahmeländern in Form von Expertenreviews bestätigt. Die Qualität der Skalen hat sich jeweils als gut erwiesen (für weitere Details siehe Tatto et al., 2012).

3.4 Datenanalysen

Im Zuge der intra-nationalen Datenanalysen galt es im Sinne der obigen Annahme, dass Individualkonstrukte ihre Bedeutung ändern, wenn sie auf nationaler Ebene analysiert werden, die hierarchische Struktur der Daten zu berücksichtigen. Dem Sampling-Design zufolge wurden in allen Teilnahmeländern zunächst zufällig Ausbildungsinstitutionen, in diesen alle vorhandenen Sekundarstufen-I-Ausbildungsgänge und innerhalb dieser

angehende Mathematiklehrkräfte im letzten Jahr ihrer Ausbildung gezogen (Blömeke et al., 2010). In den Analysen ist also zwischen einer Individualebene (Mathematiklehrkräfte), einer institutionellen Ebene (Ausbildungsgänge innerhalb der Ausbildungsinstitutionen) und einer Systemebene (TEDS-M-Teilnahmeländer) zu unterscheiden. Nur wenn diese Clusterstruktur explizit modelliert wird, ist es möglich, die Standardfehler der Parameter korrekt zu schätzen (Snijders & Bosker, 1999; Hox, 2002).

Ausbildungsgänge und Ausbildungsinstitutionen stellen differenzielle Lernumwelten dar. Wie in Deutschland werden in den meisten Ländern unterschiedliche Wege in ein Lehramt der Sekundarstufe I angeboten, die mit unterschiedlichen Lerngelegenheiten verbunden sind, sodass am Ende mit unterschiedlichen Kompetenzbefunden gerechnet werden muss. Der Grad an fachlicher Spezialisierung oder die Spannweite der Klassenstufen, für die zukünftige Lehrkräfte ausgebildet werden, stellen wie die übrigen Ausbildungsmerkmale zudem unterschiedliche Referenzrahmen für die Einschätzungen der Lehrkräfte dar, wie gut sie sich auf ihre Aufgaben vorbereitet sehen, sodass die Gruppierung in den Analysen berücksichtigt werden muss.

Auch für die Erfassung reliabler intersubjektiver Einschätzungen ist die Zuordnung der Befragten zu ähnlichen Lernumwelten notwendig. Dies wird erreicht, indem die Ausbildungsgänge innerhalb der Ausbildungsinstitutionen als Aggregateinheiten verwendet werden (in TEDS-M als *teacher preparation units,* TPU, bezeichnet). Auf diese Weise wird die doppelte Nestung der Lehrkräfte am besten berücksichtigt. Tabelle 1 gibt einen Überblick, in welchen Merkmalen sich die Ausbildungen in den einzelnen Ländern besonders stark unterscheiden.

Das mathematische und das mathematikdidaktische Professionswissen stellen in den Mehrebenenanalysen die abhängigen Variablen dar. Die subjektiven Einschätzungen werden als Prädiktoren eingeführt, um zu prüfen, inwieweit sie geeignet sind, die beiden objektiven Indikatoren vorherzusagen. Alle Prädiktoren wurden um ihren Gruppen-Mittelwert zentriert, um reine Individualeffekte zu erhalten (Snijders & Bosker, 1999). Da vorstellbar ist, dass der Zusammenhang von subjektiver Wahrnehmung und objektiver Leistung nach Land variiert, wurde ein *Random-Slope*-Modell mit Zufallseffekten auf der dritten Ebene geschätzt.

Alle Variablen wurden z-standardisiert, sodass die β-Parameter analog zu herkömmlichen Regressionsanalysen interpretiert werden können. Die vier subjektiven Indikatoren kovariieren in allen Ländern positiv und in mittlerer Stärke. Angehende Mathematiklehrkräfte nehmen sich besser auf ihren Beruf vorbereitet wahr, wenn sie häufiger entsprechende Lerngelegenheiten angeben. Dann bewerten sie auch die Qualität und die Wirksamkeit der Mathematiklehrerausbildung besser.

Zum Ausgleich unterschiedlicher Ziehungswahrscheinlichkeiten und Rücklaufquoten wurden Individual- und TPU-Gewichte verwendet, sodass die Parameterschätzungen robuste Populationswerte darstellen. Ausbildungsgänge, die in einer Institution weniger als vier Lehrkräfte im letzten Jahr ihrer Ausbildung aufwiesen, wurden von den Analysen ausgeschlossen, um stabile Schätzungen zu erhalten. Dieser Schritt reduzierte die Stichprobe der Sekundarstufen-I-Studie auf 8 098 angehende Mathematiklehrkräfte (98,9 %), die sich auf 364 TPUs in 15 Ländern verteilen.

Die Entscheidung, mit mathematischem und mathematikdidaktischem Professionswissen zwei Kriterien der Ausbildungswirksamkeit zu untersuchen, reduziert das Risiko eines „mono-operation bias“ (de Maeyer, van den Bergh, Rymenans & van Petegem, 2010) und erhöht die Validität der Untersuchung. Gleichzeitig besteht das Risiko eines erhöhten Typ-1-Fehlers (Hox, 2002), also falsch-positiver Ergebnisse aufgrund der Korrelation der beiden abhängigen Variablen. Die Spannweite der Korrelationen zwischen mathematischem und mathematikdidaktischem Wissen liegt zwischen $r = .18$ in Botswana und $r = .70$ in Deutschland (Blömeke et al., 2010). Ein multivariates Mehrebenenmodell könnte dieses Problem auffangen, lässt sich in unserem Falle aber nicht umsetzen, da bereits drei Ebenen vorliegen und das Hinzufügen einer weiteren Ebene mit instabilen Ergebnissen verbunden wäre. Da das Risiko, bedeutsame Effekte zu übersehen, zudem offensichtlich eher gering ist (de Maeyer et al., 2010), werden jeweils getrennte Modelle für das mathematische und das mathematikdidaktische Wissen geschätzt und mögliche Probleme am Ende des Beitrags diskutiert. Alle Analysen erfolgten mit HLM 6.08 für Windows.

In Ergänzung zu den Mehrebenenanalysen werden mit denselben Variablen und Stichproben dann herkömmliche Korrelationsanalysen auf Länderebene durchgeführt, um der Frage nachzugehen, ob sich durch den Wechsel der Analyseebene tatsächlich wie in TIMSS unterschiedliche Ergebnisse zeigen. In diesen Zusammenhangsanalysen wird für eine unverzerrte Ermittlung der Standardfehler die Replikationsmethode nach Fay (1989) angewandt, bei der zufällig sogenannte *Balanced Repeated Replication* (BRR)-Zonen stratifizierter Untersuchungseinheiten gebildet werden. Dieses Vorgehen kann als sehr konservativ eingeschätzt werden, sodass die Ergebnisse als hochbelastbar angesehen werden können.

4. Ergebnisse

4.1 Mehrebenenanalysen zum Zusammenhang von subjektiven Einschätzungen und tatsächlich erreichten Leistungen

Sowohl in Bezug auf das mathematische als auch in Bezug auf das mathematikdidaktische Wissen liegt ein deutlich höherer Varianzanteil zwischen den TEDS-M-Teilnahmeländern als innerhalb dieser zwischen den TPUs (siehe die Anmerkungen unter Tab. 2 und 3). In Mathematik wird mit 45,1 % fast die Hälfte der Varianz durch Unterschiede zwischen den Ländern erklärt, durch Unterschiede zwischen den TPUs dagegen nur 17,2 %. Beim mathematikdidaktischen Wissen gilt die höhere Bedeutung des Landes im Vergleich zu den TPUs ebenfalls; der größte Varianzanteil wird mit 56,9 % aber durch individuelle Unterschiede erklärt.

Die subjektive Einschätzung, wie häufig berufsvorbereitende Lerngelegenheiten während der Mathematiklehrerausbildung erfahren wurden, sagt signifikant die erzielten mathematischen und mathematikdidaktischen Testleistungen voraus (Tab. 2 und 3, Modell 1). Wie erwartet ist die Effektstärke mit $\beta = .02$ bzw. $\beta = .04$ allerdings gering.

Prädiktoren	**M1** β (SE)	**M2** β (SE)	**M3** β (SE)	**M4** β (SE)	**M5** β (SE)
Einschätzung der Häufigkeit berufsvorbereitender Lerngelegenheiten	.02 (.01)**				.01 (.00)*
Random Effect	ns				ns
Bewertung Berufsvorbereitung		ns			ns
Random Effect		−.05; .06*			ns
Bewertung Ausbildungskohärenz				ns	ns
Random Effect				−.06; .09***	−.08; .09***
Bewertung Ausbildungswirksamkeit			.04 (.01)*		.03 (.01)*
Random Effect			−.02; .09**		−.02; .09**

Anmerkungen. Varianzkomponenten des unkonditionierten Modells: Land 45.1 %, TPU 17.2 %, Lehrkraft 37.7 %. M = Modell; β = standardisierte Koeffizienten, *SE* = Standardfehler. *Random Effect:* Variation von β zwischen den Teilnahmeländern. * $p < .05$, ** $p < .01$, *** $p < .001$, *ns* = nicht signifikant.

Tab. 2: Drei-Ebenen-Modell zum Zusammenhang von subjektiven Indikatoren zur Wirksamkeit der Sekundarstufen-I-Lehrerausbildung und mathematischem Professionswissen

Prädiktoren	**M1** β (SE)	**M2** β (SE)	**M3** β (SE)	**M4** β (SE)	**M5** β (SE)
Einschätzung der Häufigkeit berufsvorbereitender Lerngelegenheiten	.04 (.00)***				.04 (.01)***
Random Effect	ns				ns
Bewertung Berufsvorbereitung		ns			−.04 (.02)*
Random Effect		ns			ns
Bewertung Ausbildungskohärenz				ns	ns
Random Effect				ns	ns
Bewertung Ausbildungswirksamkeit			.05 (.02)*		.06 (.02)*
Random Effect			−.01; .12*		−.03; .15**

Anmerkungen. Varianzkomponenten des unkonditionierten Modells = Land 31.6 %, TPU 11.5 %, Lehrkraft 56.9 %. M = Modell; β = standardisierte Koeffizienten, *SE* = Standardfehler. *Random Effect:* Variation von β zwischen den Teilnahmeländern. * $p < .05$, ** $p < .01$, *** $p < .001$, *ns* = nicht signifikant.

Tab. 3: Drei-Ebenen-Modell zum Zusammenhang von subjektiven Indikatoren zur Wirksamkeit der Sekundarstufen-I-Lehrerausbildung und mathematikdidaktischem Professionswissen

Der Effekt bleibt in beiden Fällen auch erhalten, wenn die übrigen Indikatoren im Gesamtmodell kontrolliert werden (M5).

Die globale Bewertung der Ausbildungswirksamkeit hängt ebenfalls signifikant mit dem Wissen in Mathematik und Mathematikdidaktik zusammen (M3). Auch dieser Effekt bleibt erhalten, wenn die übrigen Indikatoren kontrolliert werden (M5). Erneut ist die Effektstärke mit $\beta = .04$ bzw. $\beta = .05$ wie erwartet gering. Die Variation dieses Effekts ist zwischen den Ländern beträchtlich. In Bezug auf das mathematische Professionswissen variiert der Zusammenhang der globalen Bewertung der Ausbildungswirksamkeit mit den Testleistungen signifikant zwischen $\beta = -.02$ und $\beta = .09$; in Bezug auf das mathematikdidaktische Wissen variiert der Zusammenhang zwischen $\beta = -.01$ und $\beta = .12$. Da die Effekte fast ausschließlich im positiven Bereich liegen, bestätigt das Ergebnis den signifikant positiven Zusammenhang.

Den übrigen beiden Einschätzungen kommt keine systematische Vorhersagekraft für das fachbezogene Professionswissen zu. Die Hypothese, dass zumindest schwach-positive Zusammenhänge zwischen der Einschätzung, wie gut sich die angehenden Mathematiklehrkräfte der Sekundarstufe I auf ihren Beruf vorbereitet fühlen (M2) bzw. wie kohärent sie die Ausbildung insbesondere im Hinblick auf die Abstimmung von Theorie und Praxis wahrgenommen haben (M4), und dem mathematischen bzw. mathematikdidaktischen Wissen bestehen, wird von den TEDS-M-Daten nicht gestützt.

Zwar zeigen sich im Falle des Mathematikwissens (Tab. 2, M2 bzw. M4) ausweislich der Zufallseffekte Unterschiede zwischen den Ländern, was den Zusammenhang zwischen der Bewertung von Berufsvorbereitung und Ausbildungskohärenz mit der Testleistung angeht. Die Variation reicht mit $\beta = -.05$ bis $\beta = .06$ bzw. $\beta = -.06$ bis $\beta = .09$ aber nur von praktisch wenig bedeutsamen negativen bis zu praktisch wenig bedeutsamen positiven Korrelationen. Die geringe Bedeutsamkeit gilt auch für den schwach negativen Effekt von $\beta = -.04$ der Einschätzung zur Berufsvorbereitung auf das mathematikdidaktische Wissen im Gesamtmodell unter Kontrolle der übrigen Indikatoren (Tab. 3, M5).

4.2 Internationale Analysen auf Länderebene

Werden die Zusammenhangsanalysen zwischen subjektiven und objektiven Indikatoren auf Länderebene durchgeführt, ergibt sich ein deutlich anderes Bild als zuvor in den Mehrebenenanalysen (siehe Tab. 4). Sowohl die Angaben zur Häufigkeit berufsvorbereitender Lerngelegenheiten als auch die Bewertung der Berufsvorbereitung hängen wie erwartet signifikant negativ mit dem erworbenen Professionswissen zusammen. Je höher ihre mathematischen und mathematikdidaktischen Testleistungen ausfallen, desto seltener ($r = -.33$ bzw. $r = -.42$) und schlechter vorbereitet ($r = -.42$ bzw. $r = -.54$) nehmen sich angehende Mathematiklehrkräfte der Sekundarstufe I wahr. Die Bewertung der Ausbildungskohärenz und das Professionswissen kovariieren nicht systematisch.

Entgegen unserer Hypothese entfaltet die globale Einschätzung der Wirksamkeit der Mathematiklehrerausbildung positive Vorhersagekraft für das mathematische und das

	MCK	MPCK
Häufigkeit berufsvorbereitender Lerngelegenheiten	−.33	−.42
Bewertung Berufsvorbereitung	−.42	−.54
Bewertung Ausbildungskohärenz	ns	ns
Bewertung Ausbildungswirksamkeit	.39	.23

Anmerkungen. MCK = Mathematisches Professionswissen, MPCK = Mathematikdidaktisches Professionswissen.

Tab. 4: Zusammenhang zwischen subjektiven und objektiven Indikatoren der Wirksamkeit der Mathematiklehrerausbildung auf Länderebene

mathematikdidaktische Wissen (r = .39 bzw. r = .23). Hier stimmt das Länderergebnis also mit der Individualebene überein.

5. Zusammenfassung und Diskussion

Im vorliegenden Beitrag wurde für die Mathematiklehrerausbildung der Sekundarstufe I in 15 Ländern untersucht, inwieweit fachbezogenen Befragungen Validität in Bezug auf das tatsächlich in der Ausbildung erreichte fachbezogene Professionswissen bescheinigt werden kann. Als subjektive Indikatoren wurden solche verwendet, die typischerweise in Evaluationsstudien eingesetzt werden: eine Einschätzung der Häufigkeit berufsvorbereitender Lerngelegenheiten, eine Bewertung des Grades an Berufsvorbereitung und Ausbildungskohärenz sowie eine globale Einschätzung der Ausbildungswirksamkeit. Als objektive Maße wurden Testergebnisse zum mathematischen und mathematikdidaktischen Professionswissen verwandt, denen prädiktive Bedeutsamkeit für Schülerleistungen zukommt (Hill et al., 2005; Baumert et al., 2010).

Wird die besondere Struktur der Daten beachtet – Nestung der angehenden Mathematiklehrkräfte in Ausbildungsgängen verschiedener Ausbildungsinstitutionen und dieser wiederum in unterschiedlichen Ländern –, stützen die Analysen für die Häufigkeit berufsvorbereitender Lerngelegenheiten und die globale Einschätzung der Ausbildungswirksamkeit die Hypothese, dass der Zusammenhang zwischen Befragungs- und Testergebnissen fachbezogen signifikant positiv ist. Wer subjektiv günstigere Urteile abgibt, weist auch objektiv ein höheres mathematisches und mathematikdidaktisches Professionswissen auf. Wie erwartet sind die Effektstärken allerdings gering. Häufigkeits- und Wirksamkeitseinschätzungen aus Befragungen kann damit begrenzt prädiktive Validität für den Grad an erreichter fachbezogener Lehrerprofessionalität zugesprochen werden.

Bewertungen, wie gut sich die angehenden Mathematiklehrkräfte auf ihren Beruf vorbereitet fühlen und wie kohärent die Ausbildung war, hängen dagegen nicht mit dem erworbenen Professionswissen zusammen. Obwohl solche Einschätzungen Standard in Evaluationen der Lehrerausbildung sind und aus ihnen weitreichende Schlussfolgerun-

gen zur Lehrerprofessionalität abgeleitet werden (National Center for Education Statistics, 2003; Oser & Oelkers, 2001), stellt sich ihre prädiktive Validität auf Basis der TEDS-M-Daten als gering dar.

Diese Ergebnisse zur Lehrerausbildung stärken die zunehmenden nationalen (Braun & Hannover, 2011; König et al., 2012) und internationalen (Pascarella, 2001; Gonyea, 2005) Belege von Diskrepanzen zwischen subjektiven Einschätzungen und objektivem Kompetenzerwerb in der universitären Ausbildung. Zuletzt hat Bowman (2011) in einer Längsschnittstudie an 46 Colleges in den USA maximal gering positive Zusammenhänge für Testergebnisse und subjektive Einschätzungen des Kompetenzerwerbs in verschiedenen Domänen durch die Studierenden nachgewiesen.

Damit stellt sich die Frage, worauf die Diskrepanz beruhen könnte. Möglicherweise resultiert die geringe prädiktive Validität subjektiver Indikatoren bei angehenden Lehrkräften aus einer unterschiedlichen Konzeptualisierung von „Berufsvorbereitung" und „Ausbildungskohärenz" im Vergleich zu Experten. Unterrichtsmethodik und Handlungssicherheit spielen für die Befragten eine wichtige Rolle (Schneider & Bodensohn, 2010; Haag & Streber, 2010). Sie müssen am Ende der Ausbildung in vielen Ländern zum ersten Mal Unterricht erteilen. Die Komplexität des unterrichtlichen Geschehens richtet den Fokus kurzfristig möglicherweise eher auf das eigene „Überleben" im Klassenraum als auf das Erreichen hoher Schülerleistungen. Die subjektiven Einschätzungen zur Berufsvorbereitung wären in diesem Falle eher Indikatoren für Selbstwirksamkeitserleben als für die intendierten Merkmale. Selbige Bedeutungsdifferenz kann für die Definition von „Ausbildungskohärenz" angenommen werden, für deren Einschätzung die angehenden Lehrkräfte dem fachbezogenen Professionswissen möglicherweise untergeordnete Bedeutung zuweisen, sobald es nicht unmittelbar auf den Unterricht bezogen ist. Die Diskrepanz besteht hier dann zu einem subjektiven Erwartungshorizont an die Mathematiklehrerausbildung.

Auch wenn sich diese Ergebnisse lediglich auf die Mathematiklehrerausbildung beziehen und die Übertragbarkeit auf andere Unterrichtsfächer und die Primarstufe erst noch geprüft werden muss, mahnen die Ergebnisse zur Vorsicht, was Schlussfolgerungen aus Evaluationen der Lehrerausbildung angeht, solange sich diese nur auf subjektive Einschätzungen stützen. Mit Ausnahme einer globalen Einschätzung der Ausbildungswirksamkeit, können Befragungen objektive Maße möglicherweise nicht ersetzen, sondern nur ergänzen.

Der Aufwand, valide Hinweise zum Grad an fachbezogenem Professionswissen von Lehrkräften zu erhalten, das diese in der Ausbildung erreichen, würde mit zusätzlichen Kompetenztests allerdings deutlich steigen. Die Testungen müssten domänenspezifisch angelegt sein, sodass für jedes Unterrichtsfach und jede Schulstufe eigene Instrumente zu entwickeln, administrieren und auszuwerten wären. Befragungen sind im Vergleich dazu ökonomischer. Für die zukünftige Forschung zur Lehrerprofessionalität stellt sich damit die Aufgabe, die Genauigkeit von Selbsteinschätzungen detaillierter zu beleuchten, um falsche Schlussfolgerungen zu vermeiden. Ein wichtiger Schritt wäre, verzerrende Einflüsse zu identifizieren, um sie kontrollieren zu können.

Aufgaben ergeben sich aus den Ergebnissen auch mit Blick auf die Professionalisie-

rung von Lehrkräften. Wichtig erscheint die Schaffung eines stärkeren Bewusstseins für die Notwendigkeit einzelner Ausbildungskomponenten. Offensichtlich wird den angehenden Lehrkräften zu wenig deutlich, warum sie bestimmte fachbezogene Inhalte belegen müssen. Es sollte zum hochdidaktischen Standard gehören nachzuweisen, dass und warum spezifische fachbezogene Elemente tatsächlich wichtig für eine kompetente Berufsausübung sind, auch wenn Lehrerprofessionalität mehr als mathematisches und mathematikdidaktisches Wissen ist und beispielsweise auch Aspekte des Klassenmanagements umfasst. Diese Komplexität stellt vermutlich auch eine Erklärung dar, wieso die Effektstärken begrenzt sind.

Die Ergebnisse mahnen zudem zur Vorsicht im Umgang mit den Ergebnissen aus internationalen Vergleichen. Wie in TIMSS (Shen & Tam, 2008) muss in TEDS-M die Gefahr ökologischer Fehlschlüsse festgestellt werden, wird die falsche Analyseebene als Untersuchungseinheit gewählt. Bei Analysen auf Länderebene zeigen sich für die Häufigkeit und Bewertung der Berufsvorbereitung vermeintlich negative Zusammenhänge zum fachbezogenen Professionswissen. Angehende Mathematiklehrkräfte in Ländern mit geringem mathematischem und mathematikdidaktischem Wissen sehen sich besser auf den Beruf vorbereitet als in Ländern an der Spitze der Rangreihe.

Es ist plausibel anzunehmen, dass sich in den Aggregatergebnissen kulturell unterschiedliche Referenzrahmen niederschlagen (van de Vijver & Leung, 1997). Bei vergleichbarem mathematischem und mathematikdidaktischem Wissen können sich angehende Lehrkräfte in Ländern mit geringeren individuellen, familiären oder gesellschaftlichen Bildungsaspirationen bzw. geringeren curricular gesetzten Bildungsstandards als besser auf den Beruf vorbereitet wahrnehmen, obwohl sie weniger wissen, als Lehrkräfte in Ländern mit sehr hohen Ansprüchen und Standards. So sind die curricularen und gesellschaftlichen Erwartungen an das, was Mathematiklehrkräfte auf den Philippinen oder in Malaysia in der Schule zu leisten haben, vermutlich sehr verschieden von dem, was von Mathematiklehrkräften in Taiwan erwartet wird.

Hinzu kommen ggf. kulturell unterschiedlich ausgeprägte Tendenzen, kritische Einschätzungen vorzunehmen (Harzing, 2006). Dies wird in TEDS-M anhand von Mittelwertvergleichen deutlich, wenn man die negativen Bewertungen der Ausbildungsqualität und -wirksamkeit der mitteleuropäischen Länder Deutschland, Norwegen und Schweiz betrachtet. Entsprechende Effekte sind auf der Länderebene aufgrund der hohen Intra-Klassen-Korrelation gut zu identifizieren. Diese Ergebnisse bedeuten, dass ein direkter Vergleich von Mittelwerten, die aus Befragungen stammen, über unterschiedliche Kulturen schwierig ist. Hier muss immer besondere Vorsicht walten, ob die Daten tatsächlich Unterschiede in den erfassten Konstrukten abbilden oder ob sich nicht andere Eigenschaften niederschlagen.

Einzige Ausnahme stellt die globale Einschätzung der fachbezogenen Wirksamkeit der Mathematiklehrerausbildung dar, die auf Aggregat- und Individualebene positiv mit den Testleistungen kovariiert. Blickt man auf die konkrete Formulierung dieses Items im Vergleich zu den übrigen Einschätzungen, ist dieses Item vermutlich am wenigsten durch kulturelle Rahmenbedingungen wie curriculare oder gesellschaftliche Erwartungen geprägt, sondern eher auf die Wahrnehmung der eigenen, intra-individuellen Ent-

wicklung während der Lehrerausbildung bezogen. Diese valide einzuschätzen, scheint den angehenden Lehrkräften eher möglich zu sein.

Für die pädagogische Forschung bedeuten die Ergebnisse der vorliegenden Studie zum einen, in Untersuchungen zur Lehrerausbildung sensibel auf den Referenzrahmen zu achten. Zum anderen gilt es, Mehrebenenmodelle als Standard einzusetzen, wenn hierarchisch geschachtelte Daten vorliegen (siehe entsprechend Biedermann et al., 2012; Steinmann & Oser, 2012). Die TEDS-M-Ergebnisse machen einmal mehr auf die Berechtigung von Robinsons (1950) Warnung vor der Gefahr logischer Fehlschlüsse aufmerksam, wenn Aggregatergebnisse auf Individuen übertragen werden.

Literatur

Abs, H. J., Döbrich, P., Vögele, E., & Klieme, E. (2005). *Skalen zur Qualität der Lehrerbildung – Dokumentation der Erhebungsinstrumente: Pädagogische Entwicklungsbilanzen an Studienseminaren*. Frankfurt a. M.: Gesellschaft zur Förderung pädagogischer Forschung.

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht, 57*, 175–193.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*, 469–520.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom and student progress. *American Educational Research Journal, 47*, 133–180.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *Educational Research, 35*, 463–482.

Biedermann, H., Brühwiler, Ch., & Krattenmacher, S. (2012). Lernangebote in der Lehrerausbildung und Überzeugungen zum Lehren und Lernen. Beziehungsanalysen bei angehenden Lehrpersonen. *Zeitschrift für Pädagogik, 58*, 460–475.

Blömeke, S., Kaiser, G., & Lehmann, R. (Hrsg.) (2010). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.

Blömeke, S., Suhl, U., & Döhrmann, M. (2012). Zusammenfügen was zusammengehört. Kompetenzprofile am Ende der Lehrerausbildung im internationalen Vergleich. *Zeitschrift für Pädagogik, 58*(4), 422–440.

Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education, 62*, 154–171.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Bowman, N. A. (2011). Validity of college self-reported gains at diverse institutions. *Educational Researcher, 40*, 22–24.

Braun, E., & Hannover, B. (2011). Gelegenheiten zum Kompetenzerwerb in der universitären Lehre. Zusammenhänge zwischen den Einschätzungen Studierender und unabhängigen Beobachtungen relevanter Merkmale universitärer Lehrveranstaltungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43*, 22–28.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (S. 177–212). Göttingen: Hogrefe.

de Maeyer, S., van den Bergh, H., Rymenans, R., & van Petegem, P. (2010). Effectiveness criteria in school effectiveness studies: Further research on the choice for a multivariate model. *Educational Research Review, 5*, 81–96.

Fan, L., & Cheong, N. P. C. (2002). Investigating the sources of Singaporean mathematics teachers' pedagogical knowledge. In D. Edge & B. H. Yap (Hrsg.), *Mathematics education for a knowledge-based era 2* (S. 224–231). Singapur: AME.

Fay, R. E. (1989). Theory and Application of Replicate Weighting for Variance Calculations. *Proceedings of the Survey Research Methods Section*, 212–217. American Statistical Association.

Fend, H. (1980). *Theorie der Schule*. München: Auer.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: a classification of score adjustment procedures and review of research. *Journal of Cross-Cultural Psychology, 35*(3), 263–282.

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendation. In P. D. Umbach (Hrsg.), *Survey research: Emerging issues* (S. 73–89). San Francisco: Jossey-Bass.

Haag, L., & Streber, D. (2010). Unterrichtsvorbereitung bei Lehrern – mit System? *Lehrerbildung auf dem Prüfstand, 3*(1), 107–117.

Harzing, A.-W. (2006). Response Styles in Cross-national Survey Research. A 26-country Study. *International Journal of Cross Cultural Management, 6*(2), 243–266.

Helmke, A. (2004). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze: Kallmeyer.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430–511.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371–406.

Hofstede, G. (2001). *Culture's Consequences. Comparing Values, Behaviors, Institutions and Organizations across Nations* (2. Aufl.). Thousand Oaks: Sage.

Hox, J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah: L. Erlbaum.

Ingvarson, L., Beavis, A., & Kleinhenz, E. (2007). Factors affecting the impact of teacher education courses on teacher preparedness: Implications for accreditation policy. *European Journal of Teacher Education, 30*, 351–381.

Jäger, R. S., & Milbach, B. (1994). Studierende im Lehramt als Praktikanten – eine empirische Evaluation des Blockpraktikums. *Empirische Pädagogik, 8*, 199–234.

Joint Committee on Standards for Educational Evaluation (1994). *The Program Evaluation Standards. How to Assess Evaluations of Educational Programs*. Thousand Oaks: Sage.

Kane, M. (2006). Validation. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 17–64). Westport: National Council on Measurement in Education.

Kee, A. N. (2012). Feelings of Preparedness among Alternatively Certified Teachers: What is the Role of Program Features? *Journal of Teacher Education, 63*, 23–38.

Klieme, E., & Vieluf, S. (2009). Teaching practices, teachers' beliefs and attitudes. In OECD (Hrsg.), *Creating Effective Teaching and Learning Environments. First Results from TALIS* (S. 87–135). Paris: OECD.

König, J., Kaiser, G., & Felbrich, S. (2012). Spiegelt sich pädagogisches Wissen in den Kompetenzselbsteinschätzungen angehender Lehrkräfte? Zum Zusammenhang von Wissen und Überzeugungen am Ende der Lehrerausbildung. *Zeitschrift für Pädagogik, 58*(4), 476–491.

Kubinger, K. D. (2003). (Un-)Verfälschbarkeit. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 429–432). Weinheim/Basel: Beltz/PVU.

Lienert, G. A., & Raatz, U. (1994). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim/Basel: Beltz.

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: practical data analysis implications. *Educational and Psychological Measurement, 59*, 615–630.

Merzyn, G. (2002). *Stimmen zur Lehrerausbildung. Ein Überblick über die Diskussion*. Baltmannsweiler: Schneider-Verlag Hohengehren.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165.

National Center for Education Statistics (2003). *Public school teacher questionnaire, 2003–2004. Schools and Staffing Survey (SASS)*. http://nces.ed.gov/surveys/sass/question0304.asp [24.01.2012].

Oser, F., & Blömeke, S. (Hrsg.) (2012). Überzeugungen von Lehrpersonen. *Thementeil der Zeitschrift für Pädagogik, 58*(4).

Oser, F., & Oelkers, J. (Hrsg.) (2001). *Die Wirksamkeit der Lehrerbildungssysteme*. Zürich: Rüegger.

Pascarella, E. T. (2001). Using student self-reported gains to estimate college impact: A cautionary tale. *Journal of College Students Development, 42*, 488–492.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. In J. Sikula, T. J. Buttery & E. Guyton (Hrsg.), *Handbook of Research on Teacher Education* (S. 102–119). New York: Macmillan.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review, 15*(3), 351–357.

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

Schmidt, W. H., Blömeke, S., & Tatto, M. T. (2011). *Teacher education matters: A study of the mathematics teacher preparation from six countries*. New York: Teacher College Press.

Schneider, C., & Bodensohn, R. (2010). Entwicklung beruflicher Handlungskompetenzen in der ersten Phase der Lehrerausbildung. In J. Abel & G. Faust (Hrsg.), *Wirkt Lehrerbildung. Antworten aus der empirischen Forschung* (S. 227–234). Waxmann: Münster.

Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation, 14*(1), 87–100.

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* (S. 3–36). New York: Macmillan.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modelling*. London: Sage.

Steinmann, S., & Oser, F. (2012). Prägen Lehrerausbildende die Beliefs der angehenden Primarlehrpersonen? Shared Beliefs als Wirkungsgröße in der Lehrerausbildung. *Zeitschrift für Pädagogik, 58*(4), 441–459.

Takahashi, K., Ohara, N., Antonucci, T. C., & Akiyama, H. (2002). Commonalities and Differences in Close Relationships among the Americans and Japanese: A Comparison by the Individualism/Collectivism Concept. *International Journal of Behavioral Development, 26*(5), 453–465.

Tatto, M. T., Schwille, J., Senk, S. L., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher education and development study in mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics. Conceptual framework*. East Lansing: Teacher Education and Development International Study Center, College of Education, Michigan State University.

Tatto, M. T., Schwille, J., Senk, S. L., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., & Reckase, M. (2012). *The teacher education and development study in mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics – Findings from the IEA study of the mathematics preparation of future teachers.* Amsterdam: IEA.

Terhart, E. (2001). *Lehrerberuf und Lehrerbildung. Forschungsbefunde, Problemanalysen, Reformkonzepte*. Weinheim/Basel: Beltz.

van de Vijver, F., & Leung, K. (1997). *Methods and Data analysis of comparative research*. Thousand Oaks: Sage.

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

Abstract: Based on the international comparative study TEDS-M, the validity of surveys on teacher education is examined. On the one hand, the aim is to relate self-reported assessments of the quality of teacher training to the actual achievement of future teachers in order to thus evaluate the predicative validity of the former; on the other hand, the risk of an ecological fallacy is substantiated, which arises when the wrong analytical unit is chosen in the analysis of international data. In multi-level analyses carried out with about 8 000 future math teachers in lower secondary education from 15 different countries, four typical questions from evaluative studies on the efficacy of teacher training are related to the actually manifested mathematical and math-didactical professional competence. A global assessment of the efficacy of teacher training reveals a minor positive correlation with professional competence. To more differentiated evaluations, however, no predicative validity can be attributed. The results emphasize the need to be cautious with regard to conclusions concerning teacher training that rely solely on self-reported data. In the context of international comparisons the cultural differences between the countries have to be taken into consideration, too, since different referential frameworks for these assessments exist and the data collected change their meaning once they are aggregated on country level.

Keywords: Multi-Level Modeling, Survey, Validity, Ecological Fallacy, Comparative Study

Anschrift der Autorin

Prof. Dr. Sigrid Blömeke, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: sigrid.bloemeke@staff.hu-berlin.de